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 2005

TYPE OF REPORT: Final

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

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-09-2005	Final	1 Sep 00 – 31 Aug 05

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

5a. CONTRACT NUMBER

5b. GRANT NUMBER
DAMD17-00-1-0410

5c. PROGRAM ELEMENT NUMBER

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

E-Mail: gurd@msx.upmc.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

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. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

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

13. SUPPLEMENTARY NOTES

14. ABSTRACT

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 approach the possible impact of a unique tele-mammography 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 phases of the project encompassed the design, assembly, and technical testing of a multi-site tele-mammography system that enables the digitization, transmission, and display of wavelet compressed images, as well as associated text documents of a case combined with a "chat message and CAD results in less than 15 minutes. The possible impact of such a system was evaluated during a step-by-step assessment in a series of clinically simulated multi-site studies.

15. SUBJECT TERMS
Breast Cancer, Telemammography, Detection, CAD, Remote Management

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	131	19b. TELEPHONE NUMBER *(include area code)*

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 increasing compliance, longer life expectancy, and 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 tele-mammography 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, ultrasound) also make it desirable to enable physician "monitoring" and "management" of remote underserved locations so that some patient-management decisions can be made while the patient remains in the clinic [8-11]. In addition, a technologist who observes a possible abnormality during the performance of the study could benefit from the ability to communicate her/his suspicion, and an expert mammographer could review the specific case, together with the technologist's observation, resulting in an improved and perhaps a more timely diagnosis. Current practices result in increased patient anxiety and added practice complexity and cost. Even in practices in the urban setting, recommendations for recall are not always followed by the woman and eliminating the need to return to the clinic through implementation of this concept, in particular in remote locations, could increase overall compliance. Early attempts to develop and implement a practical tele-mammography solution to this problem failed due to several significant technical problems associated with acquisition, transmission, management, and display of the images and other related information [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 tele-mammography 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 remain limited to lower level communication capabilities. Other communication technologies, such as satellites, are being evaluated for this purpose, but it is not likely that these will displace lower level communication technologies in many underserved areas for quite some time [19-23]. 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 assembled and evaluated a unique tele-mammography 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, magnification views, ultrasound) are needed. The system was designed based on prior preliminary experience acquired in our group during ten years of research in this general area [24, 25]. It includes the use of a common carrier for communication (Plain Old Telephone System, POTS) and other "low level" communication capabilities, wavelet-based image compression for data reduction, and the optional incorporation into the transmitted information of other text information such as location of suspected abnormality, and CAD results. The main goal was to assess in a step-by-step,

clinically simulated approach whether the use of such a system could potentially reduce recall rates in the remote sites. Other objectives regarding measurements of actual practice parameters in a large academic based screening mammography practice were performed. Last, ways to improve communication between the technologist at the remote site and a radiologist at the central site, as well as creating an environment for "more active" participation of the technologist in the diagnostic process, were also explored.

Body:

Since the initiation of the project on September 1, 2000, we have been executing step by step the tasks listed in the Statement of Work, as originally submitted. As will be explained in the body of this final report, our initial findings resulted in the addition of several technical and observational tasks that were successfully performed in order to maximize our ability to learn about the practical applications being investigated in this project. As an example, during year four of the project, a significant new addition (capability) was added to the system as a result of our previous observations from the clinically simulated experiments. We incorporated the ability to submit (transmit from the remote sites) both prior mammograms (when available) as an integral part of the examination to be evaluated as well as an interactive overlay drawing of the examination with location(s) marking of suspected abnormality(s) to be reviewed by a radiologist. This required a substantial technical effort and ultimately resulted in a major software upgrade of the system. Hence, substantial parts of the last clinically simulated experiment were performed during a one year no-cost extension of the project.

Under Task 1, we performed the following:

All subtasks listed under this task were completed. We assembled and tested a multi-site tele-mammography system that met (and in several respects exceeded) our originally proposed specifications. The status of the tasks described under this category is as follows:

a) **Select and Purchase Equipment**: During year one of the project, we purchased and tested a significant amount of equipment in support of the project that was funded mainly from other sources. This includes, but is not limited to, computers, laser printers and film digitizers. During the selection phase, we performed a comprehensive side-by-side evaluation of the VIDAR and Lumisys film digitizers to assess whether or not the CCD-based VIDAR digitizer could be used for this purpose. Our assessment resulted in confirmation that the Lumisys film digitizer was significantly more robust and that the signal-to-noise ratio at high frequencies is significantly higher. In addition, the new digitizer raises the maximum optical density to ~3.8, which is a significant advantage over the older versions. As a result, we purchased three additional digitizers for the performance of this project. We also acquired (at no cost to the project) a Kodak 8600 model laser printer, tested it, and developed an interface to control it remotely.

b) **Convert Software to Windows Based:** The general design of the tele-mammography project was reconsidered, and software was written using the NT-operating system to enable significantly more flexibility for the different applications that could be implemented. This task was completed and after initial testing, refinements were performed. All communication tasks have also been tested using the modified software.

c) **Develop Interface to FFDM Acquisition System:** We designed and developed an interface on the system to accept DICOM images that were acquired on FFDM systems. We transferred FFDM images to the server and displayed these on the workstation at the central site in addition to providing printing capability. All the functionality specifications were tested, but our clinical practice postponed the transition to FFDM based screening (as we reported in years 1-3 of the project) hence other than enabling the tele-mammography system to accept FFDM based examinations (in a manner compatible with all other clinical requirements), all of our clinically simulated studies throughout the project were film based. Our own transition to a fully digital system is underway and will be completed by next September (2006). However we are a large academic center and it is not clear at all 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**: A remote site user interface was completed and tested, both subjectively (by staff members and technologists) 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 finalized.

e) **Complete Data Compression Software Module:** A compression software scheme compatible with JPEG 2000 was finalized and tested. The scheme allowed for a site-specific selectable level of compression to be used. However after the initial testing was completed, we fixed the compression levels for all sites (see below). In addition to the data compression module, the approach we incorporated in the system includes a comprehensive tissue segmentation routine followed by a wavelet transform and "dialable" data compression module. This segmentation routine enables a very efficient data reduction by eliminating non-tissue regions of the image without any loss of information.

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 also found out that the physicians' high level of acceptance of either compression level practically eliminated the need for this "selectable" option. Therefore, we proceeded using the system with a fixed level of compression (75:1), at all sites.

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

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

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

j) **Test System in Laboratory:** The system was tested extensively 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 and installation at remote sites.

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

Additional development efforts:

As our step-by-step clinically simulated experiments progressed, we kept adding functionality to the system. These changes required modifications that enabled an integrated, easy to use functionality and in year four of the project we decided to include prior images (examinations) to the case folder. Ultimately, we enabled the following tools on the system: 1) text messaging (namely, two-way "chat" between the remote technologist and central radiologists), 2) marking of suspicious locations (namely, the technologist marks suspicious regions on an image overlay), 3) CAD results, 4) prior mammography reports, and 5) prior images (when available). The reason for the additional tools was to provide the radiologist at the central location with all possible tools to enable better assessment of the examinations being sent for review (obviously, this is all done in addition to the actual mammographic images in question). Hence the last task ("high volume clinically simulated demonstration and evaluation") was performed with all tools available to the technologists (at the remote sites) and the physicians (at the central site). A major upgrade was installed and tested for this purpose in year four of the project.

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) was 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 our Breast Center. This required some construction that was completed at no cost to the project.

b) The complete system was reassembled on location at all sites.

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.

4) The design considerations and initial testing were published in comprehensive SPIE reports (see references 3 and 8 in the "Reportable Outcomes" section).

Under Task 3, we performed the following:

1) Retrospective observer performance studies:

A step by step assessment protocol was implemented in this project. Four independent observer performance studies were performed as information provided to the reader at the remote site was incrementally increased. In the first study, 306 examinations of all types (without a rigorous selection process) were sent from all sites and were read at the central site by 5 radiologists. The study included a large number of cases that were (and some that were not) suspected by the technologists at the remote site as possibly needing additional procedures. The results suggested see table 1 that a large number of additional procedures would be performed in the remote site in order to reduce recall rates (by approximately a ratio of 3:1 or 433 additional procedures would have been needed to reduce recalls by 151).

All readers combined (same cases)

		clinical read recall	no recall	total
study	recall	151	433	584
read	no recall	59	887	946
	total	210	1320	1530

	overall agreement	1038
prob-observed =	0.678	
prob-expected =	0.586	
Kappa =	0.224	

As a result, we performed three additional observer performance studies while gradually increasing the amount of information transmitted to the central site. In all three studies cases were specifically selected by the technologists when they felt during the QA review of the examinations in question that the women would likely be recalled by the radiologist for additional procedures.

A synopsis of the three observer performance studies in this area follows: registered, experienced mammography technologists from three remote imaging sites transmitted 245 screening mammography exams to a central site (radiologists), which they (the technologists) believed needed additional procedures. Four data components are transmitted from the remote site: (1) image data - current exam mammography films digitized at 50 μm pixel

dimensions; (2) text and graphic communication between the technologist and the radiologist via a "chat" box in which the technologist can describe and mark suspicious regions on integrated generic images; (3) prior patient reports when available; and (4) computer-aided detection (CAD) results. At the central site images are displayed on a workstation consisting of three high-resolution, portrait monitors. The image data with the CAD results overlaid are displayed on two monitors and the chat box and prior reports on the third monitor. Seven radiologists reviewed and rated the exams on the tele-mammography workstation and indicated: (1) if additional procedures were recommended, (2) when appropriate, which breast was involved, and (3) when appropriate, the specific recommended procedures. The performance of the radiologists on the workstation was compared with the actual clinical interpretation of the same examinations. Study 1 had two interpretation modes: (1) images only and (2) images and technologist's text message. Study 2 had two modes: (1) images and technologist's text message and (2) images, text message, and prior report. Study 3 had three modes: (1) images, technologist's text message, and prior report; (2) images, text message, prior report, and technologist's graphic location marks; and (3) images, text message, prior report, graphic marks (location), and CAD results. Amongst other analyses, we computed the potential improvements in terms of projected reduction in recall rates at the remote sites and associated "costs" in terms of "unnecessary" additional procedures.

Results: Technologists were able to identify suspicious examinations that may require additional procedures, but their "recommended" examinations amounted to a substantially larger number compared with that of a clinical interpretation by a radiologist. The screening exams were successfully transmitted, processed, reviewed, and rated. The percent of exams recalled for recommended additional procedures (termed "recall") during the actual clinical interpretation for Studies 1 (n = 130), 2 (n = 99), and 3 (n = 115) were 39.2%, 38.4%, and 42.2%, respectively. Tele-mammography Study 1; modes 1 and 2 had mean recall rates of 73.3% (+/- 17.9) and 82.5% (+/- 16.2), respectively, and mean agreements of 51.7% (+/- 5.5) and 48.7% (+/- 6.3), respectively. Study 2; modes 1 and 2 had mean recall rates of 79.6% (+/- 12.3) and 77.5% (+/- 13.8), respectively, and mean agreements of 52.3% (+/- 6.7) and 52.8% (+/- 7.0), respectively. Study 3; modes 1, 2 and 3 had mean recall rates of 72.3% (+/- 9.3), 72.3% (+/- 9.3), and 72.7% (+/- 9.2), respectively, and mean agreements of 57.4% (+/- 4.6), 57.1% (+/- 3.9), and 56.7% (+/- 3.9), respectively. However, it should be remembered that without radiologists' reviews 100 percent of these women were "recommended" for additional procedures by the technologists.

In these studies, we demonstrated that between 70 and 85 percent of recalls (as ultimately were decided during the clinical interpretation) could have been avoided, albeit at a high "cost" of performing additional procedures on these women. As we increased the information provided to the radiologist from "text message" alone to text message, prior reports and a location overlay to all of the above plus CAD results the number of "unnecessary" procedures recommended by the radiologists reduced progressively from 1.45 (246/183) to 1.26 (171/136) to 1.07 (216/202) per "saved" recall. As can be seen in the last task this number was further reduced to 0.94 (81/86) during the last experiment (see task #5 below).

2) Clinical assessment of performance levels:

a. There are several aspects of the task that are worth noting. First we were "breaking ground" in several respects that include but are not limited to the involvement of technologists in the decision-making process (namely, which cases to send over to the

central site and why), and possibly the increased "reliance" of the radiologists on the technologists' judgments. Our subjective assessments of this issue clearly indicated that both radiologists and technologists welcomed increased communication. We often heard comments like "the technologists often identify abnormalities before we do and sometimes see thing we do not". However, our practice is a very established one and it is not clear that this "reliance" and "trust" would exist in other practices.

b. As a part of this investigation, we assessed our clinical performance levels in the traditional practice (without tele-mammography). We analyzed data available in our databases concerning patient distributions and process-related information. This includes, but is not limited to, the recall rate by physician, site, type, and reason for recall. Our recall rates and cancer detection rates were found to be very stable for the group of radiologists as a whole and individual radiologists as well.

c. We also reviewed 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. One of the more interesting (and relevant) findings in this regard was the long cycle time including scheduling (average was > 20 days at the time) between the patient's call for an appointment due to a recall and the actual date of examination. This highlights the potential benefit of the use of tele-mammography to reduce recall rates (hence cycle time), in particular in busy practices such as ours. This information, which is now reviewed monthly, generated a significant effort in our system (including, but not limited to, performing diagnostic sessions during the weekends) and we have been able to make substantial improvements in this regard.

d. During the project period, we completed a large study to assess the effect of the introduction of CAD into our clinical environment and the relationship between recall rates and detection rates for our ten highest volume radiologists. One of the important issues that was raised in our group was the correlation (if any) between the recall and detection rates of radiologists. This is an important point since there is a significant pressure on radiologists to reduce their individual recall rates to below ten percent. While we recognize the tremendous value of reducing recall rates without a substantial degradation in detection rates (sensitivity), the question arises as to whether or not higher recall rates are also generally associated with higher detection rates. These studies involved the reviews of over 115,000 records and resulted in important observations that were published in JNCI and Cancer (see publications list). We strongly believe that the use of CAD will ultimately be an integral part of the diagnostic process and some of our continuing efforts to develop and improve CAD schemes were supported (only to a very minimal level) by this project (see publications list).

e. As indicated above, we were not able to assess practice parameters for screening mammography using FFDM because in our system the system had been used largely in diagnostic procedures (rather than screening).

Under Task 4, we performed the following:

a. CAD Software Module:
A software module specifically designed for the tele-mammography system was designed and written. This module is different than our CAD development efforts in that it is very flexible and any new CAD development in our research projects can be easily incorporated into the system.

b. CAD Incorporation:
During the third year of the project we completed the design, implementation and testing of the modular software set of routines that enable the incorporation of CAD into the tele-mammography system at the remote (sending) sites prior to compressing the images and transmitting the results to the central site.

c. CAD Technical Performance Evaluation:
The system was tested technically using over 100 cases, and after de-bugging, we incorporated the final module into the operations. In the last two years, all transmitted cases were processed by the CAD scheme and could be displayed on the workstation with and without the CAD results at the operator's discretion. The technical performance specifications of the system were not violated due to the incorporation of CAD because the program is faster than the digitization process and is done in parallel to all other tasks while the case (examination) is being processed.

d. CAD Operational and Clinical Use:
The operational use of CAD results was tested using a retrospective clinical review and found acceptable. The clinical aspects of this added feature were evaluated in an observer performance study.

e. Performance Analyses:
The impact of the use of CAD on radiologists' ability to make better interpretations in regard to the need for additional procedures in specific cases was assessed. The result of this effort was a reduction by ~18% (1.26/1.07) in the recommended procedures per "saved" recall as describe under task 3. As a result of this study, all cases transmitted to the central site in task #5 included the CAD results, as well (see below).

Under Task 5, Clinically simulated almost real time transmission and reporting:

Under this task we performed several pilot studies throughout the project in preparation for a simulated clinical study that took place during the fifth year (no cost extension). As indicated, the reason for the delay was the finding that remotely determined recommendations for additional procedures would remain high unless we transmit the prior examinations (when available) together with the current examination of interest. Once the system upgrade was completed, the performance of an "almost real time - high volume" demonstration of the transmission of suspected cases at the remote sites and a clinically simulated response from the central site commenced. During a period of five months when over 4,000 screening examinations were performed at the three remote sites, we asked the technologists to identify and send all cases (with all available related information) they believed would need to be recalled during their QA procedures as they

perform the screening procedure. During four real time experiments and nine simulated real time experiments, radiologists reviewed all the information sent from the remote sites and responded to the site. The real time experiments required that cases were sent simultaneously from all three sites and that during the simulated experiments cases were sent as they became available at each site. Per our protocol, the recommendations of the radiologists were not acted upon but were stored and compared with the actual clinical recommendations for the same examination at the clinic. 353 cases were sent and reviewed by radiologists in this experiment and the results are summarized below.

All readers

		clinical read		
		recall	no recall	total
study	recall	86	81	167
read	no recall	36	150	186
	total	122	231	353

overall agreement	236
prob-observed =	0.6686
prob-expected =	0.5083
Kappa =	0.3259

The details of this effort are being written for a publication at this time but the essence of the results indicated that: 1) The provision of the prior examinations improved performance significantly as compared with our previous studies. 2) At the cost of 81 additional procedures while the women could remain in the remote clinic and assuming the clinical read would ultimately result in the 36 recalls that were not recommended remotely, 86 out of the 122 actual recalls (70.5%) could have been avoided. This finding is important for remote underserved locations and the decision of whether such a practice is acceptable will depend largely on the nature of the practice at the remote site.

During the five observer performance studies alone, we performed a total of 5440 clinically simulated interpretations on the tele-mammography workstation that were each compared with the actual recommendations made during the clinical interpretations of the same examination.

Key Research Accomplishments:

During the last five years, we have been progressing according to the original plan and were able to address a large number of the technical, operational and practice related issues associated with the design, implementation, and clinically simulated testing of the multi-site tele-mammography system. The key accomplishments were:

- We designed, developed, implemented, installed and tested a unique, multi-site tele-mammography system that meets (and in many areas exceeds) the technical specifications we originally anticipated (and proposed).
- We successfully and reliably transmitted over 2,800 examinations from three remote sites to the central site (with minimal down time and technical problems). This set includes 2,432 examinations that were used in the different studies and the remainder were examinations sent (sometimes multiple times) for system testing purposes and deleted after completion of the test.
- We planned and executed a step-by-step comprehensive, technical, and clinical assessment protocol in a clinically simulated environment.
- 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.
- We carried out comprehensive reviews of the practice parameters and performance levels of our radiologists in terms of recall and cancer detection rates with and without the use of CAD.
- We continually upgraded the system as needed with a major software revision in response to radiologists' preferences during the performance of the specific task the tele-mammography system was designed for.
- We successfully reviewed a large number of cases on the workstation and generated a clinically simulated response to the remote sites.
- We completed five observer performance studies to assess both possible utility of the system as well as agreement levels between the technologists and radiologists on suspicious cases.
- We have been able to increase the communication level between technologists and physicians in regard to decision-making processes, and we are engaged in discussions concerning a more extensive use of technologists as physician extenders in several areas.
- We demonstrated that in principle one can perform effectively and efficiently remote management tasks and achieve a significant reduction in actual recall rates, with a relatively limited increase in the number of women who would receive additional procedures during their initial screening visit. This concept can be implemented in a manner that only minimally affects workflow in a busy clinical environment.

Reportable Outcomes:

1) Publications and Presentations

As we developed and tested the system, several reports were generated. Some are directly related to the design implementation and testing of the system and others are related to practice assessment tasks that were performed. The clinically simulated study which was performed during the last year is being analyzed and we are in the process of writing a comprehensive article on this topic. We anticipate submission of this article before the end of the year (2005). Published reports acknowledging this award, to date, Include:

1. 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

2. 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

3. 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

4. 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; 9:1245-1250

5. 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. Presented at the Department of Defense "Era of Hope" meeting, September 25, 2002.

6. Leader JK, Wallace LP, Hakim CM, Hertzberg TM, Hardesty LA, Sumkin JH, Cohen C, Sneddon C, Lindeman S, Craig D, and Drescher JM. Preliminary clinical evaluation of a multi-site telemammography system in a screening mammography environment. Proc SPIE 2003; 5033:273-280.

7. Zheng B, Wang XH, Wallace L, Cohen C, Hardesty LA, Hakim CM, Abrams G, Sumkin J, Gur D. Improving CAD performance in detecting masses depicted on prior images. Proc SPIE 2003; 5032:215-221

8. Drescher JM, Maitz GS, Traylor C, Leader JK, Clearfield RJ, Shah R, Ganott MA, Pugliese F, Duffner D, Lockhart J, and Gur D. A multi-site telemammography system: preliminary assessment of technical and operational issues. Proc SPIE 2003; 5033:360-369.

9. Zheng B, Hardesty LA, Poller WR, Sumkin JH, Golla S. Mammography with computer-aided detection: reproducibility assessment – initial experience, Radiology 2003; 228:58-62.

10. Zheng B, Good WF, Armfield DR, Cohen C, Hertzberg T, Sumkin JH, Gur D. Performance change of mammographic CAD schemes optimized with most-recent and prior image database, Acad Radiol 2003; 10:283-288.

11. Leader JK, Sumkin JH, Ganott MA, Hakim C, Hardesty L, Shah R, Wallace L, Klym A, Drescher JM, Maitz GS, Gur D. Subjective assessment of high-level image compression of digitized mammograms. Proc SPIE 2004; 5372:415-422.

12. Zheng B, Leader JK, Abrams G, Shindel B, Catullo V, Good WF, Gur D. Computer-aided detection schemes: The effect of limiting the number of cued regions in each case, AJR 2004; 182:579-583.

13. Gur D, Sumkin JH, Rockette HE, Ganott M, Hakim CM, Hardesty L, Poller WR, Shah R, Wallace L. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system, JNCI 2004; 96:185-190.

14. Gur D, Sumkin JH, Hardesty LA, Clearfield RJ, Cohen CS, Ganott MA, Hakim CM, Harris KM, Poller WR, Shah R, Wallace LP, Rockette HE. Recall and detection rates in screening mammography. A review of clinical experience – implications for practice guidelines. Cancer 2004; 100(8):1590-1594.

15. Gur D, Stalder JS, Hardesty LA, Zheng B, Sumkin JH, Chough DM, Shindel BE, Rockette HE, Computer-aided detection performance in mammographic examination of masses: assessment, Radiology 2004; 233:418-423.

16. Klym AH, King JL, Hardesty LA. The Effect of Routine Use of CAD System on the Practice of Breast Imagers – A Subjective Assessment. Academic Radiology 2004;11:711-713.

17. Leader JK, Chough D, Clearfield RJ, Ganott MA, Hakim C, Hardesty L, Shindel B, Sumkin JH, Drescher JM, Maitz GS, Gur D. Teleradiology and screening mammography: a telemammography system evaluation and comparison to clinical results. Proc SPIE 2005; 5749:401-412.

18. Leader JK, Clearfield RJ, Ganott MA, Hakim C, Hardesty L, Sumkin JH, Wallace L, Drescher JM, Maitz GS, Gur D. Multi–Site Telemammography System for Remote Patient Management in Screening Mammography. Era of Hope 2005 Department of Defense Breast Cancer Research Program Meeting, Philadelphia PA, June 2005.

2) <u>Other reportable outcomes:</u>

The effort supported by this project helped us generate preliminary data and perhaps more important several concepts that were used in support of three grant applications that are currently funded:

1. "Rule Based CAD of Digitized Mammograms", PI: David Gur, source: NIH, grant # CA077850

2. "Interactive CAD for Mammography", PI: Bin Zheng, source: NIH, grant # CA101733

3. "The Laboratory Effect in Breast Cancer Detection Studies", PI: David Gur, source: NIH, grant # EB003503

<u>Personnel receiving pay from this effort:</u>

David Gur, Sc.D., Joseph K. Leader, Ph.D., Glenn S. Maitz, M.S., Yuan-Hsiang Chang, Ph.D., Howard E. Rockette, Ph.D., Jules Sumkin, D.O., Xiao Hui Wang, Ph.D., Bin Zheng, Ph.D., John M. Drescher, B.S., Amy H. Klym, B.S., Jennifer S. Stalder, B.S., Christopher Traylor

All Radiologists and Technologists participating in the observer performance studies of this effort were not paid directly from the grant. Payments were made to the Department of Radiology for the services rendered by the Radiologists and Technologists.

Last, one of our investigators (Chang Y-H) in the first three years of the project has returned to Taiwan where he is employed as a faculty in the department of Electrical Engineering. Two other investigators (Drs. Joseph Leader and Xiao Hui Wang) were promoted during the project duration to Research Assistant Professors of Radiology.

Conclusions:

A comprehensive multi-task, multi-discipline applied project that involved a large team of investigators, physicians and staff was successfully executed and completed. We undertook a large number of technical and application-based tasks associated with the design, implementation, and clinically simulated evaluations of a multi-site tele-mammography system. We modified the system as needed and exceeded several of the performance goals we originally proposed. The concept of remote management of screening practices where a physician is not present was tested using a comprehensive step-by-step evaluation and was proven to be feasible which could result in improved communication between technologists and physicians at the remote and central sites. Our main observation to date is that the general concept is sound and the actual implementation resulted in an appreciation for the importance of the "comfort level" of the team (physicians and technologists) in operating and using such a system for the stated purpose. Most important perhaps is the demonstration that in principle, using this (or a similar) approach, one could achieve a significant reduction in actual recall rates for a second visit. At this time, it can only be done at some cost namely an increase in the number of women who would receive additional procedures (e.g., views) during their initial screening visit. Last, we have improved substantially our understanding of several extremely important issues related to the general practice of screening mammography (e.g. the relationship between recall rates and cancer detection rates), and the use of CAD in particular. These may have far reaching implications on this field.

So What?

The issues associated with efficient and efficacious mammographic screening in general and in remote underserved locations in particular are significant. The main goal of this project was to evaluate how the use of an "almost real-time" tele-mammography system (with or without the use of relevant information) may impact the diagnostic process in terms of complete cycle time and patients' recall rate. Our success in this project has already changed substantially our own thinking about practice issues in remote sites and we hope others will follow. We demonstrated different ways to increase communication between remote (and potentially underserved) sites and a central site. Our hope is that by using the concepts we investigated, one may be able to provide better, more timely and cost-effective

service at these sites and in the process, substantially reduce actual recall rates in remote facilities where a physician is not present. Despite significant advances in our understanding of the issues and alternatives surrounding "optimal" practices of screening mammography, many of our current clinical practice guidelines are based on limited subjective assessments and anecdotal experiences, and a significant fraction is related to operational matters in busy urban environments that are staffed by experienced radiologists. The area of optimizing remote, underserved practices has been studied only in a cursory manner. Our project is but one attempt to improve our understanding of the technical, operational, and clinical issues facing these facilities and implementing technology-based solutions that may help them provide a better service to the populations they serve. Our own institution is basing our transition strategy to a digital environment in screening mammography partially based on the observations made during this project (albeit using a PACS enabled remote management rather than tele-mammography) and we believe others should consider this or a similar approach, as well.

Background 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. Wu M, Zheng Y, North M, Pisano E. NLM tele-educational application for radiologists to interpret mammography. Proc AMIA Symposium, 2002, pg 909-913
23. Sheybani EO, Sankar R. ATMTN: a telemammography network architecture. IEEE Trans Biomed Eng 2002; 49:1438-1443
24. 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).
25. 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).

Appendices

See Attached.

APPENDICES

1. 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

2. 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

3. 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

4. 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; 9:1245-1250

5. Leader JK, Wallace LP, Hakim CM, Hertzberg TM, Hardesty LA, Sumkin JH, Cohen C, Sneddon C, Lindeman S, Craig D, and Drescher JM. Preliminary clinical evaluation of a multi-site telemammography system in a screening mammography environment. Proc SPIE 2003; 5033:273-280.

6. Zheng B, Wang XH, Wallace L, Cohen C, Hardesty LA, Hakim CM, Abrams G, Sumkin J, Gur D. Improving CAD performance in detecting masses depicted on prior images. Proc SPIE 2003; 5032:215-221

7. Drescher JM, Maitz GS, Traylor C, Leader JK, Clearfield RJ, Shah R, Ganott MA, Pugliese F, Duffner D, Lockhart J, and Gur D. A multi-site telemammography system: preliminary assessment of technical and operational issues. Proc SPIE 2003; 5033:360-369.

8. Zheng B, Hardesty LA, Poller WR, Sumkin JH, Golla S. Mammography with computer-aided detection: reproducibility assessment – initial experience, Radiology 2003; 228:58-62.

9. Zheng B, Good WF, Armfield DR, Cohen C, Hertzberg T, Sumkin JH, Gur D. Performance change of mammographic CAD schemes optimized with most-recent and prior image database, Acad Radiol 2003; 10:283-288.

10. Leader JK, Sumkin JH, Ganott MA, Hakim C, Hardesty L, Shah R, Wallace L, Klym A, Drescher JM, Maitz GS, Gur D. Subjective assessment of high-level image compression of digitized mammograms. Proc SPIE 2004; 5372:415-422.

11. Zheng B, Leader JK, Abrams G, Shindel B, Catullo V, Good WF, Gur D. Computer-aided detection schemes: The effect of limiting the number of cued regions in each case, AJR 2004; 182:579-583.

12. Gur D, Sumkin JH, Rockette HE, Ganott M, Hakim CM, Hardesty L, Poller WR, Shah R, Wallace L. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system, JNCI 2004; 96:185-190.

13. Gur D, Sumkin JH, Hardesty LA, Clearfield RJ, Cohen CS, Ganott MA, Hakim CM, Harris KM, Poller WR, Shah R, Wallace LP, Rockette HE. Recall and detection rates in screening mammography. A review of clinical experience – implications for practice guidelines. Cancer 2004; 100(8):1590-1594.

14. Gur D, Stalder JS, Hardesty LA, Zheng B, Sumkin JH, Chough DM, Shindel BE, Rockette HE, Computer-aided detection performance in mammographic examination of masses: assessment, Radiology 2004; 233:418-423.

15. Klym AH, King JL, Hardesty LA. The Effect of Routine Use of CAD System on the Practice of Breast Imagers – A Subjective Assessment. Academic Radiology 2004;11:711-713.

16. Leader JK, Chbugh D, Clearfield RJ, Ganott MA, Hakim C, Hardesty L, Shindel B, Sumkin JH, Drescher JM, Maitz GS, Gur D. Teleradiology and screening mammography: a telemammography system evaluation and comparison to clinical results. Proc SPIE 2005; 5749:401-412.

Breast Imaging

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

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

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

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

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

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

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

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

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

[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.

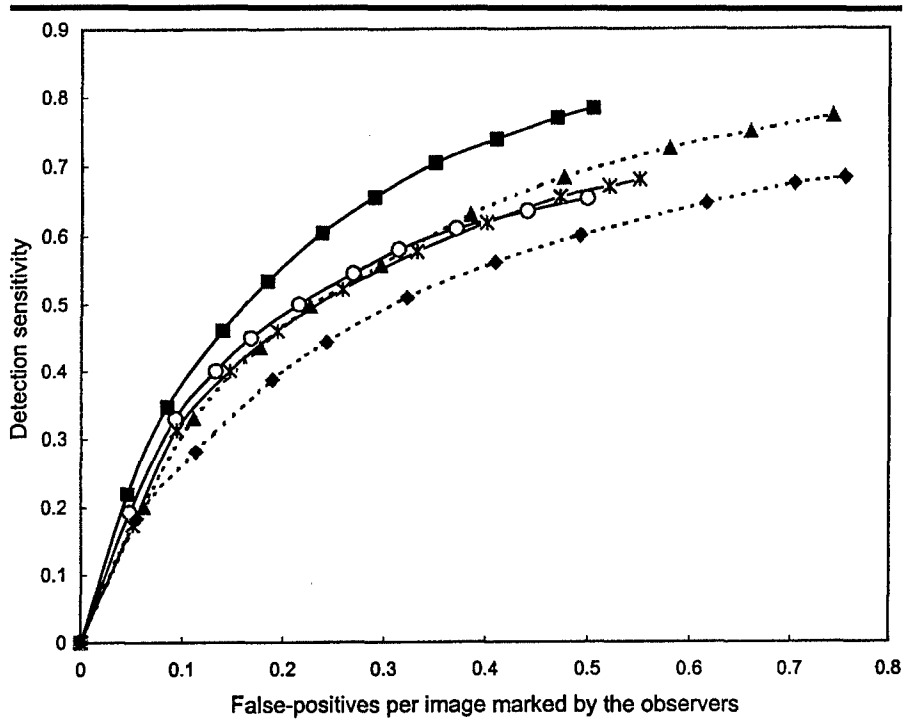

Figure 1. Free-response receiver operating characteristic curves for the average detection of mammographic abnormalities (including both masses and microcalcification clusters) by seven participating radiologists using five display modes. ○ = mode 1, ■ = mode 2, ▲ = mode 3, * = mode 4, and ◆ = mode 5.

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 ($P < .01$).

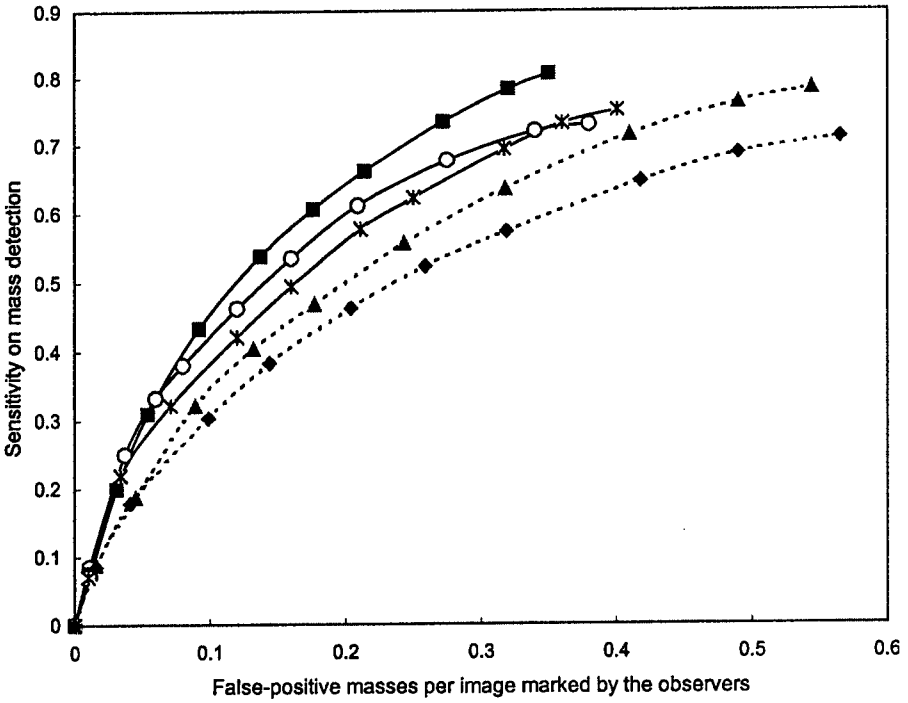

Figure 2. Free-response receiver operating characteristic curves for the average mass detection by seven radiologists using five display modes. ○ = mode 1, ■ = mode 2, ▲ = mode 3, * = mode 4, and ◆ = mode 5.

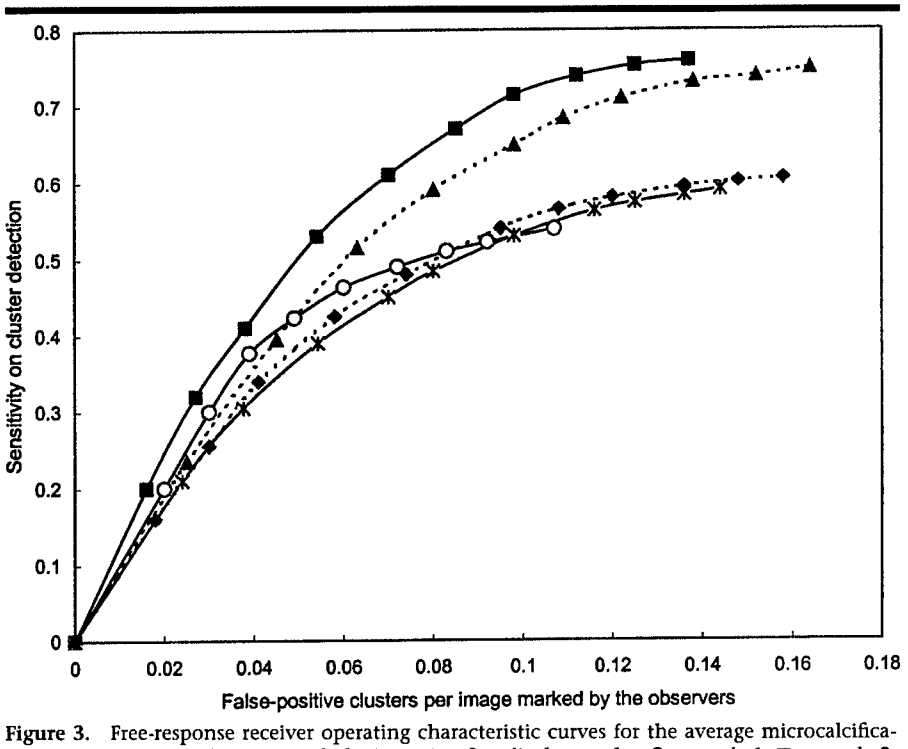

Figure 3. Free-response receiver operating characteristic curves for the average microcalcification cluster detection by seven radiologists using five display modes. ○ = mode 1, ■ = mode 2, ▲ = mode 3, * = mode 4, and ◆ = mode 5.

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.

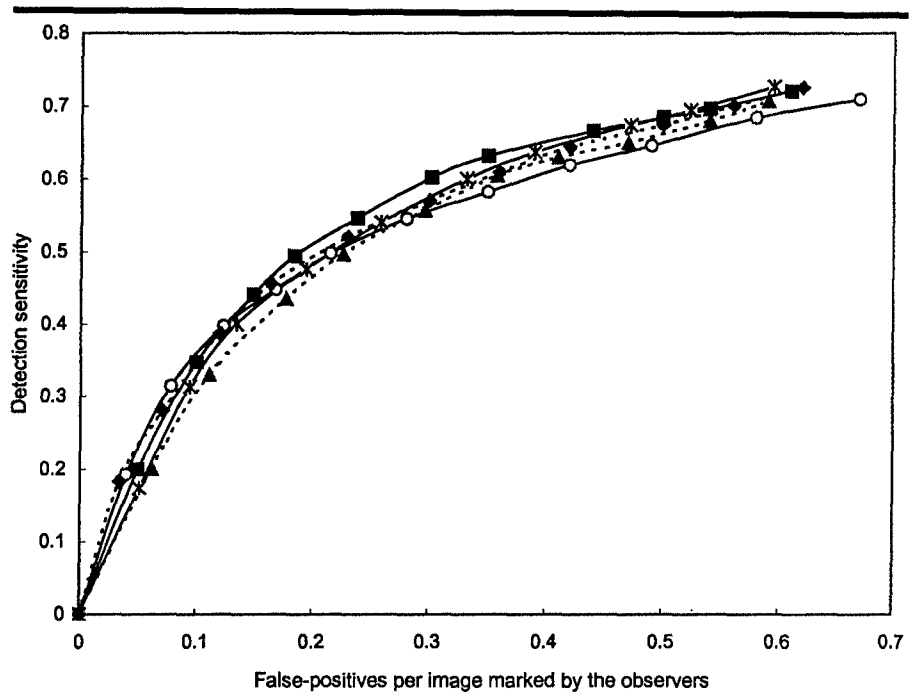

Figure 4. Free-response receiver operating characteristic curves for the average detection of abnormalities by seven radiologists as a function of the order of appearance: ○ = first time, ■ = second time, ▲ = third time, * = fourth time, and ◆ = fifth time, regardless of the reading mode.

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.

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. © *2001 American Association of Physicists in Medicine.*
[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.

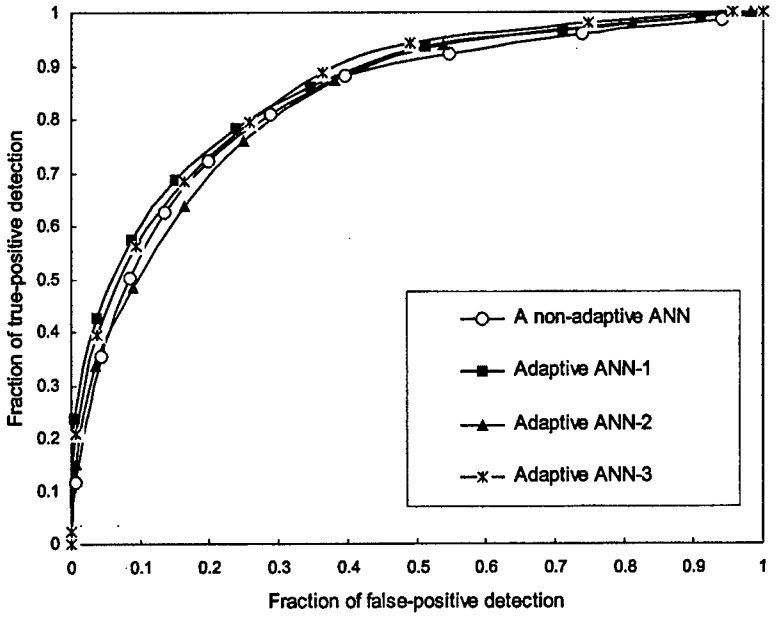

FIG. 1. ROC curves from nonadaptive ANN-1 and three sets of noncombined adaptive ANNs. The A_z values for these curves are 0.82, 0.84, 0.83, and 0.84, respectively.

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

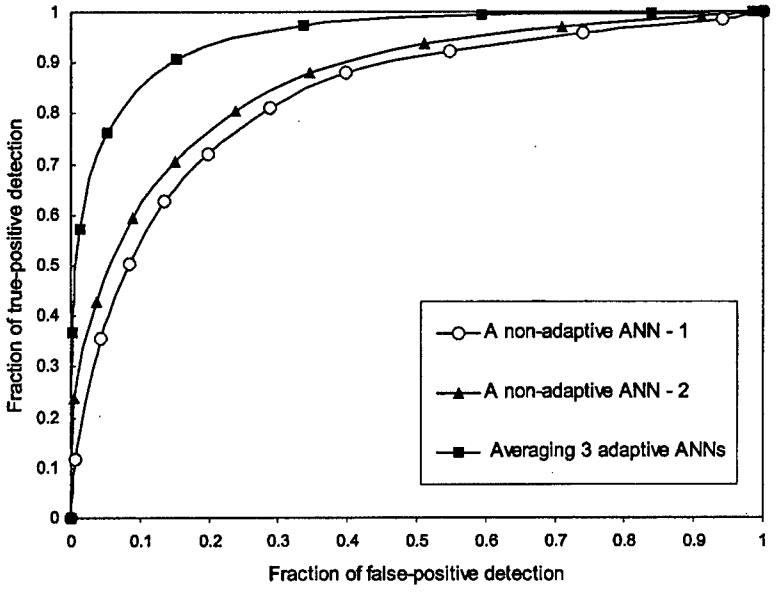

FIG. 2. ROC curves of classification results from nonadaptive schemes (ANN-1 and ANN-2) as well as after averaging scores of three sets of adaptive ANNs. The A_z values are 0.82 ± 0.01, 0.85 ± 0.01, and 0.95 ± 0.01, respectively.

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).

APPENDIX 3

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.

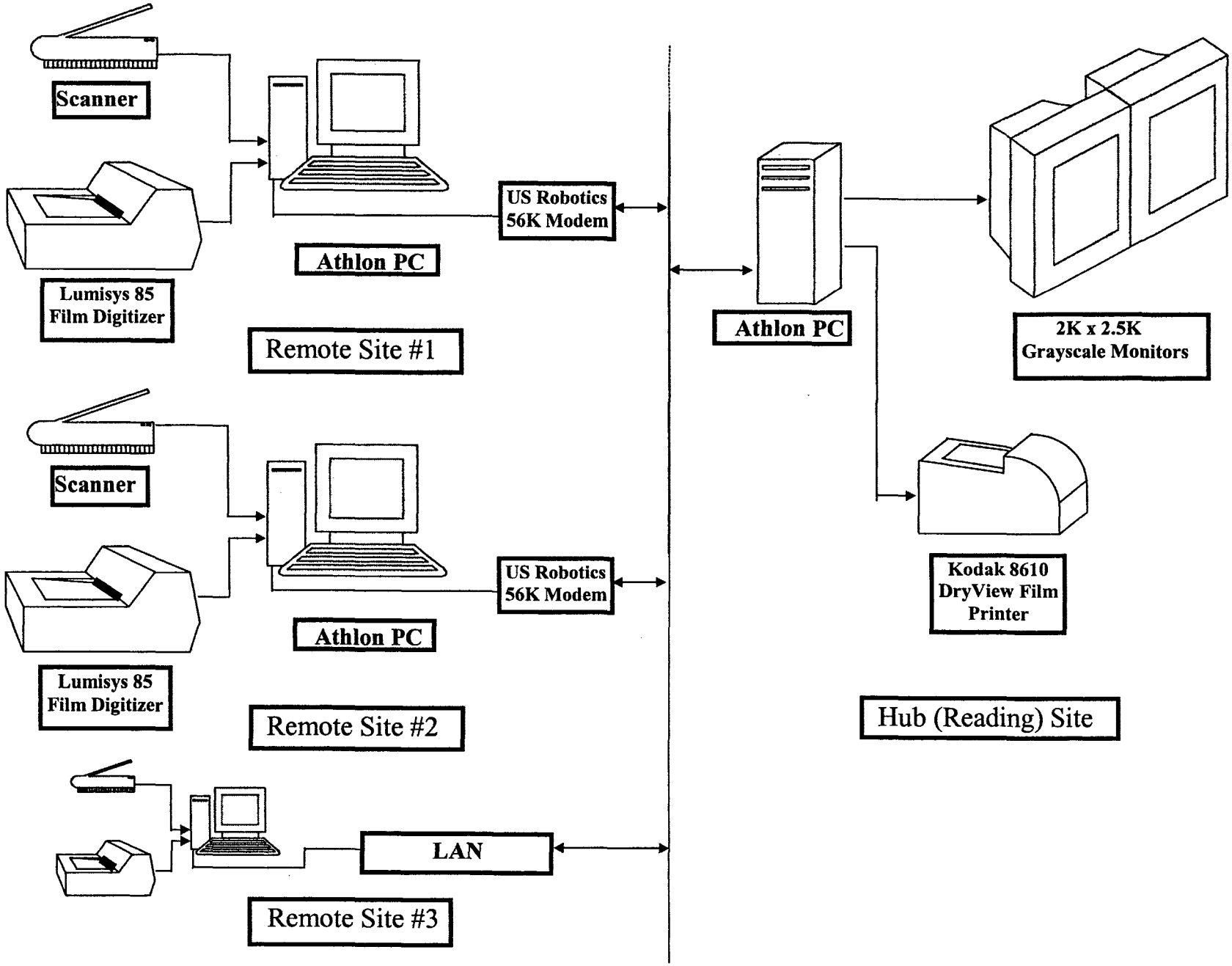

Figure 1: Schematic of the Telemammography System

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).

