Published in final edited form as:
Radiology. 2004 April ; 231(1): 208–214. doi:10.1148/radiol.2311030429.

Sensitivity of Noncommercial Computer-aided Detection System for Mammographic Breast Cancer Detection: Pilot Clinical Trial

Mark A. Helvie, MD, **Lubomir Hadjiiski, PhD**, **Erini Makariou, MD**, **Heang-Ping Chan, PhD**, **Nicholas Petrick, PhD**, **Berkman Sahiner, PhD**, **Shih-Chung B. Lo, PhD**, **Matthew Freedman, MD**, **Dorit Adler, MD**, **Janet Bailey, MD**, **Caroline Blane, MD**, **Donna Hoff, MD**, **Karen Hunt, MD**, **Lynn Joynt, MD**, **Katherine Klein, MD**, **Chintana Paramagul, MD**, **Stephanie K. Patterson, MD**, and **Marilyn A. Roubidoux, MD**
From the Department of Radiology, University of Michigan Health Systems, 1500 E Medical Center Dr, TC 2910, Ann Arbor, MI 48109-0326 (M.A.H., L.H., H.P.C., N.P., B.S., D.A., J.B., C.B., D.H., K.H., L.J., K.K., C.P., S.K.P., M.R.); and Department of Radiology, Georgetown University Hospital, Washington, DC (E.M., S.C.B.L., M.F.).

Abstract

PURPOSE—To evaluate a noncommercial computer-aided detection (CAD) program for breast cancer detection with screening mammography.

MATERIALS AND METHODS—A CAD program was developed for mammographic breast cancer detection. The program was applied to 2,389 patients' screening mammograms at two geographically remote academic institutions (institutions A and B). Thirteen radiologists who specialized in breast imaging participated in this pilot study. For each case, the individual radiologist performed a prospective Breast Imaging Reporting and Data System (BI-RADS) assessment after viewing of the screening mammogram. Subsequently, the radiologist was shown CAD results and rendered a second BI-RADS assessment by using knowledge of both mammographic appearance and CAD results. Outcome analysis of results of examination in patients recalled for a repeat examination, of biopsy, and of 1-year follow-up examination was recorded. Correct detection with CAD included a computer-generated mark indicating a possible malignancy on craniocaudal or mediolateral oblique views or both.

RESULTS—Eleven (0.46%) of 2,389 patients had mammographically detected nonpalpable breast cancers. Ten (91%) of 11 (95% CI: 74%, 100%) cancers were correctly identified with CAD. Radiologist sensitivity without CAD was 91% (10 of 11; 95% CI: 74%, 100%). In 1,077 patients, follow-up findings were documented at 1 year. Five (0.46%) patients developed cancers, which were found on subsequent screening mammograms. The area where the cancers developed in two (40%) of these five patients was marked (true-positive finding) by the computer in the preceding year. Because of CAD results, a 9.7% increase in recall rate from 14.4% (344 of 2,389) to 15.8% (378 of 2,389) occurred. Radiologists' recall rate of study patients prior to use of CAD was 31% higher than the average rate for nonstudy cases (10.3%) during the same time period at institution A.

Address correspondence to M.A.H.

From the 2002 RSNA scientific assembly.

See also the editorial by Krupinski in this issue.

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

CONCLUSION—Performance of the CAD program had a very high sensitivity of 91% (95% CI: 74%, 100%).

Index terms

Breast neoplasms, diagnosis, 00.30; Cancer screening; Computers, diagnostic aid; Diagnostic radiology, observer performance

---

Breast cancer remains the most common nonskin cancer in women. Two hundred eleven thousand women were expected to receive a diagnosis of breast cancer in 2003, with 40,200 deaths expected in the United States (1). Results of early detection with mammographic screening and physical examination demonstrated improved survival from breast cancer in randomized controlled trials (2,3). More recently, mammographic screening of large populations ("service screening") demonstrated a 40%–63% reduction in breast cancer deaths for women who underwent screening (4,5).

Because of the success of mammographic screening, efforts to improve screening methods have been actively investigated. These efforts include improvement in mammographic equipment and mammographic film, development of mammographic digital detectors, and broad application of quality improvement techniques. Radiologist interpretation skills can also be improved; through the Mammography Quality Standards Act, the federal government has mandated specific physician educational and practice standards for radiologists to achieve high-quality readings. A growing area of interest has been the use of computer-aided detection (CAD) methods to improve radiologist performance in detection of mammographic abnormalities, as well as in characterization of detected abnormalities (6–17).