APPENDIX 4

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

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

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. The latest prior examinations had been initially interpreted as negative or 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 detect a substantial 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.

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)

Acad Radiol 2002; 9:1245–1250

[1] From the Department of Radiology, Ste 4200, 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.

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

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 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 (craniocaudal [CC] and mediolateral oblique [MLO]), 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 (Eastman Kodak, Rochester, NY)

Table 1
Distribution of Selected Masses and Microcalcification Clusters

	All Cases			Malignant Cases		
Type of Abnormality	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

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 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).

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 (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.

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-proven 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 de-

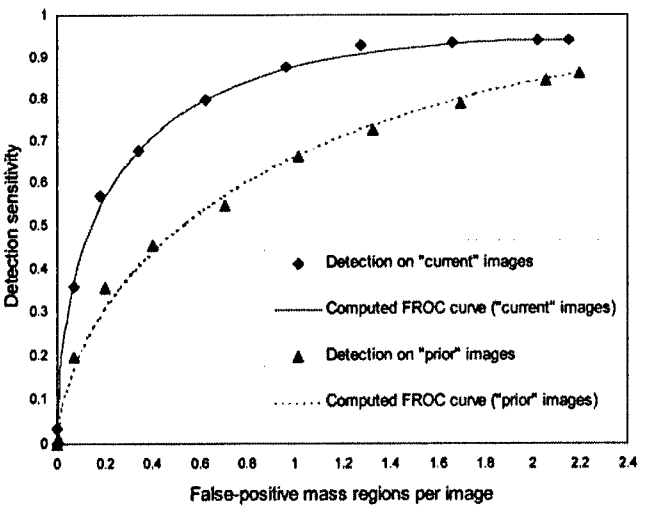

Figure 1. Comparison of case-based CAD performance for detection of masses on 200 current and prior mammographic cases. The test set included 172 masses.

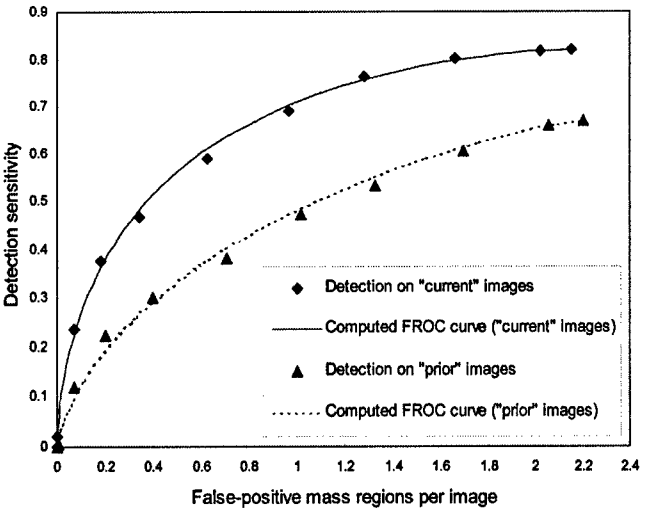

Figure 3. Comparison of region-based CAD performance for detection of masses on current and prior images. The test set included 336 mass regions.

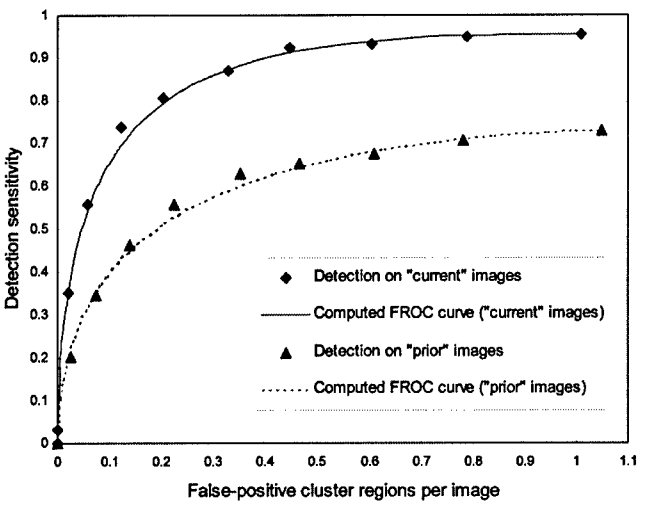

Figure 2. Comparison of case-based CAD performance for detection of microcalcification clusters on 200 current and prior mammographic cases. The test set included 128 true-positive clusters.

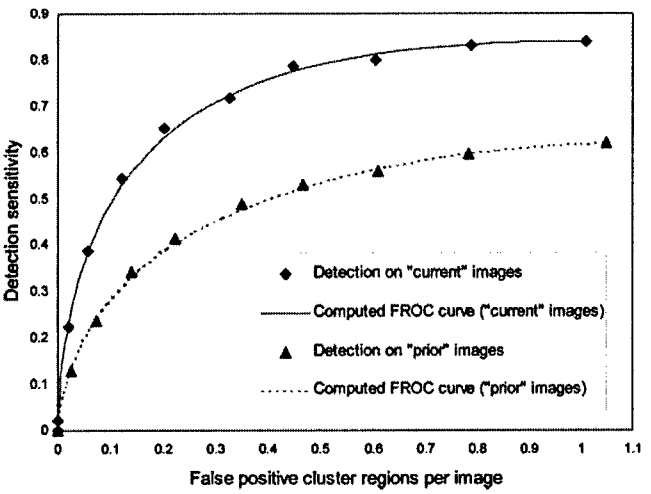

Figure 4. Comparison of region-based CAD performance for detection of microcalcification clusters on current and prior images. The test set included 248 cluster regions.

tections: 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, 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 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"). 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

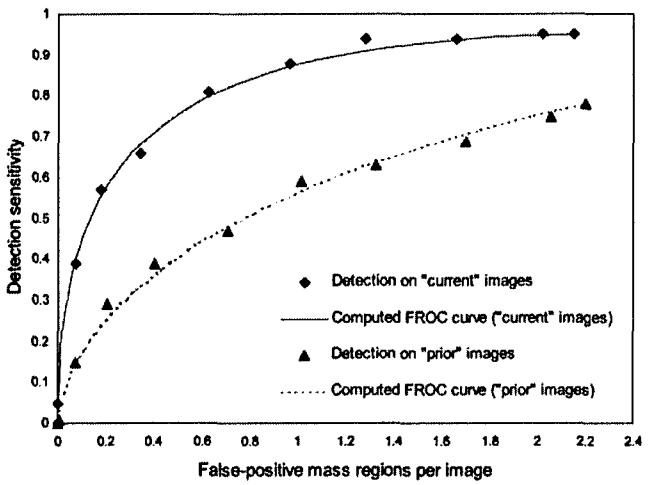

Figure 5. Comparison of case-based CAD performance for malignant mass detection. The test set included 100 malignant masses.

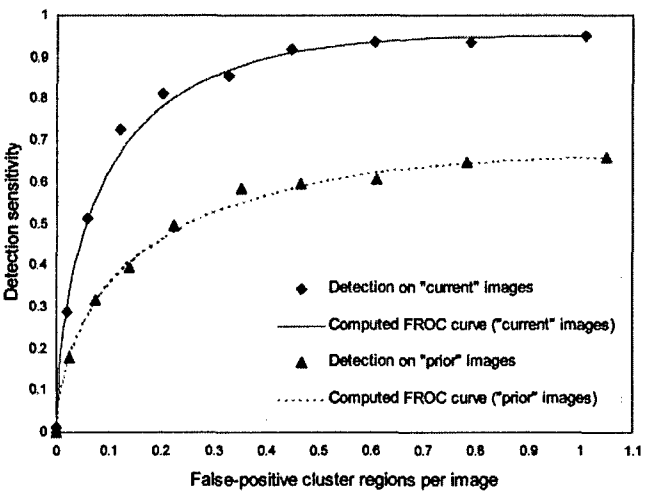

Figure 6. Comparison of case-based CAD performance for malignant microcalcification cluster detection. The test set included 79 malignant clusters.

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).

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). 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 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) 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 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

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.

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.

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.
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.

APPENDIX 5

PROGRESS IN BIOMEDICAL OPTICS AND IMAGING

SPIE—The International Society for Optical Engineering

Preliminary clinical evaluation of a multi-site telemammography system in a screening mammography environment

J. K. Leader, L. P. Wallace, C. M. Hakim, T. M. Hertzberg, L. A. Hardesty, J. H. Sumkin, C. Cohen, C. Sneddon, S. Lindeman, D. Craig, J. M. Drescher

Reprinted from

Medical Imaging 2003

PACS and Integrated Medical Information Systems: Design and Evaluation

SPIE P PROCEEDINGS SERIES

Volume 5033

18–20 February 2003
San Diego, California, USA

Preliminary clinical evaluation of a multi-site telemammography system in a screening mammography environment

J. Ken Leader[a], Luisa P. Wallace[ab], Christiane M. Hakim[ab], Todd M. Hertzberg[ab], Lara A. Hardesty[ab], Jules H Sumkin[ab], Cathy Cohen[ab], Colleen Sneddon[b], Shirley Lindeman[b], Deborah Craig[c], John M. Drescher[a]
[a]University of Pittsburgh, PA USA 15213
[b]Magee-Womens Hospital, Pittsburgh, PA USA 15213
[c]Lee Regional Hospital, Johnstown, PA USA 15904

ABSTRACT

We evaluated a telemammography system for reviewing and rating screening mammography in a clinical setting. Three remote sites transmitted 306 exams to a central site. Films were digitized at 50 micron pixel dimensions and compressed at a 50:1 ratio. At the central site images were displayed on a workstation with two high-resolution monitors. Five radiologists reviewed and rated the screens without the availability of prior images or additional information indicating: 1) if additional procedures were needed, 2) which breast was involved, and 3) when appropriate, the recommended additional procedures. During the actual clinical interpretation 13.7% (42 cases) of the patients were recalled for additional procedures. During the retrospective review radiologists 1, 2, 3, 4, and 5 recommended additional procedures for 26.1%, 29.1%, 36.3%, 45.1%, and 54.2% of the cases, respectively. The agreements between the clinical interpretation and radiologists 1, 2, 3, 4, and 5 were 77.8%, 76.1%, 69.0%, 62.7%, and 53.6%, respectively. The exceedingly high percentage of recommended additional procedures using the workstation was attributed to lack of prior images or additional information, the knowledge that case management was not affected, and the observers' expectation for an enriched case mix.

Keywords: Teleradiology, human performance, recall rate, breast cancer screening, mammography.

1. INTRODUCTION

Teleradiology can challenge typical radiology practices in areas ranging from personnel assignments to data management. In remote or underserved clinics in may be necessary to evaluate personnel qualifications in regards to deciding if teleradiology is appropriate and the necessary radiographic procedures.[1-3] Many teleradiology systems employ image processing techniques to manage the digital image data in terms of data acquistion,[4-9] transmission time (e.g., compression,[4,10,11,12] cropping,[13] image selection[14]), and image display.[7,8,10,11,13,14,15] The effects of data management techniques on diagnostic image quality are application specific. Comparisons between film-based and digitized image-based (film digitization) diagnostic radiographic interpretation have produced mixed results. In some laboratory studies the area under the receiver operating characteristic (ROC) curve, sensitivity, and accuracy have been shown to be slightly greater for film-based interpretation,[4,7,16,17] but the differences were generally not statistically significant. Reported specificity has been relatively equivalent for the two interpretation methods.[4,7,16,17]

The high-spatial resolution necessary to interpret mammographic images presents unique challenges when designing and implementing a telemammography system. Improvements in image quality of x-ray film mammography have been associated with improvements in breast cancer detection.[18-23] Therefore, it is important that the image processing techniques of a telemammography system do not degrade the diagnostic image quality of the digital (full-field digital mammography (FFDM)) or digitized (film digitization) mammographic images.

* jklst3@pitt.edu; phone (412) 641-2572; fax (412) 641-2582, University of Pittsburgh, Magee-Womens Hospital, 300 Halket Street., Suite 4200, Pittsburgh, PA 15213

Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation,
H. K. Huang, Osman M. Ratib, Editors, Proceedings of SPIE Vol. 5033 (2003)

Mammography interpretation has been reported as relatively equivalent for film mammography and digitized mammographic images. Fajardo et al.[24] (1990) found film mammography statistically superior for detecting skin and nipple abnormalities compared to digitized mammography in an ROC study, but found the two methods equivalent for detecting microcalcifications and masses. An ROC study performed by Nab et al.[25] (1992) found that the diagnostic performance of film and digitized mammography were comparable. Powell et al.[26] (1999) reported that film mammography was slightly superior to digitized mammography in several diagnostic measures (i.e., accuracy, false-positive rates, and callback rates for mammograms with normal and malignant findings), but only the callback rates for normal findings were statistically different. The callback rates for benign findings were slightly better for digitized mammography. A follow-up study by Powell et al.[27] (2000) compared film mammography to wavelet-compressed digitized mammographic images. The only statistically significant finding was that the false positive rate was lower for compressed digitized images compared to film mammography. Compressed digitized images were also slightly better (though not statistically) in terms of callback for mammograms with normal and benign findings. Film mammography was slightly better (though not statistically) for callback rates for depicting malignant abnormalities.

This manuscript presents a preliminary, retrospective clinical evaluation of an inexpensive, high-quality, multi-site telemammography system[28,29] for the review of screening mammography examinations. The study was designed to assess the effectiveness of the system for the review of breast cancer screening mammography with the objective to assess its possible use in determining the need for additional procedures (rather than primary diagnosis). The limited retrospective review was conducted using only digitized mammographic images without the benefit of prior images or any additional information. Five radiologists reviewed and rated screening exams using the telemammography system, and their results were compared to the actual clinical interpretations of the same cases regarding the need for additional procedures. It was anticipated that in this experimental protocol the number of cases recommended for additional procedures would be greater during the limited telemammography review compared to the clinical interpretation.

2. METHODS

2.1 Case selection

The 306 cases retrospectively evaluated in this study originated from patients who underwent breast cancer screening mammography at three woman's imaging centers. The mammography technologists at these centers were instructed to select an approximately equal number of cases they (the technologists) believed may and may not need additional imaging procedures for complete evaluations. Cases were selected by the technologists in a prospective mode and they did not know at the time of selection whether or not the patient would actually be recalled for additional procedures during the clinical interpretation. The mean patient age was 53.8 years ranging from 35 to 88 years old. The actual, subsequent clinical interpretation categorized each case using the Breast Imaging Reporting and Data System (BIRADS) (Table 1). The four routine screening mammographic films of the left and right craniocaudal views (LCC & RCC), and left and right mediolateral oblique views (LMLO & RMLO) were used to review and rate cases in this study.

Table 1
Distribution of BIRADS categories as a result of clinical interpretation of the cases

BIRADS Category	0	1	2	total
Number of cases	42	206	58	306

2.2 Telemammography system

The cases for this study were transmitted from the three centers (remote sites) to Magee-Womens Hospital, Pittsburgh, PA, USA (central site) using an inexpensive, high-quality, multi-site telemammography system. The operation of the system including digitization the mammographic films, digital image processing, data transmission, and image display were conducted under routine operating procedures and are described in detail by Drescher et al.[29] (2003). A brief description, as relevant to this study is provided below.

2.2.1 Central and remotes sites

The central site telemammography workstation is connected to two high-resolution (2048 x 2560) 8-bit grayscale portrait monitors at a nominal setting of 80 ftL (DS5100P, Clinton Electronics, Rockford, IL, USA). A dual 1.2 GHz

multi-processor (Athlon MP, Advanced Micro Device, Sunnyvale CA, USA) with 2 GB of RAM powers the workstation which operating under Microsoft Windows 2000 Server (Microsoft Corporation, Redmond, WA, USA). The workstation is equipped with 56K hardware modems (U.S. Robotics, Rolling Meadows, IL, USA) and an ethernet network cards (OfficeConnect 10/100 NIC, 3COM, Santa Clara, CA, USA) for communication with the remote sites.

The computers at the remote sites operate under Microsoft Windows 2000 Workstation powered by a 900MHz processor (Athlon 900, Advanced Micro Device, Sunnyvale CA, USA) with 512 MB of RAM. The mammographic films are digitized using a high-resolution, laser film digitizers (Lumiscan 85, Eastman Kodak, Rochester, NY, USA) at 50 micron pixel dimensions and 12-bit grayscale. Data communication from the remote site computers is conducted via 56K hardware modems and ethernet network cards (Integrated PRO/100 S Desktop Adapter, Intel Corporation, Santa Clara, CA, USA). Sites 1 and 2 are 15 and 20 miles from the central site, respectively, and transmit data across Plain Old Telephone System (POTS) lines. Site 3 is 90 miles from the central site and transmits data across a Local Area Network (LAN).

2.2.2 Image processing

The first image processing step was to perform an automated cropping that removed the non-tissue area surrounding the breast. Next, the image data were compressed using the irreversible (lossy), 9/7 transform, wavelet-based JPEG 2000 method at a 50:1 compression ratio. Prior to transmission from the remote sites, the data packets were encrypted using strong 128 bit Microsoft Point-to-Point Encryption (MPPE) with Microsoft Challenge Handshake Authenticate Protocol (CHAP) version 2.

Upon arrival to the central site the image data were decrypted and decompressed. The decompressed images data were minimally unsharp masked to enhance display on the workstation monitors. The image data range was maximized for display by re-scaling the image data from 0 to 4095. To facilitate image viewing default look-up table (LUT) values were automatically calculated based on the typically bimodal pixel value distribution (histogram). The images were restored to full height, but not the full width, by padding (filling) prior to image display.

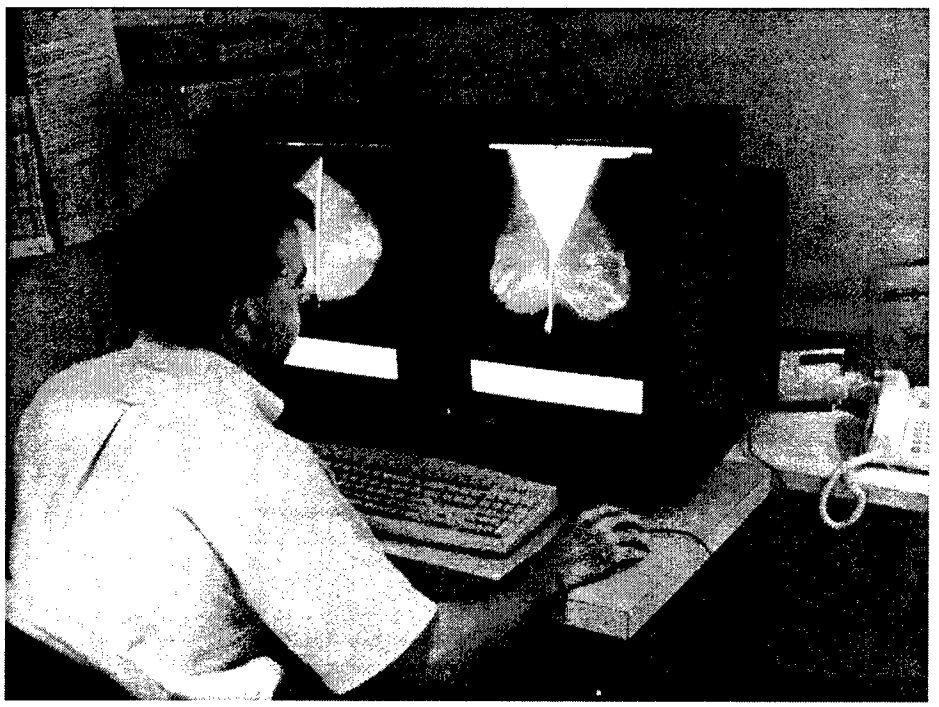

Fig. 1. Telemammography workstation at the central site pictured in the default image display format.

2.2.3 Central site image display

There are several mouse-driven image display features on the central site workstation available to the user during case review. Image display formats possible included: one image/monitor, two images/ monitor, or four images/monitor. To duplicate our standard film presentation LCC and RCC are displayed on the left monitor, and LMLO and RMLO on the right monitor as the default presentation (Fig. 1).

The typical display resolution was approximately 100 micron pixel dimensions for one image/monitor and 200 micron pixel dimensions for two images/monitor. Images can be magnified by a free-moving magnification box or quadrant panning. The magnification box size varied dependent on the image display format; for one image/monitor the box was 511 x 566 pixels and for two images/monitor the box was 204 x 266 pixels. The LUT settings could be adjusted by the user by moving the mouse horizontally or vertically. Selected LUT settings could be applied (at user's option) to all images associated with the case and could be reset to the default (automated) values at any point.

2.3 Reviewing and rating cases

Five experienced radiologists (each reading over 2000 mammograms per year) reviewed and rated each case on the telemammography workstation. Cases were randomly presented in each session. The rating form for each case was presented on the workstation monitors and completed using the computer mouse (Fig. 2). The computerized scoring form recorded: (1) if additional procedures were indicated, (2) use of prior images (disabled for this study), (3) which breast was involved, and (4) when appropriate, the specific recommended procedure. The radiologists' reviews were conducted based entirely on the four mammographic views (LCC, RCC, LMLO, & RMLO), without additional, potentially relevant information (e.g., prior images, prior reports, patient history). The radiologists were informed of the case origination, but not the case selection criteria. The written instructions to observers regarding case review were:

> In this phase of testing our telemammography system, we would like you to review cases and take a few seconds to *quickly* decide whether or not the case should be recalled for additional procedures. These cases are routine screening mammograms. You will fill out a computer form to indicate if a case should be recalled. If you choose to recall the case you must check off which additional procedures you would recommend for each breast. A "done" button on the bottom of the form will bring up the next case. The computer will automatically track the cases that you have completed and load your remaining cases; the count will be in the bottom of the right screen.

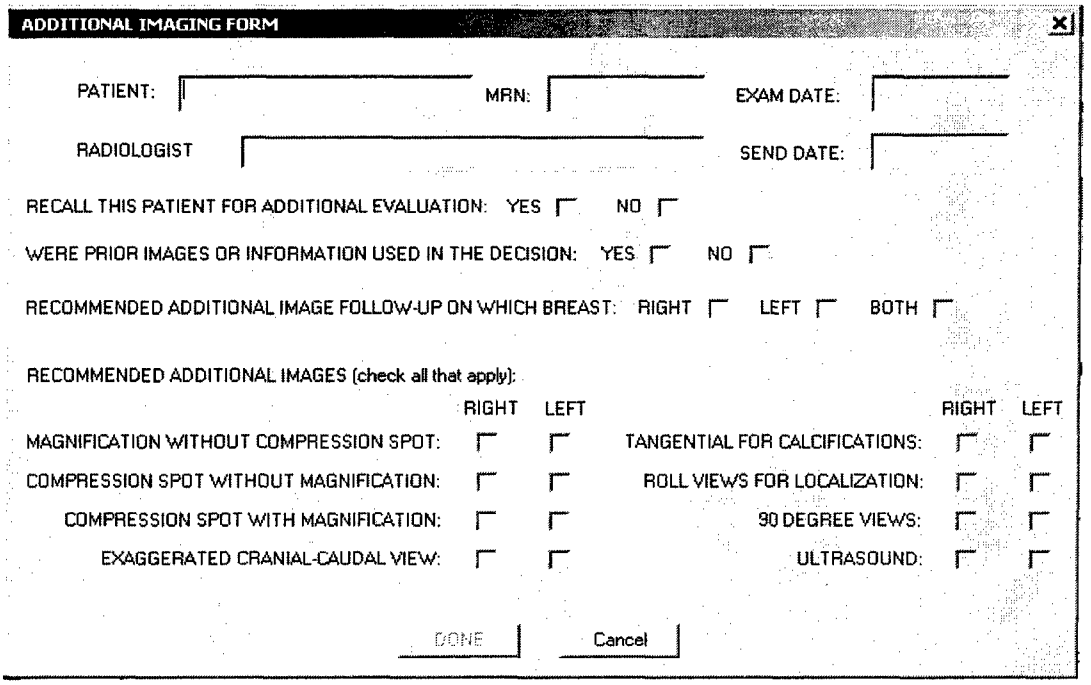

Fig. 2. Computer scoring form complete by the radiologists for each case.

2.4 Data analysis

The radiologists' recommendations using the telemammography workstation were compared with the actual clinical interpretation during the original clinical review. The comparisons were done using agreement/disagreement measures. The disagreements when clinical interpretation indicated no-recall and telemammography interpretation indicated recall were further evaluated based on the actual BIRADS ratings during the clinical interpretation.

3. RESULTS

Image quality, effects of the image processing, and features of the multi-site telemammography system were subjectively reported as more than adequate for reviewing screening mammography examinations and generally were well-received by the radiologists. The cropped images retained all breast tissue areas and were visibly appealing for image review. The automated LUT settings were normally acceptable and were changed in approximately 10% of the cases during review. Magnification allowed detailed review of the breast tissue patterns, particularly microcalcifications. Although there were some detectable differences at extremely high magnifications between non-compressed and compressed images at a 50:1 compression ratio, the images were subjectively judged to "not affect the diagnostic quality."

The preliminary assessment of the limited case review (i.e., no prior images, prior reports, or patient history) of screening exams using the multi-site telemammography system resulted in an exceedingly high recommended recall rates and modest agreement between the actual clinical interpretation and the radiologists' recommendations using the telemammography system. During the actual clinical interpretation 13.7% (42) of the cases were recalled (BIRADS = 0). Radiologists 1, 2, 3, 4, and 5 recall rates were 26.1% (80), 29.1% (89), 36.3% (111), 45.8% (138), and 54.2% (166), respectively, when using the telemammography system to determine the need for additional procedures (Table 2). The overall agreement between the clinical interpretation and the recommendations of radiologists 1, 2, 3, 4, and 5 were 77.8%, 76.1%, 69.0%, 62.7%, and 53.6%, respectively. Kappa for radiologists 1, 2, 3, 4, and 5 were 0.32, 0.32, 0.22, 0.20, and 0.13, respectively.

Table 2
Reviewing and rating screening mammography exams, telemammography workstation recommendations versus clinical interpretation

Telemammography recommendations	Clinical interpretation recall (n = 42)	no-recall (n = 264)	Total
Radiologist 1			
recall	8.8% (27)	17.3% (53)	26.1% (80)
no-recall	4.9% (15)	69.0% (211)	73.9% (226)
Radiologist 2			
recall	9.5% (29)	19.6% (60)	29.1% (89)
no-recall	4.2% (13)	66.7% (204)	70.9% (217)
Radiologist 3			
recall	9.5% (29)	26.8% (82)	36.3% (111)
no-recall	4.2% (13)	59.5% (182)	63.7% (195)
Radiologist 4			
recall	10.8% (33)	34.3% (105)	45.1% (138)
no-recall	2.9% (9)	52.0% (159)	54.9% (168)
Radiologist 5			
recall	10.8% (33)	43.5% (133)	54.2% (166)
no-recall	2.9% (9)	42.8% (131)	45.8% (140)

The cases when the recommendation using the telemammography system was "recall" and the clinical interpretation indicated "no-recall" represented a large percentage of the disagreement, and nearly one half had some type of findings reported during the clinical review. The disagreement when the clinical interpretation indicated "no-recall" and the telemammography indicated "recall" accounted for 77.9%, 82.2%, 86.3%, 92.1%, and 93.7% of the total disagreement for radiologists 1, 2, 3, 4, and 5, respectively (Table 2). Further evaluation of these disagreement cases revealed that

cases with a BIRADS category of 2 during the clinical interpretation accounted for 49.1%, 53.3%, 51.2%, 34.3%, and 36.1% of the disagreement cases for radiologists 1, 2, 3, 4, and 5, respectively (Table 3).

Table 3
Disagreement cases when the clinical interpretation was no-recall and the telemammography recommendation was recall for different BIRADS ratings during the clinical interpretation

	BIRADS category	
Disagreement cases	1 (n = 206)	2 (n = 58)
Radiologist 1 (n = 53)	50.9% (27)	49.1% (26)
Radiologist 2 (n = 60)	46.7% (28)	53.3% (32)
Radiologist 3 (n = 82)	48.8% (40)	51.2% (42)
Radiologist 4 (n = 105)	65.7% (69)	34.3% (36)
Radiologist 5 (n = 133)	63.9% (85)	36.1% (48)
Average (n = 86.6)	55.2% (49.8)	44.8% (36.8)

4. DISCUSSION

The review of breast cancer screening mammography by five experienced radiologists using the telemammography system demonstrated that the system was adequate for reviewing the mammographic image data. The limited, retrospective review of screens using the telemammography system with only mammographic image data (i.e., no prior images, prior reports, or patient history) produced modest agreement with the actual clinical interpretation. The agreement between the limited telemammography review and clinical interpretation for five radiologists ranged from 53.6% to 77.8% and Kappa ranged from 0.13 to 0.32. On average the radiologists recommended additional procedures using the limited telemammography system in 38.2% of cases which was exceedingly high compared with 13.7 % of patients actually recalled in this group during the clinical interpretation.

The majority of the disagreement between the two review formats occurred when the telemammography review resulted in a recommendation for additional procedures and the clinical interpretation did not, accounting for an average of 86.4% of the disagreement cases for the five radiologists. Of these disagreement cases (clinical no-recall and telemammography recall), on average across the radiologists 44.8% of the patients had a clinical BIRADS category of 2. That is, when findings were detected using the telemammography system under restricted conditions, but the history of the findings (i.e., new, increased, or unchanged) was unavailable, the radiologists tended to recommend additional procedures. Another potential partial explanation for the high recall rate was the radiologists' expectation of an "enriched" sample population because of their knowledge that this is a laboratory study. In addition, the mere fact that patient recall does not affect clinical management tends to produce over reading.

High recall rates were similarly observed by Elmore et al.[30] (1994), where 11-65% of patients without cancer were recommended for immediate workup. In the Elmore study, prior images were not available for any of the cases reviewed and clinical history was not available for every case. They also attributed the high recall rates to the radiologists' knowledge of an "enriched" sample population and study participation.

Although the limited, retrospective review using the telemammography system produced modest agreement with the actual clinical interpretation, the feasibility of the system use for such a review was clearly demonstrated and well-received by the radiologists. Current efforts have begun to add information such as text communication between the technologist (remote site) and radiologist (central site) to the information transmitted with each case.

ACKNOWLEDGEMENTS

This work is supported in part by the US Army Medical Research Acquisition Center, 820 Chandler Street, Fort Detrick, MD 21702-5014 under contract DAMD17-00-1-0410. The content of the information contained herein does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

REFERENCES

1. Yawn B, Krein S, Christianson J, Hartley D, and Moscovice I. "Rural radiology: who is producing images and who is reading them?" *J Rural Health* **13(2)**:136-144, 1997.
2. Benger JR. "Can nurses working in remote units accurately request and interpret radiographs?" *Emerg Med J* **19(1)**:68-70, 2002.
3. Coughlin SS, Thompson TD, Hall HI, Logan P, and Uhler RJ. "Breast and cervical carcinoma screening practices among women in rural and nonrural areas of the United States, 1998-1999". *Cancer* **94(11)**:2801-2812, 2002.
4. Bolle SR, Sund T, and Stormer J. "Receiver operating characteristic study of image processing for teleradiology and digital workstations." *J Digit Imaging* **10(4)**:152-157, 1997.
5. O'Reilly S, Spedding R, Dearden C. and Loane M. "Can *x* rays be accurately interpreted using a low cost telemedicine system?" *J Accid Emerg Med* **15(5)**:312-314, 1998.
6. Pysher L and Harlow C. "Teleradiology using low-cost consumer-oriented computer hardware and software." *AJR Am J Roentgenol* **172(5)**:1181-1184, 1999.
7. Bancroft LW, Berquist TH, Morin RL, Pietan JH, Knudson JM, and Williams HJ. "Fracture interpretation using electronic presentation: a comparison." *J Digit Imaging* **13(1)**:13-18, 2000.
8. Burgul R, Gilbert FJ, and Undrill PE. "Methods of measurement of image quality in teleultrasound." *Br J Radiol* **73**:1306-1312, 2000.
9. Davidson HC, Johnston DJ, Christian ME, and Harnsberger HR. "Comparison of radiographic image quality from four digitization devices as viewed on computer monitors." *J Digit Imaging* **14(1)**:24-29, 2001.
10. Maitz GS, Chang TS, Sumkin JH, Wintz PW, Johns CM, Ganott M, Holbert BL, Hakim CM, Harris KM, Gur D, and Herron JM. "Preliminary clinical evaluation of a high-resolution telemammography System." *Invest Radiol* **32(4)**:236-240, 1997.
11. Mitra S, Yang S, and Kustov V. "Wavelet-based vector quantization for high-fidelity compression and fast transmission of medical images." *J Digit Imaging* **11(4 Suppl 2)**:24-30, 1998.
12. Kalyanpur A, Neklesa VP, Taylor CR, Daftary AR, and Brink JA. "Evaluation of JPEG and wavelet compression of body CT images for direct digital teleradiology transmission." *Radiology* **217(3)**:772-779, 2000.
13. Lou SL, Lin HD, Lin KP, and Hoogstrate D. "Automatic breast region extraction from digital mammograms for PACS and telemammography applications." *Comput Med Imaging Graph* **24(4)**:205-220, 2000.
14. Ludwig K, Bick U, Oelerich M, Schuierer G, Puskas Z, Nicolas K, Koch A, and Lenzen H. "Is image selection a useful strategy to decrease the transmission time in teleradiology? A study of 100 emergency cranial CTs." *Eur Radiol* **8(9)**:1719-1721, 1998.
15. Lou SL, Sickles EA, Huang HK, Hoogstrate D, Cao F, Wang J, and Jahangiri M. "Full-field direct digital telemammography: Technical components, study protocols, and preliminary results." *IEEE Trans Info Technol Biomed* **1(4)**:270-278, 1997.
16. O'Sullivan DC, Averch TD, Cadeddu JA, Moore RG, Beser N, Breitenbach C, Khazan R, and Kavoussi LR. "Teleradiology in urology: comparison of digital image quality with original radiographic films to detect urinary calculi." *J Urol* **158(6)**:2216-2220, 1997.
17. Youmans DC, Don S, Hildebolt C, Shackelford GD, Luker GD, and McAllister WH. "Skeletal surveys for child abuse: comparison of interpretation using digitized images and screen-film radiographs." *AJR Am J Roentgenol* **171(5)**:1415-1419, 1998.
18. Roberts MM, Alexander FE, Anderson TJ, Chetty U, Donnan PT, Forrest P, Hepburn W, Huggins A, Kirkpatrick AE, Lamb J, Muir BB, and Prescott RJ. "Edinburgh trial of screening for breast cancer: mortality at seven years." *Lancet* **335(8684)**:241-246, 1990.
19. Sickles EA and Kopans DB. "Deficiencies in the analysis of breast cancer screening data." *J Natl Cancer Inst* **85(20)**:1621-1624, 1993.
20. Young KC, Wallis MG, and Ramsdale ML. "Mammographic film density and detection of small breast cancers." *Clin Radiol* **49(7)**:461-465, 1994.
21. Young KC, Wallis MG, Blanks RG, and Moss SM. "Influence of numbers of view and mammographic film density on the detection of invasive cancers: results from the NHS Breast Cancer Screening Programme." *Br J Radiol* **70(833)**:482-488, 1997.
22. Feig SA. "Image quality of screening mammography: effect on clinical outcome." *AJR Am J Roentgenol* **178(4)**:805-807, 2002.

23. Taplin SH, Rutter CM, Finder C, Mandelson MT, Houn F, and White E. "Screening mammography: clinical quality and the risk of interval breast cancer." *AJR Am J Roentgenol* **178(4)**:797-803, 2002.
24. Fajardo LL, Yoshino MT, Seeley GW, Hunt R, Hunter TB, Friedman R, Cardenas D, and Boyle R. "Detection of breast abnormalities on teleradiology transmitted mammograms." *Invest Radiol* **25(10)**:1111-1115, 1990.
25. Nab HW, Karssemeijer N, Van Erning LJ, and Hendriks JH. "Comparison of digital and conventional mammography: a ROC study of 270 mammograms." *Med Inform* **17(2)**:125-131, 1992.
26. Powell KA, Obuchowski NA, Chilcote WA, Barry MM, Ganobcik SN, and Cardenosa G. "Film-screen versus digitized mammography: assessment of clinical equivalence." *AJR Am J Roentgenol* **173(4)**:889-894, 1999.
27. Powell KA, Mallasch PG, Obuchowski NA, Kerczewski RJ, Ganobcik SN, Cardenosa G, and Chilcote W. "Clinical evaluation of wavelet-compressed digitized screen-film mammography." *Acad Radiol* **7(5)**:311-316, 2000.
28. Drescher JM, Maitz GS, Leader JK, Sumkin JH, Poller WR, Klaman H, Zheng B, and Gur D. "Design considerations for a multi-site, POTS-based telemammography system." *Proceedings of SPIE Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation*, 4685:416-421, San Diego, CA, February 2002.
29. Drescher JM, Maitz GS, Traylor C, Leader JK, Clearfield RJ, Shah R, Ganott MA, Pugliese F, Duffner D, Lockhart J, and Gur D. "A multi-site telemammography system: preliminary assessment of technical and operational issues." *Proceedings of SPIE Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation*, In Press, San Diego, CA, February 2003.
30. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. "Variability in radiologists' interpretations of mammograms." *N Engl J Med* **331(22)**:1493-1499, 1994.

APPENDIX 6

Improving CAD performance in detecting masses depicted on prior images

Bin Zheng [a], Xiao-Hui Wang [a], Luisa Wallace [b], Cathy Cohen [b], Lara A. Hardesty [b], Christiane M. Hakim [b], Gordon Abran [b], Jules Sumkin [b], and David Gur [a].
[a] Department of Radiology, University of Pittsburgh, Pittsburgh, PA, 15261
[b] Magee Womens Hospital, 300 Halket Street, Suite 4200, Pittsburgh, PA 15213

ABSTRACT

We investigated a new approach to improve the performance of a computer-aided detection (CAD) scheme in identifying masses depicted on images acquired earlier ("prior"). The scheme was trained using a dataset with simulated mass features. From a database with images acquired during two consecutive examinations, 100 locations matched pairs of malignant mass regions were selected in both the "current" and the most recent "prior" images. While reviewing the current images, mass regions were identified and as a result biopsies were ultimately performed. Prior images were not identified as suspicious by radiologists during the original interpretation. The same number of false-positive regions was also selected in both current and prior images. The selected regions were then randomly divided into training and testing datasets with 50 true-positive and 50 false-positive regions in each. For each selected region, five features; area, contrast, circularity, normalized standard deviation of radial length, and conspicuity; were computed. The ratios of the average difference of five feature values between current and prior mass regions in the training datasets were also computed. Multiplying these ratios by the computed values in current mass regions, we generated a new dataset of simulated features of "prior" mass regions. Three artificial neural networks (ANN) were trained. ANN-1 and ANN-2 were trained using training datasets of current and prior regions, respectively. ANN-3 was trained using simulated "prior" dataset. The performance of three ANNs was then evaluated using the testing dataset of prior images. Areas under ROC curves (A_z) were 0.613 ± 0.026 for ANN-1, 0.678 ± 0.029 for ANN-2, and 0.667 ± 0.029 for ANN-3, respectively. This preliminary study demonstrated that one could estimate an average change of feature values over time and "adjust" CAD performance for better detection of masses at an earlier stage.

Keywords: Computer-aided detection, Mammography, Mass detection, Artificial neural network

1. INTRODUCTION

Computer-Aided Detection (CAD) systems are currently used in a large number of medical institutions around the world to assist radiologists in reading and interpreting mammograms in the screening environment [1-3]. A large number of studies have been conducted to assess the possible impact of CAD systems on radiologists' performance. Although there is no general agreement on whether and how CAD systems help radiologists improve their diagnostic accuracy [3-6], several studies demonstrated that the performance of the CAD scheme itself might be an important factor to increase radiologists' confidence to accept and act on the CAD cues and help to improve their diagnostic accuracy when using such tools [6-8].

Current guidelines recommend periodic mammography screening for women over the age of 40 [9]. As compliance increases in the general population, a large fraction of patients will have undergone series of consecutive mammographic examinations. As a result, detected breast cancers will in time, "shift" on the average toward an earlier stage. In fact, retrospective review have indicated that a large fraction of breast cancers that are identified by radiologists were also visible in prior images [10]. It is expected that comparison with prior images could over time help radiologist detect

Medical Imaging 2003: Image Processing, Milan Sonka, J. Michael Fitzpatrick, Editors,
Proceedings of SPIE Vol. 5032 (2003) © 2003 SPIE · 1605-7422/03/$15.00

more subtle cancers [11,12], hence, more subtle cancers will be considered "visible" or detectable on routine mammograms. In such a changing environment, maintaining "optimal" performance of CAD schemes becomes a challenge. Although CAD schemes can detect a large number of true-positive abnormalities (e.g., masses and microcalcification clusters) depicted on prior images [7,12,13], current CAD schemes that had been optimized using a large fraction of "easy" cancers are unlikely to achieve "optimal" performance in detecting "earlier" or more "subtle" cancers. This is due to several factors: (1) performance of CAD schemes that use a feature-based machine-learning classifier heavily depends on the characteristics of training database [14,15] and (2) a large number of image features used to train CAD schemes varies differently for abnormalities as depicted on the current images as compared with prior images [16]. Several studies have demonstrated that in order to achieve optimal performance in detecting suspicious masses as depicted on prior images, a different set of image features should be selected for re-optimization of CAD schemes [17,18].

In previous studies [17,18] optimal performance in detecting masses depicted on prior images was achieved by re-training the scheme using a set of mass regions extracted from prior images. This requires a significant effort. Since there is a training database available for each CAD scheme, this database could potentially be used to re-optimize the scheme after a computational adjustment of some feature values. For this purpose, we investigated a new method to generate a simulated training database and used it to re-optimize our CAD scheme. A detailed description of our approach and preliminary experimental results follow.

2. MATERIALS AND METHODS

From an image database established in our laboratory, we selected 100 matched pairs of digitized mammograms from two consecutive (the most recent or "current" and the latest previous or "prior") examinations. There is a verified mass region depicted in each case. During the current examination, these 100 mass regions were identified by radiologists as suspicious and as a result biopsies were ultimately performed. Although in a retrospective review and with the support of available source documents, an experienced observer could identify some indication of the presence of a "mass" in the corresponding locations on prior images, these regions had not been identified as suspicious by radiologists during the original interpretation. All 100 mass regions selected for this study were associated with biopsy-proven malignancies. The locations of all masses depicted on current images and the corresponding locations on prior images were visually identified. The centers (x, y coordinate) of all verified mass regions were marked manually and saved in a reference (or "truth") file.

All 200 images (100 from current and 100 from prior examination) were processed by a CAD scheme developed previously in our laboratory [19]. To detect suspicious masses, each image is first subsampled (pixel-averaged) in both dimensions to increase pixel size from original 50 μm × 50 μm (or in some cases 100 μm × 100 μm) to 400 μm × 400 μm. The CAD scheme then uses three stages to identify suspicious regions. In the first stage, the scheme uses image subtraction and threshold results after processing by two Gaussian filters with a large difference in the kernel sizes (7 and 51 pixels) to search for the initial set of "suspicious" regions, which usually generates in the range of 10 to 30 initial "suspicious" regions per image. In the second stage, based on local contrast measurement the scheme uses an adaptive region growth algorithm to define three topographic layers. After simple intra-layer based threshold conditions on growth ratio and shape factor, this stage typically eliminates approximately 85% of regions identified in stage one, while maintaining a very high sensitivity. A set of features is computed for each detected region. During stage three the detected regions are classified based on scores (likelihood of being true-positive) generated by a nonlinear multi-layer feature-based classifier (e.g., an artificial neural network) [20]. To determine whether a detected region represents a true-positive or false-positive mass region in this study, the following criterion was used. If the distance between the center of gravity of a detected region and the center of the mass as recorded in the reference file was shorter than the radius of the longest axis of the detected region, it was considered as a true-positive identification. Otherwise, the region was considered a false-positive identification. In this experiment, all suspicious mass regions identified after the second stage of the CAD scheme became candidates for the study (namely, the classification scores in the third stage were ignored). One hundred true-positive mass regions from current images and 100 mass regions from prior images were selected. The CAD scheme detected 187 and 202 false-positive mass regions in the current and prior images as well. From these, 200 false-positive regions were randomly selected (100 from current images and 100 from prior images). Hence, 400

suspicious mass regions were selected for the study. The regions were then divided (block randomization) into training and testing datasets for both current and prior images. Each dataset included 50 true-positive and 50 false-positive mass regions.

For each region the following five features were computed:

1. Region area ($F_1 = 0.16 \times N_T$): This feature is computed by counting the number of pixels in the growth region (N_T) and then multiplying it by the size unit of each pixel (0.16 mm^2).
2. Average contrast ($F_2 = \frac{1}{N_T}\sum_{i=1}^{N_T} I_i - \frac{1}{N_B}\sum_{j=1}^{N_B} I_j$): This feature is computed by the average pixel value (I) difference between the growth region and its surrounding background.
3. Circularity ($F_3 = \frac{N_C}{N_T}$): To compute this feature, CAD scheme first computes the area of a growth region (N_T) and calculates an equivalent circle originating at the center of gravity of the region. For a circle with the same size as the growth region, the number of pixels that are located inside the growth region contour and the circle (N_C) is computed. Circularity is defined as the fraction of the growth region pixels covered by the circle.
4. Normalized standard deviation of radial length ($F_4 = \sqrt{\frac{1}{N_b}\sum_{i=1}^{N_b}(\frac{r_i - m_r}{m_r})^2}$): The radial length r_i is defined as the distance between the region center and a point (i) located on the perimeter of the region. m_r is the mean value of radial length over all points N_b in the region boundary. This feature indicates the changes in the shape of region boundary.
5. Conspicuity ($F_5 = \frac{F_2}{C_B}$): This feature is defined as "region contrast" (F_2) divided by "surrounding complexity" (C_B); where $C_B = \frac{1}{N_B}\sum_{i=1}^{N_B} |Max(I_i - I_F)|$ and $Max(I_i - I_F)$ is the maximum pixel value difference between background pixel (i) and its neighboring pixels (e.g., 24 pixels in a 5×5 square window).

Using these features, three artificial neural networks (ANN) were constructed to classify suspicious regions. The topology of all ANNs was the same. It involved five input neurons (each represented by one feature), three hidden neurons, and one output neuron. The ANN was trained using 500 iterations. The training momentum and learning rate were 0.8 and 0.01, respectively.

ANN-1 and ANN-2 were trained using training dataset of current and prior images, respectively. ANN-3 was trained using a set of simulated "prior" mass regions. To generate a simulated dataset, we computed the ratio of the average feature values for each of five features between 50 pairs of true-positive mass regions as extracted from current and prior images. Ratios were computed as follows:

$$D_k = \frac{\frac{1}{N}\sum_{i=1}^{N} F_{k,i}^{Prior}}{\frac{1}{N}\sum_{j=1}^{N} F_{k,j}^{Current}}, \quad k = 1,2,3,4,5. \text{ and } N = 50.$$

Each feature of true-positive mass region in the current training dataset was then multiplied by the ratio, such as $F'_{k,j} = F^{Current}_{k,j} \times D_k$. Hence, a set of new feature values was generated to represent each of 50 "simulated true-positive mass regions." Using these data combined with feature values of 50 original false-positive regions extracted from the current images, ANN-3 was trained. Although the 50 simulated mass regions (used in ANN-3) and 50 original prior mass regions (used in ANN-2) have identical mean values for each of the five features, the feature values for a specific region are different (i.e., $F'_{k,j} \neq F^{Prior}_{k,j}, k = 1,2,\ldots,5$). In other word, the simulated set of "prior" features does not simply duplicate the actual feature set in prior images.

The performances of three ANNs were evaluated separately using testing datasets of 50 current and 50 prior images. For each test region, the ANN generates a classification score ranged from 0 to 1, where the larger the score, the higher the computed likelihood of being a true-positive mass region. The classification scores generated for all test regions were used as input data in the ROCFIT program that generates a receiver operating characteristic (ROC) curve and computes the area under the ROC curve (A_z value) [21]. We compared performance levels when using the three ANNs to classify an independent set of suspicious mass regions as depicted on prior images.

3. RESULTS

Table 1 shows the averages of the five feature values in the two training datasets extracted from the current and prior images. Using paired chi-square test to examine the mean values of each of the five features between 50 pairs of training mass regions, the significant difference ($p < 0.05$) was found in the average value of each of the five features. Table 2 summarizes the areas under ROC curves (A_z values) for all three ANNs during training and testing. Figure 1 demonstrates three ROC curves generated by applying three ANNs to the prior testing dataset. ANN-1 yields the best performance in testing current dataset (A_z = 0.781 ± 0.019) and the worst performance in prior testing dataset (A_z = 0.613 ± 0.026) as shown in table 2. Both ANN-2 and ANN-3 yield significantly better performance than ANN-1 in classifying mass regions on prior testing dataset ($p < 0.05$). A_z values were increased by 10.6% (from 0.613 to 0.678) in ANN-2 and 8.8% in ANN-3 (from 0.613 to 0.667), respectively. The experimental results also demonstrated that there was no significant performance difference between ANN-2 and ANN-3 in testing prior dataset ($p = 0.15$).

Table 1: Average feature values and their difference ratios between 50 pairs of mass regions depicted on current and prior images.

Feature:	F_1	F_2	F_3	F_4	F_5
Average value (prior images):	78.60	34.90	0.24	0.78	4.25
Average value (current images):	122.67	42.68	0.21	0.83	5.07
Ratio:	0.70	0.82	1.14	0.94	0.84

Table 2: Areas under ROC curves (A_z values) of three ANNs during training and testing.

Network	Training	Testing current images	Testing prior images
ANN-1	0.873 ± 0.016	0.781 ± 0.019	0.613 ± 0.026
ANN-2	0.761 ± 0.021	0.709 ± 0.026	0.678 ± 0.029
ANN-3	0.779 ± 0.019	0.736 ± 0.028	0.667 ± 0.029

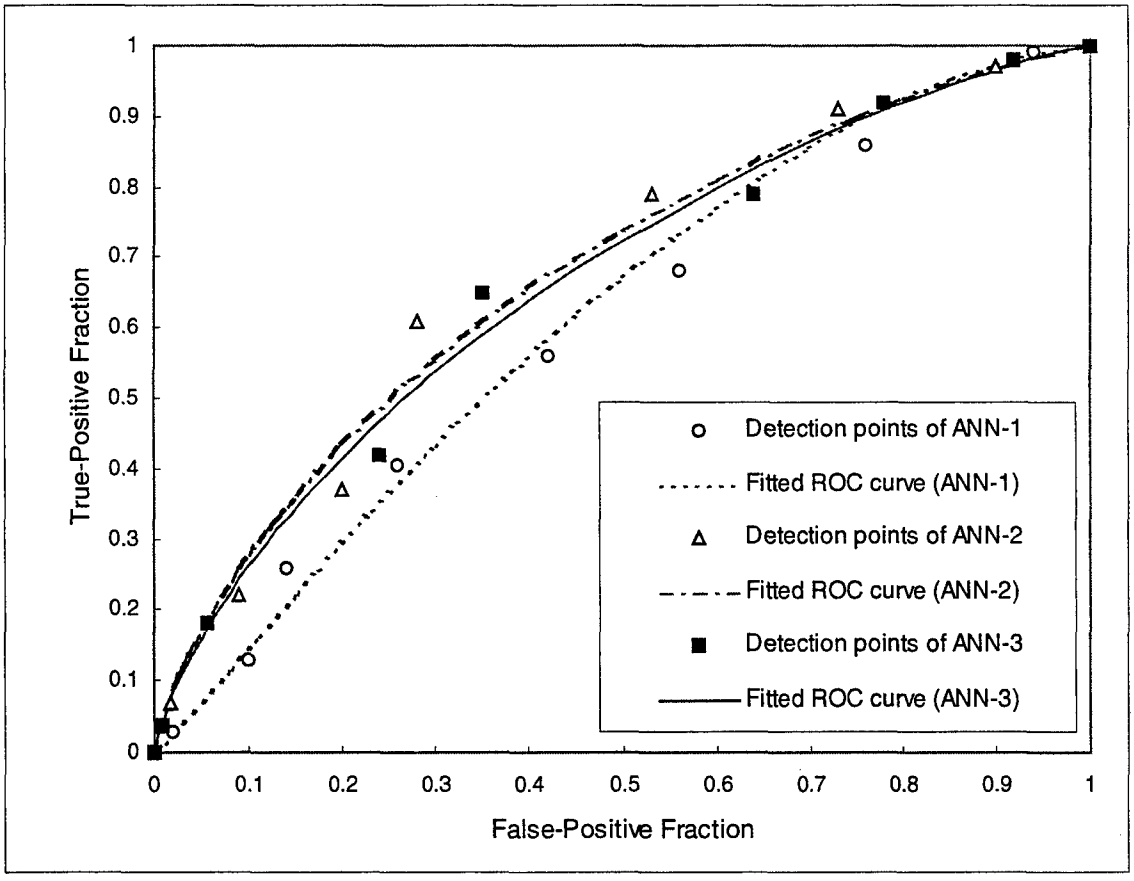

Figure 1: ROC curves of testing results when applying three ANNs to the test dataset of prior images.

4. DISCUSSION

With improvements of diagnostic technologies and increase in screening compliance of the general population, radiologists have to detect increasingly more subtle abnormalities as depicted on mammograms. As a result, CAD systems that currently provide satisfactory cueing results could face deterioration in performance over time due to a general shift in the subtleness of and stage at detection. Feature-based machine learning classifiers, such as ANNs, are widely used in final stage of the CAD schemes for identifying masses and microcalcification clusters. Since these classifiers are trained to generate "global" functions that cover the entire instance space, CAD performances heavily depend on the training databases [22]. This is true, in particular, in mammography where the size and diversity of training datasets is generally limited [14,15]. A single CAD scheme that achieves high sensitivity on both "subtle" and relatively "easy" masses at an acceptable false-positive rate can be developed, however, in reality, it is a very difficult task because image features are substantially different for suspicious mass regions extracted from the current and prior images [16,17]. In order to improve CAD performance in detecting subtle masses in an earlier stage, the schemes should be trained (or optimized) using databases involving a large fraction of subtle mass regions (e.g., new cases that had been rated originally as negative and later proven to be positive) [17,18].

However, it is a very difficult and time-consuming task to collect a large number of diverse subtle cases (e.g., the false-negative cases). This study demonstrated an alternative approach to collectively simulate such cases. By systematically adjusting the feature values extracted from current images, we generated a set of simulated "prior" mass

regions. Our results demonstrated that (1) an ANN trained using simulated prior mass regions could achieve significantly better performance in detecting the masses at an earlier stage than an ANN trained using current mass regions and (2) there is no significant difference in the performance between the ANNs trained using either real or simulated prior mass regions. As a result, by estimating the change over time of some important features, one can adjust CAD performance for better detection of masses at an earlier stage. Since this is a very preliminary study involving a limited database and a small set of features, the concept need to be further investigated. If this approach is validated with significantly larger image databases and larger number of features, it may provide a simple and efficient method to periodically update (or re-optimize) CAD schemes.

ACKNOWLEDGEMENTS

This work is supported in part by Grants CA77850, CA85241 and CA80836 from the National Cancer Institute of National Institutes of Health and also by the US Army Medical Research Acquisition Center, 820 Chandler Street, Fort Detrick, MD 21702-5014 under Contract DAMD17-00-1-0410. The content of the information contained here does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

REFERENCES

1. A. Malich, C. Marx, M. Facius, T. Boehm, M. Fleck, W.A. Kaiser, "Tumour detection rate of a new commercially available computer-aided detection system," *Eur Radiol* 11:2454-2459 (2001).
2. M. Lechner, M. Nelson, E. Elvecrog, "Comparison of two commercially available computer-aided detection (CAD) systems," *Appl Radiol* 31:31-35 (2002).
3. T.W. Freer, M.J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center," *Radiology* 220:781-786 (2001).
4. K. Moberg, N. Bjurstam, B. Wilczek, L. Rostgard, E. Egge, C. Muren, "Computed assisted detection of interval breast cancers," *Eur J. Radiology* 39:104-110 (2001).
5. W.R. Hendee, "In the next decade automated computer analysis will be an accepted sole method to separate "normal" from "abnormal" radiological images," *Med Phys* 26:1-4 (1999).
6. B. Zheng, M.A. Ganott, C.A. Britton, C.M. Hakim, L.A. Hardesty, T.S. Chang, H.E. Rockette, D. Gur, "Soft-copy mammographic reading with different computer-assisted detection cueing environments: Preliminary findings," *Radiology* 221:633-640 (2001).
7. G.M. te Brake, N. Karssemeijer, J.H. Hendriks, "Automated detection of breast carcinomas not detected in a screening program," *Radiology* 207:465-471 (1998).
8. A. Malich, T. Azhari, T. Bohm, M. Fleck, W.A. Kaiser, "Reproducibility – an important factor determining the quality of computer aided detection (CAD) systems," *Eur J. Radiology* 36:170-174 (2000).
9. S.A. Feig, C.J. D'Orsi, R.E. Hendrick, "American College of Radiology guidelines for breast cancer screening," *Am J Roentgenol* 171:29-33 (1998).
10. L.J. Warren 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, R.A. Castellino, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," *Radiology* 215:554-562 (2000).
11. L.W. Bassett, B. Shayestehfar, I. Hirbawi, "Obtaining previous mammograms for comparison: usefulness and cost," *Am J Roentgenol* 163:1083-1086 (1994).
12. M.P. Callaway, C.R. Boggis, S.A. Astley, "Influence of pervious films on screening mammographic interpretation and detection of breast carcinoma," *Clin Radiol* 52:527-529 (1997).
13. R.L. Birdwell, D.M. Ikeda, K.F. O'Shaughnessy, E.A. Sickles, "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection," *Radiology* 219:192-202 (2001).
14. B. Zheng, Y.H. Chang, W.F. Good, D. Gur, "Adequacy testing of training set sample sizes in the development of a computer-assisted diagnosis scheme," *Acad Radiol* 4:497-502 (1997).
15. M.A. Kupinski, M.L. Giger, "Feature selection with limited database," *Med Phys* 26:2176-2182 (1999).

16. B. Zheng, R. Shah, L. Wallace, C. Hakim, M.A. Ganott, D. Gur, "Computer-aided detection in mammography: An assessment of performance on current and prior images," *Acad Radiol* in press (2002).
17. L. Hadjiiski, B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, M. Gurcan, "Analysis of temporal changes of mammographic features: Computer-aided classification of malignant and benign breast masses," *Med Phys* 28:2309-2317 (2001).
18. B. Zheng, W.F. Good, D.R. Armfield, C. Cohen, T. Hertzberg, J.H. Sumkin, D. Gur, "Performance changes of mammographic CAD schemes optimized using most recent and prior image databases," *Acad Radiol* in review (2002).
19. B. Zheng, Y.H. Chang, D Gur, "Computerized detection of masses in digitized mammograms using single image segmentation and a multi-layer topographic feature analysis," *Acad Radiol* 2:959-966 (1995).
20. B. Zheng, J.H. Sumkin, W.F. Good, G.S. Maitz, Y.H. Chang, D. Gur, "Applying computer-assisted detection schemes to digitized mammograms after JPEG data compression: An assessment," *Acad Radiol* 7:595-602 (2000).
21. C.E. Metz, B.A. Herman, J.H. Shen, "Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," *Stat Med* 17:1033-1053 (1998).
22. T.M. Mitchell, *Machine learning,* WCB/McGraw-Hill, Boston, MA, (1997).

PROGRESS IN BIOMEDICAL OPTICS AND IMAGING

SPIE—The International Society for Optical Engineering

Multi-site telemammography system: preliminary assessment of technical and operational issues

J. M. Drescher, G. S. Maitz, C. Traylor, J. K. Leader, R. J. Clearfield, R. Shah, M. A. Ganott, F. Pugliese, D. Duffner, J. Lockhart, D. Gur

Reprinted from

Medical Imaging 2003

PACS and Integrated Medical Information Systems: Design and Evaluation

SPIE P PROCEEDINGS SERIES

Volume 5033

18–20 February 2003
San Diego, California, USA

A multi-site telemammography system: preliminary assessment of technical and operational issues

John M. Drescher*[a], Glenn S. Maitz[a], Christopher Traylor[a], J. Ken Leader[a], Ronald J. Clearfield[ab], Ratan Shah[ab], Marie A. Ganott[ab], Francine Pugliese[b], Dian Duffner[b], Janet Lockhart[b], David Gur[a]
[a]University of Pittsburgh, Pittsburgh, PA USA 15213
[b]Magee-Womens Hospital, Pittsburgh, PA USA 15213

ABSTRACT

Our goal was to develop an inexpensive, high-quality, multi-site telemammography system, implemented with low-level data connections that provided a communication link for an "almost real-time" response from a radiologist (central site) to remote "underserved" sites. The remote sites digitize mammographic films using high-resolution, laser digitizers. Images are automatically cropped, compressed (wavelet-based), and encrypted prior to transmission. At the central site images are decrypted, decompressed, unsharp masked, and displayed using automatically determined LUTs. The sites communicate instantly via a "chat box." Remote sites 1, 2, and 3 are 15, 20, and 90 miles from the central site, respectively, and connected by POTS (sites 1 and 2) and LAN (site 3). Only minimal noticeable difference at compression levels of 50:1 and 75:1 could be identified unless magnified to extreme levels. Two experienced observers rated the LUTs for 200 images as "acceptable" to "excellent." Average cycle times to digitize, transmit and receive cases (four films each) at 75:1 compression were 5.97, 6.85, and 5.77 min/case from sites 1, 2, and 3, respectively. Unique data-handling schemes significantly decrease the image file size and allow successful transmission in a reliable, timely manner. Over 1000 cases have been transmitted to date. Messaging was found to be easy to use.

Keywords: Teleradiology, breast cancer screening, image decision making, mammography.

1. INTRODUCTION

The benefits of breast cancer screening mammography of asymptomatic women have been extensively studied and reported in the recent literature.[1-6] Mammographic screening will continue to be widely used worldwide, despite periodic reports of limited or no benefits from such practices.[7-9] Management of mammographic screening in terms of public perception and compliance,[10-12] radiologist's practice and performance,[13-15] and personnel shortages[11,16] could be improved in both rural and urban clinics. The use of teleradiology is one approach that could assist in this regard.

The high-spatial resolution required by mammography necessitates the use of commercial digitizers and high-resolution monitors to sufficiently preserve image quality.[17] Transmission time of large amounts of mammographic image data (35-55 MBytes per image) is frequently dependent on the communication link. Low-level data connections (i.e., Plain Old Telephone System (POTS)) may require data processing to decrease the image file size to enable transmission of large amounts of data in a timely manner.

This manuscript presents preliminary assessment of technical and operational issues regarding a multi-site telemammography system using low-level data connections. This study is a continuation of an ongoing effort over the past several years.[18,19] The system was designed on the concept of distributed acquisition/centralized review and to facilitate communication between a radiologist at a central site and a technologist at a remote "underserved" site. For the purpose of this project, "underserved" means a location where a physician is not physically present when the screening examinations are conducted. The technical features described were designed and implemented using a low-cost approach to transmit data across low-level data connections in a timely manner and maintain a high-level of image quality. Issues evaluated included: look-up table settings (window and level), image cropping, image compression,

* drescherjm@msx.upmc.edu; phone (412) 641-2563; fax (412) 641-2582, University of Pittsburgh, Magee-Womens Hospital, 300 Halket Street., Suite 4200, Pittsburgh, PA 15213

Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation, H. K. Huang, Osman M. Ratib, Editors, Proceedings of SPIE Vol. 5033 (2003)

transmission time, and workstation display features. We expect to demonstrate that the combination of efficient data handling, intelligent image processing, and easy to use messaging can be implemented to produce an inexpensive, high quality telemammography system capable of an "almost real-time" response from the central site radiologist to remote site technologist.

2. METHODS

2.1 Central and remote sites

The central site is staffed by experienced radiologists and located at Magee-Womens Hospital, Pittsburgh, PA, USA. The telemammography workstation at the central site is powered by a dual 1.2 GHz multi-processor (Athlon MP, Advanced Micro Device, Sunnyvale CA, USA) with 2 GB of RAM operating under Microsoft Windows 2000 Server (Microsoft Corporation, Redmond, WA, USA). The workstation display consists of three high-resolution (2048 x 2560) 8-bit grayscale portrait monitors at a nominal setting of 80 ftL (DS5100P, Clinton Electronics, Rockford, IL, USA). For data communication, the workstation uses 56K hardware modems (U.S. Robotics, Rolling Meadows, IL, USA) and ethernet network cards (OfficeConnect 10/100 NIC, 3COM, Santa Clara, CA, USA). A Kodak Dryview film printer (Eastman Kodak, Rochester, NY, USA) is connected to the workstation for film printing as necessary (Fig. 1).

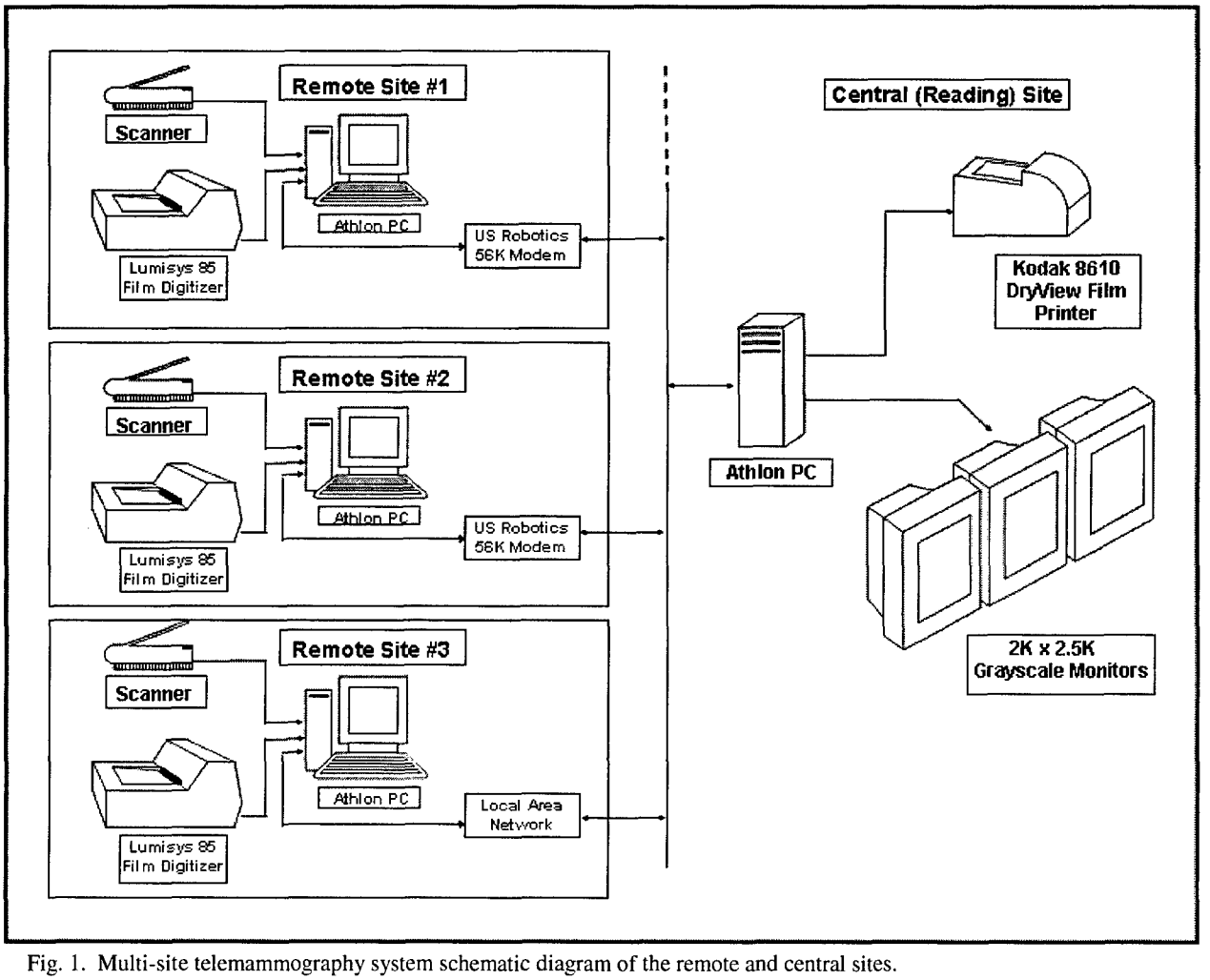

Fig. 1. Multi-site telemammography system schematic diagram of the remote and central sites.

The remote sites are staffed by mammography technologists. The computer hardware at the remote sites operates under Microsoft Windows 2000 Workstation powered by a 900 MHz processor (Athlon 900, Advanced Micro Device, Sunnyvale CA, USA) with 512 MB of RAM. High-resolution, laser film digitizers (Lumiscan 85, Eastman Kodak, Rochester, NY, USA) are connected to the remote computers via SCSI interface and equipped with a film feeder capable of holding six films as large as 10 x 12 inches. Mammographic films are digitized at 50 micron pixel dimensions and 12-bit grayscale. The remote site computers also have 56K hardware modems and ethernet network cards (Integrated PRO/100 S Desktop Adapter, Intel Corporation, Santa Clara, CA, USA) for data communication. Prior patient reports or history are transmitted along with the images by inserting them into an attached page scanner (hp scanjet 5490C, Hewlett-Packard, Palo Alto, CA, USA). Sites 1 and 2 transmit data across Plain Old Telephone System (POTS) lines and are located 15 and 20 miles from the central site, respectively (Fig.1). Site 3 is 90 miles from the central site and transmitted data across a Local Area Network (LAN).

2.2 Software Design

The software architecture at the central and remote sites is a multithreading design that allows independent task assignment with simultaneous response to user input. A message dispatch mechanism synchronizes bi-directional communication between all the main threads, except for the Time Manager (Fig. 2). The Time Manager periodically dispatches elapsed time messages to the other main threads without receiving messages. Each main thread acts on only messages associated with its function and may spawn subordinate (worker) threads that share data objects to accomplish tasks. A Reader/Writer lock, derived from Microsoft Windows synchronization primitives, prevents corruption of the shared data. The Reader/Writer lock permits access to the shared data to any number of readers simultaneously.

Central site main threads:

- Time manager - periodically indicates elapsed time.
- Archive manager - manages disk space by loading images, saving images, managing cases, and deleting archived cases when disk space is limited.
- Case manager - creates cases, assigns data, and performs database functions.
- Display manager - displays images and forwards messages to the main application window.
- Distribution manager - receives, transmits, and processes data.

Remote site main threads:

- Digitization manager - manages film digitizing.
- Case manager
- Display manager
- Distribution manager

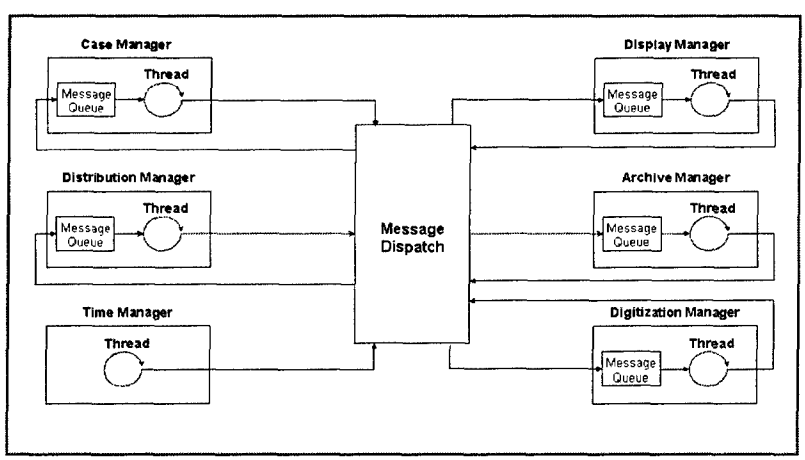

Fig. 2. Main threads and intra-process communication. Time manager does not receive messages.

2.3 Image processing

The first step in the series of the image processing procedures is designed to automatically crop each image to decrease the non-tissue area surrounding the breast (Fig. 3). The automated cropping algorithm begins by sub-sampling the image at an 8:1 ratio. The standard deviation (STD) of a 7 x 7 pixel mask is calculated at each sub-sampled pixel (STD of the sub-sampled image). Next, a threshold is applied to the STD image to separate tissue and non-tissue regions where a high STD indicated tissue regions. A region growing algorithm based on 4-neighbor connectivity is used to identify breast tissue as the largest region in the image. Finally, rudimentary logic is used to determine the cropping parameters based on the orientation of the tissue regions which is applied to the original image.

Following image cropping, the image data are compressed using the irreversible (lossy), 9/7 transform, wavelet-based JPEG 2000 method. Prior to transmission from the remote sites, the data packets are encrypted using strong 128 bit Microsoft Point-to-Point Encryption (MPPE) with version 2 authenticate Microsoft Challenge Handshake Authenticate Protocol (CHAP). The first steps at the central site are decryption and decompression of the image data.

Image display on the workstation monitors at the central site is enhanced by minimal unsharp masking of the decompressed image data prior to display. To begin unsharp masking, the image data are first smoothed with a 2-D 129 mean kernel. The weighted (0.10) smoothed image is subtracted from the decompressed image. The resulting pixel values of the image data are then re-scaled from 0 to 4095.

To minimize the need for manual adjustment during image viewing, default look-up table (LUT) values are automatically calculated based on the pixel value distribution (histogram). The typical pixel value distribution is bimodal. The window value (contrast) is set as the span of the two modes, and the level value (brightness) is set as the center between the two modes. The final stage of the image processing prior to image display is to pad (fill) the images to restore the full height of the image, but not the full width (Fig. 3).

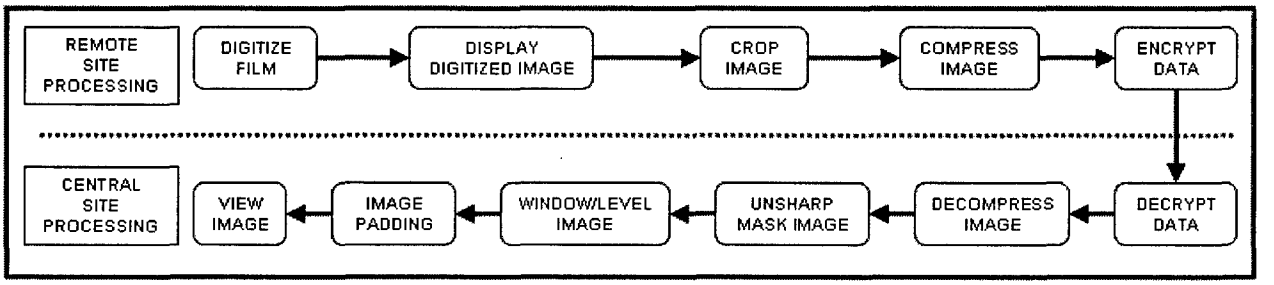

Fig. 3. Data flow of the telemammography system illustrating the order of the image processing tasks and where (remote or central) the process is performed.

2.4 Workstation display functions and features

To allow user-specific preferences to be used during case review, display options on the workstation are flexible with all features being mouse-driven. The default display is left and right craniocaudal views (LCC & RCC) on the left monitor, and left and right mediolateral oblique views (LMLO & RMLO) on the center monitor to be similar to our conventional clinical film presentation (Fig. 4). However, a large number of display options are available to users. If a film is digitized in an incorrect orientation, the user has the ability to flip images (top to bottom or left to right) and rotate images 180 degrees. Communication from the remote site is displayed on the right monitor (Fig. 4).

Two forms of image magnification are available on the workstation display. Typically, the normal display scale with a single image per monitor is approximately 100 micron pixel dimensions and with two images per monitor it is approximately 200 micron pixel dimensions. A scrollable image magnification box provides a true 1:1 presentation (monitor pixel:digitized pixel) resulting in 50 micron pixel dimensions. The size of the box varies from 511 x 566 pixels for one image/monitor, and 408 x 566 pixels for two images/monitor, and 204 x 266 pixels for four images/monitor. It is also possible to pan across the image quadrant-by-quadrant.

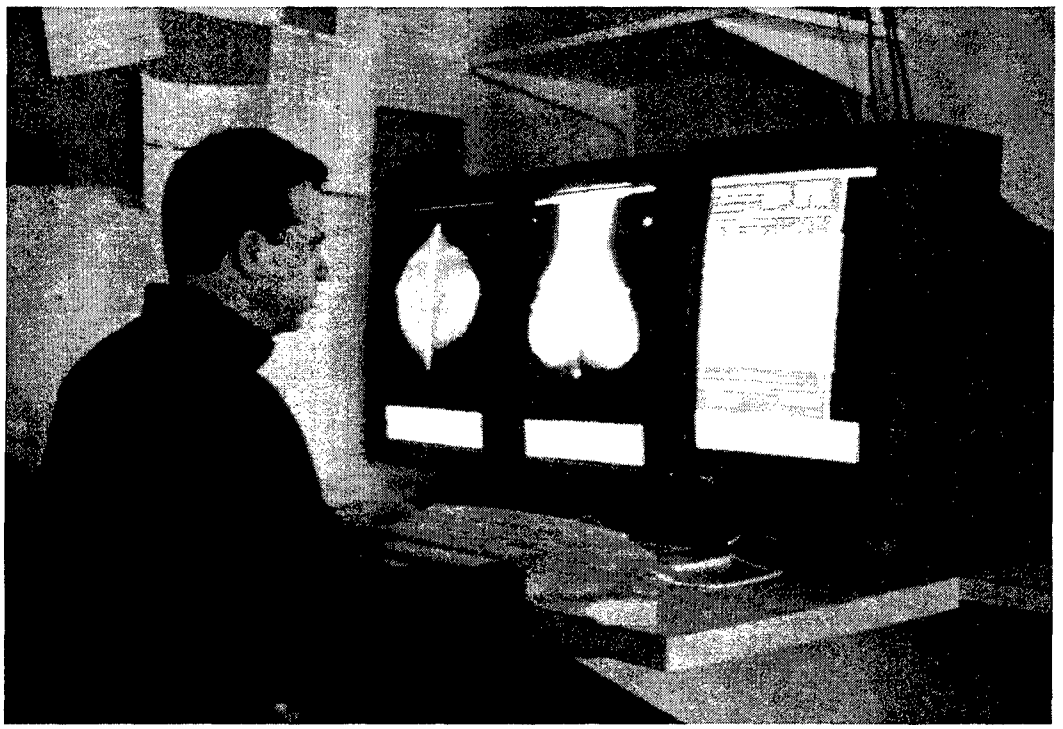

Fig. 4. Telemammography workstation at the central site pictured in the default image display format.

The automated LUT values can be manually adjusted per observer's preference. The window and level values are determined based on the mouse position (movement), and the image display is instantly updated as the mouse is moved. Once the desired values are determined, these can be applied to the individual image or all images associated with the case. The LUT values can be reset to the automated (default) values at anytime during viewing.

2.5 Inter-site communication

To facilitate effective communication between the technologists (remote site) and radiologists (central site), a "chat box" type messaging function was implemented. The "chat message" can be sent with each case and it provides a real-time, interactive communication tool between the sites. During the initial phase of evaluating the system, communication is performed in one cycle. The technologist sends a chat message with each case, and the radiologist responds directly to the message. The chat boxes on both sides contained four general areas: (1) patient demographics, (2) message display area, (3) pull-down menus, and (4) free text area (Fig. 5). There are five pull-down menus on the technologist chat box to focus communication on possible actionable items. These indicate: (1) breast: left or right; (2) view: craniocaudal and/or mediolateral oblique; (3) finding: mass or calcifications; (4) comparison with prior exam: baseline, new, or change in findings; and (5) possible additional procedure needed: additional views and/or ultrasound. The radiologists can reply after reviewing each case. His/her response includes: (1) do recommended procedure as suggested; (2) no additional procedures necessary; and (3) do not do the procedure recommended, but do X, Y, and Z.

2.6 Technical and operational evaluation

In the preliminary technical assessment phase, three processes of the telemammography system were evaluated. First to assess the user's acceptance of the automated LUT values for image review without the need to adjust display parameters, 50 cases (200 images) sent from all sites were subjectively rated by two experienced observers on a scale of 1 to 4. The experiment was designed to assess acceptability of default values for the purpose of reviewing each case and determining the need (or not) for additional procedures. In all of our studies we evaluated the system under normal operating conditions. As a result, intra- and inter-site measured variability reflect what could be expected in an "on-line" clinical operation. Second, the implementation of high-level image compression in mammographic imaging was evaluated during subjective Just Noticeable Difference (JND) studies. The studies compared images at no compression,

50:1, and 75:1 compression levels. Third, the average cycle time from initiation of digitization to availability for display at the central site was evaluated. This involved transmission of a series of four cases (back to back) each consisting of four images per case (all images were 8 x 10 inches).

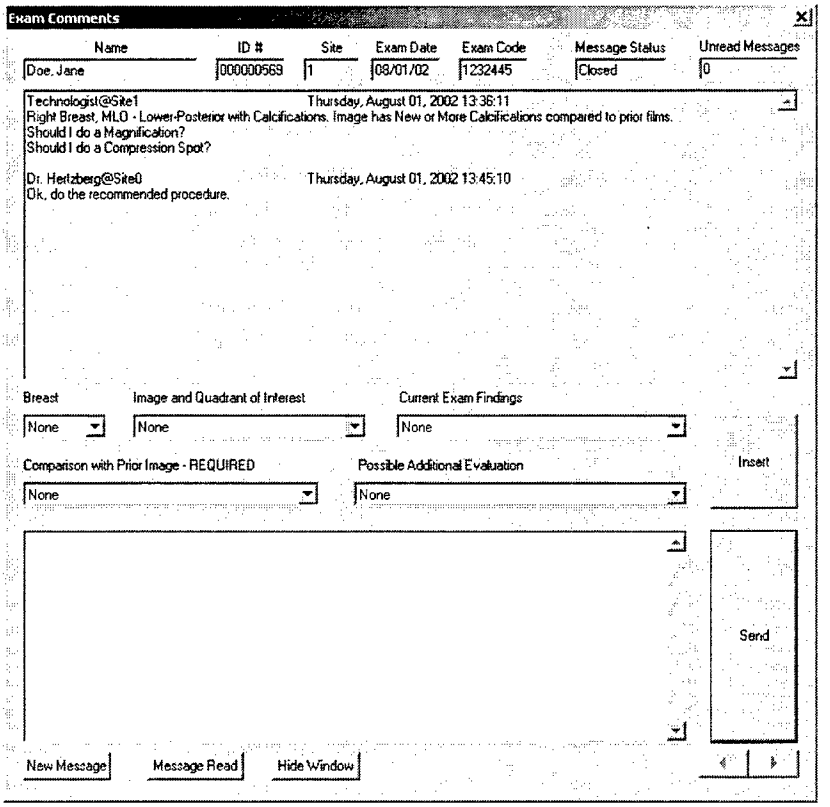

Fig. 5. "Chat box" for the remote site technologists.

3. RESULTS

The evaluation of the technical and operational processes was favorable in all areas. The automated LUT settings, the image cropping, the high level of image compression, and the cycle time to transmit and receive cases were all acceptable for implementation of the telemammography system for the designed purpose. The initial impressions of the inter-site communication, "chat messaging," indicate that it can facilitate effective communication between the technologist at remote sites and the radiologist at the central site. Although the technical issues with regard to scanning and transmitting patient reports with each case have been resolved, the practice of has not been implemented to date.

The automatically calculated LUT settings were reported as "acceptable" to "excellent" by two experienced mammography researchers. On a scale of 1 to 4 (1 = unusable, 2 = need minor adjustments, 3 = acceptable, and 4 = excellent), the two observers had mean ratings for 200 automatically computed LUT settings of 2.64 (STD = 0.57) and 3.51 (STD = 0.53). After minor adjustments were made as the result of the above experiment, all observers including clinicians using the workstation to test different aspects of the system accepted automatically set values in over 90% of cases. Consequently, window and/or level manipulations are being performed in less than 10% of cases during retrospective and simulated prospective case reviews.

For review of non-magnified or moderately magnified images, 50:1 and 75:1 data compression levels were comparable and acceptable when evaluated on either laser-printed films or the telemammography workstation. Subjective JND studies were conducted using laser-printed films as well as images displayed side-by-side on workstation monitors. The studies indicated that at extreme magnifications, differences were detected, but did not necessarily result in degradation

of perceived diagnostic quality. For example, the "visibility" and "clarity" of microcalcifications in the digital images were judged as "almost equivalent" between the full-scale, non-compressed images and images compressed at a 75:1 ratio (Figs. 6 and 8). Comparable results were obtained with magnification (Figs. 7 and 9). The automated cropping did not remove breast tissue in any of our cases to date, and it produces "aesthetically pleasing" images.

The time to transmit and receive four films (8 x 10 inches each) was reliably less than 7 minutes/case for each site using 75:1 data compression (Table 1). The combination of image cropping and 75:1 data compression ratio decreased image file size to allow cycle times that were adequate for implementation of the telemammography concept and met our planned technical specifications. Sites 1 and 2 were connected via 56K modems that dialed a four digit telephone number (i.e., connected via an in-house telephone line) and a ten digit telephone number (i.e., connected via an outside telephone line), respectively. Consistent bandwidths of sites 1 and 2 were approximately 33 Kbits/second and 21 Kbits/second, respectively. The digitization process (approximately 50 seconds/film) was the limiting factor at site 3 which was connected via LAN. Site 2 had communication problems (decreased bandwidth) during the first measurement that have been largely resolved.

TABLE 1
Experimentally Measured Average Cycle Time for Digitizing, Transmitting and Receiving a Case with 4 Films (8 x 10 inches each)

Image format	Site 1 - POTS* (min/case)	Site 2 - POTS (min/case)	Site 3 - LAN (min/case)
50:1 compression, not cropped, and not encrypted	13.22	24.42	5.38
50:1 compression, cropped, and encrypted	6.47	13.13	5.65
75:1 compression, cropped, and encrypted	5.97	6.85	5.77

*in-house POTS

4. DISCUSSION

The "proof of concept" to design an inexpensive, high-quality, multi-site telemammography system implemented with low-level data connections has been established to facilitate the concept of "almost real-time" distributed acquisition/centralized review. The technical feasibility of the concept was demonstrated by: (1) the digitization of films acquired during clinical breast cancer screening mammography; (2) the timely transmission of the digitized images across low-level data connections (less than 7 minutes/case); and (3) the efficient archiving, retrieving, and viewing of image data at the central site. The short cycle time of the system was realized because of the image file size reduction due to automated image cropping and image data compression and the efficient multi-tasking software approach based on a synchronized multi-threading design. Image processing methods were fundamental to the success of the telemammography system. The automated cropping and compression produced images without a significant degradation of the diagnostic image quality, which were well-received by the radiologists. Although the automated window and level calculations were found to be acceptable, in approximately ten percent of cases, radiologists manually employed window and level settings during an individual case review. The high-resolution image display of the telemammography workstation was rated acceptable for reviewing screening mammographic images for the purpose of determining the need for additional procedures.

To date, over 1000 screening exams have been successfully transmitted using the telemammography system. The preliminary results suggest that the telemammography system could accomplish the goals to increase effective communication between remote "underserved" sites and the central location, and permit experienced radiologists to remotely monitor and facilitate some decision making while the patient remains in the clinic. The addition of two key components to the telemammography system should improve the system's capability and effective utilization. First, scanned prior patient reports will be added to the information transmitted with each case. Second, Computer Aided Detection (CAD) schemes will be incorporated into the system and the results will be displayed at the central site.

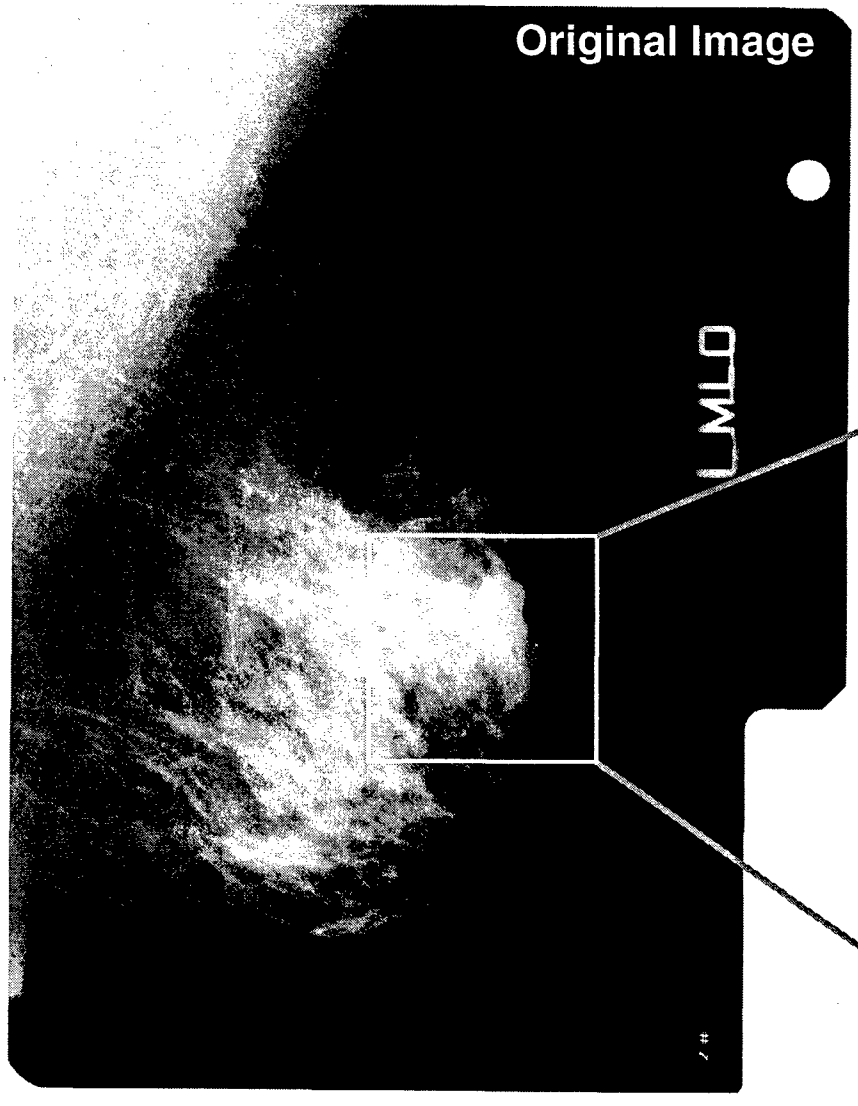

Fig. 6. Original left medial lateral oblique image of patient #569. Image is not cropped, compressed, or unsharp masked.