Use of a second reader of mammograms has been shown to increase cancer detection (18,19). CAD offers the potential to act as a second reader. Although most CAD systems have shown improvement in sensitivity (ability to detect more cancers at mammography), this improvement has been accompanied by many false-positive marks made by CAD that are evaluated by the radiologist. For example, if a CAD system is allowed one mark per image (four marks for a typical two-view bilateral mammogram), for every 1,000 screening bilateral mammograms, 4,000 marks (4,000 images multiplied by one mark per image) will be generated. Since the cancer incidence in women who undergo screening annually is only three per 1,000, then three to six (0.1%) of 4,000 marks will be true-positive. The majority of marks, 3,994 –3,997 (99.9%) of 4,000, will be false-positive. With most CAD systems, a mark is considered true-positive if it is placed on the craniocaudal or mediolateral oblique views or both. The radiologist becomes the discriminator used by CAD to separate true-positive from false-positive marks. The long-term clinical importance of these CAD trends in mammographic reading and interpretation has yet to be fully worked out. High sensitivity is generally perceived as a desirable feature of CAD systems, and the radiologist therefore accepts many false-positive marks that must be assessed.

The purpose of our study was to evaluate a noncommercial CAD program for breast cancer detection with screening mammography.

MATERIALS AND METHODS

Institutional review board approval was obtained prior to the commencement of this investigation. Individual informed consent was obtained from all subjects.

CAD System

A noncommercial CAD system (M-vision; Department of Radiology, University of Michigan, Ann Arbor) was developed for detection of mammographically depicted carcinoma. The goal

of use of this system was performance better than that of existing commercial systems. The CAD system included programs for detection of masses and of microcalcifications. The computer vision techniques used in these detection programs and their performance have been discussed elsewhere (20,21).

Briefly, for the detection of masses, the digitized mammograms were preprocessed with a nonlinear density-weighted contrast enhancement filter to accentuate mammographically depicted structures. Edge detection was then used to define the borders of the enhanced structures. The local maxima of these structures were used as seed locations to identify seed objects. The object definition was refined by using the k-means clustering algorithm of feature vectors of the pixels in a background-corrected region of interest that enclosed each seed object. Morphologic and textural features were then extracted from the refined objects. Rule-based and linear discriminant classifiers were applied to the feature space to distinguish masses from normal structures. The free-response receiver operating characteristic performance of the mass detection program was evaluated with independent test sets (20).

For the detection of microcalcifications, a linear band-pass filter was used to enhance the signals and suppress the low-frequency background on the digitized mammograms. Potential signals were identified by means of global gray-level thresholding and subsequent adaptive local thresholding. The number of false-positive marks was reduced by using rule-based classification, with morphologic features extracted from each potential signal. An artificial neural network was then applied to the remaining signals to further distinguish true- and false-positive microcalcifications. Finally, regional clustering was used to locate clustered microcalcifications that were suspected of being cancerous (21).

Subjects and Imaging

Subjects were recruited from two academic medical centers in different parts of the United States. At institution A about 32,000 breast examinations per year were performed, and at institution B about 11,000 examinations per year were performed. Eligibility criteria for subjects allowed inclusion of women of any age who were undergoing routine screening mammography with a delayed reading, which was generally a reading on the following day. Patients were excluded who had palpable abnormalities, current clinical concerns, prior breast cancer, and breast implants or who underwent follow-up mammography for probably benign findings. Patients were recruited by the mammographic technologist at the time of screening mammography and, if they were interested, were given a detailed patient information packet for informed consent. If the individual agreed, she was entered into the study. There were 2,389 subjects; 59 patients were examined again at a subsequent annual examination. Patients ranged in age from 26 to 85 years, with a mean age of 52 years.

Mammographic images were obtained by using Mammography Quality Standards Act–approved systems. At institution A, one type of equipment (DMR; GE Medical Systems, Milwaukee, Wis) was used with a screen-film system (Kodak 2000; Eastman Kodak, Rochester, NY). At institution B, a different type of equipment (M-IV; Lorad, Danbury, Conn) with a screen-film system (AD; Fuji, Stamford, Conn) was used. Routine mediolateral oblique and craniocaudal views of each breast were obtained.

Digitization and CAD Evaluation

For the recruited subjects, the screening mammograms at institution A were digitized with a laser scanner (model 85; Lumisys, Los Altos, Calif) at a pixel size of 0.05 × 0.05 mm and 12-bit gray levels. The pixel size of these images was automatically increased to 0.1 × 0.1 mm by means of averaging adjunct 2 × 2 pixels and subsampling before the images were inputted to the CAD system. The screening mammograms at institution B were digitized with another laser

scanner (model 150; Lumisys) at a pixel size of 0.1 × 0.1 mm and 12-bit gray levels. The optical density in the film was digitized linearly to a pixel value at a calibration of 0.001 optical density unit per pixel by using both digitizers.