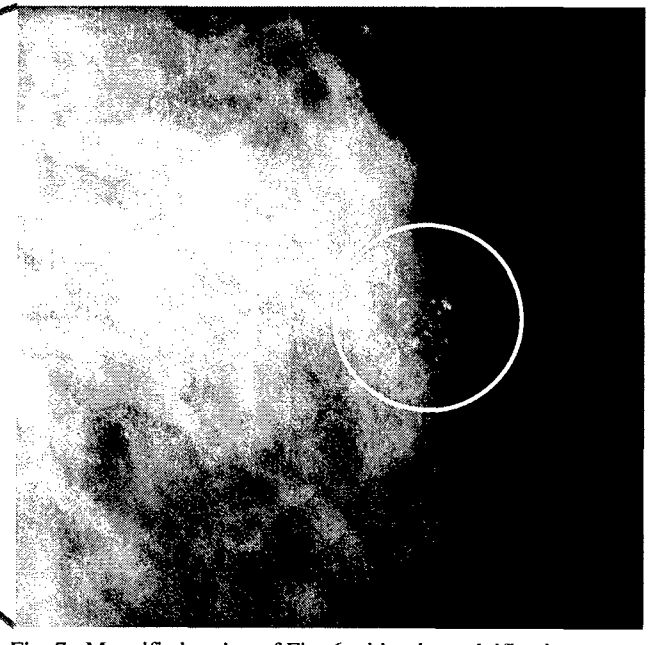

Fig. 7. Magnified region of Fig. 6 with microcalcifications marked.

Fig. 8. Processed left medial lateral oblique image of patient #569. Image is cropped, compressed at a 75:1 ratio, and unsharp masked.

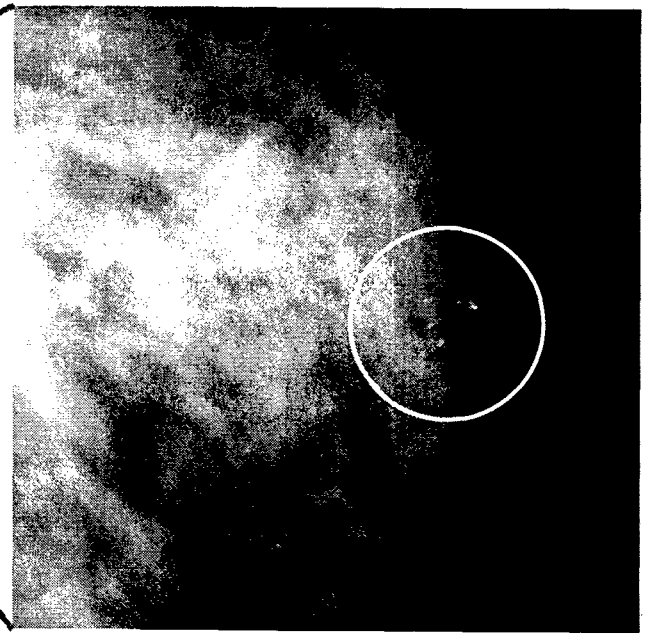

Fig. 9. Magnified region of Fig. 8 with microcalcifications marked.

ACKNOWLEDGEMENTS

This work is supported in part by the US Army Medical Research Acquisition Center, 820 Chandler Street, Fort Detrick, MD 21702-5014 under contract DAMD17-00-1-0410. The content of the information contained herein does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

REFERENCES

1. Duffy SW, Tabar L, Chen HH, Holmqvist M, Yen MF, Abdsalah S, Epstein B, Frodis E, et al. "The impact of organized mammography service screening on breast carcinoma mortality in seven Swedish Counties." *Cancer* **95(3)**:458-469, 2002.
2. Feig SA. Current status of screening mammography. *Obstet Gynecol Clin North Am* 2002;29(1):123-136.
3. Huphrey LL, Helfand M, Chan BKS, and Woolf SH. "Breast cancer screening: a summary of the evidence for the U.S. preventive services task force." *Ann Intern Med* **137(5, Part 1)**:E347-E367, 2002.
4. Lee CH. "Screening mammography: proven benefit, continued controversy." *Radiol Clin North Am* **40(3)**:395-407, 2002.
5. Nystrom L, Andersson I, Bjurstam N, Frisell J, Nordenskjold B, and Rutqvist LE. "Long-term effects of mammography screening: update overview of the Swedish randomised trials." *Lancet* **359(9310)**:909-919, 2002.
6. Tabar L, Vitak B, Chen HHT, Yen MF, Duffy SW, and Smith RA. "Beyond randomized clinical trials: organized mammographic screening substantially reduces breast carcinoma mortality." *Cancer* **91(9)**:1724-1731, 2002.
7. Gotzsche PC and Olsen O. "Is screening for breast cancer with mammography justifiable?" *Lancet* **355**:129-134, 2000.
8. Olsen O and Gotzsche PC. "Cochrane review on screening for breast cancer with mammography." *Lancet* **358**:1340-1342, 2001.
9. Miller AB, To T, Baines CJ, and Wall C. "The Canadian Nation Breast Screening Study-1: nreast cancer mortality after 11 to 16 years of follow-up: a randomized screening trial of mammography in women age 40 to 49 years." *Ann Intern Med* **137(5, Part 1)**:E305-E315, 2002.
10. Chamot E and Perneger TV. "Misconception about efficacy of mammography screening: a public health dilemma." *J Epidemiol Community Health* **55(11)**:799-803, 2001.
11. Coughlin SS, Thompson TD, Hall HI, Logan P, and Uhler RJ. "Breast and cervical carcinoma screening practices among women in rural and nonrural areas of the United States, 1998-1999." *Cancer* **94(11)**:2801-2812, 2002.
12. Michaelson J, Satija S, Moore R, Weber G. Halpern E, Garland A, Puri D, and Kopans DB. "The pattern of breast cancer screening utilization and its consequences." *Cancer* **94(1)**:37-43, 2002.
13. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. "Variability in radiologists' interpretations of mammograms." *N Engl J Med* **331(22)**:1493-1499, 1994.
14. Warren RML and Duffy SW. "Comparison of single reading with double reading of mammograms and change in effectiveness with experience." *Br J Radiol* **68(813)**:958-962, 1995.
15. Hulka CA, Slanetz PJ, Halpern EF, Hall DA, McCarthy KA, Moore R, Boutin S, and Kopans DB. "Patients' opinion of mammography screening services: immediate results versus delayed results due to interpretation by two observers." *AJR Am J Roentgenol* **168**:1085-1089, 1997.
16. Yawn B, Krein S, Christianson J, Hartley D, and Moscovice I. "Rural radiology: who is producing images and who is reading them?" *J Rural Health* **13(2)**:136-144, 1997.
17. Lou SL, Lin HD, Lin KP, and Hoogstrate D. "Automatic breast region extraction from digital mammograms for PACS and telemammography applications." *Comput Med Imaging Graph* **24(4)**:205-220, 2000.
18. Maitz GS, Chang TS, Sumkin JH, Wintz PW, Johns CM, Ganott M, Holbert BL, Hakim CM, Harris KM, Gur D, and Herron JM. "Preliminary clinical evaluation of a high-resolution telemammography System." *Invest Radiol* **32(4)**:236-240, 1997.
19. Drescher JM, Maitz GS, Leader JK, Sumkin JH, Poller WR, Klaman H, Zheng B, and Gur D. "Design considerations for a multi-site, POTS-based telemammography system." *Proceedings of SPIE Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation*, 4685:416-421, San Diego, CA, February 2002.

Radiology

Bin Zheng, PhD
Lara A. Hardesty, MD
William R. Poller, MD
Jules H. Sumkin, DO
Sara Golla, MD

Index terms:
Breast neoplasms, diagnosis, 00.119
Breast radiography, technology, 00.119
Computers, diagnostic aid, 00.119

Published online before print
10.1148/radiol.2281020489
Radiology 2003; 228:58–62

Abbreviation:
CAD = computer-aided detection

[1] From the Departments of Radiology, University of Pittsburgh and Magee-Womens Hospital, Imaging Research, Suite 4200, 300 Halket St, Pittsburgh, PA 15213. Received April 29, 2002; revision requested June 21; final revision received September 23; accepted October 23. 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, Fort Detrick, Md, under contract DAMD17-00-1-0410. **Address correspondence to** B.Z. (e-mail: *zhengb@msx.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.

Author contributions:
Guarantor of integrity of entire study, B.Z.; study concepts, B.Z.; study design, B.Z., J.H.S.; literature research, B.Z., S.G.; clinical and experimental studies, B.Z., W.R.P.; data acquisition, B.Z., L.A.H., S.G.; data analysis/interpretation, B.Z.; statistical analysis, B.Z.; manuscript preparation, B.Z., S.G., L.A.H.; manuscript definition of intellectual content, B.Z.; manuscript editing, B.Z., S.G., L.A.H.; manuscript revision/review and final version approval, all authors.

Mammography with Computer-aided Detection: Reproducibility Assessment—Initial Experience[1]

PURPOSE: To examine the performance and reproducibility of a commercially available computer-aided detection (CAD) system with a set of mammograms obtained in 100 patients who had undergone biopsy after positive findings at mammography.

MATERIALS AND METHODS: One hundred positive mammographic examinations (four views each), depicting 96 masses and 50 microcalcification clusters, were scanned and analyzed three times by the CAD system. Reproducibility of detection sensitivity and the individual CAD-generated cues in the three images were examined. Both abnormality- and region-based detection sensitivities were compared.

RESULTS: Forty-eight (96.0%) of 50 microcalcification clusters were marked on all three images in the abnormality-based analysis. Of the remaining two clusters, one was marked in two images and one was marked in only one. The abnormality-based sensitivity for mass detection ranged from 66.7% (64 of 96) to 70.8% (68 of 96). The system generated identical patterns (including images with and those without cues) for all three images in 53.3% (213 of 400) of images. For true-positive cluster regions, 88.9% (80 of 90) were marked at the same location in all images. For true-positive mass regions, 69.5% (82 of 118) were marked at the same locations in all images. In false-positive detections, only 44.0% (81 of 184) of false-positive mass regions and 31.9% (38 of 119) of false-positive cluster regions were marked at the same locations on all three images.

CONCLUSION: Reproducibility of marked regions generated by the CAD system is improved from that reported previously, largely as a result of the substantial reduction in the false-positive detection rates. Reproducibility of true-positive identification of masses remains an important issue that may have methodologic and clinical practice implications.

Mammography is a common and effective method with which to screen for early detection of breast cancer, to interpret mammograms, and particularly to identify subtle masses and microcalcification clusters surrounded by complex breast tissue patterns, but it is a difficult and time-consuming task. Findings in studies show that from 10% to 30% of breast cancers that are visible on mammograms during retrospective readings are missed during the original interpretations for various reasons (1–3). One well-documented method to reduce false-negative rates in mammography is the use of an independent double-reading approach (4,5). However, this approach is both inefficient and costly. As a result, after intensive research and substantial improvements in the past 2 decades, computer-aided detection (CAD) systems have been developed to provide radiologists with a "second opinion" when they identify suspicious regions for masses or microcalcification clusters. In the current study, we used one of three commercially available CAD systems that have been approved by the U.S. Food and Drug Administration and are used for this purpose.

Because of the potential importance of CAD systems in the clinical environment, several studies (6–10) have been conducted recently to evaluate the performance of CAD systems

alone and their possible effect on diagnostic performance of radiologists under a variety of clinical conditions. In one recent study involving 12,860 patients in a community breast center, use of CAD resulted in a 19.5% increase in the number of cancers detected without undue effect on the recall rate (from 6.5% to 7.7%) (6). In another large retrospective study, a false-negative rate of 21% was found when 14 radiologists interpreted mammograms, and the CAD system correctly marked 77% of these missed cases (7). Thus, researchers claim that CAD cueing could potentially reduce this false-negative rate by as much as 77% without an increase in the recall rate (8). On the other hand, findings in a different study showed that despite high (and clinically viable) sensitivity, the CAD system had no effect on radiologist performance (including sensitivity and specificity) (9). These researchers suggested that perhaps the many false-positive markings influenced the radiologists not to have sufficient confidence in the CAD results to alter their original interpretations (9). Results in another retrospective study demonstrated that the performance of a CAD system could affect the performance of radiologists in the detection of masses and microcalcification clusters. Highly performing CAD schemes with high sensitivity and a low false-positive rate could improve radiologists' performance significantly, while poorly performing CAD schemes could significantly ($P < .01$) decrease readers' performance (10).

An important issue related to the use of CAD is the reproducibility of results. In one study, an early version of ImageChecker (R2 Technology, Los Altos, Calif) was evaluated, and the authors suggested that its reproducibility may be insufficient for the routine clinical environment (11). Recently, a new version of the software was used, which improves the detection sensitivity and specificity (12). In the version used in the current study (ImageChecker, version 2.0), the stated detection sensitivity for the cancer cases was increased from 83.7% to 90.4% (including an increase in mass detection from 74.7% to 85.7% and an essentially unchanged performance for microcalcification detection of more than 98%). At the same time, the false-positive rate was reduced substantially from approximately 1.0 per image to 0.5 per image (or 4.1–2.06 false-positive cues per four views in true-negative cases) (12). The purpose of our study was to examine the performance and reproducibility of a commercially available CAD system by using a set of mammograms acquired in 100 patients who had undergone biopsy after positive findings at mammography.

MATERIALS AND METHODS

Cases

During the past several years, a large database (>1,000 cases) of digitized images and associated diagnostic results has been established and managed in our laboratory under an approved institutional review board protocol (informed consent was waived). For the purpose of third-party, we asked a staff member not otherwise related to this current investigation to randomly select 100 mammographic cases (four views each) from the biopsy records of our institution during the years 1999–2001. We requested that 25 of the cases depict microcalcification clusters and 75 cases depict masses as a primary detection finding. At least two-thirds of the cases were to be selected from those proven to be malignant. With the exception of these conditions, cases were selected solely by the staff member from the biopsy records. The selection process did not involve a previous review of any of the images. Therefore, there was no preselection (and potentially biasing) process as related to the average tissue density or the subtlety of the abnormalities depicted in the images.

Each case could involve one or more abnormalities (mass, microcalcification cluster, or both). In these 100 cases, 51 depicted only masses (43 depicted one mass and eight depicted two masses), 12 depicted only microcalcification clusters (11 depicted one cluster and one depicted two clusters), and 37 depicted both masses and clusters (one mass and one cluster). There were no cases with more than three abnormalities depicted. The data set involved 96 verified masses and 50 verified microcalcification clusters. Sixty-five of the 96 masses were malignant, and 31 were benign. Thirty-one of the 50 microcalcification clusters were associated with malignancy, and 19 were benign. By examining all source documents (including pathology reports), the locations of all abnormalities were specified by radiologists.

CAD Evaluation

These 400 images were scanned through the CAD system three times within a period of 3 weeks. After digitization and computation, suspicious masses and microcalcification clusters identified by the CAD system were marked on the output paper images by using the standard identification scheme. The CAD system does not outline the entire mass region or individual microcalcifications in a cluster, only a small star or a triangle is superimposed on the image to indicate the presence of a suspicious region for a mass or a cluster, respectively. The boundaries of masses and clusters were identified visually on the images by a researcher (B.Z.), who consulted with radiologists in cases of ambiguity. If the star was located anywhere inside a true-positive mass region in the image, this mass was considered to be identified correctly by the CAD system. Similarly, as long as a triangle was overlapping any of the microcalcification areas, the mark was considered to represent a true-positive detection. Otherwise, the cue was considered to identify a false-positive region. The processing of each case resulted in three sets of output images.

Data Analysis

The sensitivity, false-positive rate, and reproducibility of the CAD system with these 100 cases (or 400 images) were analyzed for abnormality- and region-based values. In the abnormality-based analysis, the sensitivity is assessed on the basis of the correct marking of at least one true-positive region in either view (craniocaudal, mediolateral oblique, or both), which included 96 masses (65 malignant) and 50 calcifications (31 malignant) in the 100 cases. In cases with more than one abnormality, each was considered to be independent of the others. In the region-based analysis, the abnormality depicted in each view (either craniocaudal or mediolateral oblique) was considered an independent true-positive finding. Sensitivity was then computed on the basis of the number of correctly detected true-positive regions (rather than abnormalities). This approach included 292 positive findings—namely, 96 masses and 50 clusters, each visible on two views. To compare the differences in proportions of correctly detected abnormalities among replicated images, the pairwise McNemar test was applied to the data set.

RESULTS

Tables 1 and 2 summarize the performance of the CAD system with respect to mass and microcalcification cluster detection in each of the three scans. Abnormality-based sensitivity for mass detection ranged from 66.7% (64 of 96) to 70.8% (68 of 96). Although scan 2 yielded highest sensitivity for mass detection (68 of 96), scan 1 depicted the highest number of malignant masses (47 of 65). For microcalcification cluster detection, 48 of 50 clusters were

TABLE 1
Mass Detection Performance of CAD System during Each Scan

Scan No.	Sensitivity (all cases)		Sensitivity (malignant cases only)		False-Positive Rate per Image
	Abnormality Based (%)	Region Based (%)	Abnormality Based (%)	Region Based (%)	
1	69.8 (67 of 96)	52.1 (100 of 192)	72.3 (47 of 65)	54.6 (71 of 130)	0.33 (130 of 400)
2	70.8 (68 of 96)	52.6 (101 of 192)	70.8 (46 of 65)	52.3 (68 of 130)	0.33 (131 of 400)
3	66.7 (64 of 96)	51.0 (98 of 192)	69.2 (45 of 65)	51.5 (67 of 130)	0.31 (125 of 400)

TABLE 2
Microcalcification Cluster Detection Performance of CAD System during Each Scan

Scan No.	Sensitivity (all cases)		Sensitivity (malignant cases only)		False-Positive Rate per Image
	Abnormality Based (%)	Region Based (%)	Abnormality Based (%)	Region Based (%)	
1	96.0 (48 of 50)	85.0 (85 of 100)	93.5 (29 of 31)	85.5 (53 of 62)	0.17 (69 of 400)
2	98.0 (49 of 50)	87.0 (87 of 100)	96.8 (30 of 31)	87.1 (54 of 62)	0.19 (77 of 400)
3	100 (50 of 50)	86.0 (86 of 100)	100 (31 of 31)	87.1 (54 of 62)	0.20 (79 of 400)

TABLE 3
Number of Times a Mass (or a Region) was Detected

No. of Times Detected	True-Positive Masses	Malignant Masses	True-Positive Mass Regions	Malignant Mass Regions	False-Positive Mass Regions	Total Marked Mass Regions
3 (%)	58 (77.3)	41 (78.8)	82 (69.5)	58 (71.6)	81 (44.0)	163 (54.0)
2 (%)	8 (10.7)	4 (7.7)	17 (14.4)	8 (9.9)	40 (21.7)	57 (18.9)
1 (%)	9 (12.0)	7 (13.5)	19 (16.1)	14 (17.3)	63 (34.3)	82 (27.1)
Total	75	52	118	81	184	302

TABLE 4
Number of Times a Microcalcification Cluster (or a Region) was Detected

No. of Times Detected	True-Positive Clusters	Malignant Clusters	True-Positive Cluster Regions	Malignant Cluster Regions	False-Positive Cluster Regions	Total Marked Cluster Regions
3 (%)	48 (96.0)	29 (93.5)	80 (88.9)	50 (89.3)	38 (31.9)	118 (56.5)
2 (%)	1 (2.0)	1 (3.2)	8 (8.9)	5 (8.9)	30 (25.2)	38 (18.2)
1 (%)	1 (2.0)	1 (3.2)	2 (2.2)	1 (1.8)	51 (42.9)	53 (25.3)
Total	50	31	90	56	119	209

detected by the CAD system on all three images. Two malignant clusters were missed in two of three scans (scans 1 and 2), and one of these clusters was missed in scan 2. With the pairwise McNemar test, no significant ($P > .3$) differences were found in the detection results between any pair of the three scans.

For region-based sensitivity, mass detection ranged from 51.0% (98 of 192) to 52.6% (101 of 192). The total number of masses detected ranged from 98 to 101 in each of the three scans. However, the actual difference in the individual mass regions detected was larger. For example, scan 1 depicted 100 regions and scan 2 depicted 101 regions. However, only 88 of these regions were detected in both images. For the detection of microcalcification clusters, the region-based sensitivity ranged from 85.0% (85 of 100) to 87.0% (87 of 100) for individual cluster regions and from 85.5% (53 of 62) to 87.1% (54 of 62) for malignant clusters.

Although Tables 1 and 2 show that the total number of regions detected in this set of images is relatively constant with all three scans, the locations of the regions detected (in particular, false-positive regions) could differ from scan to scan. In 213 of 400 images, the output results for all three scans were identical, which represents an overall reproducibility of 53.3%. Among these images, 37.6% (80 of 213) had no cues (including neither true-positive nor false-positive cues) in all three scans. For the remaining 320 images, the CAD system marked 511 regions (1.6 cues per image) in three scans (including true-positive cues). Of these 511 marked regions, 281 were identified on all three scans (55% region-based reproducibility).

Tables 3 and 4 summarize the number of true-positive and false-positive masses and microcalcification clusters (including both abnormalities and regions) that were identified in all three scans, two scans, or only one scan. The results show that the reproducibility for the true-positive regions (those identified in all three scans) is substantially higher than that for the false-positive regions. For the true-positive mass regions, the CAD system generated 118 cues in three scans, and 82 (69.5%) of them were marked at the same locations. For the true-positive cued cluster regions, 88.9% (80 of 90) of cues were in the same locations for all three scans. On the other hand, the reproducibility of the false-positive cues was much lower, with a higher fraction of different cues being generated in each scan. Only 44.0% (81 of 184) of the false-positive mass regions and 31.9% (38 of 119) of the false-positive microcalcification cluster regions were marked at the same locations in all three scans.

DISCUSSION

In a previous study, 38.5% (77 of 200) of images had CAD cues that were located congruently in all three scans (11). In the current study, the CAD system generated identical results on 53.0% (213 of 400) of the images. The improvement in reproducibility may be largely a result of the substantial decrease in the false-positive detection rate (from approximately 1.0 to 0.5 per image) (12). When we exclude 80 images that had no CAD cues, the reproducibility in the remaining 320 images was reduced to 41.6% (133 of 320). However, the reproducibility in detecting specific true-positive masses and microcalcification clusters is perhaps more important than the more general image-based reproducibility. It is generally difficult to directly compare the detection performance in two experiments, because different image databases were used and the results depend heavily on the difficulty of the selected cases (13). However, some comparative information can be ascertained. In a previous report, the CAD system performed better for mass detection (86.9% abnormality-based sensitivity) than for microcalcification cluster detection (76.6%) (11), while in the current study, sensitivity for the detection of microcalcification clusters was higher than 96%, and the sensitivity for mass detection was in the range of 70%. These results may indicate that the microcalcification clusters depicted in our data set were easier to detect, and masses depicted in our database were more subtle. The case selection protocol we used should have reduced biases; however, the results presented herein with a small database may not represent the actual performance of the system in a clinical setting. Findings in the current study demonstrated clearly that the issue of reproducibility of image-based CAD systems needed to be investigated further.

It should be noted that we obtained somewhat different results in absolute terms for the benign and malignant cases, but the pattern for the two groups remained similar. All cases in our study were sufficiently suspicious to ultimately warrant a recommendation for biopsy. We believe that at this stage, CAD schemes should be designed and optimized to identify this group of cases, including those that ultimately prove to be benign. It is well known that repeated scanning of the same image results in a slightly different digital value matrix for a variety of technical reasons. In current CAD systems, a binary threshold is typically used to generate detection marks. Each marked region has a computed score that is above a predetermined threshold; hence, lesions with computed scores that are near the threshold are vulnerable to small changes and may be detected in one image and missed in another. Findings in the present study show that the reproducibility of false-positive cues was much lower than that of true-positive cues (Tables 3 and 4), because the detection scores may be close to the threshold. We did not perform a complete long-term follow-up to confirm that all false-positive cues actually represented negative regions. Should any false-positive detection prove to be a true abnormality, the computed reproducibility level would be lower than that reported herein.

Note that the databases used in this and a previous (11) study were small; hence, the results may not represent the actual reproducibility of CAD systems in the screening environment. Despite this limitation, findings in the two studies highlight an important finding. Current CAD schemes are sensitive to small variations in the digital value matrices that result from repeated scanning of the same images. This may have methodologic and clinical practice implications that need to be addressed. The fact that all abnormalities depicted in the present study were visible on both views indicates that the cases were not particularly subtle and that the findings we report herein, including possible implications, may be magnified in cases that are more difficult to identify visually or when the abnormality is visible only on one view. We suspect that this sensitivity to minor changes in the matrices is not unique to the CAD system evaluated in the current study. Full-field digital mammography systems are rapidly becoming available (14,15). By definition, once an image is acquired, the CAD detection result will be 100% reproducible when the same CAD scheme is applied repeatedly to such an image. To be optimal, however, current CAD schemes may have to be reengineered and reoptimized by using digitally acquired images before these schemes can be applied optimally to full-field digital mammography systems. An investigation on possible effects of repeated image acquisition of the same breast on CAD results is beyond the scope of the present study.

Findings in our preliminary study suggest that sensitivity for the detection of microcalcification clusters is high; as a result, reproducibility is also high. These results are achieved at a low false-positive detection rate; hence, it is a useful tool during the diagnostic process. Our results raise the important question about the possible need to maintain records of CAD cues as available during the interpretation of the individual cases. This may become an even more important issue as cancer detection continues to progress toward an earlier stage (hence, a more subtle appearance) on the average. Detailed documentation of all available information at the time of diagnosis is not always done, particularly since information is often provided verbally. In the case of screening mammographic interpretation, however, the presence of a malignancy that was visible (in retrospect) on a previous mammogram and in which a follow-up scan of the original images in a CAD system may produce a true-positive identification, could present a medicolegal problem. It will be difficult to argue that the abnormality in question was not identified as suspicious on the original image. Findings in our preliminary study suggest that this may be the case in a noticeable fraction of mass cases (approximately 20%, as shown in Table 3).

The current practice associated with the use of CAD in the mammographic environment is not clear on whether a record of the CAD results used during the case interpretation should be retained. Until mass detection is substantially improved, results in our study suggest that such a practice should be considered. Interestingly, although largely impractical, our study findings clearly suggest that at this level of performance, multiple repeated scans of each case could be acquired to improve the performance of CAD schemes.

References

1. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. Radiology 1992; 184:613–617.
2. Harvey JA, Fajardo LL, Innis CA. Previous mammograms in patients with impalpable breast carcinomas: retrospective vs blinded interpretation. AJR Am J Roentgenol 1993; 161:1167–1172.
3. Goergen SK, Evans J, Cohen GP, MacMillan JH. Characteristics of breast carcinomas missed by screening radiologists. Radiology 1997; 204:131–135.
4. Thurfjell EL, Lernevail KA, Taube AAS. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191:241–244.
5. Hendee WR, Beam C, Hendrick E. Proposition: all mammograms should be double-read. Med Phys 1999; 26:115–118.
6. Freer TW, Ulissey MJ. Screening mammography with computer-aided detec-

tion: prospective study of 12,860 patients in a community breast center. Radiology 2001; 220:781–786.
7. Warren 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.
8. Birdwell RL, Ikeda DM, O'Shauhhnessy 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.
9. Moberg K, Bjurstam N, Wilczek B, Rostgard L, Egge E, Muren C. Computed assisted detection of interval breast cancers. Eur Radiol 2001; 39:104–110.
10. 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.
11. Malich A, Azhari T, Bohm T, Fleck M, Kaiser WA. Reproducibility: an important factor determining the quality of computer aided detection (CAD) systems. Eur Radiol 2000; 36:170–174.
12. Castellino RA, Roehrig JR, Zhang W. Improved computer-aided detection (CAD) algorithms for screening mammograms (abstr). Radiology 2000; 217(P): 400.
13. Nishikawa RM, Giger ML, Doi K, et al. Effect of case selection on the performance of computer-aided detection schemes. Med Phys 1994; 21:265–269.
14. Lewin JM, Hendric RE, D'Orsi CJ, et al. Comparison of full-field digital mammography with screen-film mammography for caner detection: results of 4,945 paired examinations. Radiology 2001; 218:873–880.
15. 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.

Performance Change of Mammographic CAD Schemes Optimized with Most-Recent and Prior Image Databases[1]

Bin Zheng, PhD, Walter F. Good, PhD, Derek R. Armfield, MD, Cathy Cohen, MD
Todd Hertzberg, MD, Jules H. Sumkin, DO, David Gur, ScD

Rationale and Objectives. The authors evaluated performance changes in the detection of masses on "current" (latest) and "prior" images by computer-aided diagnosis (CAD) schemes that had been optimized with databases of current and prior mammograms.

Materials and Methods. The authors selected 260 pairs of matched consecutive mammograms. Each current image depicted one or two verified masses. All prior images had been interpreted originally as negative or probably benign. A CAD scheme initially detected 261 mass regions and 465 false-positive regions on the current images, and 252 corresponding mass regions (early signs) and 471 false-positive regions on prior images. These regions were divided into two training and two testing databases. The current and prior training databases were used to optimize two CAD schemes with a genetic algorithm. These schemes were evaluated with two independent testing databases.

Results. The scheme optimized with current images produced areas under the receiver operating characteristic curve of 0.89 ± 0.01 and 0.65 ± 0.02 when tested with current images and prior images, respectively. The scheme optimized with prior images produced areas under the receiver operating characteristic curve of 0.81 ± 0.02 and 0.71 ± 0.02 when tested with current images and prior images, respectively. Performance changes for both current and prior testing databases were significant ($P < .01$) for the two schemes.

Conclusion. CAD schemes trained with current images do not perform optimally in detecting masses depicted on prior images. To optimize CAD schemes for early detection, it may be important to include in the training database a large fraction of prior images originally reported as negative and later proven to be positive.

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

Mammography is considered the most reliable and cost-effective screening method for the early detection of breast cancers, which could lead to early treatment and substantially reduce associated mortality and morbidity (1,2). The large volume of mammograms obtained and the low cancer detection rates in a mammographic screening environment could result in radiologists missing as many as 10%–30% of cancers rated "visible" during retrospective reviews (3,4). To assist radiologists in detecting more cancers at screening, computer-aided detection (CAD) systems are being used in many medical institutions around the world (5,6). A number of studies have been conducted to assess their possible effect on radiologists' performance. Although there is no general agreement on whether and how CAD systems help radiologists

Acad Radiol 2003; 10:283–288

[1] From the Department of Radiology, University of Pittsburgh and Magee-Womens Hospital, 300 Halket St, Suite 4200, Pittsburgh, PA 15213-3180. Received October 10, 2002; revision requested November 25; revision received and accepted December 10. Supported in part by grants CA85241, CA77850, and CA80836 from the National Cancer Institute, National Institutes of Health, and also by the U.S. Army Medical Research Acquisition Center under contract DAMD17-00-1-0410. **Address correspondence to** B.Z.

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

improve their diagnostic accuracy (7,8), a number of studies have demonstrated that the performance of the particular CAD scheme (including sensitivity, false-positive rate, and reproducibility) could be important in this regard (9–11).

Current guidelines recommend periodic mammographic screening for women over age 40 years (12). As compliance increases in the general population, a large fraction of patients will have undergone a series of mammographic examinations. As more of the most easily detected cancers are identified during the initial examination with the incorporation of CAD into the diagnostic process, detected breast cancers will be shifted, on average, toward an earlier stage. In other words, more subtle cancers will be considered visible or detectable on routine mammograms. This will occur also in part because of the availability of previous images for comparison, which could help radiologists detect more subtle cancers (13,14). In this changing environment, it is not clear whether current CAD schemes optimized with a large number of easily detected cancers are best suited for the detection of earlier or more subtle cancers. This may become an important issue in developing and evaluating new CAD schemes. In our experiment, an artificial neural network (ANN) previously used in our own CAD scheme for mass detection was reoptimized separately by means of mass regions depicted on "current" images (from the most recent examination, at which the mass was actually reported, leading to biopsy) and those depicted on the corresponding "prior" images (originally interpreted as negative). Hence, two different schemes were used. The changes in their performance were then evaluated when they were applied to independent sets of cases with masses depicted on both current and prior images.

MATERIALS AND METHODS

We searched our database for verified cases in which both current and prior images had been collected and digitized. Inclusion criteria required that at least one mass had been identified by a radiologist on the current images and that biopsy had been performed as a result. In addition, during a retrospective review and with the support of available source documents, an experienced observer (B.Z.) had to be able to identify a mass at the corresponding locations on the prior images. In each case, the most recent prior image had been interpreted as negative or "not highly suspicious."

As a result, 134 cases were selected for this study. The mass was visible on both views (craniocaudal and mediolateral oblique) in 126 cases and on only one view in eight cases. Hence, 260 pairs of images, with each pair consisting of one current image and one prior image, were included in the study. On these images 270 distinct mass regions were identified (10 images depicted two mass regions), 220 of which were associated with biopsy-proved malignancy (50 were benign). The locations of all masses depicted on current images and the corresponding regions on prior images were visually identified as confirmed by the diagnostic reports and pathology results. The centerpoint (x,y coordinate) of each verified mass region was marked manually and saved in a reference (or "truth") file.

All 520 images (260 current and 260 prior) were processed by a CAD scheme developed previously in our laboratory to identify and classify suspicious regions (15). The scheme includes three stages. First, it uses image subtraction and threshold results after processing by two Gaussian filters with a large difference in kernel sizes (7 pixels and 51 pixels) to search for the initial set of suspicious regions, a process that usually results in the identification of 10–30 suspicious regions per image. In the second stage, on the basis of local contrast measurement, the scheme uses an adaptive region growth algorithm to define three topographic layers for each region. Through the imposition of threshold conditions of growth ratio and shape factor for each layer in the regions identified as potential lesions, this stage eliminates approximately 85% of identified regions from consideration, while maintaining high sensitivity. A set of features is computed for each detected region. During the third stage, the remaining regions are classified according to scores generated by a nonlinear multilayer feature–based classifier, defining the likelihood of there being true-positive findings in those regions (16).

In this experiment, all remaining regions identified as suspicious mass regions after the second stage of the CAD scheme were selected for further consideration (the classification scores in the third stage were ignored). As a result, 726 suspicious regions on the 260 current images and 723 suspicious regions on the 260 prior images were selected. If the location of a selected region matched that of a verified mass, the region identification was considered true-positive. Specifically, the distance between the center of gravity of a region, as detected automatically by the CAD scheme, and the center of the mass, as recorded in the reference file, had to be shorter than the radius of

Number of Suspicious Mass Regions in Each Data Set

Images	Training Data Set		Testing Data Set	
	True-Positive	False-Positive	True-Positive	False-Positive
Current	131 (103)	233	130 (108)	232
Prior	126 (100)	236	126 (104)	235

Note.—Numbers in parentheses indicate the regions associated with malignant masses.

the longest axis of the detected region. Otherwise, the region was considered a false-positive identification.

The locations of 261 of the 726 selected regions on current images matched those of verified masses, compared with 252 of the 723 regions on the prior images. All true-positive and false-positive regions were then randomly divided into four mutually exclusive data sets, two for current images and two for prior images. To minimize potential bias, true-positive regions of the same mass (depicted on craniocaudal and mediolateral oblique views) were assigned to the same data set (either training or testing), and when a mass region was assigned to the training (or testing) subset in current images, its corresponding regions as depicted on prior images were also assigned to the training (or testing) subset. The Table summarizes the number and distribution of true-positive regions and false-positive regions in each of the four data sets.

Training data sets from the current and prior images were used to optimize two feature-based ANNs independently as substitutes for the third stage in our CAD scheme (16). Previous studies have demonstrated that the feature distributions were different for mass regions depicted on current images and those depicted on prior images and that different feature sets should be used for optimal classification results (17,18). Therefore, we applied a genetic algorithm to search separately for optimal sets of features on current images and on prior images, using the genetic algorithm software and optimization protocol that had been used in our previous studies to optimize both Bayesian belief networks (19) and ANNs (20).

In brief, a binary coding method is applied to create a chromosome used in the genetic algorithm. Each extracted feature corresponds to a gene (that is, either to 0 or to 1). To determine the optimal number of neurons in the second (hidden) layer of the ANN, we include four additional genes in the chromosome. Hence, the chromosome has a fixed length of 40 genes, of which the first 36 represent extracted image features and the last four indicate the binary-coded number of hidden neurons (eg, 0101 is the code for five hidden neurons) (20). To set up initial parameters in the genetic algorithm software, we included a population size of 100 and assigned the crossover rate, the mutation rate, and the generation gap to 0.6, 0.001, and 1.0, respectively. To minimize overfitting and increase robustness of the ANN performance, we adopted a limited number of training iterations (1,000), as well as a large ratio between the momentum (0.8) and learning rate (0.01) in the ANN. The output of the ROCFIT software program (University of Chicago, Ill) (21) was interfaced with the fitness function of the genetic algorithm, and A_z values computed by the program were defined as fitness criteria in the genetic algorithm. The genetic algorithm was terminated when it either converged to the "highest" A_z value (with no further improvement accomplished in the new generation) or reached a predetermined number of generations (eg, 100).

Using this approach, we generated two optimal ANNs, each using a different training data set. ANN-1 was trained with the suspicious mass regions extracted solely from the current images, and ANN-2 was trained with regions extracted solely from the prior images. Then we applied each of the ANNs to the two mutually exclusive testing data sets of regions extracted from both current and prior images. The classification scores in each test were used to generate four receiver operating characteristic (ROC) curves. The four A_z values were compared. We defined the threshold as a false-positive detection rate similar to that of the leading commercial CAD products—approximately 0.4 false-positive mass regions per image (7). At this level, we found the corresponding detection sensitivity levels and computed the expected number of detected true-positive regions (130 in the data set of current images, and 126 in that of prior images). Thus, we compared the change in expected true-positive detection levels with the use of ANN-1 and ANN-2 for current and prior images at an operating point currently accepted in clinical CAD.

RESULTS

From the genetic algorithm and training data sets of current images and of prior images, two optimal ANNs were generated. ANN-1 included 13 features, and ANN-2 included 11 (Fig 1); four features were common to both. Many of the features are not orthogonal, which is not unique to our scheme. The highest A_z values achieved for the training data sets were 0.92 ± 0.01 for ANN-1 and

Figure 1. Features selected by means of the genetic algorithm for ANN-1 and ANN-2. Those in boldface are common to both ANNs.

ANN-1	ANN-2
1. **Region size (1st layer)**	1. **Region size (1st layer)**
2. Contrast (1st layer)	2. Minimum pixel value inside the region
3. Standard deviation of pixel values (2nd layer)	3. Size growth ratio between 2nd and 3rd layers
4. Circularity (2nd layer)	4. Skewness of pixel values (3rd layer)
5. Region size (3rd layer)	5. Standard deviation of pixel values in background
6. Contrast (3rd layer)	6. Region perimeter divided by size (3rd layer)
7. **Standard deviation of radial length (3rd layer)**	7. **Standard deviation of radial length (3rd layer)**
8. **Circularity (3rd layer)**	8. **Circularity (3rd layer)**
9. Ratio between the maximum and minimum radial lengths (3rd layer)	9. Skewness of pixel values of background
10. Difference of minimum pixel values inside and outside of the growth region (3rd layer)	10. Average local pixel value fluctuation (within a 5 x 5 frame) of the segmented breast area
11. **Region conspicuity (3rd layer)**	11. **Region conspicuity (3rd layer)**
12. Standard deviation of pixel values (3rd layer)	
13. Standard deviation of pixel values in the segmented breast area	

0.76 ± 0.02 for ANN-2. When ANN-1 was applied to the testing data sets, the A_z values were 0.89 ± 0.01 and 0.65 ± 0.02 for current and prior images, respectively. Figure 2 shows three ROC curves for training and two testing results. When ANN-2 was applied to the same data sets, the A_z values were 0.81 ± 0.02 for current and 0.71 ± 0.02 for prior images. Figure 3 shows the corresponding ROC curves for ANN-2.

The test results differed significantly ($P < .01$) between ANN-1 and ANN-2 for both the current and prior image testing data sets. As shown in Figure 4, A_z values were reduced by 9.0% (from 0.89 with ANN-1 to 0.81 with ANN-2) for the current testing data set and increased by 9.2% (from 0.65 to 0.71) for the prior testing data set. In addition, at an operating point of 0.4 false-positive detections per image, the sensitivity levels represented by the two ROC curves in Figure 2 are 0.82 and 0.40. In Figure 3, the corresponding sensitivity levels are 0.68 and 0.52. If we convert these levels to an expected number of detected true-positive mass regions, ANN-1 would detect 18 additional mass regions in the current testing data set, while ANN-2 would detect 15 additional mass regions in the prior testing data set.

The results are not substantially different when benign masses are excluded from the analysis. ANN-1 yielded performance levels (A_z) of 0.88 ± 0.02 and 0.63 ± 0.02 for current and prior images, respectively; the comparable values for ANN-2 were 0.81 ± 0.02 and 0.70 ± 0.03.

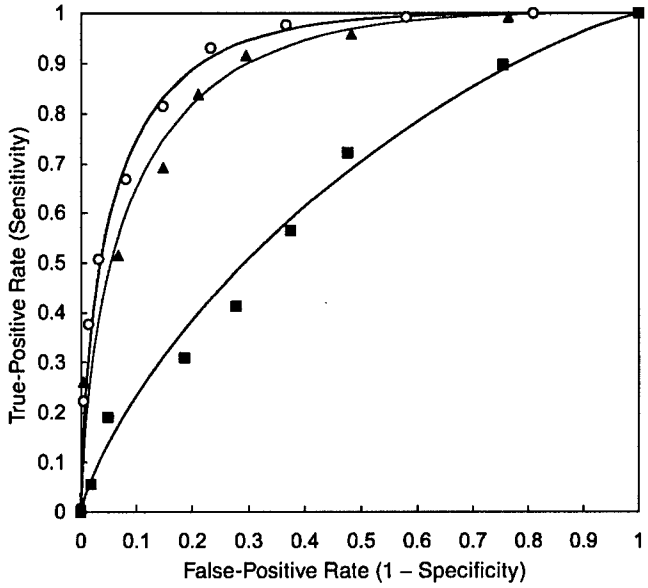

Figure 2. ROC curves showing the performance of ANN-1 during training with the current image data set (○) and during testing with the current image data set (▲) and the prior image data set (■).

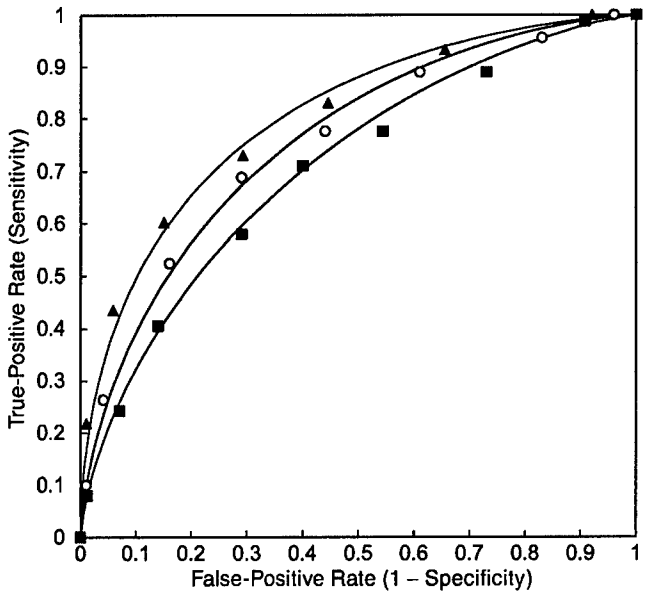

Figure 3. ROC curves showing the performance of ANN-2 during training with the prior image data set (○) and during testing with the current image data set (▲) and the prior image data set (■).

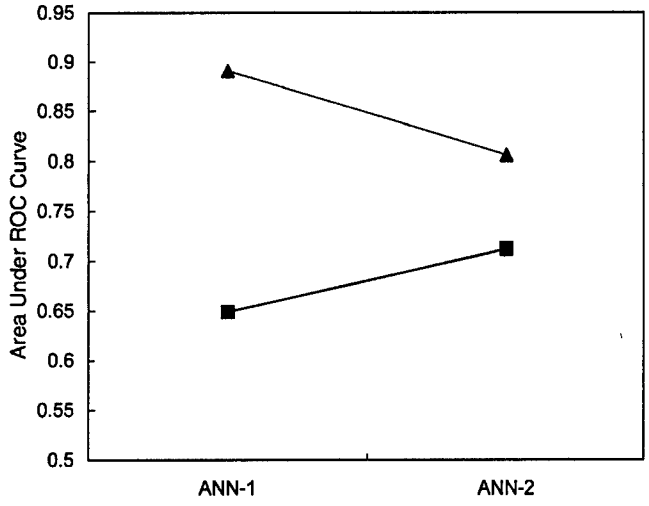

Figure 4. Differences in area under the ROC curve (A_z) for ANN-1 and ANN-2 when tested with the current image data set (▲) and the prior image data set (■).

DISCUSSION

Feature-based machine learning classifiers, such as ANNs, are widely used in CAD schemes as a final stage in identifying and classifying abnormalities. Since these classifiers are trained to generate a "global" function to cover the entire instance space (22), their performance depends heavily on the training databases. This is particularly true in mammography, for which the size and diversity of training data sets are often limited (23,24). Optimal feature sets such as those selected by the genetic algorithm could differ for different limited-size training databases. Hence, the features selected in this study for the current images were very similar but not identical to those selected in our previous studies (16,18). A single CAD scheme that achieves high sensitivity for both subtle and relatively easy-to-detect masses at an acceptable false-positive rate can be developed if a large and diverse image database is available. However, the creation of such a database is very difficult, because image features (including texture- and morphology-based features) are substantially different for suspicious mass regions extracted from current and prior images, as previous studies have demonstrated (17,18).

The CAD scheme trained with the current image data set did not perform optimally when tested with the prior image data set, and vice versa. On the one hand, it is important for a CAD scheme to detect more subtle masses, because most radiologists can identify the easily detected ones. On the other hand, users may lose confidence in a scheme if it frequently misses masses that should be easy to detect. Without such confidence, radiologists will most likely be reluctant to accept CAD cuing on subtle masses or make any changes in their initial interpretation (8), preventing the full benefit of CAD schemes from being realized in clinical environments. When ANN-2, which had been trained with the prior image data set, was tested with the current image data set, the testing results were better (higher A_z) than the training results, demonstrating the general robustness of the scheme (Fig 3).

Like most commercially available CAD systems, our CAD scheme was designed to detect, not classify, suspicious abnormalities. Therefore, we believe that the scheme should be highly sensitive to all suspicious mass regions considered "actionable" by radiologists (eg, recommended for follow-up or biopsy), even if some regions later prove benign. One of our previous studies suggested that radiologists' performance in classifying abnormalities as benign or malignant was not affected by the performance of CAD cuing for detection purposes (11). In any event, the inclusion of the benign mass regions as true-positive cases in this experiment did not affect our results and conclusions.

With improvements in diagnostic technology and increasing compliance with screening recommendations among women generally, radiologists have to detect increasingly subtle abnormalities depicted on mammograms.

As a result, the performance of a CAD system that initially provided satisfactory cuing results when optimized could deteriorate substantially over time. Therefore, it may be beneficial to update training data sets periodically and reoptimize the schemes by using a large fraction of new cases originally rated negative and later found positive. An alternative approach could be to provide two types of cues, one trained with current and one with prior images ("early signs"). We believe that our experimental results are not unique to our own image database, our CAD scheme, or ANN-based CAD schemes but should apply to all types of CAD schemes in which feature-based machine learning classifiers are used.

REFERENCES

1. Tabar L, Vitak B, Chen HH, et al. Beyond randomized trials: organized mammographic screening substantially reduces breast cancer mortality. Cancer 2001; 91:1724–1731.
2. Feig S. Increased benefit from shorter screening mammography intervals for women ages 40–49 years. Cancer 1997; 80:2035–2039.
3. Yankaskas BC, Schell MJ, Bird RE, Desrochers DA. Reassessment of breast cancers missed during routine screening mammography: a community-based study. AJR Am J Roentgenol 2001; 177:535–541.
4. Birdwell RL, Ikeda DM, O'Shaughnessy KF, Sickles EA. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. Radiology 2001; 219:192–202.
5. Malich A, Marx C, Facius M, Böhm T, Fleck M, Kaiser WA. Tumour detection rate of a new commercially available computer-aided detection system. Eur Radiol 2001; 11:2454–2459.
6. Lechner M, Nelson M, Elvecrog E. Comparison of two commercially available computer-aided detection (CAD) systems. Appl Radiol 2002; 31:31–35.
7. 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.
8. 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.
9. te Brake GM, Karssemeijer N, Hendriks JH. Automated detection of breast carcinomas not detected in a screening program. Radiology 1998; 207:465–471.
10. Malich A, Azhari T, Böhm 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.
11. Zheng B, Ganott MA, Britton CA, et al. Soft-copy mammographic reading with different computer-assisted detection cuing environments: preliminary findings. Radiology 2001; 221:633–640.
12. Feig SA, D'Orsi CJ, Hendrick RE. American College of Radiology guidelines for breast cancer screening. AJR Am J Roentgenol 1998; 171:29–33.
13. Bassett LW, Shayestehfar B, Hirbawi I. Obtaining previous mammograms for comparison: usefulness and cost. AJR Am J Roentgenol 1994; 163:1083–1086.
14. Callaway MP, Boggis CR, Astley SA. Influence of previous films on screening mammographic interpretation and detection of breast carcinoma. Clin Radiol 1997; 52:527–529.
15. 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.
16. 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.
17. Hadjiiski L, Sahiner B, Chan HP, Petrick N, Helvie MA, Gurcan M. Analysis of temporal changes of mammographic features: computer-aided classification of malignant and benign breast masses. Med Phys 2001; 28:2309–2317.
18. 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; 9:1245–1250.
19. Zheng B, Chang YH, Wang XH, Good WF, Gur D. Feature selection for computerized mass detection in digitized mammograms by using a genetic algorithm. Acad Radiol 1999; 6:327–332.
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. 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.
22. Towell G, Shavlik J. An approach to combining explanation-based and neural learning algorithms. Connection Sci 1989; 1:233–255.
23. 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.
24. Kupinski MA, Giger ML. Feature selection with limited datasets. Med Phys 1999; 26:2176–2182.

Subjective assessment of high-level image compression of digitized mammograms

J. Ken Leader*[a], Jules H. Sumkin[ab], Marie A. Ganott[ab], Christiane Hakim[ab], Lara Hardesty[ab], Ratan Shah[ab], Luisa Wallace[ab], Amy Klym[a], John M. Drescher[a], Glenn S. Maitz[a], David Gur[a]
[a]University of Pittsburgh, Pittsburgh, PA USA 15213
[b]Magee-Womens Hospital, Pittsburgh, PA USA 15213

ABSTRACT

This study was designed to evaluate radiologists' ability to identify highly-compressed, digitized mammographic images displayed on high-resolution, monitors. Mammography films were digitized at 50 micron pixel dimensions using a high-resolution laser film digitizer. Image data were compressed using the irreversible (lossy), wavelet-based JPEG 2000 method. Twenty images were randomly presented in pairs (one image per monitor) in three modes: mode 1, non-compressed versus 50:1 compression; mode 2, non-compressed versus 75:1 compression; and mode 3, 50:1 versus 75:1 compression with 20 random pairs presented twice (80 pairs total). Six radiologists were forced to choose which image had the lower level of data compression in a two-alternative forced choice paradigm. The average percent correct across the six radiologists for modes 1, 2 and 3 were 52.5% (+/-11.3), 58.3% (+/-14.7), and 58.3% (+/-7.5), respectively. Intra-reader agreement ranged from 10 to 50% and Kappa from -0.78 to -0.19. Kappa for inter-reader agreement ranged from -0.47 to 0.37. The "monitor effect" (left/right) was of the same order of magnitude as the radiologists' ability to identify the lower level of image compression. In this controlled evaluation, radiologists did not accurately discriminate non-compressed and highly-compressed images. Therefore, 75:1 image compression should be acceptable for review of digitized mammograms in a telemammography system.

Keywords: Image compression, data compression, JPEG 2000, telemammography

1. INTRODUCTION

Breast cancer screening mammography is widely practiced and increasingly challenging to manage in the clinical environment, but there is potential for improvement.[1-7] Teleradiology is an approach that may provide more timely patient management. Image compression,[8-13] image cropping,[12-14] and image selection[15] are commonly used in teleradiology to facilitate the timely transmission of data. The high-spatial resolution required for mammography complicates the design and implementation of a telemammography system. The large mammographic image file size (33-55 MBytes per image) is one obstacle to timely transmission of data, especially across low-level data connections. High-level image compression may assist in overcoming this obstacle and can only be realized with lossy image compression techniques, which necessitates the loss of some image information and a degree of image degradation.

The use of high-level image compression in medical applications is frequently met with skepticism because of the potential degradation of the depiction of objects under investigation. Human observer performance studies designed to evaluate wavelet compression of medical images for clinical applications have reported acceptable compression levels ranging from 8:1 to 100:1.[16-26] Wavelet-based compression, the trend in medical image compression, is reported to be superior to the original JPEG compression based on the direct cosine transform in terms of image quality at high-levels of image compression.[16,17]

* jklst3@pitt.edu; phone (412) 641-2572; fax (412) 641-2582, University of Pittsburgh, Magee-Womens Hospital, 300 Halket Street, Suite 4200, Pittsburgh, PA 15213

Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment, edited by Dev P. Chakraborty, Miguel P. Eckstein, Proceedings of SPIE Vol. 5372 (SPIE, Bellingham, WA, 2004) · 1605-7422/04/$15 · doi: 10.1117/12.533201

From our perspective the effect of image degradation from lossy compression of medical image interpretation remains unresolved, particularly regarding mammography. Observer studies reported that 8:1[22] and 10:1[27] compression ratios are acceptable for mammography applications using both wavelet and the original JPEG compression methods. Visualization of calcifications depicted on digitized mammograms was subjectively rated as excellent for wavelet compression ratios as high as 56:1.[19] Uncompressed digitized mammographic images were rated to be comparable to images compressed at 30:1 using wavelet compression.[20] These studies are indeed promising, and high-levels of image compression may be ultimately clinically acceptable in mammography.

Powell et al.[22] (2000) conducted a clinical evaluation that compared film mammography to digitized images compressed at 8:1 using wavelet based compression. The accuracy for detecting malignancy was not statistically different when depicted on film or digitized images in a receiver operating characteristics (ROC) study. The false positive rate at a fixed sensitivity of 0.90 was significantly lower (better) using digitized images as compared with film. Compressed digitized images were also slightly better (though not statistically) than film in terms of recall rate for negative mammograms and those depicting benign findings. The recall rate for mammograms depicting malignant abnormalities was slightly better (though not statistically) when original films were used as compared with digitized images.

The objective of this study was to determine an acceptable level of image compression in a telemammography application. The ability of radiologists to discriminate high-levels of image compression as applied to digitized mammograms was evaluated. Image pairs of different compression levels were randomly presented and viewed side-by-side on two high-resolution monitors. Six radiologists were forced to choose the lower level of image compression and rate the relative utility of the images for use in a screening mammography environment.

2. METHODS

2.1 Case selection

This study used twenty breast cancer screening examinations randomly selected from a larger telemammography project, which was designed to evaluate the ability telemammography to reduce the number of patients being recalled for additional imaging procedures. One image view from each case (i.e., twenty images total) was selected to represent each examination. The verified findings depicted in these examinations included masses and calcification clusters (Table 1). The dataset for this retrospective study was assembled and analyzed under University of Pittsburgh Institutional Review Board approved protocol, and the image data was anonymized.

Table 1
Image views and depicted abnormalities

	Abnormality depicted on image			
View	Mass	Calcifications	Mass & calcifications	No finding
MLO	3	2	2	3
CC	3	3	1	3

MLO - mediolateral oblique
CC - craniocaudal

2.2 Image processing

Mammographic films were digitized at 50 micron pixel dimensions and 12-bit grayscale using a high-resolution, laser film digitizer (Lumiscan 85, Eastman Kodak, Rochester, NY, USA). Each digitized mammographic image was automatically cropped to decrease the non-tissue area surrounding the breast. The cropped image data were compressed using the irreversible (lossy), 9/7 transform, wavelet-based JPEG 2000 method at compression ratios of 50:1 and 75:1 and subsequently decompressed prior to display. A total of sixty images were generated for the study, the twenty original digitized images plus two compressed images at 50:1 and 75:1 ratios for each of these (or a total of sixty images).

2.3 Image display

The images were displayed on two calibrated, high-resolution (2048 x 2560), 8-bit grayscale, portrait monitors at a nominal setting of 80 ftL (DS5100P, Clinton Electronics, Rockford, IL, USA). Typically, when a single image displayed on the monitor the display scale was approximately 100 micron per pixel. Minimal unsharp masking was employed. In short, image data were first smoothed with a 2-D 129 mean kernel, and subsequently the weighted (0.10) smoothed image was subtracted from the original image. Finally, the resulting pixel values were re-scaled from 0 to 4095. Image magnification and window/level adjustments were not permitted during the study.

Fixed look-up table (LUT) values are automatically calculated based on the pixel value distribution (histogram). In short, the typical pixel value distribution of digitized mammographic images is bimodal. The center between the two modes was set as the level value (brightness), and the span of the two modes was set as the window value (contrast). Additionally, the cropped images were padded (filled) prior to display to restore the full height of the image.

2.4 Study protocol

Six experienced radiologists participated in the study. They were presented image pairs (one image per monitor) that consisted of the same image at different levels of compression (Fig. 1). The images were paired in three modes: mode 1, non-compressed versus 50:1 compression; mode 2, non-compressed versus 75:1 compression; and mode 3, 50:1 versus 75:1 compression. The sixty image pairs were randomly presented with 20 randomly selected pairs presented a second time to evaluate intra-observer variability (or a total of eighty pairs). Compression levels were also randomly assigned between the two monitors for counterbalancing.

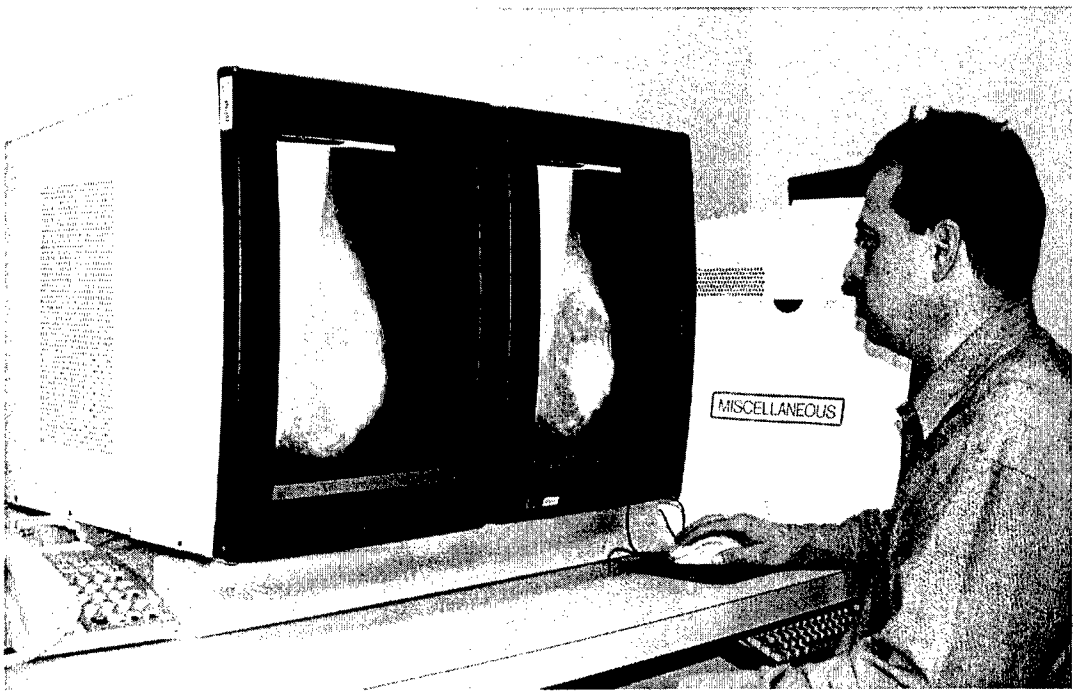

Fig. 1. Telemammography workstation used for the study.

In a 2-AFC paradigm the radiologists were forced to choose the image (i.e., right or left monitor) that had the lower level of data compression. In addition, they compared and rated the clinical utility between the two images presented in each pair. After image review, two questions were presented on a computer scoring form and answered using the computer mouse (Fig. 2). The radiologists were given written instructions regarding the protocol:

You will be presented with 80 pairs of images, one image on each monitor. The window and level values for the monitor display will be fixed. Magnification features will not be available during this study. One image will contain less information than the other as a result of data compression. The monitor that displays the less compressed image will be randomly selected. The same image pairs will appear multiple times throughout the study. After you have reviewed the images, the "eval case" button on the bottom task bar will bring up two questions to be answered.

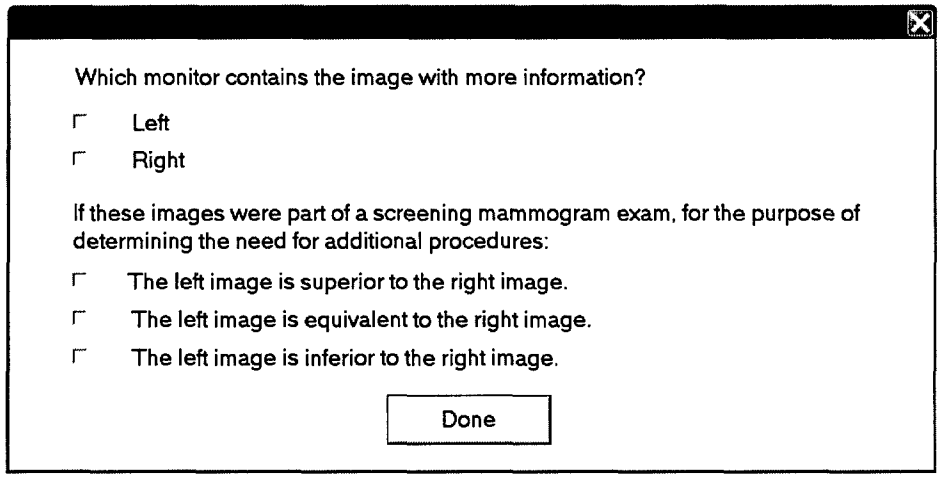

Fig. 2 Computerized scoring form completed for each image pair.

2.5 Data analysis

The average percent correct decisions across the six readers for discriminating the lower level of image compression was compared with a random (chance) selection using a one-sample T-test for each mode and each monitor. Friedman Two-Way Analysis of Variances by Ranks was used to test if there was a difference between modes. Kappa was used to evaluate intra-reader agreement for the twenty repeated pairs of images and inter-reader agreement for each mode. To determine if a learning effect was present the percent correct decision for the first, second, and third presentations of pairs of images was tested for trend using the Page Test for Ordered Alternatives. All images were presented a minimum of three times with the twenty repeated pairs randomly selected. The percent of image pairs rated as clinically equivalent for both the correct or incorrect decisions for identifying the lower level of image compression were compared to random (chance) selection using a one-sample T-test for each mode and each monitor.

3. RESULTS

The subjective appearance of the compressed images was extremely similar to the original uncompressed image. The task of discriminating the more compressed image in each pair was reported to be difficult by all readers. The smoothing effect of wavelet compression did not produce distinguishable image features such as blocking artifacts characteristic of high-level original JPEG compression.

Readers' ability to correctly discriminate the lower level of image compression was only slightly better than chance and was of the same order of magnitude as the "monitor effect" (Table 2). Readers' performance levels were not significantly different across the three presentation modes ($p > 0.05$). However, the readers correctly identified images compressed at 50:1 ratio as lower than 75:1 image compression at a rate significantly greater than chance ($p < 0.05$). On average the readers performed better when the lower level of compression was presented on the left monitor for all three modes, but the "monitor effect" (left versus right) was not significant.

Table 2
Average percent correct for discriminating the lower compression level for all image pairs when the correct image was on the right monitor and the left monitor

	mode 1[ad]	mode 2[b]	mode 3[c]
All images	52.5 (11.3)	58.3 (14.7)	58.3 (7.5)[e]
Images on right monitor	45.7 (25.3)	43.2 (25.8)	47.8 (26.5)
Images on left monitor	62.5 (14.1)	73.2 (25.3)	69.0 (23.1)

[a] mode 1 - non-compressed & 50:1 compression
[b] mode 2 - non-compressed & 75:1 compression
[c] mode 3 - 50:1 & 75:1 compression
[d] group mean and standard deviation in ( )
[e] $p < 0.05$ one sample T-test

Intra- and inter-reader agreements for discriminating the lower level of data compression were poor for the individual as well as between readers (Tables 3 and 4). Kappa for intra-reader agreement for readers 1, 2, 3, 4, 5, and 6 were -0.25, -0.39, -0.30, -0.19, -0.78, and -0.30, respectively. No two readers consistently agreed across the three presentation modes. Inter-reader Kappa for discriminating the lower level of image compression for the six readers ranged from -0.47 to 0.26, -0.36 to 0.37, and -0.30 to 0.30 for modes 1, 2, and 3, respectively (Table 4).

Table 3
Comparison between the first and second reads of the twenty repeated image pairs

		second read	
reader	first read	correct[a]	incorrect
1	correct	10 (2)	30 (6)
	incorrect	30 (6)	30 (6)
2	correct	15 (3)	30 (6)
	incorrect	40 (8)	15 (3)
3	correct	20 (4)	30 (6)
	incorrect	35 (7)	15 (3)
4	correct	5 (1)	25 (5)
	incorrect	25 (5)	45 (9)
4	correct	5 (1)	40 (8)
	incorrect	50 (10)	5 (1)
5	correct	10 (2)	50 (10)
	incorrect	20 (4)	20 (4)

[a] percentage and number in ( )

Table 4
Kappa for inter-reader agreement for the six readers and the three presentation modes

		reader				
mode	reader	2	3	4	5	6
1[a]	1	-0.471	-0.042	0.043	-0.200	-0.038
	2		0.118	0.223	-0.100	-0.237
	3			0.255	-0.200	0.151
	4				-0.100	-0.101
	5					-0.300
2[b]	1	0.175	-0.359	-0.354	-0.300	0.368
	2		-0.284	-0.023	0.100	-0.177
	3			0.018	0.100	-0.217
	4				-0.200	0.125
	5					-0.300
3[c]	1	-0.099	-0.300	0.121	0.100	-0.237
	2		-0.100	-0.099	0.100	0.175
	3			0.100	-0.200	0.100
	4				0.300	-0.031
	5					-0.100

[a] mode 1 - non-compressed & 50:1 compression
[b] mode 2 - non-compressed & 75:1 compression
[c] mode 3 - 50:1 & 75:1 compression

A slight learning effect was observed in the average reader's ability to select the lower level of image compression during the first three presentations (Table 5). The mean percent for correctly discriminating the lower level of image compression showed an increasing trend across the three presentations that was not significant ($p > 0.05$). Reader 6 was an outlier, and, although the trend was not significant, excluding this reader from the analysis removed the increasing trend across the three presentations.

Table 5
Percent correct for selecting the less compressed image during the first, second, and third presentations

reader	first (n= 20)	second (n = 20)	third (n = 20)
1	65.0	50.0	60.0
2	55.0	55.0	60.0
3	50.0	50.0	55.0
4	60.0	65.0	70.0
5	50.0	50.0	50.0
6	35.0	80.0	65.0
mean	52.5	58.3	60.0[a]
std	10.4	12.1	7.1

[a] $p > 0.05$

Images correctly identified as less compressed by the readers were rated as "clinically equivalent" at relatively the same rate as images incorrectly identified (Table 6). However, on the left monitor the readers rated correctly selected images as "clinically equivalent" more often than random selection ($p < 0.05$). The average number of image pairs rated as clinically equivalent by the six radiologist were 14.2 (± 4.8), 14.2 (± 4.1), and 13.3 (± 5.5) out of the twenty possible pairs for modes 1, 2, and 3, respectively.

Table 6
Percent of image pairs rated "clinically equivalent" for correct and incorrect selection of lower compression level for either monitor, the right monitor, and the left monitor

	correct choice of lower compression level			incorrect choice of lower compression level		
mode	either monitor[d]	right monitor	left monitor	either monitor	right monitor	left monitor
1[a]	48.3 (20.1)	24.0 (13.6)	24.3 (15.2)	51.7 (20.1)	34.2 (24.9)	17.5 (10.5)
2[b]	62.3 (19.2)	18.9 (15.6)	43.4 (17.8)[e]	37.7 (19.2)	26.3 (15.9)	11.4 (12.9)[e]
3[c]	53.1 (18.6)	19.2 (15.4)	33.9 (17.0)	46.9 (18.6)	27.1 (20.7)	19.8 (14.7)

[a] mode 1 - non-compressed & 50:1 compression
[b] mode 2 - non-compressed & 75:1 compression
[c] mode 3 - 50:1 & 75:1 compression
[d] group mean and standard deviation in ( )
[e] $p < 0.05$ one sample T-test

4. DISCUSSION

In this controlled evaluation, image compression achieved with wavelet-based JPEG 2000 was not reliably discriminated and rated by radiologists and, therefore, could be considered applicable for telemammography applications. Radiologists did not accurately or reliably select the lower level of image compression between image pairs when presented side-by-side with non-compressed images and those compressed at 50:1 and 75:1 compression levels. Interestingly, the "monitor effect" (left versus right) was of the same order of magnitude as the radiologists' ability to discriminate the lower level of image compression. As a group the readers' ability to identify the lower level of data compression slightly improved across the readings, but not significantly. The majority of image pairs, which were compressed at different ratios, were rated as "clinically equivalent" for use in a screening environment independent of whether the readers selected correctly or incorrectly the less compressed image.

The images in our study were presented on separate, side-by-side monitors with magnification, pan zoom, and window/level features disabled. Permitting magnification and window/level may (or may not) have improved discrimination. A similar 2-AFC study by Slone et al.[17] (2000) evaluated wavelet and original JPEG compression of

posteroanterior chest digital radiographs and reported that image degradation was detected at compression levels greater than 11:1 for both compression methods. At a compression level of 75:1 the lower compressed image was correctly identified approximately 95 % of the time for both the wavelet and the JPEG compression methods. The images were presented on a single monitor, and the readers were permitted to magnify and toggle between images, which they acknowledged was conservative and tested the reader's temporal sensitivity.

Since radiologists could not accurately or reliably discriminate non-compressed and highly-compressed mammographic images, their interpretation using either non-compressed or highly-compressed images is not likely to differ substantially. We also note that diligent monitor calibration may be critical to image fidelity.

ACKNOWLEDGEMENTS

This work is supported in part by the US Army Medical Research Acquisition Center, 820 Chandler Street, Fort Detrick, MD 21702-5014 under contract DAMD17-00-1-0410. The content of the information contained herein does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

REFERENCES

1. Chamot E and Perneger TV. "Misconception about efficacy of mammography screening: a public health dilemma." *J Epidemiol Community Health* **55(11)**:799-803, 2001.
2. Coughlin SS, Thompson TD, Hall HI, Logan P, Uhler RJ. Breast and cervical carcinoma screening practices among women in rural and nonrural areas of the United States, 1998-1999. *Cancer* **94(11)**:2801-2812, 2002.
3. Michaelson J, Satija S, Moore R, Weber G. Halpern E, Garland A, Puri D, Kopans DB. "The pattern of breast cancer screening utilization and its consequences." *Cancer* **94(1)**:37-43, 2002.
4. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. "Variability in radiologists' interpretations of mammograms." *N Engl J Med* **331(22)**:1493-1499, 1994.
5. Warren RML and Duffy SW. "Comparison of single reading with double reading of mammograms and change in effectiveness with experience." *Br J Radiol* **68(813)**:958-962, 1995.
6. Hulka CA, Slanetz PJ, Halpern EF, Hall DA, McCarthy KA, Moore R, Boutin S, Kopans DB. "Patients' opinion of mammography screening services: immediate results versus delayed results due to interpretation by two observers." *AJR Am J Roentgenol* **168**:1085-1089, 1997.
7. Yawn B, Krein S, Christianson J, Hartley D, Moscovice I. "Rural radiology: who is producing images and who is reading them?" *J Rural Health* **13(2)**:136-144, 1997.
8. Bolle SR, Sund T, Stormer J. "Receiver operating characteristic study of image processing for teleradiology and digital workstations." *J Digit Imaging* **10(4)**:152-157, 1997.
9. Maitz GS, Chang TS, Sumkin JH, Wintz PW, Johns CM, Ganott M, Holbert BL, Hakim CM, Harris KM, Gur D, Herron JM. "Preliminary clinical evaluation of a high-resolution telemammography System." *Invest Radiol* **32(4)**:236-240, 1997.
10. Mitra S, Yang S, Kustov V. "Wavelet-based vector quantization for high-fidelity compression and fast transmission of medical images." *J Digit Imaging* **11(4 Suppl 2)**:24-30, 1998.
11. Kalyanpur A, Neklesa VP, Taylor CR, Daftary AR, Brink JA. "Evaluation of JPEG and wavelet compression of body CT images for direct digital teleradiology transmission." *Radiology* **217(3)**:772-779, 2000.
12. 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." *Proceedings of SPIE Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation* **4685**:416-421, February 2002.
13. Drescher JM, Maitz GS, Traylor C, Leader JK, Clearfield RJ, Shah R, Ganott MA, Pugliese F, Duffner D, Lockhart J, Gur D. "A multi-site telemammography system: preliminary assessment of technical and operational issues." *Proceedings of SPIE Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation*, **5033**:360-369, February 2003.
14. Lou SL, Lin HD, Lin KP, Hoogstrate D. "Automatic breast region extraction from digital mammograms for PACS and telemammography applications." *Comput Med Imaging Graph* **24(4)**:205-220, 2000.
15. Ludwig K, Bick U, Oelerich M, Schuierer G, Puskas Z, Nicolas K, Koch A, Lenzen H. "Is image selection a useful strategy to decrease the transmission time in teleradiology? A study of 100 emergency cranial CTs." *Eur Radiol* **8(9)**:1719-1721, 1998.

16. Ricke J, Maass P, Hänninen EL; Liebig T, Amthauer H, Stroszczynski C, Schauer W, Boskamp T, Wolf M. "Wavelet versus JPEG (Joint Photographic Expert Group) and fractal compression: impact on the detection of low-contrast details in computed radiographs." *Invest Radiol* **33(8)**:456-463, 1998.
17. Slone RM, Foos DH, Whiting BR, Muka E, Rubin DA, Pilgram TK, Kohm KS, Young SS, Ho P, Hendrickson DD. "Assessment of visually lossless irreversible image compression: comparison of three methods by using an image-comparison workstation." *Radiology* **215**:543-553, 2000.
18. Goldberg MA, Pivovarov M, Mayo-Smith WW, Bhalla MP, Blickman JG, Bramson RT, Boland GW, Llewellyn HJ, Halpern E. "Application of wavelet compression to digitized radiographs." *AJR Am J Roentgenol* **163**:463-468, 1994.
19. Lucier BJ, Kallergi M, Qian W, DeVore RA, Clark RA, Saff EB, Clarke LP. "Wavelet compression and segmentation of digital mammograms." *J Digit Imaging* **7(1)**:27-38, 1994.
20. Perlmutter SM, Cosman PC, Gray RM, Olshen RA, Ikeda D, Adams CN, Betts BJ, Williams MB, Perlmutter KO, Li J, Aiyer A, Fajardo L, Birdwell R, Daniel BL. "Image quality in lossy compressed digital mammograms." *Signal Processing* **59**:189-210, 1997.
21. Persons KR, Palisson PM, Manduca A, Charboneau WJ, James EM, Charboneau NT, Hangiandreou NJ, Erickson BJ. "Ultrasound grayscale image compression with JPEG and wavelet techniques." *J Digit Imaging* **13(1)**:25-32, 2000.
22. Powell KA, Mallasch PG, Obuchowski NA, Kerczewski RJ, Ganobcik SN, Cardenosa G, Chilcote W. "Clinical evaluation of wavelet-compressed digitized screen-film mammography." *Acad Radiol* **7(5)**:311-316, 2000.
23. Terae S, Miyasaka K, Kudoh K, Nambu T, Shimizu T, Kaneko K, Yoshikawa H, Kishimoto R, Omatsu T, Fujita N. "Wavelet compression on detection of brain lesions with magnetic resonance imaging." *J Digit Imaging* **13(4)**:178-179, 2000.
24. Trapnell CJ, Scarfe WC, Cook JH, Silveira AM, Regennitter FJ, Haskell BS. "Diagnostic accuracy of film-based, TIFF, and wavelet compressed digital temporomandibular joint images." *J Digit Imaging* **13(1)**:38-45, 2000.
25. Zheng LM, Sone S, Itani Y, Wang Q, Hanamura K, Asakura K, Li F, Yang ZG, Wang JC, Funasaka T. "Effect of CT digital image compression on detection of coronary artery calcification." *Acta Radiol* **41(2)**:116-121, 2000.
26. Zalis ME, Hahn PF, Arellano RS, Gazelle GS, Mueller PR. "CT colonography with teleradiology: effect of lossy wavelet compression on polyp detection-initial observations." *Radiology* **220**:387-392, 2001.
27. Good WF, Maitz GS, Gur D. "Joint Photographic Experts Group (JPEG) compatible data compression of mammograms." *J Digit Imaging* **7(3)**:123-132, 1994.

APPENDIX 11

Computer-Aided Detection Schemes: The Effect of Limiting the Number of Cued Regions in Each Case

Bin Zheng[1]
Joseph K. Leader
Gordon Abrams
Betty Shindel
Victor Catullo
Walter F. Good
David Gur

OBJECTIVE. We assessed performance changes of a mammographic computer-aided detection scheme when we restricted the maximum number of regions that could be identified (cued) as showing positive findings in each case.

MATERIALS AND METHODS. A computer-aided detection scheme was applied to 500 cases (or 2,000 images), including 300 cases in which mammograms showed verified malignant masses. We evaluated the overall case-based performance of the scheme using a free-response receiver operating characteristic approach, and we measured detection sensitivity at a fixed false-positive detection rate of 0.4 per image after gradually reducing the maximum number of cued regions allowed for each case from seven to one.

RESULTS. The original computer-aided detection scheme achieved a maximum case-based sensitivity of 97% at 3.3 false-positive detected regions per image. For a detection decision score set at 0.565, the scheme had a 79% (237/300) case-based sensitivity, with 0.4 false-positive detected regions per image. After limiting the number of maximum allowed cued regions per case, the false-positive rates decreased faster than the true-positive rates. At a maximum of two cued regions per case, the false-positive rate decreased from 0.4 to 0.21 per image, whereas detection sensitivity decreased from 237 to 220 masses. To maintain sensitivity at 79%, we reduced the detection decision score to as low as 0.36, which resulted in a reduction of false-positive detected regions from 0.4 to 0.3 per image and a reduction in region-based sensitivity from 66.1% to 61.4%.

CONCLUSION. Limiting the maximum number of cued regions per case can improve the overall case-based performance of computer-aided detection schemes in mammography.

Computer-aided detection systems are routinely used in a number of medical institutions around the world to assist radiologists in the detection of abnormalities depicted on mammograms. The number of mammograms scanned through commercial computer-detection systems has been rapidly increasing. Although no general agreement has been reached on how computer-aided detection affects radiologists' performance in terms of sensitivity and specificity [1–4], there are indications that the performance of the computer-aided detection scheme itself has an impact on radiologists' performance in detecting abnormalities [5, 6], and observer confidence levels in accepting the cues generated by these systems increases with higher performance levels of the scheme [7, 8]. Several commercial computer-aided detection systems have been approved by the United States Food and Drug Administration, and the relative performance levels of such systems have been compared [9, 10]. All commercial computer-aided detection systems use specific threshold values to determine whether an identified suspicious region is ultimately cued as a positive finding, and the performance of these systems is frequently evaluated on the basis of the case-based sensitivity achieved at a given false-positive detection rate. In a case-based (or a breast-based) analysis, sensitivity is based on the correct detection of at least one true-positive region on either the craniocaudal or mediolateral oblique mammographic view or on both [1].

Evaluation of computer-aided detection performance is not a simple matter. Previous studies have shown that performance can vary widely depending on which scoring method is used, and there is no general agreement on which scoring method should be used for this purpose [11, 12]. One study showed that at approximately the same false-positive rate (e.g., 1.5 per image), the

Received May 7, 2003; accepted after revision September 11, 2003.

The information contained in this article does not necessarily reflect the position or the policy of the United States government, and no official endorsement should be inferred.

Supported in part by grants CA85241, CA77850, and CA80836 from the National Cancer Institute of the National Institutes of Health, and by contract DAMD17-00-1-0410 from the United States Army Medical Research Acquisition Center at Fort Detrick, MD.

[1]All authors: Department of Radiology, Imaging Research, Magee-Women's Hospital, University of Pittsburgh, 300 Halket St., Ste. 4200, Pittsburgh, PA 15213-3180. Address correspondence to B. Zheng (zhengb@msx.upmc.edu).

AJR 2004;182:579–583

0361–803X/04/1823–579

measured sensitivity for the detection of microcalcification clusters ranged between 45% and 85% depending on which of three different assessment methods were used [11].

In addition, computer-aided detection performance depends on the composition of the image database used [13]. In general, computer-aided detection schemes may identify a large number of suspicious regions on some images (e.g., images depicting dense tissue patterns), but only a few suspicious regions on other images (e.g., images dominated by fatty tissue) [14]. Therefore, limiting the maximum number of suspicious regions allowed to be cued for one case could potentially reduce the false-positive rate with a relatively small decrease in sensitivity. This approach is used in commercially available systems, but to the best of our knowledge, the effect of implementing the approach on image- and case-based sensitivity and false-positive detection rates has not been described in detail. This study was performed to assess this issue.

Materials and Methods

We selected 500 cases (or 2,000 digitized mammograms) from a large image database available in our laboratory. Among these cases, verified malignant masses were depicted in 300 cases, and the remaining 200 were negative findings. In all cases with positive findings, a panel of radiologists identified the locations of the mass regions on the images using the original diagnostic and biopsy reports. The central coordinates (x and y) of each mass region were visually identified, marked, and saved in a "truth file." In this data set, mass regions were visible on both the craniocaudal and mediolateral oblique mammographic views in 270 cases and were only visible on one of the two views in 30 cases. Thus, 570 mass regions were identified on the images in this study. Figure 1 shows the size distribution of the 300 masses in the data set.

A computer program determined the size of each mass region by counting the total number of pixels inside the identified boundary contour of the region (multiplied by 0.0016 cm^2 per pixel). The size of a mass was represented by a large computed area on either the craniocaudal or mediolateral oblique mammogram. For each identified mass region, the panel of radiologists assigned a subjective rating of subtlety using a 5-point rating scale that ranged from 1 (very easily visible) to 5 (very subtly visible). Figure 2 shows the distribution of assigned subtlety ratings in this data set. Subtlety of a mass was represented by the lower rating assigned to either the craniocaudal or mediolateral oblique mammographic view. We verified all cases with negative (or benign) findings by reviewing the available diagnostic information and the data from a follow-up examination with negative results, confirming a minimum of one disease-free year.

A computer-aided detection scheme developed previously in our laboratory [15] was applied to the 2,000 images in the data set. Because we only examined computer-aided detection performance for mass detection in this study, each image was first reduced by pixel averaging (a factor of 8 in both x and y directions), increasing the effective pixel size from 50×50 μm in the original digitized image to 400×400 μm. The mass detection scheme then identified between 10 and 30 suspicious regions in each image depending on the regional tissue patterns. For each identified region, a multilayer regional growth algorithm [16] was applied to define the contours of the region as depicted in the image. If the region met simple growth criteria, a set of features from the interior and surrounding background of the region was computed by the scheme. Otherwise, the region was considered to have negative findings and was deleted. Finally, a feature-based artificial neural network classified each suspicious region as showing positive or negative findings by assigning a detection (or probability) score. In a manner similar to the commercial computer-aided detection products, our detection scheme identified a region as having a positive finding if the detection score exceeded a predetermined threshold. If the detection score did not exceed the threshold, the region was not cued and was considered to be a negative finding.

After processing all images, we compared the regions with detected positive findings with the results saved in the truth file. To determine whether a detected region was considered a true-positive finding, we applied the following criterion: If the distance between the computed center of a detected region and the visually marked coordinate on a mammogram was shorter than the effective radius (the average radial length computed by the computer-aided detection scheme), the region was considered to be a match to a true-positive mass. Otherwise, the region was considered a false-positive case.

To show the original performance of the computer-aided detection scheme when applied to this data set, we plotted free-response receiver operating characteristic curves for both case-based and region-based scores. In the case-based performance curve, sensitivity was assessed on the basis of the correct marking of at least one true-positive region in either (or both) of the two mammographic views, and if two regions were detected, the higher score was se-

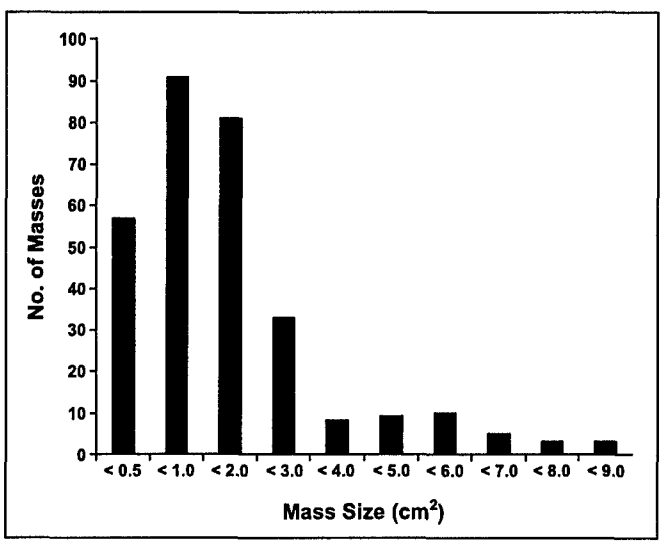

Fig. 1.—Bar graph shows size distribution of 300 masses depicted in data set. Mass size is represented by larger depicted area on either craniocaudal or mediolateral oblique mammographic view.

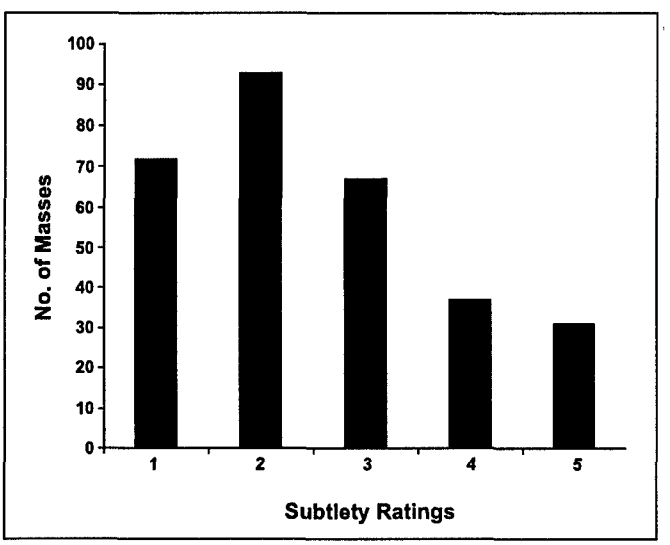

Fig. 2.—Bar graph shows distribution of subjectively rated subtlety of 300 masses depicted in data set. Subtlety of each identified mass was rated on 5-point scale, ranging from 1 (very easily visible) to 5 (very subtly visible). Mass subtlety is represented by lower-rated depiction on either craniocaudal or mediolateral oblique mammographic view.

lected to represent the mass. In the region-based performance curve, if the same mass was depicted on both craniocaudal and mediolateral oblique views, we considered these two images to represent two independent regions.

We applied a threshold score to the artificial neural network results to evaluate the sensitivity of the scheme at different false-positive rates. We also adjusted the threshold value to produce a false-positive rate comparable to that of the leading commercial computer-aided detection systems (e.g., a false-positive rate of 0.4 regions per image [2]). By changing the total number of cued regions permitted in each case to anywhere from seven to one, we compared the change in performance levels (including both sensitivity and false-positive rate). The scores generated by the artificial neural networks for all detected regions were sorted by value from the highest to the lowest, and the regions with higher scores were selected sequentially until the predetermined limit of cued regions per case was reached. In addition, we kept the case-based sensitivity constant by reducing the detection threshold and assessed the changes in false-positive rates and image-based sensitivity as the total number of allowed cues per case was reduced from seven to two.

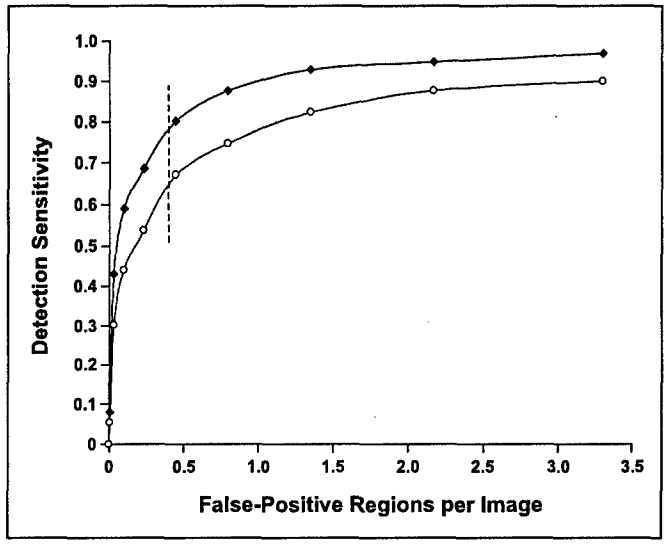

Fig. 3.—Graph illustrates overall performance of computer-aided detection scheme when applied to database of 2,000 mammograms (500 cases) with no limitation on number of cued regions. Detection decision threshold line is represented by dotted line. ◆ = case-based free-response receiver operating characteristic curve, O = image-based free-response receiver operating characteristic curve.

Results

Figure 3 shows two computed free-response receiver operating characteristic curves after the application of our computer-aided detection scheme to this data set. One is a case-based free-response receiver operating characteristic performance curve; the other is a region-based curve. Setting the threshold value of the artificial neural network detection scores at 0.565 generated a decision threshold line, as shown in Figure 3. At this level, the computer-aided detection scheme identified 79% of the malignant masses with 0.4 false-positive regions per image being cued. At this threshold, the scheme did not detect any false-positive regions in 33.2% (166/500) of the cases.

Table 1 provides the performance levels of the computer-aided detection scheme when we limited the maximum number of cued regions allowed in one case at this threshold level (0.565). The false-positive detection rate decreased substantially faster than the case-based sensitivity. For example, when we limited the maximum number of cued regions to two per case, the detection sensitivity decreased by 7.2% (from 237/300 to 220/300 cases), whereas the false-positive detection rate decreased by 47.3% (from 0.40 to 0.21 per image). In 65% of the true-positive cases, the region with the highest artificial neural network score was the malignant mass region (Table 1).

Figure 4 shows five free-response receiver operating characteristic curves generated when the maximum allowed number of cues per case was limited to between seven and two. As the maximum number of allowed cues was reduced, the free-response receiver operating characteristic curves tended to become steeper. Table 2 summarizes the results after limiting the maximum number of cued regions and changing the threshold value of the artificial neural network detection scores to maintain a 79% case-based sensitivity. The table shows that we were able to reduce the false-positive rates while maintaining a constant sensitivity. For example, by limiting the maximum allowed number of cues to two per case and adjusting the artificial neural network threshold to 0.36, we reduced the false-positive rate from 0.4 to 0.3 regions per image.

One interesting finding was that the 17 (of the 237) masses detected using these two scoring methods were not identical. When the maximum number of cued regions was limited to two per case, 17 masses with artificial neural network scores higher than 0.565 (range, 0.57–0.77) were eliminated. Reducing the

TABLE 1 **Performance Levels of Computer-Aided Detection as a Function of the Maximum Number of Cued Regions Allowed per Case**

Maximum No. of Cued Regions Allowed per Case	Sensitivity[a]				False-Positive Regions[b]	
	Case-Based		Region-Based			
	No.[c]	%	No.[d]	%	No.	Per-Image Rate
No limit	237	79.0	377	66.1	803	0.40
7	237	79.0	376	66.0	795	0.40
5	236	78.7	370	64.9	753	0.38
4	233	77.7	364	63.9	695	0.35
3	227	75.7	351	61.6	588	0.29
2	220	73.3	316	55.4	423	0.21
1	195	65.0	195	34.2	224	0.11

Note.—Artificial neural network threshold vaue was set at 0.565.