After digitization, the CAD program was applied to the mammograms. The decision threshold for detection of masses was chosen to operate at high sensitivity, with a maximum of three marks allowed per image. The decision threshold for detection of microcalcifications was chosen to operate at a high sensitivity, with one false-positive cluster of microcalcifications allowed per image, on average.

Radiologists

Radiologists were aware that not all CAD marks would be true-positive for cancer. They were not aware of the exact sensitivity of the system but knew it would be similar to that of commercial systems. Individual radiologists were assigned daily to the preexisting department of radiology breast imaging schedule of the study institution, and their frequency of assignment was proportional to their hours worked in breast imaging. All 13 radiologists were certified according to the Mammography Quality Standards Act and had experience ranging from 3 to 24 years (mean, 9.7 years) in mammographic interpretation. At both centers, the study radiologists generally worked at least half of their time in breast imaging and interpreted findings in an average of 3,300 examinations per year; 77% were fellowship trained in breast imaging.

Radiologists' Performance

Screening mammographic images obtained in subjects in the CAD study were placed on mammographic viewing alternators routinely used for screening. Images were displayed with other screening images that were not part of the CAD study. Approximately 30 –75 images were read daily, and a minority (range, 1–10) of them were obtained in subjects in the CAD study. The reader was aware that a particular image was part of the CAD study because there was a notation next to the patient's name. Comparison images, when available, were displayed at the same time. Therefore, reading conditions for the images in the CAD study were similar to the reading conditions for the images that were not part of the CAD study during the same time period. These methods were used to minimize potential bias of the radiologists.

Once a mammographic image had been read in the routine manner, the radiologist was asked to interact with an adjacent personal computer–based CAD workstation. The CAD workstation has graphical user interface and was developed in conjunction with the radiologists. A training session with 20 previous screening images from patient files was given to each participating radiologist to familiarize him or her with its operation prior to the beginning of the clinical experiment. The geographic position of the display facilitated easy physician access to the keyboard and the mouse. The radiologist's first action was to record any potential mass on the video display of the four views (craniocaudal and mediolateral oblique views of each breast) that he or she had detected during the initial clinical reading. All abnormalities were considered "masses" or "calcifications" for the purpose of CAD. The "mass" category included masses, as well as architectural distortion and focal densities. The radiologist next selected the appropriate Breast Imaging Reporting and Data System (BI-RADS) assessment category for potential masses. If there was no mass, a BI-RADS 1 or 2 assessment category was recorded. He or she was forced to commit and register his or her interpretation prior to visualization of the CAD results. The radiologists were blocked from "looking ahead" at the CAD result without registering an initial prospective interpretation. Next, similar action was performed for any potential calcification clusters.

After the radiologists inputted their initial BI-RADS interpretations and registered their marks at the CAD workstation, the CAD system displayed images with marks at the sites of potential masses on the video monitor. The radiologist then reviewed the original screening mammograms, with consideration of annotations on the CAD image, and made a second decision regarding the presence or absence of a mass and its BI-RADS category based on a review of the mammograms and the CAD information. For example, if two potential mass lesions were marked on the CAD system images and the physician thought both were false-positive, the physician would ignore the CAD information and code the mammogram as BI-RADS category 1 or 2. Alternatively, if the computer identified two potential masses and the radiologist confirmed one of the two masses to represent a true finding, this mass would be marked on the computer screen as a true mass by the radiologist, and he or she would enter a BI-RADS category of 0, with a need for recalling the patient. If the radiologist was concerned about an area not marked by the CAD system, he or she would annotate this area on the screen. Next, the same process was repeated with microcalcifications and required the physician to prospectively render his or her opinion before CAD display and after CAD detection. The markings and action categories of the radiologists both before and after CAD display were recorded in a database file. Radiologist performance was an aggregate of the performance results of all those who read images.

The standard practice of radiologists was to withhold assignment of a final BI-RADS assessment category for all recalled patients until further diagnostic mammography had been performed. The category 3, 4, and 5 assessments were not rendered until after recall for diagnostic imaging. This was the practice for both routine clinical and study cases. Recall rate was defined as the number of category 0 (incomplete, requires additional imaging) screening mammograms divided by the total number of screening mammograms. Institutional recall rate (excluding study cases) was calculated in the same manner.

Follow-up information regarding all patients in whom images were assigned BI-RADS assessment categories of 0, 3, 4, and 5 was reviewed. Results of diagnostic recall evaluations and biopsies were recorded. In addition, data in all patients with images assigned classifications of category 3, probably benign, were documented at the recommended short-term follow-up examination of 4 –6 months. These data included the mammographic, physical examination, and pathologic analysis results for those patients who subsequently received a diagnosis of cancer. Positive predictive values for biopsy were calculated for the radiologist alone and for the radiologist and CAD combined. The positive predictive value for the radiologist alone was defined as the number of cancers found divided by the number of biopsies recommended if CAD had not been used. The positive predictive value for the radiologist and CAD combined was defined as the number of cancers found divided by the number of biopsies recommended after the radiologist had reviewed CAD data. Results were obtained from institutional pathologists who rendered the pathologic diagnosis.