[a]Detected true-positive cases.

[b]Detected false-positive regions.

[c]Cases.

[d]Regions.

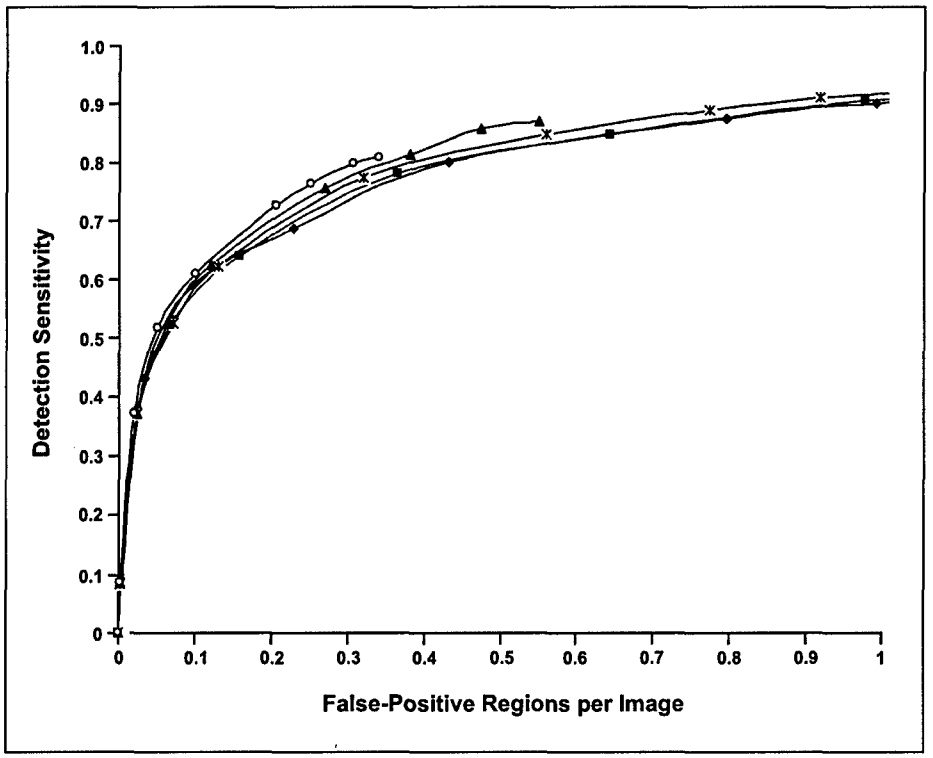

Fig. 4.—Graph shows five plots depicting free-response receiver operating characteristic curves generated by different maximum numbers of cued regions allowed per case. Maximum number of cued regions indicated by ◆ = no limit, ■ = ≤7, ✳ = ≤5, ▲ = ≤3, ○ = ≤2.

threshold score to 0.36 resulted in the identification of 17 different masses with artificial neural network scores in the range between 0.36 and 0.51. Figure 5 shows the distribution of mass sizes and subtlety ratings of the 34 masses missed by both scoring methods. The results suggest that the 17 masses that were detected only when the number of allowed cues was limited to two per case and the threshold was lowered tended to be somewhat small. All 34 masses were actually positive findings. At this time, the follow-up period on these patients has not been long enough to assess the difference (if any) in clinical impact of the two approaches.

Discussion

Case distributions and rating methods could have a significant effect on the evaluation of computer-aided detection performance levels [11–13]. In this study, we tested a simple scoring method that alters measured performance. The method of limiting the maximum number of cued regions allowed per case is commonly used in commercial computer-aided detection products. However, the actual scores for each region are not available to users. Therefore, several related issues—such as the effect of this approach on overall performance and on the detection (or the missed detection) of specific masses—have not, to our knowledge, been described in the past.

Our study showed that by limiting the maximum number of allowed regions to be cued in each case, a substantial fraction of false-positive regions can be eliminated with only a small decrease in sensitivity. If one wishes to maintain sensitivity, threshold values can be appropriately adjusted for this purpose. Because most masses were visible on both the craniocaudal and mediolateral oblique mammograms and because the detection performance of computer-aided detection systems is commonly evaluated using case-based sensitivity, our results are quite encouraging. It appears that this approach could reduce the false-positive detection rate of the scheme and possibly eliminate some true-positive region-based detections while retaining the initial (unrestricted number of cues) case-based sensitivity. Although the sensitivity can be maintained using this approach (changing the threshold levels for detection), one does not detect exactly the same true-positive masses. We found that limiting the maximum number of cues allowed per case and adjusting the thresh-

TABLE 2 **Performance Levels of Computer-Aided Detection with Constant Sensitivity of 79% as a Function of the Maximum Number of Cued Regions Allowed per Case**

Maximum No. of Cued Regions Allowed per Case	Region-Based Sensitivity[a]		False-Positive Rate[b]		Detection Decision Value of Artificial Neural Network Scores
	No.[c]	%	No.	Per-Image Rate	
No limit	377	66.1	803	0.40	0.565
5	371	65.1	773	0.39	0.560
4	378	66.3	902	0.45	0.500
3	375	65.8	781	0.39	0.470
2	350	61.4	604	0.30	0.360

[a]Detected true-positive cases.

[b]Detected false-positive regions.

[c]Regions.

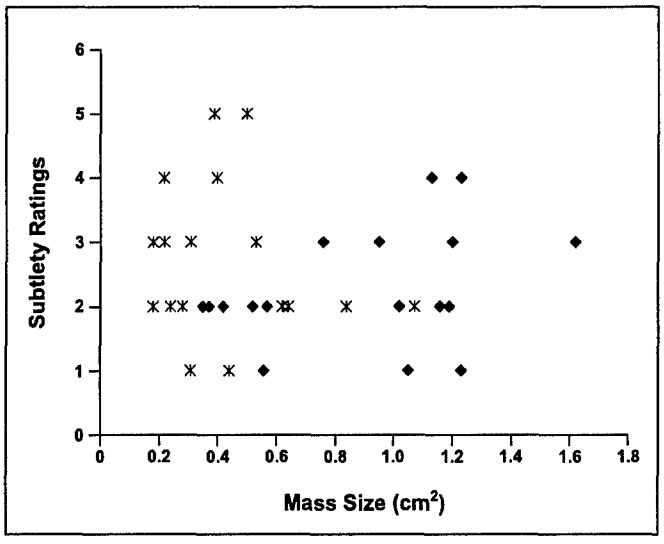

Fig. 5.—Scatterplot shows sizes and subtlety ratings distributions for 34 masses that were undetected by both case-based and image-based scoring methods. ◆ = no limit to number of regions in each case that may be cued as showing positive findings, ✳ = maximum number of regions that may be cued is ≤ 2.

old appropriately increased computer-aided detection sensitivity in the subset of smaller masses. In general, this effect is desirable in that it could reduce the number of regions that have to be ruled out by the radiologist. We caution that the use of this approach may not yield improvements of similar magnitude in the clinical environment with a substantially different distribution of truly positive and truly negative cases.

It should be noted that the size and subtlety ratings of masses in the data set were somewhat conservative. In Figures 1 and 2, we used the larger of the sizes computed for a mass from the two mammographic views and presented the less subtle rating for the same mass. Hence, distribution based on image or region would show a somewhat smaller average mass size and a more subtle data set.

Only malignant masses were considered true-positive identifications in this study. In visually assessing the false-positive regions with higher scores (e.g., > 0.7), we found that 19% (40/213) of these regions represented well-defined benign masses (i.e., round benign masses with high contrast and relatively sharp margins). Considering the detection of benign masses as either true-positive or false-positive may have a substantial impact on the evaluation of computer-aided detection performance levels. Because of the approach we used to reduce the number of cued regions per case and because of the size and diversity of the data set used, we believe that our results are not unique to our own computer-aided detection scheme.

References

1. Warren 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
2. 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
3. Brem RF, Schoonjans JM. Radiologist detection of microcalcifications with and without computer-aided detection: a comparative study. *Clin Radiol* 2001;56:150–154
4. Ciatto S, Turco MR, Risso G, et al. Comparison of standard reading and computer aided detection (CAD) on a national proficiency test of screening mammography. *Eur J Radiol* 2003;45:135–138
5. 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
6. Zheng B, Ganott MA, Britton CA, et al. Soft-display mammographic reading with different computer-assisted detection cueing environments: preliminary findings. *Radiology* 2001;221:633–640
7. Moberg K, Bjurstam N, Wilczek B, Rostgard L, Egge E, Muren C. Computer assisted detection of interval breast cancers. *Eur J Radiol* 2001;39:104–110
8. D'Orsi CJ. Computer-aided detection: there is no free lunch. *Radiology* 2001;221:585–586
9. Malich A, Marx C, Facius M, Boehm T, Fleck M, Kaiser WA. Tumor detection rate of a new commercially available computer-aided detection system. *Eur Radiol* 2001;11:2454–2459
10. Hoffmeister JW, Rogers SK, DeSimio MP, Brem R. Determining efficacy of mammographic CAD systems. *J Digit Imaging* 2002;15:198–200
11. Nishikawa RM, Yarusso LM. Variations in measured performance of CAD schemes due to database composition and scoring protocol. *Proc SPIE* 1998;3338:840–844
12. Kallergi M, Carney GM, Gaviria J. Evaluating the performance of detection algorithms in digital mammography. *Med Phys* 1999;26:267–275
13. Nishikawa RM, Giger ML, Doi K, et al. Effect of case selection on the performance of computer-aided detection schemes. *Med Phys* 1994;21:265–269
14. Zheng B, Chang YH, Gur D. Adaptive computer-aided diagnosis scheme of digitized mammograms. *Acad Radiol* 1996;3:806–814
15. 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
16. Zheng B, Chang YH, Gur D. Computerized detection of masses in digitized mammograms using single image segmentation and multilayer topographic feature analysis. *Acad Radiol* 1995; 2:959–966

ARTICLES

Changes in Breast Cancer Detection and Mammography Recall Rates After the Introduction of a Computer-Aided Detection System

David Gur, Jules H. Sumkin, Howard E. Rockette, Marie Ganott, Christiane Hakim, Lara Hardesty, William R. Poller, Ratan Shah, Luisa Wallace

Background: **Computer-aided mammography is rapidly gaining clinical acceptance, but few data demonstrate its actual benefit in the clinical environment. We assessed changes in mammography recall and cancer detection rates after the introduction of a computer-aided detection system into a clinical radiology practice in an academic setting.** ***Methods:*** **We used verified practice- and outcome-related databases to compute recall rates and cancer detection rates for 24 Mammography Quality Standards Act–certified academic radiologists in our practice who interpreted 115 571 screening mammograms with (n = 59 139) or without (n = 56 432) the use of a computer-aided detection system. All statistical tests were two-sided.** ***Results:*** **For the entire group of 24 radiologists, recall rates were similar for mammograms interpreted without and with computer-aided detection (11.39% versus 11.40%; percent difference = 0.09, 95% confidence interval [CI] = −11 to 11; $P = .96$) as were the breast cancer detection rates for mammograms interpreted without and with computer-aided detection (3.49% versus 3.55% per 1000 screening examinations; percent difference = 1.7, 95% CI = −11 to 19; $P = .68$). For the seven high-volume radiologists (i.e., those who interpreted more than 8000 screening mammograms each over a 3-year period), the recall rates were similar for mammograms interpreted without and with computer-aided detection (11.62% versus 11.05%; percent difference = −4.9, 95% CI = −21 to 4; $P = .16$), as were the breast cancer detection rates for mammograms interpreted without and with computer-aided detection (3.61% versus 3.49% per 1000 screening examinations; percent difference = −3.2, 95% CI = −15 to 9; $P = .54$).** ***Conclusion:*** **The introduction of computer-aided detection into this practice was not associated with statistically significant changes in recall and breast cancer detection rates, both for the entire group of radiologists and for the subset of radiologists who interpreted high volumes of mammograms.** **[J Natl Cancer Inst 2004;96:185–90]**

A mounting body of evidence suggests that early detection of breast cancer through periodic mammography screening reduces the morbidity and mortality associated with this disease *(1,2)*. Mammography screening is rapidly gaining acceptance worldwide, and the number of mammography procedures performed continues to increase *(3,4)*. However, mammography screening has a relatively low cancer detection rate of only two to six cancers per 1000 mammograms after the first 2 years of screening *(5)*.

The performance levels among radiologists who read and interpret mammograms vary widely. Several factors may account for this variability. These include, but are not limited to, the low incidence of breast cancer, the difficulty in identifying suspicious (i.e., potentially malignant) regions in the surrounding breast tissue, and the tedious and somewhat repetitious nature of the task of reading mammograms *(5–7)*.

In recent years, a major effort has been expended to develop computer-aided detection systems to assist radiologists with the diagnostic process. The hope is that these computer-aided detection systems will improve the sensitivity of mammography without substantially increasing mammography recall rates, in addition to possibly decreasing inter-reader variability. These systems are intended for the early detection of breast cancer and, accordingly, are designed to assist the radiologist in the identification (i.e., detection) of suspicious regions (i.e., findings), such as clustered microcalcifications and masses *(8–10)*. Computer-aided diagnosis (discrimination) systems are currently being developed to help radiologists determine whether an identified suspicious region is likely to represent a benign or a malignant finding *(11–13)*.

The U.S. Food and Drug Administration (FDA) has approved several computer-aided detection systems for clinical use, and Medicare and many insurance companies have approved reimbursement for the use of these systems in clinical practice. The initial FDA approval process for these systems included retrospective interpretations of select groups of cases in a laboratory environment *(9,14,15)*. Results of these studies *(9,15)* suggest that the use of computer-aided detection systems can potentially increase cancer detection rates by approximately 20% without substantially increasing recall rates. However, there are only limited data on the impact of such systems when used prospectively in a clinical environment *(16–19)*. We used large, prospectively ascertained databases to evaluate the recall and cancer detection rates in our clinical breast imaging practice in an

Affiliations of authors: Department of Radiology, University of Pittsburgh, and Magee-Womens Hospital of the University of Pittsburgh Medical Center, Pittsburgh, PA (DG, JHS, HER, MG, CH, LH, WRP, RS, LW); Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA (HER).

Correspondence to: David Gur, ScD, Imaging Research, Suite 4200, Department of Radiology, University of Pittsburgh, 300 Halket St., Pittsburgh, PA 15213 (e-mail: gurd@upmc.edu).

See "Notes" following "References."

DOI: 10.1093/jnci/djh067

academic setting for a 3-year period during which a computer-aided diagnosis system was introduced.

METHODS

Subjects and General Procedures

All screening mammography examinations performed in our facilities at Magee-Womens Hospital of the University of Pittsburgh Medical Center (Pittsburgh, PA) and its five satellite breast imaging clinics during 2000, 2001, and 2002 were included in this study. Our study was carried out under an institutional review board–approved protocol.

The data sources for our analysis were databases that contained information on procedure scheduling, procedure completion, radiology reporting, and procedure-related outcomes as determined from relevant pathology reports. These databases were assembled from the original reports for quality assurance purposes, as required by the Mammography Quality Standards Act (MQSA) *(20)*, among other reasons. The same computerized reporting system was in use throughout the study period.

In the second quarter of 2001, we introduced a computer-aided detection system (R2 Technologies, Los Altos, CA) into our clinical practice at the main facility, where most of the screening mammograms in our practice were read in batch mode. By the third quarter of 2001, more than 70% of the screening mammograms were interpreted with use of the computer-aided detection system. By the fourth quarter of 2001, more than 80% of the screening mammograms were interpreted with the assistance of the computer-aided detection system. The radiologists in our practice could not select which mammograms would be interpreted with or without the computer-aided detection system. After training on the computer-aided detection system was completed (June 2001), all screening mammograms interpreted in our main facility were processed by and interpreted with the assistance of the computer-aided detection system. Radiologists at the five satellite clinics sometimes reviewed screening mammograms if time allowed, but the number of these cases was small, and there was no selection process that could bias the analyses performed in this study. Knowing the schedule for radiologists' presence at the remote sites, we assembled a batch of serially acquired mammograms for them to read in the same way they would be read at the central facility, and those mammograms were interpreted and reported in the same manner (with the exception of the use of computer-aided detection). This set of mammograms was not specifically selected because of suspicious findings by the technologists. To reduce possible biases, an individual not involved in this investigation was asked to examine summaries of time-dependent recall rates for all radiologists in our practice for the study period. A different team examined all cancers detected throughout our practice as a result of screening mammography during the same period.

During the study period, our practice performed a total of 115 571 screening examinations that were interpreted by 24 radiologists, 18 of whom interpreted more than 1000 mammograms each. All radiologists were members of the Breast Imaging Section of the Department of Radiology and would be considered breast imaging specialists in an academic practice. We also repeated our analysis by using only data for the seven highest volume radiologists, all of whom read more than 8000 mammograms each over a 3-year period. These seven radiologists, who were with our institution throughout the study period, performed the most readings, both with and without computer-aided detection assistance.

For the purpose of computing recall rates, mammograms were considered to be positive if recall for additional imaging evaluation was recommended (i.e., mammograms classified as Breast Imaging Reporting and Data System [BI-RADS] category 0) and negative if a 1-year follow-up was recommended (i.e., mammograms classified as either BI-RADS category 1 or 2) *(21)*. Radiologists at these facilities did not use BI-RADS assessment categories 3, 4, or 5 for screening examinations. Positive outcome was defined as breast cancer detected as a result of the diagnostic work-up initiated by a positive screening mammogram.

Computation of Mammography Recall Rates

Recall rates for each radiologist and for the group of 24 radiologists were computed directly from mammographic interpretation records. In all of our analyses, we excluded recommendations for recall that were due to technical reasons, such as image artifacts (<1%). Recalls due to palpable findings identified during clinical breast examinations performed on all women by the technologist were included in our analyses because the majority of these findings were also marked on the mammograms. Such recalls amounted to approximately 1% of the screening examinations; hence, the underlying rates attributable to mammography interpretations alone are approximately 1% lower than those reported here. The women in this group of recalls are not the same as the group of women with palpable findings discovered by the woman herself or by a physician during a breast physical examination. Women in the latter group were scheduled for diagnostic examinations and were not included in our study. In our practice, palpable findings that are discovered by the technologists are noted during the physical examination and the procedure continues as a screening examination (including the use of computer-aided detection). The interpreting radiologists are aware of the technologists' findings and recall the women for additional procedures as needed. We recognize that this practice may not be a common one. We assumed that the effects of recalling this group of women due to palpable findings, if any, on the recall rates of individual radiologists would be proportional to the overall volume of mammograms read by each radiologist; hence, it should not substantially affect the results.

A small percentage (<4%) of the examinations in our practice classified as BI-RADS category 0 were scheduled for an interpretation at a later date because the needed comparison films were missing during the originally scheduled interpretation. Those cases were distributed proportionally to the volume of mammograms read by each radiologist and were included in the recall rates because it was not clear how many of them would have been recalled anyway.

Each mammography examination was identified in our database as to whether computer-aided detection was used during the interpretation. We therefore analyzed the data according to whether cases were interpreted with computer-aided detection.

Computation of Breast Cancer Detection Rates

Breast cancer detection rates were computed as follows: For every breast cancer detected, we found the most recent screening

mammogram that identified a finding that led to a diagnostic follow-up and ultimately resulted in a biopsy that was positive for cancer. Only the interpreter of the original screening mammogram that led to the detection of breast cancer was credited with the finding (i.e., invasive and ductal carcinoma *in situ*). Findings of lobular carcinoma *in situ* were not attributed to the interpreting radiologist as a cancer detected in the analyses. If a woman was recommended for a biopsy directly as a result of the screening examination, the interpreter was credited with the finding as well. Cases were excluded from the analysis if the most recent screening mammogram prior to biopsy had been performed more than 180 days before the biopsy or if the original interpreter had not recommended a recall (i.e., false-negative cases). We chose a cutoff of 180 days because we have found that, in the vast majority of cases, women are lost to follow-up or ignore the recall recommendation altogether if the recommended follow-up diagnostic procedure is not scheduled within 90 days or performed within 180 days of the original mammogram. We attributed any subsequent findings associated with recalls for diagnostic work-ups that did not take place within 180 days of the original mammogram to the subsequent examination. We included all examinations that that had been originally scheduled as screening procedures but were diagnosed during the same visit and during which a diagnosis was made that resulted in a positive outcome (i.e., converted into a diagnostic procedure that led to a finding of cancer). However, these cancer cases (n = 30) were excluded from the computed breast cancer detection rates in our analysis (both nominator and denominator) because they were all diagnosed by a radiologist without the use of computer-aided detection, and we therefore could not determine whether these cases would have been detected had they undergone routine interpretation (with or without computer-aided detection) as a routine screening procedure. In addition, all breast cancer patients who were referred to us from other facilities and for whom the diagnosis did not originate from a screening examination done at one of our facilities were excluded from the analysis.

Statistical Methods

Recall and detection rates with and without computer-aided detection were compared by using a generalized estimating equations (GEE) logistic regression model that accounts for clustering of findings within each reader *(22)*. In addition, we asked an independent team of investigators to evaluate the numbers of cancer cases that were detected with and without computer-aided detection by the type of abnormality(s) noted in the original report. Those findings were assigned to one of the following categories: 1) mass(es) only; 2) clustered microcalcifications only; 3) mass(es) and clustered microcalcifications; and 4) other findings. Because the performance levels of computer-aided detection systems are generally outstanding for detecting microcalcifications *(16)*, we used the GEE model to analyze our findings with respect to possible changes in the percentage of cancer detections attributable to microcalcification clusters associated with the use of computer-aided detection. In addition, all analyses were repeated using a mixed-effect logistic regression model in which readers were considered a random effect, and modality (i.e., with or without computer-aided detection) was considered a fixed effect *(23)*. We also examined data from the seven high-volume radiologists (i.e., those who interpreted more than 8000 mammograms each during the study period). Because of the serial nature of the analysis (namely, this was not a randomized study), we repeated the analyses with respect to the timing of the major use of computer-aided detection in our practice by comparing the results for all cases interpreted without computer-aided detection from January 1, 2000, through June 30, 2001, when computer-aided diagnosis was used in only a small percentage of cases (<0.2%) at our facilities, with results for all cases interpreted with computer-aided detection from October 1, 2001, through December 31, 2002, when most (>93%) of the cases at our facilities were interpreted with computer-aided detection. All statistical tests were two-sided.

Results

The mean age of the screened population (n = 115 571) during the study period was 50.05 years (standard deviation = 11.17 years). During the study period, the percentage of women who were screened for the first time gradually decreased from approximately 40% in 2000 to 30% in the last quarter of 2002, whereas the percentage of women who had repeated screenings gradually increased.

Table 1 summarizes our data for the 24 radiologists who interpreted screening mammograms at our facility with and without the use of a computer-aided detection system. Among the 115 571 examinations in our database, 56 432 (48.8%) were interpreted without the use of the computer-aided detection system and 59 139 (51.2%) were interpreted with the use of the computer-aided detection system. Recall rates for the entire group of 24 radiologists were 11.39% for mammograms interpreted without computer-aided detection and 11.40% for mammograms interpreted with it (percent difference = 0.09, 95% confidence interval [CI] = −11 to 11; $P = .96$). Recall rates for the 18 radiologists who interpreted more than 1000 mammograms each during the study period ranged from 7.7% to 17.2% (data not shown). Recall rates for the seven high-volume radiologists who interpreted more than 8000 mammograms each during the study period ranged from 7.7% to 14.9% (data not shown). Among this latter group of radiologists, there was no

Table 1. Mammography recall rates and breast cancer detection rates for 24 radiologists performing screening mammograms without and with computer-aided detection*

Type of interpretation	No. of mammograms read	No. of recalls	No. of breast cancers detected	Recall rate, %	Breast cancer detection rate per 1000 mammograms read
Without computer-aided detection	56 432	6430	197	11.39	3.49
With computer-aided detection	59 139	6741	210	11.40	3.55
Total	115 571	13 171	407	11.40	3.52

*The analysis excluded 30 conversion (screening to diagnostic) cancer cases, all of which were interpreted without computer-aided detection.

statistically significant correlation (rho = −0.21, P = .64) between recall rate and the total number of screening mammograms interpreted by individual radiologists. In our practice, approximately 3.0% of the cases recommended for recall are typically lost to follow-up because the woman either undergoes re-screening at another institution or ignores our recommendations. This group remained relatively constant as a percentage of recalled women over the period in question.

Table 2 summarizes our data for the seven high-volume radiologists who interpreted more than 8000 screening mammograms each with and without the use of a computer-aided detection system. During the study period, these radiologists interpreted a total of 82 129 screening mammograms and were credited with the detection of 292 breast cancers as a result of these screening procedures. In this group, the recall rates decreased from 11.62% for mammograms interpreted without computer-aided detection to 11.05% for mammograms interpreted with computer-aided detection (percent difference = −4.9, 95% CI = −21 to 4; P = .16).

Breast cancer detection rates for the entire group of 24 radiologists were 3.49 per 1000 screening examinations for mammograms interpreted without computer-aided detection and 3.55 per 1000 screening examinations for mammograms interpreted with it (percent difference = 1.7, 95% CI = −11 to 19; P = .68) (Table 1). Breast cancer detection rates for the seven high-volume radiologists were 3.61 per 1000 screening examinations for mammograms interpreted without computer-aided detection and 3.49 per 1000 screening examinations for mammograms interpreted with computer-aided detection (percent difference = −3.2, 95% CI = −15 to 9; P = .54) (Table 2).

The cancer detection rates associated with recalls due to the detection of clustered microcalcifications alone were 1.35 per 1000 mammograms interpreted without computer-aided detection and 1.44 per 1000 mammograms interpreted with computer-aided detection (P = .66) (data not shown). We observed no trend in breast cancer detection rates over time when we reviewed average detection rates for all 24 radiologists by calendar quarter (data not shown). We repeated our analyses using a random-effects logistic regression model and found that there were no statistically significant changes in recall rates or detection rates for all measurements presented above. Our results were not substantially affected when we compared only mammograms interpreted without computer-aided detection prior to July 1, 2001, with only those interpreted with computer-aided detection after October 1, 2001.

Discussion

The introduction of computer-aided detection into our practice was not associated with statistically significant changes in recall and breast cancer detection rates for the entire group of radiologists as well as for the subset of seven radiologists who interpreted high volumes of mammograms. The magnitudes of the improvements we observed were substantially less than those reported in the literature as the range of possible improvements based on retrospective analyses and limited prospective data *(9,17,18)*. The improvements we observed may be attributable to the better detection of clustered microcalcifications associated with malignancy. Our findings are consistent with the range of improvement in detection rates estimated and reported by others *(9,16–18)*. However, our large confidence intervals reflect the relatively low number of breast cancers detected with and without computer-aided detection and the large inter-reader variability among the radiologists in our practice. Because there were no repeat measures in this database—that is, each of the examinations was interpreted only once by one radiologist—we could not assess intra-reader variability.

It should be noted that we could not provide detailed information for individual radiologists without providing individually traceable data because each staff radiologist knows his or her reading volume and approximate recall rate. Our data are not adjusted for any learning effect: namely, the majority of interpretations made without computer-aided detection occurred chronologically prior to those made with computer-aided detection. We also did not account for any effect that may have resulted from a continuous effort to improve performance (in particular, sensitivity) by group reviews of all false-negative cases or from the steps undertaken to reduce recall rates through various actions, such as monthly performance reviews and direct consultation with interpreters who had higher-than-average recall rates.

Although one could argue that some or all of the reduction in recall rates we observed for the high-volume radiologists may be attributable to the use of computer-aided detection, the corresponding decrease in cancer detection rates we observed among the radiologists in this group is not easily explained by expected practice variations. An assessment of whether the small improvement we observed in cancer detection is due to learning effects—namely, that our radiologists had substantially more overall experience interpreting mammograms without computer-aided detection than with computer-aided detection—is beyond the scope of this investigation.

This investigation covered a period during which conventional film mammography was performed in all of our screening procedures. Hence, we cannot comment on the possible effect of computer-aided detection in a digital mammography environment. In our study, we did not account for women who had decided to follow up on our recommendations elsewhere. However, because compliance in patient follow-up was relatively constant during the study period, any bias in the results due to changes in patient loss to follow-up is likely to be small.

There are limited reported data concerning the actual effect of computer-aided detection on breast cancer detection and mam-

Table 2. Mammography recall rates and breast cancer detection rates for the seven high-volume radiologists performing screening mammograms without and with computer-aided detection

Type of interpretation	No. of mammograms read	No. of recalls	No. of breast cancers detected	Recall rate, %	Breast cancer detection rate per 1000 mammograms read
Without computer-aided detection	44 629	5188	161	11.62	3.61
With computer-aided detection	37 500	4145	131	11.05	3.49
Total	82 129	9333	292	11.36	3.56

mography recall rates. The prospective data reported by Freer and Ulissey *(16)*, which suggested a substantial improvement (19.5%) in breast cancer detection rates associated with the use of computer-aided detection systems, may have been affected by the fact that the results of mammographic interpretations without and with computer-aided detection were reported on the same cases (i.e., mammograms were read in one sitting, first without computer-aided detection then immediately afterward with the use of a computer-aided detection system). Another prospective study performed in a similar manner reported a 12% improvement in detection rates associated with the use of a computer-aided detection system *(18)*. This type of protocol, namely reading mammograms without computer-aided detection followed immediately by readings of the same mammograms with the use of a computer-aided detection system and a reassessment of the original finding without computer-aided detection, may have introduced a lower level of vigilance among radiologists during the initial interpretation without computer-aided detection, because they knew that computer-aided detection would be available to them for the final recommendation and that the initial interpretation did not constitute a formal clinical recommendation.

Results of the only study similar to ours, albeit on a substantially smaller group of patients and under a different set of circumstances, suggested that computer-aided detection was associated with a 13% improvement in breast cancer detection rates *(17)*. One of the advantages of the approach taken in our investigation is that the radiologists' interpretations were performed and recorded prospectively in a clinical setting and data were collected primarily for quality-assurance purposes *(24)*.

Our results for the interpretations made with computer-aided detection may be marginally biased because the outcomes of as many as nine recommendations for recalls and three recommendations for biopsies during the last quarter of 2002 are not yet available. Although some of these follow-up procedures or biopsies may ultimately be performed at our institution, we assume that the women who underwent the original mammograms have been lost to follow-up. However, on the basis of our typical recall-to-cancer-detection ratios (approximately 1 of 32 cases) and biopsy-to-confirmed cancer ratios (approximately 1 of 5 cases), we suspect that this bias would not substantially affect our findings or conclusions. It is possible that the gradually increasing fraction of women who had prior screening examinations created a bias in our results. Repeat screening examinations have a slightly lower number of cancers present as more are detected during the first screen, and on average, cancers detected on repeat mammograms may be more "difficult" to detect because more of the "easier" (e.g., larger) cancers are detected during the initial screen. Repeat mammograms have a lower recall rate, as the radiologists have prior films for comparison, to help inform their decision. The availability of prior examinations for comparison (in the repeat examinations) should have aided in the interpretation of these mammograms and offset the possible effect (if any) on the interpretations due to an increase in the "average case difficulty." The fact that our recall rates and detection rates remained virtually constant over time suggests that the possible bias due to a gradual increase in repeat examinations is not a statistically significant factor. We suspect that this increasing availability of prior examinations for comparison is a general phenomenon that is observed by most mammography screening practices and that there is not a simple way to account for it in an analysis such as the one we performed. When we included the 30 examinations that had been originally scheduled as screening procedures but were diagnosed during the same visit and resulted in a positive outcome in the estimation, our actual cancer detection rate attributable to screening was 3.8 per 1000 examinations, which is reasonable for a population in which the majority of women had undergone several screening procedures prior to the study period *(19)*.

On the basis of published performance levels of other computer-aided detection systems *(25)*, we believe that our results are not unique to the specific computer-aided detection system that is used at our institution. It is possible, however, that in clinical practices with substantially lower recall rates than ours, computer-aided detection would have larger effects on mammography recall rates and detection rates than what we observed. Such an improvement in detection rates would be consistent with results of a study *(17)* that reported lower recall rates without computer-aided detection (8.02%) than with computer-aided detection (8.43%).

The financial implications of our findings are beyond the scope of this work. However, a simple assessment of the additional estimated cost of using computer-aided detection per additional cancer detected in our practice (approximately $150 000 per additional detected cancer, assuming a reimbursement rate of $10 per case for professional and technical components combined) clearly indicates that more rigorous evaluations of the cost effectiveness of this practice are needed.

Our observations with respect to recall and detection rates may be exceptions (stemming from large inter-practice variations) that highlight the need for additional recall and detection rate data from multiple clinical practices and different reading environments. However, until such data clearly demonstrate that our experience is indeed an exception, these results represent an important first step.

This analysis of our practice was designed to assess the changes, if any, that occurred in recall and breast cancer detection rates with the introduction of computer-aided detection. Our results suggest that, in our practice, neither recall rates nor breast cancer detection rates changed with the introduction of this technology at its current level of performance, particularly as related to the detection of abnormalities other than clustered microcalcifications. Due to large confidence intervals, our results are statistically consistent with the possibility of large improvements in cancer detection rates with computer-aided detection. Yet, actually observed changes in our practice were substantially lower than expected. This is not to say that the use of computer-aided detection would not be beneficial or cost-effective in other practices. Rather, we suggest that, at its current level of performance, computer-aided detection may not improve mammography recall or breast cancer detection rates (especially as related to the detection of masses) in academic practices similar to ours that employ specialists for interpreting screening mammograms.

REFERENCES

(1) Humphrey LL, Helfand M, Chan BK, Woolf SH. Breast cancer screening: a survey of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med 2002;137:347–60.

(2) Tabár T, Duffy SW, Vitak B, Chen HH, Prevost TC. The natural history of breast carcinoma: what have we learned from screening? Cancer 1999;86: 449–62.

(3) Hendrick RE, Klabunde C, Grivegnee A, Pou G, Ballard-Barbash R. Technical quality control practices in mammography screening programs in 22 countries. Int J Qual Health Care 2002;14:219–26.
(4) Ng EH, Ng FC, Tan PH, Low SC, Chiang G, Tan KP, et al. Results of intermediate measures from a population-based, randomized trial of mammographic screening prevalence and detection of breast carcinoma among Asian women: the Singapore Breast Screening Project. Cancer 1998;82: 1521–8.
(5) Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology 2002;224:861–9.
(6) Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994;331: 1493–9.
(7) Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. J Natl Cancer Inst 2003;95:282–90.
(8) Doi K, Giger ML, Nishikawa RM, Schmidt RA. Computer-aided diagnosis of breast cancer on mammograms. Breast Cancer 1997;4:228–33.
(9) Warren Burhenne LJ, Wood SA, Orsi CJ, Feig SA, Kopans DB, O'Shaughnessy KF, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000;215: 554–62.
(10) Brem RF, Schoonjans JM. Radiologist detection of microcalcifications with and without computer-aided detection: a comparative study. Clin Radiol 2001;56:150–4.
(11) Leichter I, Fields S, Nirel R, Bamberger P, Novak B, Lederman R, et al. Improved mammographic interpretation of masses using computer-aided diagnosis. Eur Radiol 2000;10:377–83.
(12) Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE, et al. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. Radiology 1999;212:817–27.
(13) Jiang Y, Nishikawa RM, Schmidt RA, Toledano AY, Doi K. Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications. Radiology 2001;220: 787–94.
(14) Wagner RF, Beiden SV, Campbell G, Metz CE, Sacks WM. Assessment of medical imaging and computer-assist systems: lessons from recent experience. Acad Radiol 2002;9:1264–77.
(15) Brem RF. Enhancement of mammographic interpretation with computer-aided detection (CAD): a multi-institutional trial. Presented at the 87th Scientific Assembly and Annual Meeting of the Radiological Society of North America, November 25–30, 2001, Chicago, IL. Radiology 2001; 221(P):472.
(16) Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. Radiology 2001;220:781–6.
(17) Cupples TE. Impact of computer-aided detection (CAD) in a regional screening mammography program. Presented at the 87th Scientific Assembly and Annual Meeting of the Radiological Society of North America, November 25–30, 2001, Chicago, IL. Radiology 2001;221(P):520.
(18) Bandodkar P, Birdwell RL, Ikeda DM. Computer aided detection (CAD) with screening mammography in an academic institution: preliminary findings. Presented at the 88th Scientific Assembly and Annual Meeting of the Radiological Society of North America, December 1–6, 2002, Chicago, IL. Radiology 2002;225(P):458.
(19) Young WW, Destounis SV, Bonaccio E, Zuley ML. Computer-aided detection in screening mammography: can it replace the second reader in an independent double read? Preliminary results of a prospective double blinded study. Presented at the 88th Scientific Assembly and Annual Meeting of the Radiological Society of North America, December 1–6, 2002, Chicago, IL. Radiology 2002;225(P):600.
(20) Food and Drug Administration. Quality Standards and Certification Requirements for Mammography Facilities (21 CFR Part 900). Federal Register 1993;58:67565.
(21) American College of Radiology (ACR). Breast imaging reporting and data system (BI-RADS). Reston (VA): American College of Radiology, 1998. Available at http://www.acr.org/departments/stand_accred/birads/contents.html. [Last accessed: 12/9/03.]
(22) Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.
(23) Brown H, Prescott R. Applied mixed models in medicine. Chichester (NY): J. Wiley & Sons; 1999, pages 104–11
(24) Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R. Association of recall rates with sensitivity and positive predictive values of screening mammography. AJR Am J Roentgenol 2001;177:543–9.
(25) Hoffmeister JW, Rogers SK, DeSimio MP, Brem RF. Determining efficacy of mammographic CAD systems. J Digit Imaging 2002;15 Suppl 1:198–200.

NOTES

Supported in part by Public Health Service grants CA85241, CA67947, and CA77850 (to the University of Pittsburgh) from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services and by the U.S. Army Medical Research Acquisition Center (Fort Detrick, MD) contract DAMD17-00-1-0410 (to the University of Pittsburgh).

We thank Jennifer Herrmann, Jill King, Amy Klym, Christopher Traylor, and Andriy Bandos for their diligent and tireless work on this project.

Manuscript received May 14, 2003; revised October 5, 2003; accepted November 26, 2003.

APPENDIX 13

Recall and Detection Rates in Screening Mammography

A Review of Clinical Experience—Implications for Practice Guidelines

David Gur, Sc.D.[1]
Jules H. Sumkin, D.O.[1]
Lara A. Hardesty, M.D.[1]
Ronald J. Clearfield, M.D.[1]
Cathy S. Cohen, M.D.[1]
Marie A. Ganott, M.D.[1]
Christiane M. Hakim, M.D.[1]
Kathleen M. Harris, M.D.[1]
William R. Poller, M.D.[1]
Ratan Shah, M.D.[1]
Luisa P. Wallace, M.D.[1]
Howard E. Rockette, Ph.D[1,2]

[1] Department of Radiology, University of Pittsburgh and Magee-Womens Hospital, Pittsburgh, Pennsylvania.

[2] Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania.

See related editorial on pages 1549–52, this issue.

Supported in part by Grants CA77850 and CA67947 from the National Cancer Institute, National Institutes of Health, and also by the U.S. Army Medical Research Acquisition Center under Contract DAMD17-00-1-0410.

The authors thank Jennifer Herrmann, Jill King, Amy Klym, and Christopher Traylor for their diligent and tireless work on this project.

Address for reprints: David Gur, Sc.D., Imaging Research, Suite 4200, Department of Radiology, University of Pittsburgh, 300 Halket Street, Pittsburgh, PA 15213-3180; Fax: (412) 641-2582; E-mail: gurd@msx.upmc.edu

Received October 15, 2003; revision received December 1, 2003; accepted December 3, 2003.

BACKGROUND. The authors investigated the correlation between recall and detection rates in a group of 10 radiologists who had read a high volume of screening mammograms in an academic institution.

METHODS. Practice-related and outcome-related databases of verified cases were used to compute recall rates and tumor detection rates for a group of 10 Mammography Quality Standard Act (MQSA)-certified radiologists who interpreted a total of 98,668 screening mammograms during the years 2000, 2001, and 2002. The relation between recall and detection rates for these individuals was investigated using parametric Pearson (r) and nonparametric Spearman (rho) correlation coefficients. The effect of the volume of mammograms interpreted by individual radiologists was assessed using partial correlations controlling for total reading volumes.

RESULTS. A wide variability of recall rates (range, 7.7–17.2%) and detection rates (range, 2.6–5.4 per 1000 mammograms) was observed in the current study. A statistically significant correlation ($P < 0.05$) between recall and detection rates was observed in this group of 10 experienced radiologists. The results remained significant ($P < 0.05$) after accounting for the volume of mammograms interpreted by each radiologist.

CONCLUSIONS. Optimal performance in screening mammography should be evaluated quantitatively. The general pressure to reduce recall rates through "practice guidelines" to below a fixed level for all radiologists should be assessed carefully. *Cancer* **2004;100:1590–4.** *© 2004 American Cancer Society.*

KEYWORDS: mammography, screening, tumor detection rates, recall rates.

As periodic mammographic screening is rapidly gaining acceptance, our understanding of many strategic, operational, and financial issues related to this practice is improving as well. Several performance indices have been used to define "optimal" practice parameters in screening mammography. These include, but are not limited to, sensitivity, specificity, positive predictive value (PPV), and cost per detected tumor.[1,2] Clearly, the focus of screening for early detection should primarily be on improved sensitivity. At the same time, the large number of patients being recalled for additional procedures as a result of an initial review is a recognized problem for the very same reasons (operational and financial), with the added concern of the well documented increase of anxiety levels in women who are recalled.[3,4] Therefore, there is a belief that through a variety of actions including but not limited to specific and targeted training, one can augment observer performance levels, including the reduction of recall rates in screening mammography.[5,6] Although not specifically

DOI 10.1002/cncr.20053
Published online 8 March 2004 in Wiley InterScience (www.interscience.wiley.com).

regulated, there is a publicly stated goal to reduce recall levels to < 10%.[5,7] The question of what effect, if any, does a forced reduction in recall rates have on detection rates remains somewhat controversial. Some studies suggest that recall and detection rates are not highly correlated (particularly at high recall rates); hence, a reduction in the former does not necessarily affect the latter.[6,8] Other researchers believe that, after appropriate training, highly experienced radiologists individually operate largely along a single receiver operating characteristic curve; hence, pressuring them to reduce their recall rate may result in a corresponding reduction in the detection rates as well.[2,9] Because of the well documented variability among radiologists, the latter effect and its possible magnitude have to our knowledge been investigated only recently.[10–13] This type of an investigation is not easy to perform, because the expected yield (detection of actually positive cases that result from the screening) has been reported to be quite low in a population of women who already have been screened in the past.[14,15] Therefore, one generally needs to evaluate detection rates from the data of large groups of individual radiologists pooled together or have access to sufficient data from radiologists who each have interpreted a large number of mammograms. In this article, we present an analysis of the latter type of investigation.

MATERIALS AND METHODS

Screening mammography examinations performed in the study facilities at Magee-Womens Hospital (of the University of Pittsburgh Medical Center) and its five satellite breast imaging clinics during the years 2000, 2001, and 2002 were reviewed under an Institutional Review Board-approved protocol. Mammograms that had been interpreted by the 10 highest volume mammographers at the study institution during this period were included in the current study.

The data sources used in the current analysis were databases of procedure scheduling, procedure completion, radiology reporting, and procedure-related outcomes as determined from pathology reports. These databases have been assembled from original reports for several reasons, including quality assurance purposes that are required by the Mammography Quality Standard Act (MQSA).[16,17] The computerized reporting system and data entry protocols used in our practice remained the same throughout the study period. Because the number of positive findings leading to the detection of tumors by each individual were low, the records of all mammograms read by each of the participating radiologists "with" and "without" the availability of results from a commercial Computer-Assisted Detection (CAD) system were pooled for the purpose of this analysis. Our clinical practice for screening mammography during this period was film based, and most screening mammograms were read at the main facility in a batch mode. We included in the current analysis the results from the interpretations of the 10 highest volume radiologists in our practice, most of whom were with the study institution throughout much of the period in question. Each has performed > 3500 interpretations of screening mammography examinations.

Recall rates for each radiologist were computed directly from mammography interpretation records (Breast Imaging Reporting and Data System Atlas [BI-RADS® Atlas; American College of Radiology, Reston, VA] rating of 0). We excluded recommendations for recall due to technical reasons ("technical recalls"). These account for approximately 1% of cases. However, recalls resulting from palpable findings during clinical breast examinations were included because the majority of these findings also were depicted in the mammograms. These findings amount to < 1% of examinations; therefore, the underlying rates attributable to mammography interpretations alone are accordingly somewhat lower than those reported in the current study. The effect of "palpable" findings on individual radiologists is expected to be distributed proportionally to their overall volume.

In our practice, the interpretation of some examinations (< 4%) is delayed because of missing comparison films during the initial interpretation. These generally are distributed proportionally to the volume read by each radiologist and are included in the recall rates because it is not clear how many of these cases would have been actually recalled in any case.

Tumor detection rates were computed as follows. We identified the latest screening examination for each detected tumor that resulted in a diagnostic follow-up (recall) and ultimately resulted in pathologically verified carcinoma. The radiologist who interpreted the screening mammogram that led to the detection of breast carcinoma was credited with the finding for the purposes of the current analysis. Cases were excluded from the analysis if the latest screening mammogram prior to biopsy had been performed > 180 days earlier. In our experience, these women generally are "lost" to follow-up at other institutions or ignore the recommendations for a diagnostic workup (recall) altogether. Cancer patients who were referred to us from other facilities and for whom the diagnosis did not originate from a screening examination in one of our facilities were excluded. Women who originally were presented as screening procedures but were diagnosed using additional radio-

graphic procedures or other modalities (e.g., ultrasound) during the same visit ("conversion" cases from screening to diagnostic) were accounted for and were included in the current analysis. However, because a substantial number of these may originally have been identified as "potentially abnormal" by a technologist (who personally shows the case to a radiologist) during a quality assurance review of the images, we repeated the analysis after excluding this group of cases. For the purpose of these analyses, we assume that any effect due to the performance level of the radiologists who were performing and interpreting the diagnostic procedures during the follow-up visit are distributed in a manner that does not affect the study conclusions. The radiologists could not select the examinations they interpreted in our practice.

The correlation between recall and detection rates was evaluated using both the parametric Pearson (r) and the nonparametric Spearman (rho) correlation coefficients. We also examined the results after partial correction for the total volume of mammograms interpreted by each radiologist during the period in question.

RESULTS

Recall and detection rates for the 10 radiologists whose data were analyzed in the current study were computed. Each performed > 3500 interpretations (range, 3605–16,128 interpretations) during the period in question. We were unable to publish detailed information for individual radiologists without providing individually traceable data because each staff radiologist is aware of the approximate volume of screening examinations they interpreted and their approximate recall rate. These 10 radiologists interpreted a total of 98,668 cases during this time and detected 368 cases of carcinoma. Twenty-six "conversion" cases were included in the analysis. These cases originally were presented as a screening procedure but the patients underwent "follow-up" procedures (e.g., ultrasound) during the same visit (because of a physician being present on site at the time of the visit). A wide range of recall rates (range, 7.7–17.2%) and detection rates (range,2.6–5.4 per 1000 mammograms) was observed. Despite the low number of radiologists (10), when recall and detection rates were compared using the parametric Pearson (r) correlation coefficient, the correlation between the recall and detection rates was significant (r = 0.76; $P = 0.01$). Similarly, a significant correlation was observed in the group of radiologists using the nonparametric Spearman correlation coefficient (rho = 0.72; $P = 0.02$). A linear least square fit between the recall and detection rates for the group in which each radiologist represents a single "operating point" is presented in Figure 1. Despite significant interreader variability, the slope indicates an average of 0.22 additional detections per 1% increase in recall rates (95% confidence interval on the slope is +0.068 to +0.378). The correlation between recall and detection rates remained significant ($P < 0.05$) after accounting for the total volume read by each radiologist using partial correlations. Repeated analyses after the exclusion of the 26 "conversion" cases indicated no substantial difference in the correlations reported herein. The correlations remained significant when the analysis was repeated for the 7 ($P = 0.05$), 8 ($P < 0.05$), and 9 ($P < 0.05$) highest volume radiologists. These results demonstrate that, in general, in our practice, the higher the recall rates, the higher the detection rates. This increase in detection rate was found to persist over the range of observed recall rates and extended beyond the currently recommended practice guideline of 10%.

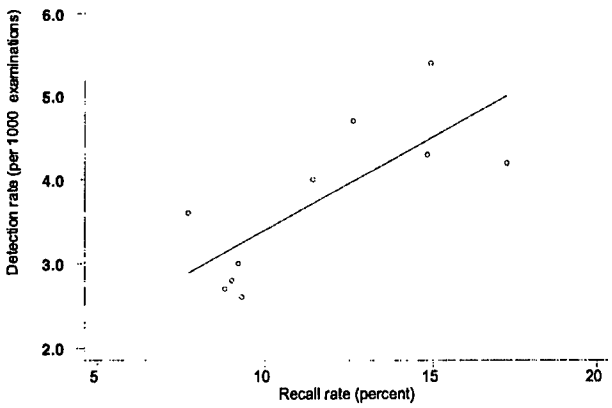

FIGURE 1. A linear fit of detection rates as a function of recall rates for the 10 radiologists in the current study.

DISCUSSION

There is little doubt that continuing education and training are important factors in the ability of radiologists to be consistent in interpreting mammograms and to improve their overall performance. However, to our knowledge, there are no conclusive data published to date regarding to what extent improvement continues beyond a certain level of training or experience.[12] Although there are questions with regard to whether volume and experience affect performance,[12] the general belief has been that one can reduce recall rates relatively easily without a significant impact on detection rates. As a result, there is an ongoing significant effort to do so, particularly in practices similar to ours with recall

rates that are in the higher range (≥ 10%). PPV as a result of screening has been of great interest as one of the indicators of the performance level of radiologists in this area.[8] However, if sensitivity is affected by recall rates, particularly in a group of well trained, high-volume radiologists whose recall rates are relatively high, the fundamental question of whether to continually pressure them to reduce their recall rates following currently accepted practice guidelines remains. This stems from the fact that the detection of "earlier tumors" with higher recall rates may be as or perhaps more important than actually reducing the recall rates or improving the PPV somewhat. It is interesting to note that an important review of several related issues suggested observations that were similar to those of the current study.[10] Unfortunately, to our knowledge the radiology community has not objectively addressed this potentially important matter to date.

Similar to the findings reported by Yankaskas et al.[8], the results of the current study suggest that detection rates generally are affected by recall rates in the lower range. However, unlike the observations of Yankaskas et al.,[8] the effect in our group of 10 highly trained radiologists, who individually read a reasonably high volume of mammograms, persisted over the entire range of observed recall rates (as high as 17%). In the higher range of recall rates (≥ 7%), Yankaskas et al.[8] showed no correlation between the recall and detection rates. Therefore, their results could suggest that, in this critical range, a reduction in recall rates should not affect the detection rates. It is possible that this difference arises from the fact that the current study took place in a "reasonably stable" screening population in whom the majority of "prevalence (or "baseline") carcinomas" had been detected already. Another possible explanation may be the number of mammograms interpreted by individual radiologists in the two studies. Clearly, more data are needed in this regard.

The total number of mammography screening interpretations by the radiologist with the lowest screening volume reported herein over a 3-year period was relatively low. However, our regionwide referral base was found to result in a large number of other diagnostic and interventional breast-imaging procedures that typically amount to approximately 50% of the screening examinations. Hence, our radiologists should be considered as "specialists" in breast imaging.

It should be noted that in our practice the average recall rates (≈11 percent) are generally relatively high compared with some reports,[18,19] and they are in better agreement with, and in some cases lower than, others.[15,20,21] We have no simple explanation for this observation. The results of the current study are in agreement with the findings of Beam et al.[12] and others in that there is a large variability in the performance of the radiologists in this area. We did not detect a significant correlation between the volume read by the individual radiologists during the period in question and their performance level, although the radiologists in the current study all can be considered high volume, "well trained" readers with significant experience. There are several arguments one can raise with regard to why the estimated recall and detection rates in the current study may not be precise in terms of absolute values. These include but are not limited to the inclusion of palpable cases and incomplete follow-up of cancer patients who may be lost to other institutions. The fact that our primary area of interest is the relative performance levels of the radiologists (rather than absolute) makes the results valid despite these limitations, as long as one does not bias the interpretation process by selectively assigning a specific subset to be interpreted by one radiologist or another (e.g., all "high risk" women or all examinations of women with dense breasts are assigned to "conservative" or "high-volume" radiologists). This was clearly not the case in our practice. Therefore, one would expect that any related corrections as a result of these limitations would be largely proportional to the volume of cases interpreted by each radiologist in the course of their routine clinical practice. The correlation between detection rates and outcome or even "average stage of disease " at the time of detection is beyond the scope of this project because the number of tumors detected by an individual radiologist was too small and the follow-up time after detection too short to meaningfully assess differences, if any, in outcome.

The results of the current study suggest that before we unilaterally pressure radiologists to reduce their recall rates because of a notion that this will improve our practices (and reduce overall management costs), we need to carefully evaluate the impact such an effort may have on early (and perhaps even "earlier") detection. If we believe that screening should focus primarily on maximizing early detection, and the earlier the better, one has to consider whether there may be an individualized optimal operating level that should be considered, rather than a "globally" recommended practice guideline of a maximum "acceptable" recall rate that applies to all screening mammographers. This view may be supported by women who appear to strongly prefer a small increase in detection rates, even at the expense of higher recall rates and the associated impact in terms of cost and added

anxiety.[22–24] The current limited study included a group of 10 academic radiologists practicing at 1 institution under 1 set of practice conditions. Clearly, more data are required before one can generalize the findings reported herein to the population of radiologists who interpret screening mammography in this country. At the same time, the number and type of examinations used in the current analysis may be generalizable to the screening population in a large number of academic practices around the U.S.

Conclusions

The performance level of a radiologist in the screening environment is a complex, multifactorial issue that cannot and should not be simplified. Reducing recall rates by "decree" (through the enforcement of recommended practice guidelines) may result in a corresponding reduction in the detection rates, hence the associated delays. The impact of external pressure on individual radiologists to reduce their recall rates should be evaluated carefully.

REFERENCES

1. Linver MN. Audits measure practice quality of mammography. *Diagn Imaging.* 2000;22:57–61.
2. Burnside E, Belkora J, Esserman L. The impact of alternative practices on the cost and quality of mammographic screening in the United States. *Clin Breast Cancer.* 2001; 2:145–152.
3. Brett J, Austoker J. Women who are recalled for further investigation for breast screening: psychological consequences 3 years after recall and factors affecting re-attendance. *J Public Health Med.* 2001;23:292–300.
4. Sandin B, Chorot P, Valiente RM, Lostao L, Santed MA. Adverse psychological effects in women attending a second-stage breast cancer screening. *J Psychosom Res.* 2002;52:303–309.
5. Feig SA. Economic challenges in breast imaging. A survivor's guide to success. *Radiol Clin North Am.* 2000;38:843–852.
6. Sickles EA. Successful methods to reduce false-positive mammography interpretations. *Radiol Clin North Am.* 2000; 38:693–700.
7. U.S. Department of Health and Human Services, Agency for Health Care Policy and Research. Clinical practice guideline number 13: quality determinants of mammography. AHCPR Pub. No. 95-0632. Washington, DC: . U.S Department of Health and Human Services, Agency for Health Care Policy and Research, 1994:78–86.
8. Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R. Association of recall rates with sensitivity and positive predictive values of screening mammography. *AJR Am J Roentgenol.* 2001;177: 543–549.
9. Kopans DB. The accuracy of mammographic interpretation [editorial]. *N Engl J Med.* 1994;331:1521–1522.
10. Moskowitz M. Retrospective reviews of breast cancer screening: what do we really learn from them. *Radiology.* 1996;199:615–620.
11. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med.* 1994;331:1493–1499.
12. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. *J Natl Cancer Inst.* 2003;95: 282–290.
13. Elmore JG, Miglioretti DL, Reisch LM, et al. Screening mammograms by community radiologists: variability in false-positive rates. *J Natl Cancer Inst.* 2002;94:1373–1380.
14. Frankel SD, Sickles EA, Curpen BN, Sollitto RA, Ominsky SH, Galvin HB. Initial versus subsequent screening mammography: comparison of findings and their prognostic significance. *AJR Am J Roentgenol.* 1995;164:1107–1109.
15. Young WW, Destounis SV, Bonaccio E, Zuley ML. Computer-aided detection in screening mammography: Can it replace the second reader in an independent double read? Preliminary results of a prospective double blinded study. Presented at the 88th Scientific Assembly and Annual Meeting of the Radiological Society of North America. *Radiology.* 2002;225(P):600.
16. Food and Drug Administration. Quality Standards and Certification Requirements for Mammography Facilities (21 CFR Part 900) *Federal Register* 58, no. (December 21, 1993): 67565.
17. Linver MN, Osuch JR, Brenner RJ, et al. The mammography audit: a primer for the mammography quality standards act (MQSA). *AJR Am J Roentgenol.* 1995;165:19–25.
18. Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology.* 2002;224:861–869.
19. 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.
20. Harvey SC, Geller B, Oppenheimer RG, Pinet M, Riddell L, Garra B. Increase in cancer detection and recall rates with independent double interpretation of screening mammography. *AJR Am J Roentgenol.* 2003;180:1461–1467.
21. Fletcher SW, Elmore JG. Clinical practice. Mammographic screening for breast cancer. *N Engl J Med.* 2003;348:1672–1680.
22. Schwartz LM, Woloshin S, Sox HC, Fischhoff B, Welch H. US women's attitudes to false positive mammography results and detection of ductal carcinoma in situ: cross sectional survey. *BMJ.* 2000;320:1635–1640.
23. Nekhlyudov L, Ross-Degnan D, Fletcher SW. Beliefs and expectations of women under 50 years old regarding screening mammography: a qualitative study. *J Gen Intern Med.* 2003;18:182–189.
24. Silverman E, Woloshin S, Schwartz LM, Byram SJ, Welch HG, Fischhoff B. Women's views on breast cancer risk and screening mammography: a qualitative interview study. *Med Decis Making.* 2001;21:231–240.

Radiology

David Gur, ScD
Jennifer S. Stalder, BS
Lara A. Hardesty, MD
Bin Zheng, PhD
Jules H. Sumkin, DO
Denise M. Chough, MD
Betty E. Shindel, MD
Howard E. Rockette, PhD

Index terms:
Breast neoplasms, diagnosis, 00.32
Cancer screening
Computers, diagnostic aid

Published online before print
10.1148/radiol.2332040277
Radiology 2004; 233:418–423

Abbreviation:
CAD = computer-aided detection

[1] From the Departments of Radiology (D.G., J.S.S., L.A.H., B.Z, J.H.S., D.M.C., B.E.S) and Biostatistics (H.E.R.) and Magee-Womens Hospital (D.G., L.A.H., J.H.S., D.M.C., B.E.S.), University of Pittsburgh, 300 Halket St, Suite 4200, Pittsburgh, PA 15213-3180. Received February 16, 2004; revision requested April 20; revision received May 5; accepted May 24. Supported in part by grants CA77850 and CA84241 from the National Cancer Institute, National Institutes of Health, and also by the U.S. Army Medical Research Acquisition Center under contract DAMD17-00-1-0410. **Address correspondence to** D.G. (e-mail: *gurd@upmc.edu*).

Authors stated no financial relationship to disclose.

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

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

Computer-aided Detection Performance in Mammographic Examination of Masses: Assessment[1]

PURPOSE: To compare performance of two computer-aided detection (CAD) systems and an in-house scheme applied to five groups of sequentially acquired screening mammograms.

MATERIALS AND METHODS: Two hundred nineteen film-based mammographic examinations, classified into five groups, were included in this study. Group 1 included 58 examinations in which verified malignant masses were detected during screening; group 2, 39 in which all available latest examinations were performed prior to diagnosis of these malignant masses (subset of 39 women from group 1); group 3, 22 in which findings were interpreted as negative but were verified as cancer within 1 year from the negative interpretation (missed cancers); group 4, 50 in which findings were negative and patients were not recalled for additional procedures; and group 5, 50 in which patients were recalled for additional procedures and findings were negative for cancer. In all examinations, images were processed with two Food and Drug Administration–approved commercially available CAD systems and an in-house scheme. Performance levels in terms of true-positive detection rates and number of false-positive identifications per image and per examination were compared.

RESULTS: Mass detection rates in positive examinations (group 1) were 67%–72%. Detection rates among three systems were not significantly different ($P > .05$). In 50 negative screening examinations (group 4), false-positive rates ranged from 1.08 to 1.68 per four-view examination. Performance level differences among systems were significant for false-positive rates ($P = .008$). Performance of all systems was at levels lower than publicly suggested in some retrospective studies. False-positive CAD cueing rates were significantly higher for negative examinations in which patients were recalled (group 5) than they were for those in which patients were not recalled (group 4) ($P \leq .002$).

CONCLUSION: Performance of CAD systems for mass detection at mammography varies significantly, depending on examination and system used. Actual performance of all systems in clinical environment can be improved.

An increasing body of evidence suggests that early detection of breast cancer through periodic screening is beneficial (1,2). Mammographic screening is rapidly gaining acceptance worldwide, and the number of procedures performed continues to increase (3,4). The difficulty in identification of some subtle suspicious regions depicted on mammograms, particularly those related to masses and asymmetric densities, the repetitious and sometimes tedious nature of the task, and the shortage of experienced radiologists who specialize in breast imaging and who routinely read high volumes of images in examinations have resulted in a wide variability in observer performance levels, as well as in relatively high recall rates for additional procedures (5–7). The effectiveness of mammographic screening programs depends on many factors. These factors include, but are not limited to,

the expertise and judgmental ability of the radiologist who reads the mammogram. Variability among radiologists can actually be useful insofar as studies show that double reading, namely, having two radiologists read the same mammogram independently, could increase detection by as much as 15% (8). Even if there were no shortage of experienced radiologists, however, the cost of true double reading as a standard practice is prohibitive for most facilities.

In recent years, major efforts have been expended to develop computer-aided detection (CAD) systems that will help radiologists with breast cancer detection. The hope is that these systems will serve as a second reader and will help improve sensitivity without a substantial increase in recall rates and at the same time possibly decrease reader variability, as well. These systems are currently aimed at the early detection of cancer and are accordingly designed to assist the radiologist in detection of suspicious regions depicted as clustered microcalcifications and masses (9–11). Computer-aided diagnosis systems are also being developed to assist radiologists in the classification task, namely, the determination of whether or not an identified finding is likely to represent a malignancy (11–13). The Food and Drug Administration has approved several detection systems for routine clinical use, and Medicare and other insurance companies have approved reimbursement for their use in clinical practice.

Results of studies (14,15) suggest that the use of CAD systems could potentially increase cancer detection rates by as much as 20% without a significant increase in recall rates. To date, there are limited data on the actual effect of the prospective use of such systems in the clinical environment (16,17). There is some evidence that the performance of radiologists, at least in the laboratory setting, is affected by the performance of the CAD scheme itself (18). Hence, a high level of performance is an important factor in the ultimate clinical success of CAD.

Data for comparison of the performance of CAD systems applied to the same set of cases are limited (19–22). The purpose of our study, therefore, was to compare the performance of two FDA-approved commercially available CAD systems and an in-house–developed scheme in five groups of sequentially acquired screening mammograms.

MATERIALS AND METHODS

Screening Examination Groups

Screening mammographic examinations performed at Magee-Womens Hospital, University of Pittsburgh Medical Center, and at its five satellite breast-imaging clinics during 2002 were included in this study. These examinations were classified into five groups. This study was conducted with an institutional review board–approved protocol. Informed consent was waived. Images in all examinations included in this study were acquired with film (MIN-R-2000; Eastman Kodak, Rochester, NY) and were clinically interpreted with CAD (ImageChecker; R2 Technologies, Sunnyvale, Calif) as a part of our routine practice.

The data sources for the selection of examinations were databases of procedure scheduling, procedure completion, radiology reporting, and procedure-related outcomes as determined from relevant pathology reports.

Group 1 included 58 examinations performed in women with biopsy-proved cancer that initially had been identified as a mass by a radiologist in our group during a screening examination in 2002. Images were selected sequentially from our procedure-related outcome database by a staff member (J.S.S.) who did not have any prior knowledge of the specific details about the patient or of the visual characteristics of the depicted mass.

In addition, there was an interest in the performance of CAD systems applied to examinations performed 1 year prior to observation of a positive finding. Group 2 hence included 39 available latest negative prior examinations (subset of 39 women from group 1 who underwent a different examination formed group 2) performed during or prior to 2001 that had been performed before the screening examination that led to a finding positive for cancer.

Group 3 included 22 consecutive false-negative examinations in which images depicted masses in retrospect. In 21 examinations, one mass in each was depicted on images, and in one examination two masses were depicted, which produced a total of 23 masses. Findings in these examinations were defined in our practice as false-negative interpretations. Findings in these examinations had been interpreted as negative or benign (Breast Imaging Reporting and Data System category 1 or 2) during the screening examination and were biopsy proved as positive for cancer, with a mass depicted on subsequent mammograms obtained within 1 year of the negative examination. These examinations constitute a different set of cases and are not a subset of the 39 prior examinations described previously as group 2.

Group 4 included 50 verified negative examinations (Breast Imaging Reporting and Data System category 1 or 2) that were selected randomly by the same staff member who selected those in group 1 from the examinations performed during two preselected dates in 2002 (March 1 and 2, 2002). Findings in all of these examinations were verified with findings at a 1-year follow-up screening examination that were interpreted as negative. A 1-year follow-up examination was the latest available examination in these women.

Group 5 included 50 consecutive examinations in which patients had been recalled during April 2002 (Breast Imaging Reporting and Data System category 0). Results of the diagnostic work-up that followed were negative or benign (Breast Imaging Reporting and Data System category 1 or 2), and results of the work-up for the annual examination in 2003 were negative, as well.

As a result, a total of 219 examinations in 180 women were included in the study. The median age of the women whose examinations were used in this study was 54.5 years, with a range of 38–87 years.

Evaluation of Masses

All examinations were reviewed by several investigators (D.G., J.H.S., L.A.H., J.S.S.) together with source documents to generate a truth file that included depicted findings for the examinations in question. The boundaries of the masses were drawn subjectively and conservatively (approximately 5 mm larger than the depicted masses in all directions) on the image obtained at the examination performed in 2002 that resulted in the finding and on the corresponding areas on the images obtained at the prior examinations, when applicable. If masses were depicted with spiculations, these were included in the mass region. Hence, the allowed target was larger in all directions than was the depicted mass. This selection for the increased size of the target was arbitrary and increased the marked regions, in some cases substantially, because mass contours with the expectation that any identification (detection) by the CAD system close to the actual mass would not be disregarded by

TABLE 1
Fraction of Detected Masses according to Breast and Image for Biopsy-proved Cancers

Group and Examination Type	Detection Fraction according to Breast: No. of Breasts with Malignant Masses	Second Look (%)	ImageChecker (%)	In-house Scheme (%)	Detection Fraction according to Image: No. of Visible Malignant Masses	Second Look (%)	ImageChecker (%)	In-house Scheme (%)
1, True-positive	58	72 (42)	71 (41)	67 (39)	114*	56 (64)	55 (63)	51 (58)
2, Prior to true-positive	39	23 (9)	26 (10)	15 (6)	78†	14 (11)	14 (11)	9 (7)
3, False-negative	23	35 (8)	39 (9)	30 (7)	45‡	27 (12)	27 (12)	18 (8)

Note.—Only the breast in which cancer was found was included in the calculations. Numbers in parentheses were used to calculate the percentages.
* One patient had two malignant masses in the same breast. In four examinations, the malignant mass was visible on only one mammographic view.
† One patient had two malignant masses in the same breast. In two examinations, the malignant mass was visible on only one mammographic view during the later examination.
‡ One patient had one malignant mass in both breasts. In one examination, the malignant mass was visible on only one mammographic view during the later examination.

the interpreting radiologists. It also allowed position changes at the prior examination to be more conservatively accounted for because of the larger allowed target for detection.

For each examination, processing was performed with three CAD systems. One system (ImageChecker M1000, version 3.1; R2 Technologies) was used routinely in our clinical practice and was the system with which processing had been performed in all of the examinations during the original clinical interpretation. Another system (Second Look, version 6.0 Beta; CADx Systems, Beavercreek, Ohio) was used to process all images as well. A third system was an in-house–developed scheme, and its use has been reported in the past (23–25).

To ensure that there was no bias in the results, with the exception of the fact that the initial selection may have been affected somewhat by the use of the system that we used during the initial clinical interpretation, we fixed the detection threshold for determination of suspicious regions on the in-house system. This was done to provide a binary output in our own scheme (identified regions were either marked or not marked), which was similar to that of the commercial system, rather than a continuous output (0–1). Hence, we provided an automated operation (no operator decisions or options) to an experienced staff member (J.S.S.) who had processed images in several thousands of examinations with both commercial systems during the past 3 years and who processed all the images used in this study with all three systems.

The digitized images (model 861; Howtek, Hudson, NH) obtained with the Second Look system were then transferred to the in-house scheme and processed in exactly the same manner. A true-positive finding detected by the CAD system was attributed to each mark (cued region) noted by the CAD system if the center of the marked region was overlapping in any way (within the boundary of the conservatively drawn contour) with the recorded mass area in the manually drawn truth file. Otherwise, the CAD system markings were considered false-positive findings. This task was performed by one staff person (the same experienced staff person mentioned previously) to avoid interoperator biases. Biases, if any, were assumed to be consistent for all three systems, and this assumption enabled a relative comparison among them, even if there were some biases in absolute terms.

Statistical Analysis

True and false findings were tabulated for all examinations. Both breast-based (on either of the mammographic views) and image-based (each image considered as an independent examination) detections were recorded, and detection rates per breast and per image, as well as false-positive rates per examination (all four mammographic views), were computed. The three systems were compared for detection levels (sensitivity) by using a repeated-measures binary-response model in which there were three replicates, one for each patient according to each of the three modalities. The average of false-positive cues provided among the three systems was compared by using Friedman two-way analysis of variance. The number of false-positive findings that were detected in negative screening examinations and those in examinations for which patients were recalled were compared by using the Mann-Whitney *U* test. All analyses were performed with software (SAS, version 8.2; SAS Institute, Cary, NC). For each modality, the difference in false-positive rates between negative screening examinations and those in which patients were recalled was compared, assuming independent Poisson distributions. All statistical tests were two sided, and a difference with $P < .05$ was considered significant.

RESULTS

Table 1 summarizes the findings in groups 1–3 according to breast and image. Table 1 also demonstrates that the detection rates of the three systems in detection of true-positive masses in group 1 (58 breasts with malignant masses) were 72% (42), 71% (41), and 67% (39) for Second Look, ImageChecker, and the in-house scheme, respectively. Table 1 further demonstrates the results of processing the latest prior examinations with positive findings (group 2). These were acquired between 1 year and 2 years 4 months prior to the subsequent positive examinations, with an average time difference of 1 year 4 months. As expected, detection rates were substantially lower in the same patients when images obtained in the latest prior examinations were processed. Although in 39 breasts with malignant masses, 23% (nine), 26% (10), and 15% (six) of masses were detected retrospectively on images obtained at prior examinations with CAD, the images in the examinations were read as not suspicious enough to result in a recall of the patient during the original clinical interpretations. CAD detection rates in the false-negative group (group 3) with 23 breasts with malignant masses were 35% (eight),

TABLE 2
False-Positive Cueing Rate per Patient, per Region, and per Image for All Examinations

Group and Examination Type	Total No. of Examinations	False-Positive Rate according to Patient			False-Positive Rate according to Region*			Total No. of Images	False-Positive Rate according to Image		
		Second Look	Image-Checker	In-house Scheme	Second Look	Image-Checker	In-house Scheme		Second Look	Image-Checker	In-house Scheme
1, True-positive	58	1.53 (89)	1.05 (61)	1.05 (61)	1.28 (74)	0.88 (51)	0.98 (57)	228	0.39 (89)	0.27 (61)	0.27 (61)
2, Prior to true-positive	39	1.64 (64)	1.13 (44)	1.33 (52)	1.41 (55)	1.00 (39)	1.10 (43)	156	0.41 (64)	0.28 (44)	0.33 (52)
3, False-negative	22	1.18 (26)	1.00 (22)	1.50 (33)	0.86 (19)	0.82 (18)	1.36 (30)	84	0.31 (26)	0.26 (22)	0.39 (33)
4, Screening mammography	50	1.68 (84)	1.08 (54)	1.20 (60)	1.40 (70)	0.96 (48)	1.06 (53)	200	0.42 (84)	0.27 (54)	0.30 (60)
5, Mammography with recalled patients	50	2.70 (135)	2.16 (108)	2.86 (143)	2.28 (114)	1.82 (91)	2.68 (134)	200	0.68 (135)	0.54 (108)	0.72 (143)

Note.—All positive and negative images obtained in each patient were used to calculate the false-positive rate. Numbers in parentheses were used to calculate the false-positive rates. False-positive rates according to patient and region were based on total number of examinations.

* If a region was cued on two mammographic views, it was counted as one marked false-positive region.

39% (nine), and 30% (seven) for the three systems, respectively.

Table 2 shows that the false-positive rates in negative examinations (group 4) were 1.68, 1.08, and 1.20 per examination (four images) for the same three systems, respectively. For the examinations in which patients were recalled but findings were later verified as negative (group 5), the false-positive rates were 2.70, 2.16, and 2.86, respectively (Table 2). In Table 2, we also provide the number of cued but negative regions, which is less than the total number of false-positive cues, after adjustment for regions that were cued on both mammographic views and after counting of these matched cues as one false-positive region. When we compared the performance of the three systems, the differences were not significant ($P = .63$) for detection in actually positive examinations that led to the detection of cancer, and the differences in false-positive rates were significant ($P = .008$) for the number of false-positive identifications in the negative screening examinations. The differences in average false-positive rates between negative screening examinations and examinations in which patients had been recalled were significant ($P = .002$, $P < .001$, $P < .001$ for the three systems).

DISCUSSION

Several studies about the performance levels of each of the systems in question have been published (14,16,26,27), but the differences in patient selection and study design make any direct comparison difficult. The comparison of basic performance levels can only be performed appropriately when the systems are tested on the same set of examinations with a sample that is large enough and, preferably, contains as representative a sample as possible (eg, sequentially ascertained examinations) so that results can be generalized to the screening population of women.

It is not reasonable to expect, especially on a worldwide level, that film images will quickly be totally replaced by digital images. For that reason, most CAD systems currently in use must provide a method to digitize images, and this process in combination with differences in CAD algorithms may lead to problems in regard to standardization and reproducibility of results even when applied to a single system (28–30). There is little doubt that differences in performance among CAD systems will remain. If we want to collect data that allow radiologists to improve the practice of screening mammography, it is important that we understand the possible effects that may result from using different CAD systems. There are few data about a comparison of performance of different CAD systems when applied to the same sets of examinations. As systems continue to evolve and improve, the results of such comparisons are valid only for the experimental conditions being implemented with the specific systems (eg, digitizer and software versions) that were studied. For that reason, the results of such studies, while interesting and possibly suggestive of the effects of system differences at a given time and for a specific distribution of examinations, may be obsolete within a short period. It is important, however, to recognize that there are differences (frequently substantial) in the performance levels of different CAD systems. If such differences affect radiologists during clinical interpretations of findings of screening mammographic examinations, one should be aware of them (18,31).

Lechner et al (19) compared two Food and Drug Administration–approved CAD devices, ImageChecker M1000 (R2 Technologies) and Second Look. They found that 90% and 89% of abnormalities associated with cancers in 120 examinations were detected by the ImageChecker and Second Look systems, respectively. While 100% and 90% of the ten examinations with both masses and microcalcification clusters were detected with the two systems, respectively, only 84% and 82% of the 67 masses without clusters were identified with the two systems. Similar performance levels were reported in other studies (26), albeit no comparisons with other systems were made. A review of the findings from these studies, as well as of the Food and Drug Administration approval process, suggests the following: The performance of the two commercial systems is reasonably comparable for all practical purposes. If differences exist, they are small and would require large sample sizes to quantify them (32).

Our study is somewhat different in the examination selection process. We attempted to select a sequentially acquired, and potentially representative, sample of each type of examination to allow generalizability, at least to our own screening population. Recently, investigators in a study (20) reported that the patient-based sensitivity for detection of "actionable architectural distortion" with these two systems when applied to 45 exami-

nations (in 43 patients) was less than 50% for either system. In another study of retrospectively reviewed prior examinations with findings that suggested "evidence of cancer on prior mammograms," approximately 50% sensitivity for mass detection (eight of 19 with Second Look and 12 of 19 with ImageChecker) on prior images was indicated (22).

Although our study is similar to that of Shile and Guingrich (22) in that we attempted to select a representative population of examinations, it differs in several respects. First, we included a series of all available sequentially acquired sets of examinations.

Second, our false-positive rate was computed from a set of negative examinations rather than from the same examinations in which a mass was found.

Third, our assessment of CAD performance in the five sets of examinations allows one to have a better perspective of the possible effect of CAD on clinical practices with each type of mammogram. In our study, performance of all systems was at somewhat lower levels than expected. This could be the result of several factors. These factors included, but were not limited to, the difficulty of detection of the "average" cancer with our screening program. The conservatively defined mass regions (targets) reduced the possibility of biases that would result from exact marking. The use of only one experienced person, who was not involved in our CAD development team, to rate the correct markings ensured consistency in the scoring. This should have decreased, if not completely eliminated, any biases in the relative comparison among the systems. At this level of performance, we showed that experienced radiologists do not substantially improve their mass detection performance levels in the laboratory (18), and we suspect that this might be the case in the clinical environment, as well (17). Interestingly, the false-positive rates for examinations in which patients were recalled but that later proved to be negative examinations (group 5) were higher than were the rates for negative screening examinations. This finding suggests that these mammograms are more difficult for the CAD, as well as for the human observer, to analyze correctly. The performance of all three CAD systems was not very high in both the sets of examinations with false-negative interpretations and prior examinations with actually positive interpretations. This finding suggests that, at least in our environment, the potential improvements in earlier detection of masses with the use of current CAD systems is perhaps somewhat limited. Although seemingly unimportant as long as detection rates are comparable, the false-positive rates may affect general radiologists' reliance on the CAD results. High false-positive rates may result in low reader confidence in the CAD marking, since many cues have to be reviewed and discarded as negative findings (18).

In addition, there are some indications that performance in the noncued areas may be affected by the false-positive rate, as well (18). Because of the substantial difference in medicolegal liability between false-negative and false-positive interpretations, the effect of the CAD-generated false-positive cueing rate on noncued cancers may be an important issue to consider.

As to the lower performance of our own in-house scheme for CAD, we note that the scheme was originally designed and optimized for images digitized with a different digitizer (18,25), which has substantially different signal and noise characteristics. Also, our current scheme does not limit the total number of regions identified as suspicious per examination, as do other systems (33). Despite these limitations, it performed reasonably well in a direct comparison with two commercial systems.

Our study had several limitations. First, as previously indicated, our selection protocol may have been somewhat biased in favor of the ImageChecker system in that the images obtained in these examinations (with the exception of group 2) had been processed with this system during the initial clinical interpretation, and this bias possibly influenced the results of these examinations. However, our experience to date indicates that in our practice the changes were minor at best, particularly with respect to the detection of masses (17).

Second, the verification of negative examinations was based on findings at the subsequent screening examination. Although not optimal, this was the most recent available examination at the time, and we assumed that errors in this regard, if any, were not likely to affect the relative performance comparisons we described.

Third, the study was limited to the mammograms acquired at one institution and the masses detected by one group of radiologists. However, we do not believe that this limitation affected the results in a manner that would substantially affect similar comparisons at other institutions.

Fourth, our conservative approach to generation of the targets (ie, drawing the mass regions) may have affected the results. However, we verified that this effect was not substantial (<5% in this set of cases) and did not affect the comparison of relative performance levels of the three CAD systems.

Fifth, it could be argued that one of the limitations of the study was that we tested complete systems and not the software scheme alone. Hence, the comparison could have been affected by the digitizers in the two commercial systems we used. The fact is that a commercial CAD system is integrated, and these systems were tested largely as they would be used in a clinical environment. In this study, we cannot comment on a comparison that would be based on testing of the software alone.

Last, our study focused on the detection of masses. The significantly higher performance of CAD systems in the detection of microcalcifications may be sufficient to warrant the routine use of these systems alone. Other nondetection issues, such as the assessment of possible efficiency improvements in the reading of mammograms because of the high performance in the detection of microcalcifications, were clearly beyond the scope of this study.

In summary, we observed somewhat lower than expected case-based and image-based detection rates with CAD for all three systems. This is not to indicate that CAD cannot help the radiologist, even at these levels of performance, in different clinical environments, particularly radiologists with less experience in the interpretation of screening mammograms. However, the level of improvement is not likely to be what had been estimated from retrospective studies in a laboratory environment. Results of this study clearly indicate that marked improvements in CAD performance levels for mass detection are both desired and possible, and continuing efforts should be expanded in this area.

Acknowledgments: The authors thank Jill King, MS, Glenn Maitz, MS, John Drescher, BS, and Christopher Traylor for their support in this project and CADx Systems for providing equipment used in the experiment performed in this study.

References

1. Tabar L, Dean PB. Mammography and breast cancer: the new era. Int J Gynaecol Obstet 2003; 82:319–326.

2. Freedman GM, Anderson PR, Goldstein LJ, et al. Routine mammography is associated with earlier stage disease and greater eligibility for breast conservation in breast carcinoma patients age 40 years and older. Cancer 2003; 98:918–925.
3. Elmore JG, Nakano CY, Koepsell TD, Desnick LM, D'Orsi CJ, Ransohoff DF. International variation in screening mammography interpretations in community-based programs. J Natl Cancer Inst 2003; 95:1384–1393.
4. Leung GM, Lam TH, Thach TQ, Hedley AJ. Will screening mammography in the East do more harm than good? Am J Public Health 2002; 92:1841–1846.
5. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample. Arch Intern Med 1996; 156:209–213.
6. Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology 2002; 224:861–869.
7. Smith-Bindman R, Chu PW, Miglioretti DL, et al. Comparison of screening mammography in the United States and the United Kingdom. JAMA 2003; 290:2129–2137.
8. Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191:241–244.
9. Tourassi GD, Vargas-Voracek R, Catarious DM Jr, Floyd CE Jr. Computer-assisted detection of mammographic masses: a template matching scheme based on mutual information. Med Phys 2003; 30:2123–2130.
10. Gurcan MN, Chan HP, Sahiner B, Hadjiiski L, Petrick N, Helvie MA. Optimal neural network architecture selection: improvement in computerized detection of microcalcifications. Acad Radiol 2002; 9:420–429.
11. Roque AC, Andre TC. Mammography and computerized decision systems: a review. Ann N Y Acad Sci 2002; 980:83–94.
12. Kouskos E, Markopoulos C, Revenas K, Koufopoulos K, Kyriakou V, Gogas J. Computer-aided preoperative diagnosis of microcalcifications on mammograms. Acta Radiol 2003; 44:43–46.
13. Leichter I, Buchbinder S, Bamberger P, Novak B, Fields S, Lederman R. Quantitative characterization of mass lesions on digitized mammograms for computer-assisted diagnosis. Invest Radiol 2000; 35: 366–372.
14. Brem RF, Baum J, Lechner M, et al. Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial. AJR Am J Roentgenol 2003; 181:687–693.
15. Warren 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. 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.
17. Gur D, Sumkin JH, Rockette HE, et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. J Natl Cancer Inst 2004; 96:185–190.
18. Zheng B, Ganott MA, Britton CA, et al. Soft-copy mammographic readings with different computer-assisted detection cuing environments: preliminary findings. Radiology 2001; 221:633–640.
19. Lechner M, Nelson M, Elvecrog W. Comparison of two commercially available computer-aided detection (CAD) systems. Appl Radiol 2002; 31(suppl 4):31–35.
20. Baker JA, Rosen EL, Lo JY, Gimenez EI, Walsh R, Soo MS. Computer-aided detection (CAD) in screening mammography: sensitivity of commercial CAD systems for detecting architectural distortion. AJR Am J Roentgenol 2003; 181:1083–1088.
21. Nelson MT, Lechner MC. Comparison of three commercially available FDA approved computer-aided detection (CAD) systems (abstr). Radiology 2002; 225(P): 600.
22. Shile PE, Guingrich JA. Detecting "missed" breast cancers: a comparison. Appl Radiol 2002; 31(suppl 10):2–4.
23. 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.
24. 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.
25. 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.
26. 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.
27. Castellino RA. Computer-aided detection: an overview. Appl Radiol 2001; 30(suppl 11):5–8.
28. Taylor CG, Champness J, Reddy M, Taylor P, Potts HW, Given-Wilson R. Reproducibility of prompts in computer-aided detection (CAD) of breast cancer. Clin Radiol 2003; 58:733–738.
29. 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.
30. Zheng B, Hardesty LA, Poller WR, Sumkin JH, Golla S. Mammography with computer-aided detection: reproducibility assessment—initial experience. Radiology 2003; 228:58–62.
31. D'Orsi CJ. Computer-aided detection: there is no free lunch. Radiology 2001; 221:585–586.
32. Hoffmeister JW, Rogers SK, DeSimio MP, Brem RF. Determining efficacy of mammographic CAD systems. J Digit Imaging 2002; 15(suppl 1):198–200.
33. Zheng B, Leader JK, Abrams G, et al. Computer-aided detection schemes: the effect of limiting the number of cued regions in each case. AJR Am J Roentgenol 2004; 182:579–583.

The Effect of Routine Use of a Computer-Aided Detection System on the Practice of Breast Imagers: A Subjective Assessment[1]

Amy H. Klym, BS, Jill L. King, MS, Lara Hardesty, MD

The optimal use of any technological tool, such as computer-aided detection (CAD), requires the user to both understand the strengths and limitations of the technology and feel at ease in adapting it into his or her practice. Much previous and ongoing effort has been directed to the study of the impact of CAD systems, in terms of device performance (eg, digitizer, detection algorithms) and clinical impact (eg, detection of cancers, recall rates) after implementation (1–5). However, to date, there is no published information about the perception of breast imagers with substantial experience in using mammography CAD with regard to what impact it had on their own practice. This is an important issue because breast imagers could ignore the CAD results altogether if they felt uncomfortable with the cueing results. Reimbursement for CAD would then in effect become an unnecessary expense. In addition, there may be an increase in liability for breast imagers in cases where CAD cues are actually correct but cancers were missed by the radiologist.

Acad Radiol 2004; 11:711–713

[1] From the Department of Radiology, University of Pittsburgh, Pittsburgh, PA (A.H.K., J.L.K., L.H.); and Magee Womens Hospital, University of Pittsburgh Medical Center, Pittsburgh, PA(L.H.). Received February 3, 2004; accepted February 3, 2004. Supported in part by Public Health Service grant nos. CA67947 and CA77850 to the University of Pittsburgh from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, and by the US Army Medical Research Acquisition Center (Fort Detrick, MD) under contract DAMD17-00-1-0410 to the University of Pittsburgh. The content of the information contained herein does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. **Address correspondence to** A.H.K. e-mail: klymah@msx.upmc.edu

doi:10.1016/j.acra.2004.02.007

Current mammography CAD schemes provide false-positive cues in the majority of cases, and therefore a primary concern expressed about the widespread incorporation of CAD into screening mammography practices is the potential for "over-reading," namely, recalling too many women for additional breast imaging procedures (6,7). If breast imagers themselves believe that they are recalling too many women or taking too much time to interpret examinations to rule out the false-positive cues, they may ignore the CAD results altogether. To evaluate the breast imager's perceptions of differences in their practice with respect to recall rate, or the time required to interpret a mammogram when CAD is used, we surveyed 12 highly experienced breast imagers who had at least 2 years of practice at Magee Womens Hospital of the University of Pittsburgh Medical Center (Pittsburgh, PA).

In June of 2001, our facility began using a commercially available CAD system (Image Checker; R2 Technology, Sunnyvale, CA) for most screening and diagnostic mammograms acquired in our facilities. The screen films are digitized and analyzed by the CAD system and examinations are interpreted using an alternator that displays the CAD cues on a monitor placed below the displayed films. A large fraction of our diagnostic examinations performed at the hospital breast center are performed using a Full Field Digital Mammography System (Senographe 2000; GE Medical Systems, Waukesha, WI). This system uses an algorithm provided by R2, and CAD cues are displayed on the system's dedicated workstation.

To assess how breast imagers subjectively perceived the effect of CAD on their own practice, if any at all, we administered a voluntary short survey. This survey was performed approximately 1 year after the introduction of

Table 1
Distribution of Answers to Each of the Survey Questions

Questions of how Practice has Changed in the last 3 Years	Did not Change	Changed	Direction of Change	
			Increased	Decreased
1. My general practice:	6	6		
2. My recall rates during interpretations of screening mammograms	7	5	4	1
3. My recommendations for biopsy as a fraction of diagnostic interpretation	12	0	0	0
4. My recommendations for the use of US as a fraction of the total number of diagnostic procedures	11	1	1	0
5. The time spent reading each screening examination has	3	9	4	5
6. The time spent reading each diagnostic examination has	8	4	3	1
7. The use of CAD affected the way I practice	3	9		

CAD into our screening practice. The survey questions asked the breast imagers about any perceived changes in his or her practice in the last 3 years. It must be noted that this survey was administered before any of the participants became aware of the results we obtained assessing the actual impact of CAD on recall and detection rates in our practice, which showed only minimal changes in recall rates (3). We wish to emphasize that participants were told the survey was being conducted primarily to assess their subjective feelings about changes in their practice as a result of the use of CAD.

The distribution of the answers to each question in our survey is shown in the Table. Although only six of the respondents (50%) indicated that their practice had changed in the last 3 years (question 1), nine of 12 answered positively to the question whether or not CAD had changed the way they practiced (question 7). These nine radiologists perceived a number of CAD-related changes. Five perceived a change in reading time; one in reading time and recall rate; one in recall rate only; one in reading time, recall rate, and rate of recommendation for breast ultrasound; and one did not mark any specific changes listed in the survey (questions 2–6). Of the three radiologists who perceived that there was *no change* in their practice because of CAD, two indicated general changes in their practice. One indicated a decrease in recall rate and an increase in screening reading time, the other indicated an increase in recall rate and a decrease in interpretation time. Therefore, only one respondent indicated no specific change in response to all questions in the survey.

Five respondents perceived that their screening recall rate had changed (four thought that it had increased, one indicated that it had decreased), and one reader perceived his/her rate of recommending breast ultrasound had increased. None of the 12 readers perceived any change in their rate of recommending biopsy as a result of using CAD during the interpretation of diagnostic procedures. Unsolicited comments from four readers suggested changes were considered to be temporary, a "learning effect." There was one unsolicited comment written questioning the overall usefulness of CAD in a diagnostic setting.

This survey was not intended as a proof of actual changes in the practice of breast imagers, if any, as a result of incorporation of CAD into the diagnostic process. Rather it was designed to assess their perceptions in this regard. The fact that all but one reader recorded some kind of practice change suggests that the CAD results are not simply being ignored. The wide distribution of the answers, indicating fairly equally perceived increases and decreases in interpretation times and recall rates, suggests that they adopted the practice without any significant difficulties. Interestingly, their assessment of the impact on recall rates generally agrees with actual observation we made during a review of over 100,000 examinations when interpreted with and without the use of CAD result. In summary, as a group, the breast imagers practicing in our facility have incorporated the use of CAD with little concern that their practice has been substantially altered with regard to recall rates or time required for the interpretation of mammograms.

REFERENCES

1. Baydush AH, Catarious DM, Abbey CK, et al. Computer aided detection of masses in mammography using subregion Hotelling observers. Med Phys 2003; 30:1781–1787.
2. Paquerault S, Petrick N, Chan HP, et al. Improvement of computerized mass detection on mammograms: fusion of two-view information. Med Phys 2002; 29:238–247.
3. Gur D, Sumkin JH, Rockette HE, et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. J Natl Cancer Inst 2004; 96:185–190.
4. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in community breast center. 2001; 220:781–786.
5. Karssemeijer N, Otten JD, Verbeek AL, et al. Computer-aided detection versus independent double reading of masses on mammograms. Radiology 2003; 227:192–200.
6. Zheng B, Ganott MA, Britton CA, et al. Soft-copy mammographic readings with different computer-assisted detection cuing environments: preliminary findings. Radiology 2001; 221:633–640.
7. D'Orsi CJ. Computer-aided detection: there is no free lunch. Radiology 2001; 221:585–586.

Teleradiology and screening mammography: a telemammography system evaluation and comparison to clinical results

J. Ken Leader*[a], Denise Chough[ab], Ronald J. Clearfield[ab], Marie A. Ganott[ab], Christiane Hakim[ab], Lara Hardesty[ab], Betty Shindel[ab], Jules H. Sumkin[ab], John M. Drescher[a], Glenn S. Maitz[a], David Gur[a]

[a]University of Pittsburgh, Pittsburgh, PA USA 15213
[b]Magee-Womens Hospital, Pittsburgh, PA USA 15213

ABSTRACT

Radiologists' performance reviewing and rating breast cancer screening mammography exams using a telemammography system was evaluated and compared with the actual clinical interpretations of the same interpretations. Mammography technologists from three remote imaging sites transmitted 245 exams to a central site (radiologists), which they (the technologists) believed needed additional procedures (termed "recall"). Current exam image data and non-image data (i.e., technologist's text message, technologist's graphic marks, patient's prior report, and Computer Aided Detection (CAD) results) were transmitted to the central site and displayed on three high-resolution, portrait monitors. Seven radiologists interpreted ("recall" or "no recall") the exams using the telemammography workstation in three separate multi-mode studies. The mean telemammography recall rates ranged from 72.3% to 82.5% while the actual clinical recall rates ranged from 38.4% to 42.3% across the three studies. Mean Kappa of agreement ranged from 0.102 to 0.213 and mean percent agreement ranged from 48.7% to 57.4% across the three studies. Eighty-seven percent of the disagreement interpretations occurred when the telemammography interpretation resulted in a recommendation to recall and the clinical interpretation resulted in a recommendation not to recall. The poor agreement between the telemammography and clinical interpretations may indicate a critical dependence on images from prior screening exams rather than any text based information. The technologists were sensitive, if not specific, to the mammography features and changes that may lead to recall. Using the telemammography system the radiologists were able to reduce the recommended recalls by the technologist by approximately 25 percent.

Keywords: Teleradiology, human performance, recall rate, breast cancer screening, mammography

1. INTRODUCTION

Screening for breast cancer using mammography is and will continue to be practiced worldwide with extensive research supporting the benefits of screening,[1-6] despite sporadic studies reporting limited or no benefit from screening mammography.[7-9] The ubiquitous practice and growing population of candidates for screening mammography present many challenges to the practitioners creating possible opportunity to improve screening mammography. Some elements of screening mammography that have the potential to be improved include radiologist's practice and performance, personnel shortages, and public perception and compliance.[10-16]

The practice of teleradiology may potentially improve the management of screening mammography in particular in remote or underserved locations where physicians are not physically present, but the high-spatial demands of mammography present challenges to effective implementation of telemammography based practices. There are several image processing techniques commonly used in teleradiology to facilitate handling large amounts of data, which include image compression, image cropping, image selection, and display format.[17-24] In addition, the qualifications of personnel necessary for the successful implementation may need to be evaluated.[25-27] We have design and tested a telemammography system capable of handling the large data requirements of mammography that is operated by mammography technologists at the remote sites and experienced radiologists who review transmitted examinations at the central site.[28-30]

* leaderjk@upmc.edu; phone (412) 641-2572; fax (412) 641-2582, University of Pittsburgh, Magee-Womens Hospital, 300 Halket Street, Suite 4200, Pittsburgh, PA 15213

Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment,
edited by Miguel P. Eckstein, Yulei Jiang, Proceedings of SPIE Vol. 5749
(SPIE, Bellingham, WA, 2005) · 1605-7422/05/$15 · doi: 10.1117/12.592823

In this study, we evaluated radiologists' performance during off-line reviewing and rating screening mammography exams using the telemammography system and compared their performance to the clinical interpretations of the same examinations. The incremental addition of information was analyzed in three separate multi-mode studies to determine the independent effect of each type of information. The motivation was to determine what information is necessary and sufficient in a telemammography system implemented to reduce the number of patients recalled for additional procedures as part of the screening exam. The long-term objective is to decrease the number of patients recalled by remote management, particularly in underserved areas.

2. METHODS

2.1 Telemammography system

Cases used in this retrospective study were accrued using a high-quality, multi-site telemammography system that consists of one central site (Magee-Womens Hospital, Pittsburgh, PA, USA) and three remote sites (satellite woman's imaging centers of Magee-Womens Hospital). Cases for this study were acquired under normal operating procedures. The specific technical information such as software design, image processing, workstation features, and inter-site communication are described in detail by Drescher et al.[30] (2003). The system is described briefly as necessary below.

In short, a technologist at a remote site digitizes mammographic films, composes a message (text and graphic) to describe and locate their impression, and scans the patient's prior report (when available). This information and the results of Computer Aided Detection (CAD) scheme to detected suspicious regions are transmitted to the central site. At the central site, a radiologist reviews the mammographic image data, technologists message (text and graphic), and patient reports using a three monitor custom workstation.

2.1.1 Mammographic image digitization and processing

The first step in the image acquisition pipeline at the remote sites is to digitize the mammographic films using a high-resolution, laser film digitizers (Lumiscan 85, Eastman Kodak, Rochester, NY, USA) at 50 micron pixel dimensions and 12-bit grayscale. Next, the digital images are automatically cropped to reduce the non-tissue areas surrounding the breast, which significantly reduces the image size. A CAD scheme is then executed on the cropped images. Next, the images are compressed at a ratio of 75:1 using the irreversible (lossy), 9/7 transform, wavelet-based JPEG 2000 method. Finally, the image data are parsed into data packets and encrypted using strong 128 bit Microsoft Point-to-Point Encryption (MPPE) with Microsoft Challenge Handshake Authenticate Protocol (CHAP) version 2. The data packets and CAD results are transmitted to the central site.

At the central site, the mammographic image data are decrypted and decompressed. Image display on the workstation is enhanced through minimal unsharp masking. Look-up table (LUT) values are automatically calculated to aid image viewing. To reduce the visual effects of cropping images are restored to full height, but not to full width, by padding (filling) prior to image display. The CAD results are presented as an overlay on the images with regions suspicious for masses outlined and regions suspicious for microcalcification circled.

2.1.2 Remote Site

The computer hardware at the three remote sites is an Athlon 900 machine with a 900 mHz processor and 512 MB of RAM (Advanced Micro Device, Sunnyvale CA, USA) operating under Microsoft Windows 2000 Workstation (Microsoft Corporation, Redmond, WA, USA). They are equipped with both 56K hardware modems and ethernet network cards (Integrated PRO/100 S Desktop Adapter, Intel Corporation, Santa Clara, CA, USA). Sites 1, 2 and 3 are 15, 20, and 15 miles from the central site, respectively. However, we successfully tested the system in the past at a site located 90 miles from the central site. Sites 1 and 2 transmit data across Plain Old Telephone System (POTS) lines. Site 3 transmits data across the Local Area Network (LAN).

The technologists scan the patient's prior report or history using hp Scanjet 5470c scanners (Hewlett-Packard Company, Palo Alto, CA, USA) that are equipped with automatic document feeders. Prior to transmission to the central site the reports are converted to one bit per pixel portable network graphic (PNG) images.

2.1.3 Central Site

The central site telemammography workstation is powered by an Athlon MP dual 1.2 GHz multi-processor with 2 GB of RAM (Advanced Micro Device, Sunnyvale CA, USA), which operates under Microsoft Windows 2000 Server (Fig. 1). The workstation display consist of three high-resolution (2048 x 2560), 8-bit grayscale, portrait monitors at a nominal setting of 80 ftL; two Dome C5i flat-panel monitors (Planar Systems, Beaverton, OR, USA) for image display and one Clinton DS5100P cathode ray tube monitor (Clinton Electronics, Rockford, IL, USA) for text. The workstation communicates with the remote sites via 56K hardware modems (U.S. Robotics, Rolling Meadows, IL, USA) and ethernet network cards (OfficeConnect 10/100 NIC, 3COM, Santa Clara, CA, USA).

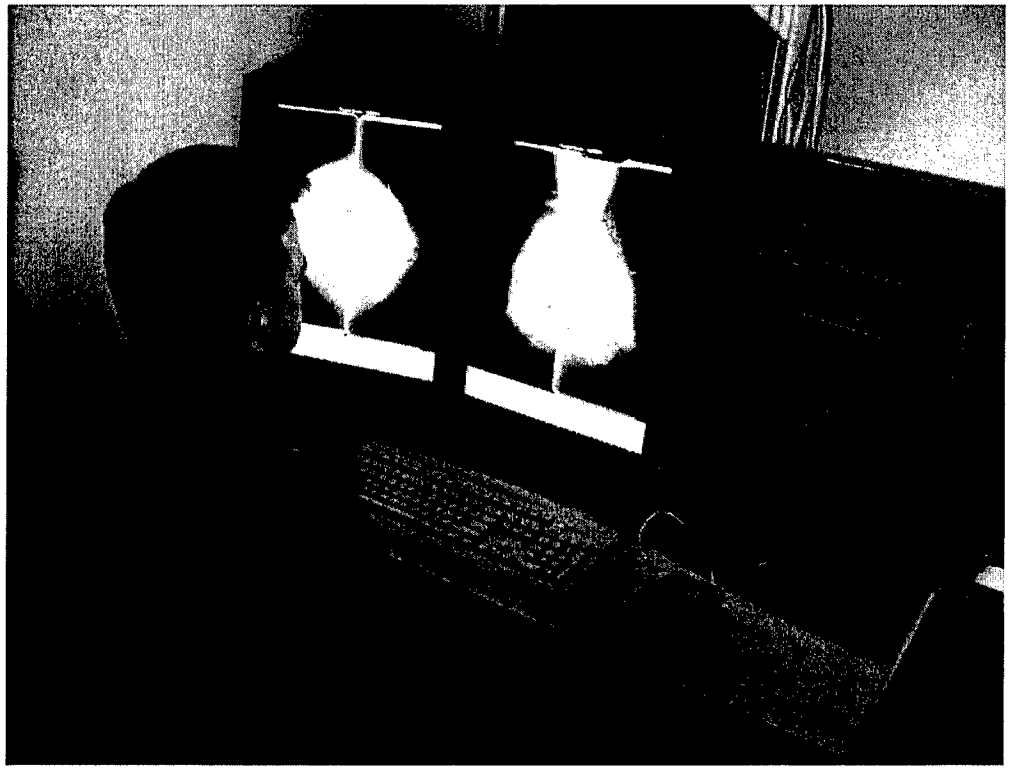

Fig 1. Telemammography workstation at the central site in the default viewing format.

The key display features available on the workstation include manual LUT adjustments, magnification, quadrant viewing (images viewed one quadrant at a time), and multiple display formats, which are all mouse-driven. Possible image display formats include: one image/monitor, two images/monitor, or four images/monitor. The typical display resolution was approximately 100 micron pixel dimensions for one image/monitor and 200 micron pixel dimensions for two images/monitor. Images can be magnified by a free-moving magnification box or quadrant panning. The magnification box size varies depending on the image display format; for one image/monitor the box is 511 x 566 pixels and for two images/monitor the box is 204 x 266 pixels. The left and center monitors display the image data with the CAD results overlaid, and the right monitor displays the message windows, prior reports, case lists, etc.

2.1.4 Inter-site communication

The technologists (remote site) and radiologists (central site) communicate effectively using a message window that features free text and interactive graphic windows and operates in almost real-time (Fig. 2). Typically a message window is sent with each case with communication performed in one cycle. The technologist sends a message with each case, and the radiologist responds directly to the message. The message window at the remote and central sites both contains five areas: (1) patient demographics, (2) message display area, (3) pull-down menus, (4) interactive generic image of breast, and (5) free text area. There are five pull-down menus on the technologist message window to focus communication on possible actionable items that indicate: (1) breast: left or right; (2) view: craniocaudal

and/or mediolateral oblique; (3) finding: mass, calcifications, tissue asymmetry, palpable lump, or nodule; (4) comparison with prior exam: baseline exam, new finding, slight change, moderate change, or remarkable change; (5) other findings stable, and (6) possible additional procedure needed: additional views and/or ultrasound. The interactive generic image of the breast allows the technologists to place an "X" mark precisely on the region of suspicion. The radiologists can reply after reviewing each case. His/her response includes: (1) do recommended procedure as suggested; (2) no additional procedures necessary; and (3) do not do the procedure recommended, but do X, Y, and Z. If the radiologists recommends additional procedures, the interactive generic image of the breast allows the radiologist to place a "square" mark precisely on the region that requires the additional work-up.

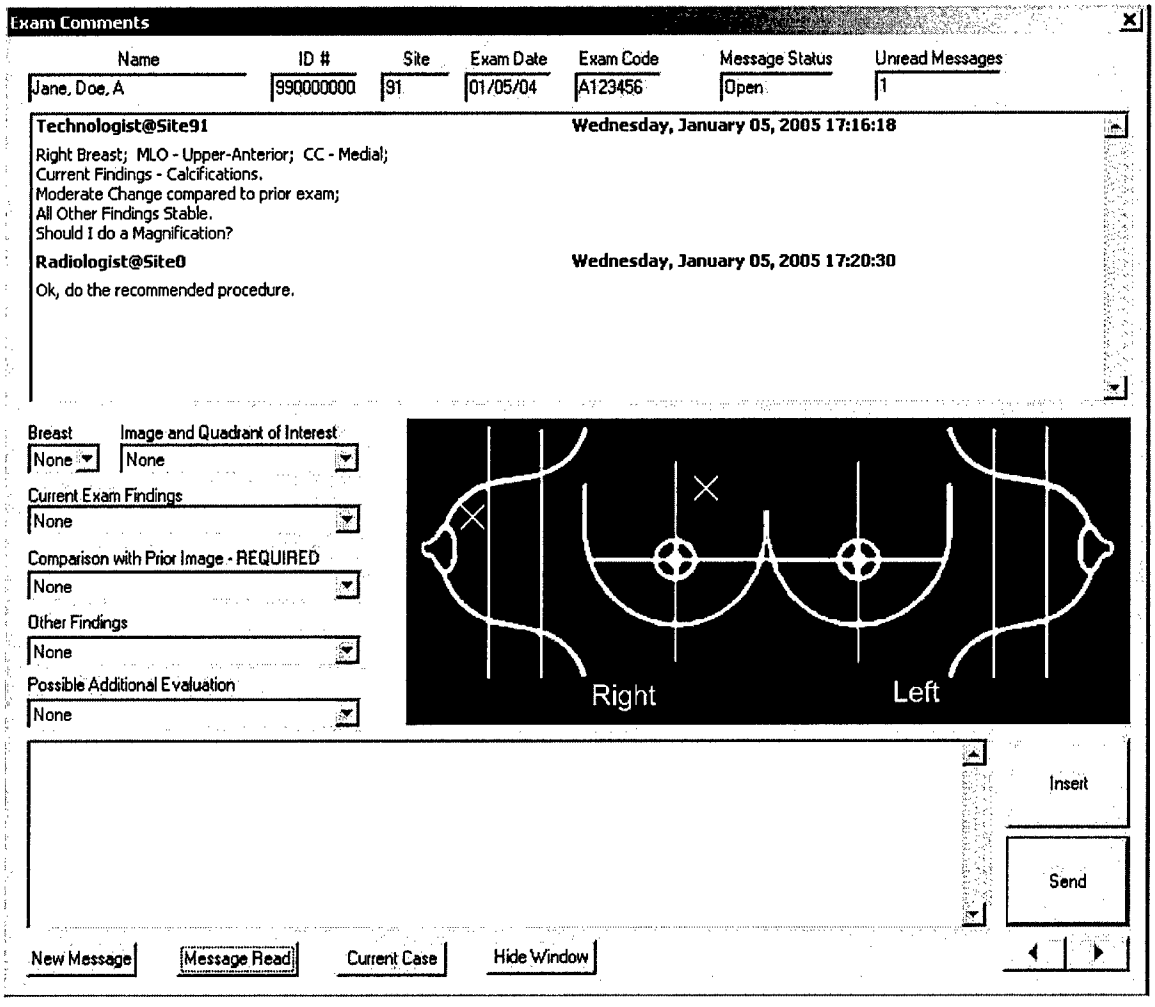

Fig 2. Remote site message window used by the technologists to communicate with the central site (radiologists).

2.1.5 Telemammography system operation

The system has been operational for greater than two years and to date over 2000 cases have been transmitted. The image quality, image display, effects of the image processing, and telemammography system features are generally well-received and considered more than adequate for reviewing screening mammography examinations by participating radiologists. The magnification features provide a detailed review of the breast tissue patterns, particularly microcalcifications. The automated LUT settings were acceptable in nearly 90% of the cases during review. The breast tissue was completely retained following the automated cropping producing images that were visibly appealing for review. There may be some detectable differences at extremely high magnifications between non-compressed digitized mammographic images and images compressed at a 75:1 ratio, but based on several assessments these differences should not affect the diagnostic image quality. In a two-alternative forced choice

discrimination experiment radiologists could not accurately or reliably discriminate between non-compressed images and those compressed at 50:1 and 75:1 compression levels when displayed side-by-side.[24]

2.2 Case selection

Two hundred and forty-five breast cancer screening mammography exams were retrospectively evaluated during this study that were acquired using the telemammography system from three remote woman's imaging centers. Registered mammography technologists from the remote imaging centers transmitted screening exams to the central site (radiologists) that they (the technologists) believed needed additional imaging procedures. The technologists selected the exams prospectively and were unaware at the time of selection whether or not the patient would ultimately undergo additional procedures during the actual clinical interpretation. One hundred and thirty cases were used in Study 1. Study 2 consisted of 99 cases that were a subset of the cases used in Study 1. One hundred and fifteen different cases were used in Study 3. The actual, subsequent clinical interpretation categorized each case using the Breast Imaging Reporting and Data System (BIRADS) (Table 1). The four routine mammographic films acquired at our centers during a breast cancer screening exam include the left and right craniocaudal views (LCC & RCC), and left and right mediolateral oblique views (LMLO & RMLO).

Table 1
Distribution of BIRADS categories as a result of clinical interpretation

BIRADS Category	0	1	2	total
Study 1	51	34	45	130
Study 2	38	25	36	99
Study 3	47	41	27	115

2.3 Study design and data analysis

This study was composed of three separate multi-mode studies in which information was incrementally presented progressively during each of the individual modes (Table 2). All modes were completed during a single reading. Five components of information were presented during the three studies: (1) four mammographic images; (2) technologist's text message detailing the region of suspicions in terms of type of finding (e.g., mass, microcalcifications), location, comparison to prior exams (when available), and their (the technologists) recommended additional procedures; (3) patient's report from the prior mammography exam (when available); (4) technologist's graphic marks on a generic breast image to highlight the region of suspicion; and (5) CAD results.

Table 2
Information presented for case interpretation during each mode of the three studies

Study	Mode 1	Mode 2	Mode 3
1	mammographic images only	mammographic images & technologist's message	n/a
2	mammographic images & technologist's message	mammographic mages, message, & prior report	n/a
3	mammographic images, technologist's message, & prior report	mammographic images, message, prior report, & technologist's graphic marks	mammographic images, message, prior report, graphic marks, & CAD

Seven board certified radiologists specializing in mammography participated as readers in this study who were informed of the exam origination and case selection criteria, but not the mix of "recall" and "no-recall" cases. They reviewed and rated the screening mammography exams on the telemammography workstation and indicated: (1) if additional procedures were recommended, (2) when appropriate, which breast was involved, and (3) when appropriate, the specific recommended procedures. During the telemammography interpretation the ratings were recorded via a computerized scoring form displayed on the workstation monitor using the computer mouse (Fig. 3). The full functionality of the workstation (e.g., window and level, magnification, quadrant viewing) was available during case review. The experience of the radiologists ranged from 6 to 33 years with each performing or reading over 2000 breast imaging procedures per year. Two radiologists participated in all three studies. Two radiologists participated in Studies 1 and 2. Three radiologists participated only in Study 3. (Reader order was scrambled).

The performance of the radiologists when using the telemammography system was compared with the actual clinical interpretation of the same screening mammography examinations. Performance was evaluated in terms of the percent of exams recommended recalled for additional procedures (termed "recall"), the percent agreement, and the Kappa of agreement during both types of interpretations (i.e., telemammography and clinical).

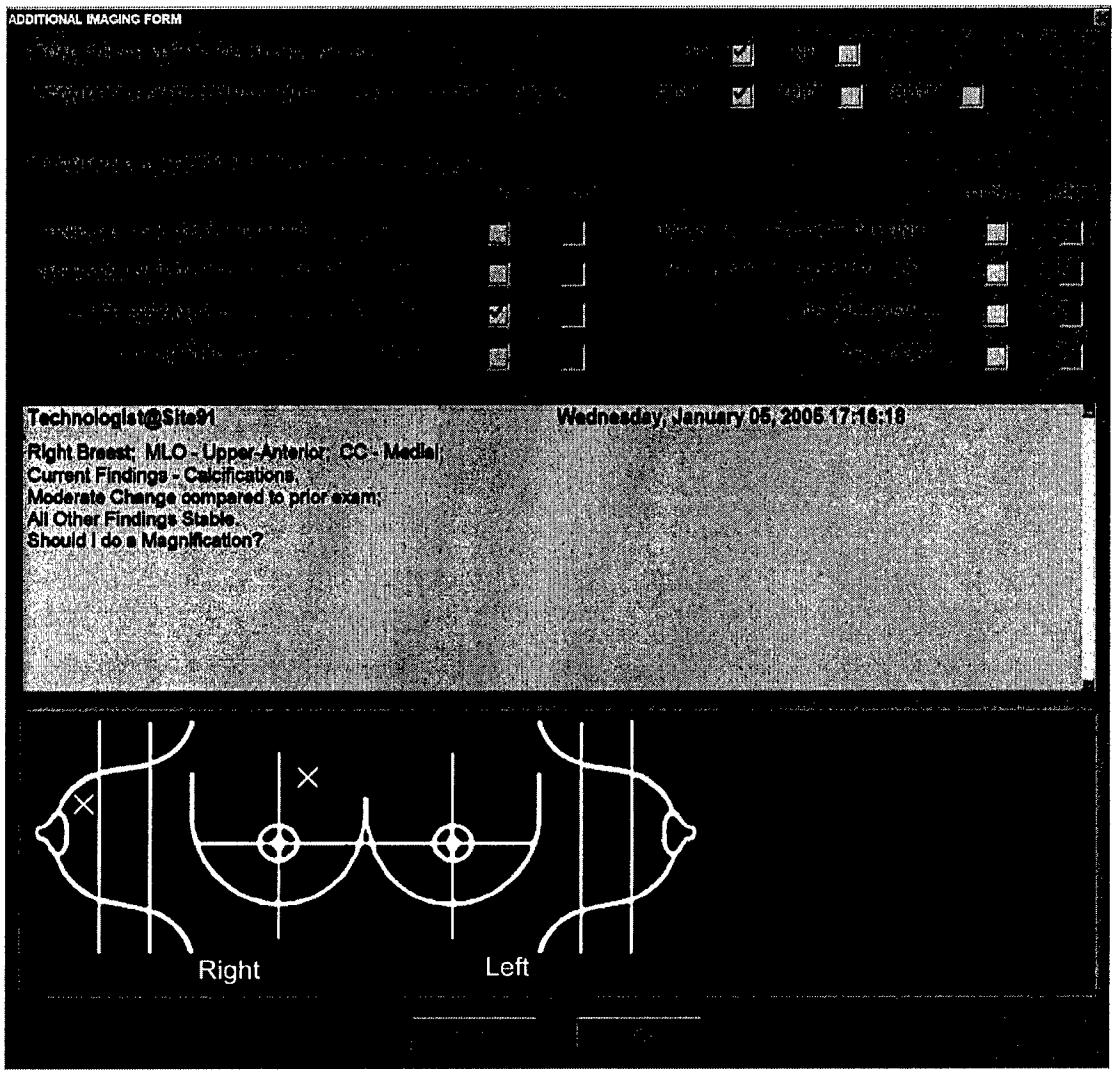

Fig 3. Scoring form used by the radiologists as it would appear during mode 3 of Study 3.

3. RESULTS

Technologists were able to identify suspicious examinations that may require additional procedures, but their "recommended" examinations amounted to a substantially larger number compared with that of the actual clinical interpretation by a radiologist. The percent of exams recommended for recalled for additional procedures (termed "recall") during the actual clinical interpretation for Studies 1, 2 (a subset of Study 1), and 3 were 39.2% (51/130), 38.4% (38/99), and 40.9% (47/115), respectively. The screening exams sent by the technologists were those cases that they (the technologists) believed need additional imaging procedures to complete the exam. The 245 exams were successfully transmitted, processed, reviewed, and rated.

The recall rates for all radiologists during the telemammography interpretations all three multi-mode studies were significantly higher than the actual clinical interpretations (Tables 3 - 9). As a result, there was poor agreement between the two interpretations types (telemammography and clinical) for all studies. The majority of disagreement

interpretations resulted when the telemammography interpretation resulted in a recommendation to recall and the actual clinical interpretation resulted in a recommendation not to recall. There were a total of 1635 disagreement interpretations across all three multi-mode studies between the telemammography and actual clinical interpretation and of these disagreements 86.7% (1418/1635) occurred when the telemammography interpretation resulted in recall and the clinical interpretation resulted in a recommendation not to recall.

In Study 1, the recall rates of all the radiologists during the telemammography interpretations significantly increased from mode 1 (images only) to mode 2 (images and technologist's text message) while the agreement between the telemammography and actual clinical interpretations decreased from mode 1 to mode 2 for three out of four radiologists (Tables 3 and 4). Modes 1 and 2 of Study 1 had mean Kappa of 0.125 (+/- 0.041) and 0.102 (+/-0.059), respectively, mean agreements of 51.7% (+/- 5.5) and 48.7% (+/- 6.3), respectively, and mean recall rates of 73.3% (+/- 17.9) and 82.5% (+/- 16.2), respectively. The mean number of disagreement interpretations between the telemammography and actual clinical interpretations were 62.8 and 66.8 for modes 1 and 2, respectively. The mean percentage of these disagreements occurring when the telemammography interpretation resulted in a recommendation to recall and the clinical interpretations resulted in a recommendation not to recall were 83.9% (53.5/62.8) and 91.2% (61.5/66.8) for modes 1 and 2, respectively.

Table 3
Study 1, mode 1 (images only): telemammography workstation interpretations compared to clinical interpretations

Telemammography recommendations	Clinical interpretation recall (n = 51)	no-recall (n = 79)	Total (n=130)	Kappa
Radiologist 1				0.097
recall	38.5% (50)	52.3% (68)	90.8% (118)	
no-recall	0.8% (1)	8.5% (11)	9.2% (12)	
Radiologist 2				0.127
recall	31.5% (41)	40.0% (52)	71.5% (93)	
no-recall	7.7% (10)	20.8% (27)	28.5% (37)	
Radiologist 3				0.182
recall	23.8% (31)	25.4% (33)	49.2% (64)	
no-recall	15.4% (20)	35.4% (46)	50.8% (66)	
Radiologist 4				0.093
recall	34.6% (45)	46.9% (61)	81.5% (106)	
no-recall	4.6% (6)	13.8% (18)	18.5% (24)	

Table 4
Study 1, mode 2 (images and technologist's text message): telemammography workstation interpretations compared to clinical interpretations

Telemammography recommendations	Clinical interpretation recall (n = 51)	no-recall (n = 79)	Total (n=130)	Kappa
Radiologist 1				0.020
recall	39.2% (51)	59.2% (77)	98.5% (128)	
no-recall	0.0% (0)	1.5% (2)	1.5% (2)	
Radiologist 2				0.103
recall	36.2% (47)	48.5% (63)	84.6% (110)	
no-recall	3.1% (4)	12.3% (16)	15.4% (20)	
Radiologist 3				0.159
recall	27.7% (36)	32.3% (42)	60.0% (78)	
no-recall	11.5% (15)	28.5% (37)	40.0% (52)	
Radiologist 4				0.124
recall	37.7% (49)	49.2% (64)	86.9% (113)	
no-recall	1.5% (2)	11.5% (15)	13.1% (17)	

The differences between modes 1 (images and technologist's text message) and 2 (images, technologist's text message, and prior report) of Study 2 were a slight decrease in the recommended recall rates for three of the four radiologists from mode 1 to mode 2 as well as small changes in agreement across the radiologists between the two modes (Tables 5 and 6). Modes 1 and 2 of Study 2 had mean Kappa of 0.163 (+/- 0.077) and 0.165 (+/- 0.081), respectively, mean agreements of 52.3% (+/- 6.7) and 52.8% (+/- 7.0), respectively, and mean recall rates of 79.6% (+/- 12.3) and 77.5% (+/- 13.8), respectively. The mean number of disagreement interpretations between the telemammography and actual clinical interpretations were 47.3 and 46.8 for modes 1 and 2, respectively. The mean percentage of these disagreements occurring when the telemammography interpretation resulted in a recommendation to recall and the clinical interpretation resulted in a recommendation not to recall were 92.5% (44.0/47.3) and 90.8% (42.8/46.8) for modes 1 and 2, respectively.

Table 5
Study 2, mode 1 (images and technologist's text message): telemammography workstation interpretations compared to clinical interpretations

Telemammography recommendations	Clinical interpretation recall (n = 38)	no-recall (n = 61)	Total (n=99)	Kappa
Radiologist 1				0.219
recall	32.3% (32)	36.4% (36)	68.7% (68)	
no-recall	6.1% (6)	25.3% (25)	31.3% (31)	
Radiologist 2				0.208
recall	37.4% (37)	44.4% (44)	81.8% (81)	
no-recall	1.0% (1)	17.2% (17)	18.2% (18)	
Radiologist 3				0.174
recall	32.3% (32)	39.4% (39)	71.7% (71)	
no-recall	6.1% (6)	22.2% (22)	28.3% (28)	
Radiologist 4				0.051
recall	38.4% (38)	57.6% (57)	96.0% (95)	
no-recall	0.0% (0)	4.0% (4)	4.0% (4)	

Table 6
Study 2, mode 2 (images, technologist's text message, and prior report): telemammography workstation interpretations compared to clinical interpretations

Telemammography recommendations	Clinical interpretation recall (n = 38)	no-recall (n = 61)	Total (n=99)	Kappa
Radiologist 1				0.206
recall	30.3% (30)	34.3% (34)	64.6% (64)	
no-recall	8.1% (8)	27.3% (27)	35.4% (35)	
Radiologist 2				0.237
recall	37.4% (37)	42.4% (42)	79.8% (79)	
no-recall	1.0% (1)	19.2% (19)	20.2% (20)	
Radiologist 3				0.167
recall	31.3% (31)	38.4% (38)	69.7% (69)	
no-recall	7.1% (7)	23.2% (23)	30.3% (30)	
Radiologist 4				0.051
recall	38.4% (38)	57.6% (57)	96.0% (95)	
no-recall	0.0% (0)	4.0% (4)	4.0% (4)	

The difference between modes 1 (images and technologist's text message), 2 (images, technologist's text message, and prior report), and 3 (images, technologist's text message, prior report, and CAD) in Study 3 were relatively small with a slight decreased in Kappa between the three modes (Tables 7, 8 and 9). Modes 1, 2 and 3 of Study 3 had mean Kappa of 0.213 (+/- 0.072), 0.206 (+/- 0.060), and 0.201 (+/- 0.061), respectively, mean agreements of 57.4% (+/-

4.6), 57.1% (+/- 3.9), and 56.7% (+/- 3.9), respectively, and mean recall rates of 72.3% (+/- 9.3), 72.3% (+/- 9.3), and 72.7% (+/- 9.2), respectively. The mean number of disagreement interpretations between the telemammography and actual clinical interpretations were 49.0, 49.4, and 49.8 for modes 1, 2 and 3, respectively. The mean percentage of these disagreements occurring when the telemammography interpretation resulted in a recommendation to recall and the clinical interpretation resulted in a recommendation not to recall was 82.7% (40.6/49.0), 82.4% (40.8/49.4) and 81.9% (40.8/49.8) for modes 1, 2, and 3, respectively.

Table 7
Study 3, mode 1 (images, technologist's text message, and prior report): telemammography workstation interpretations compared to clinical interpretations

Telemammography recommendations	Clinical interpretation recall (n = 47)	no-recall (n = 68)	Total (n=115)	Kappa
Radiologist 1				0.255
recall	38.3% (44)	38.3% (44)	76.5% (88)	
no-recall	2.6% (3)	20.9% (24)	23.5% (27)	
Radiologist 2				0.152
recall	39.1% (45)	46.1% (53)	85.2% (98)	
no-recall	1.7% (2)	13.0% (15)	14.8% (17)	
Radiologist 3				0.318
recall	33.9% (39)	28.7% (33)	62.6% (72)	
no-recall	7.0% (8)	30.4% (35)	37.4% (43)	
Radiologist 4				0.157
recall	30.4% (35)	33.9% (39)	64.3% (74)	
no-recall	10.4% (12)	25.2% (29)	35.7% (41)	
Radiologist 5				0.182
recall	34.8% (40)	38.3% (44)	73.0% (84)	
no-recall	6.1% (7)	20.9% (24)	27.0% (31)	

Table 8
Study 3, mode 2 (images, technologist's text message, prior report, and technologist's graphic marks): telemammography workstation interpretations compared to clinical interpretations

Telemammography recommendations	Clinical interpretation recall (n = 47)	no-recall (n = 68)	Total (n=115)	Kappa
Radiologist 1				0.255
recall	38.3% (44)	38.3% (44)	76.5% (88)	
no-recall	2.6% (3)	20.9% (24)	23.5% (27)	
Radiologist 2				0.152
recall	39.1% (45)	46.1% (53)	85.2% (98)	
no-recall	1.7% (2)	13.0% (15)	14.8% (17)	
Radiologist 3				0.285
recall	33.0% (38)	29.6% (34)	62.6% (72)	
no-recall	7.8% (9)	29.6% (34)	37.4% (43)	
Radiologist 4				0.157
recall	30.4% (35)	33.9% (39)	64.3% (74)	
no-recall	10.4% (12)	25.2% (29)	35.7% (41)	
Radiologist 5				0.182
recall	34.8% (40)	38.3% (44)	73.0% (84)	
no-recall	6.1% (7)	20.9% (24)	27.0% (31)	

Table 9
Study 3, mode 3 (images, technologist's text message, prior report, technologist's graphic marks, and CAD): telemammography workstation interpretations compared to clinical interpretations

Telemammography recommendations	Clinical interpretation recall (n = 47)	no-recall (n = 68)	Total (n=115)	Kappa
Radiologist 1				0.241
recall	38.3% (44)	39.1% (45)	77.4% (89)	
no-recall	2.6% (3)	20.0% (23)	22.6% (26)	
Radiologist 2				0.152
recall	39.1% (45)	46.1% (53)	85.2% (98)	
no-recall	1.7% (2)	13.0% (15)	14.8% (17)	
Radiologist 3				0.285
recall	33.0% (38)	29.6% (34)	62.6% (72)	
no-recall	7.8% (9)	29.6% (34)	37.4% (43)	
Radiologist 4				0.143
recall	30.4% (35)	34.8% (40)	65.2% (75)	
no-recall	10.4% (12)	24.3% (28)	34.8% (40)	
Radiologist 5				0.182
recall	34.8% (40)	38.3% (44)	73.0% (84)	
no-recall	6.1% (7)	20.9% (24)	27.0% (31)	

4. DISCUSSION

In this controlled study, the percentage of breast cancer screening mammography exams recommended for additional procedures ("recall") by the radiologists when interpreting exams suspected by technologists to require additional procedures using the telemammography system was significantly higher than the actual clinical interpretations of the same exams. Adding non-mammographic image information (i.e., technologist's text message to describe suspicious regions, prior patient reports or history, technologist's graphic marks to highlight suspicious regions, and CAD) to the telemammography system did not significantly change the radiologists' interpretations compared with mammographic image only interpretations. The majority of disagreement occurred when the telemammography interpretation recommended recall and clinical interpretations recommended no-recall.

The significantly high recall rates (nearly double) interpreting screening mammography exams using the telemammography system as compared to the actual clinical interpretation is in agreement with our previous study [29] and similar to Elmore et al.[13] (1994). The mean telemammography recall rates ranged from 72.3% to 82.5% while the actual clinical recall rates ranged from 38.4% to 42.3% across the three separate multi-mode studies.

This study indicates that technologists are sensitive, if not specific, to the mammography features and changes that may lead to a recall. There were 245 unique screening mammography exams used in this study and 60.0% (147/245) were not recommended for recall for additional procedures during the actual clinical interpretation. Of these 147 exams not recalled 49.0% (72/147) had a clinical BIRADS category of 2. Therefore, the technologists were able to detect abnormal findings during the mammography exam, but were not skilled at differentiating whether the findings may represent potential disease.

The radiologists' high recall rates for additional procedures using the telemammography system suggests a critical dependence on prior images when making management decisions albeit other factors may have had some effect. Therefore, our telemammography system was recently modified to incorporate images from the patient's prior screening exam (when available). The modified telemammography system to include prior image data was tested operationally and a clinically simulated study to evaluate the impact of this final modification on recommended recall rates is underway.

There are several limitations to the current study that may have caused the high recall rate during the telemammography interpretation. First, the telemammography interpretation in this study constituted a limited review due to the lack of images from the prior screening mammography exam for comparison. Second, the lack of prior

images combined with the technologists skill at detecting abnormalities in the mammography exams and their (the technologist) description of the abnormality (e.g., new finding, moderate change in finding) may have influenced the radiologists to recommend additional procedures to further evaluate the "abnormal" findings suggested by the technologist. Third, the participating radiologists may have expected an "enriched" sample population because they knew that this was a laboratory study. A similar explanation for high recall rates was reported in Elmore et al.[13] (1994). Finally, this retrospective study did not affect clinical management of the patient, so this knowledge may prompted the observed over-reading.

The seven experienced radiologists who participated in this study confirmed the feasibility of our telemammography system to provide remote patient "management" when a physician is not present in the clinic. Particularly, our effort to reduce the number of patients recommended for recalled for additional procedures as part of breast cancer screening mammography through the identification of these patients while they remain at the remote clinic, hence reducing patient anxiety associated with recall. The limited information provided to the radiologists (i.e., no images from the prior exam) enabled a moderate reduction in the number of recommended recalls by the technologists by approximately 25 percent. Inclusion of the final pertinent information component, mammographic images from the prior exam, is expected to further reduce the number of recommended procedures and significantly improve the agreement between the telemammography and actual clinical interpretations.

ACKNOWLEDGEMENTS

This work is supported in part by the US Army Medical Research Acquisition Center, 820 Chandler Street, Fort Detrick, MD 21702-5014 under contract DAMD17-00-1-0410. The content of the information contained herein does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

REFERENCES

1. Duffy SW, Tabar L, Chen HH, Holmqvist M, Yen MF, Abdsalah S, Epstein B, Frodis E, et al. "The impact of organized mammography service screening on breast carcinoma mortality in seven Swedish Counties." *Cancer* **95(3)**:458-469, 2002.
2. Feig SA. Current status of screening mammography. *Obstet Gynecol Clin North Am* 2002;29(1):123-136.
3. Humphrey LL, Helfand M, Chan BKS, and Woolf SH. "Breast cancer screening: a summary of the evidence for the U.S. preventive services task force." *Ann Intern Med* **137(5, Part 1)**:E347-E367, 2002.
4. Lee CH. "Screening mammography: proven benefit, continued controversy." *Radiol Clin North Am* **40(3)**:395-407, 2002.
5. Nystrom L, Andersson I, Bjurstam N, Frisell J, Nordenskjold B, and Rutqvist LE. "Long-term effects of mammography screening: update overview of the Swedish randomised trials." *Lancet* **359(9310)**:909-919, 2002.
6. Tabar L, Vitak B, Chen HHT, Yen MF, Duffy SW, and Smith RA. "Beyond randomized clinical trials: organized mammographic screening substantially reduces breast carcinoma mortality." *Cancer* **91(9)**:1724-1731, 2002.
7. Gotzsche PC and Olsen O. "Is screening for breast cancer with mammography justifiable?" *Lancet* **355**:129-134, 2000.
8. Olsen O and Gotzsche PC. "Cochrane review on screening for breast cancer with mammography." *Lancet* **358**:1340-1342, 2001.
9. Miller AB, To T, Baines CJ, and Wall C. "The Canadian Nation Breast Screening Study-1: nreast cancer mortality after 11 to 16 years of follow-up: a randomized screening trial of mammography in women age 40 to 49 years." *Ann Intern Med* **137(5, Part 1)**:E305-E315, 2002.
10. Chamot E and Perneger TV. "Misconception about efficacy of mammography screening: a public health dilemma." *J Epidemiol Community Health* **55(11)**:799-803, 2001.
11. Coughlin SS, Thompson TD, Hall HI, Logan P, and Uhler RJ. "Breast and cervical carcinoma screening practices among women in rural and nonrural areas of the United States, 1998-1999." *Cancer* **94(11)**:2801-2812, 2002.
12. Michaelson J, Satija S, Moore R, Weber G. Halpern E, Garland A, Puri D, and Kopans DB. "The pattern of breast cancer screening utilization and its consequences." *Cancer* **94(1)**:37-43, 2002.
13. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. "Variability in radiologists' interpretations of mammograms." *N Engl J Med* **331(22)**:1493-1499, 1994.

14. Warren RML and Duffy SW. "Comparison of single reading with double reading of mammograms and change in effectiveness with experience." *Br J Radiol* **68(813)**:958-962, 1995.
15. Hulka CA, Slanetz PJ, Halpern EF, Hall DA, McCarthy KA, Moore R, Boutin S, and Kopans DB. "Patients' opinion of mammography screening services: immediate results versus delayed results due to interpretation by two observers." *AJR Am J Roentgenol* **168**:1085-1089, 1997.
16. Yawn B, Krein S, Christianson J, Hartley D, and Moscovice I. "Rural radiology: who is producing images and who is reading them?" *J Rural Health* **13(2)**:136-144, 1997.
17. Bolle SR, Sund T, and Stormer J. "Receiver operating characteristic study of image processing for teleradiology and digital workstations." *J Digit Imaging* **10(4)**:152-157, 1997.
18. Maitz GS, Chang TS, Sumkin JH, Wintz PW, Johns CM, Ganott M, Holbert BL, Hakim CM, Harris KM, Gur D, and Herron JM. "Preliminary clinical evaluation of a high-resolution telemammography System." *Invest Radiol* **32(4)**:236-240, 1997.
19. Mitra S, Yang S, and Kustov V. "Wavelet-based vector quantization for high-fidelity compression and fast transmission of medical images." *J Digit Imaging* **11(4 Suppl 2)**:24-30, 1998.
20. Kalyanpur A, Neklesa VP, Taylor CR, Daftary AR, and Brink JA. "Evaluation of JPEG and wavelet compression of body CT images for direct digital teleradiology transmission." *Radiology* **217(3)**:772-779, 2000.
21. Lou SL, Lin HD, Lin KP, and Hoogstrate D. "Automatic breast region extraction from digital mammograms for PACS and telemammography applications." *Comput Med Imaging Graph* **24(4)**:205-220, 2000.
22. Ludwig K, Bick U, Oelerich M, Schuierer G, Puskas Z, Nicolas K, Koch A, and Lenzen H. "Is image selection a useful strategy to decrease the transmission time in teleradiology? A study of 100 emergency cranial CTs." *Eur Radiol* **8(9)**:1719-1721, 1998.
23. Lou SL, Sickles EA, Huang HK, Hoogstrate D, Cao F, Wang J, and Jahangiri M. "Full-field direct digital telemammography: Technical components, study protocols, and preliminary results." *IEEE Trans Info Technol Biomed* **1(4)**:270-278, 1997.
24. Leader JK, Sumkin JH, Ganott MA, Hakim C, Hardesty L, Shah R, Wallace L, Klym A, Drescher JM, Maitz GS, Gur D. "Subjective assessment of high-level image compression of digitized mammograms." *Proceedings of SPIE Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment* **5372**:415-422, February 2004.
25. Yawn B, Krein S, Christianson J, Hartley D, and Moscovice I. "Rural radiology: who is producing images and who is reading them?" *J Rural Health* **13(2)**:136-144, 1997.
26. Benger JR. "Can nurses working in remote units accurately request and interpret radiographs?" *Emerg Med J* **19(1)**:68-70, 2002.
27. Coughlin SS, Thompson TD, Hall HI, Logan P, and Uhler RJ. "Breast and cervical carcinoma screening practices among women in rural and nonrural areas of the United States, 1998-1999". *Cancer* **94(11)**:2801-2812, 2002.
28. 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." *Proceedings of SPIE Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation* **4685**:416-421, February 2002.
29. Leader JK, Wallace LP, Hakim CM, Hertzberg TM, Hardesty LA, Sumkin JH, Cohen C, Sneddon C, Lindeman S, Craig D, and Drescher JM. "Preliminary clinical evaluation of a multi-site telemammography system in a screening mammography environment." *Proceedings of SPIE Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation* **5033**:273-280, February 2003.
30. Drescher JM, Maitz GS, Traylor C, Leader JK, Clearfield RJ, Shah R, Ganott MA, Pugliese F, Duffner D, Lockhart J, Gur D. "A multi-site telemammography system: preliminary assessment of technical and operational issues." *Proceedings of SPIE Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation*, **5033**:360-369, February 2003.