Detailed follow-up data were available for 1,077 (65%) subjects at 1 year at institution A. Such results were not available at institution B. Results of mammography and biopsy were recorded. For those patients who had breast cancer at 1-year follow-up, a retrospective review of images and findings was undertaken by one author (M.A.H.) to determine if the area (within 2 cm) where cancer was found had been marked by the CAD program on the study mammogram from 1 year earlier. Next, a subjective decision was rendered by the same author in regard to whether the CAD mark on the initial mammogram should have been considered actionable a year earlier (a prospective opinion was rendered for all such CAD marks). These marks were deemed actionable if recall should have occurred or not actionable if the mark should have been ignored.

Statistical Data Analysis

A CAD mark was considered true-positive when the CAD system marked an area on either the mediolateral oblique or the craniocaudal mammogram and subsequent biopsy findings demonstrated invasive carcinoma or ductal carcinoma in situ. Lobular carcinoma in situ was considered a benign histologic entity for the purposes of this study. A false-positive mark was defined as an area that was marked on the initial mammogram that was eventually deemed negative for cancer, benign, or probably benign by the radiologist. True-negative marks were defined as areas where no mark occurred and that were deemed negative for cancer, benign, or probably benign by the radiologist. A false-negative mark was defined as an area on the mammogram that was not marked on either the craniocaudal or mediolateral oblique view, with cancer found by the radiologist or palpable cancer detected within 1 year. The CIs were determined with software (Excel; Microsoft, Bothell, Wash) by the normal approximation to the binomial distribution at the 95% level.

RESULTS

Identification of Cancers

Eleven (0.46%) of 2,389 patients had mammographically detected nonpalpable breast cancers. Ten (91%) of 11 cancers were correctly identified with CAD. Sensitivity of the CAD system was 91% (95% CI: 74%, 100%) (Table 1). The cancer that was not detected with CAD was a 7-mm mass (invasive ductal carcinoma) that projected over the pectoral muscle and was imaged only on the mediolateral oblique projection. The radiologist's overall sensitivity without CAD was 91% (10 of 11; 95% CI: 74%, 100%). One cancer that was not observed by the radiologist was a 10 × 5-mm cluster of microcalcifications that was detected only after CAD analysis. At pathologic examination, this proved to be ductal carcinoma in situ (Table 1).

Recommendations

Mammographic appearance of cancers and their CAD detection are presented in Table 2. Overall, biopsy was recommended in 40 (1.67%) of 2,389 patients. The positive predictive value for biopsy for the radiologist and CAD combined was 28% (11 of 40). The theoretic positive predictive value of the radiologist alone (without CAD) would have been 10 (27%) of 37. Of the 40 patients in whom biopsy was recommended, 36 (90%) were identified with the CAD system. In addition, fine-needle aspiration biopsy was recommended in 14 patients, and in all of them results were benign. Twelve (86%) of 14 patients in whom fine-needle aspiration biopsy was recommended were identified with the CAD system (Table 3). Short-term follow-up at 6 months for probably benign cases was recommended in 83 patients. CAD was used to identify 68 (82%) of these patients (Table 3).

Cancer Characteristics

Of 11 cancers discovered, five (45%) were invasive ductal, one (9%) was not otherwise specified, and five (45%) were ductal carcinoma in situ. The size of the invasive cancers ranged from 7 to 30 mm (median, 16 mm). The size range of the ductal carcinoma in situ tumors was 10–17 mm (median, 14 mm). The age range of patients with cancer was 45–80 years (mean, 60 years).

With the CAD system, one false-negative result (one [9%] of 11) occurred. Results in three (75%) of four biopsies recommended by the radiologist but deemed negative for cancer with the CAD system proved to be benign. In both cases in which fine-needle aspiration biopsy was recommended by the radiologist but were not recommended on the basis of CAD results, findings proved to be benign. Results of one (25%) of four biopsies recommended by the radiologist but were not recommended on the basis of CAD results proved to be malignant.

Recall Rate

Combined recall rate of radiologists who interpreted study mammograms prior to reviewing CAD information at both institutions was 14.4%. This recall rate increased to 15.8% after the radiologist reviewed the CAD results. This is an increase of 1.4% in recall rates, but proportionately, it reflected an increase of 9.7% (1.4% of 14.4%). At institution A, where detailed records of recall rates were obtained of nonstudy and study subjects during this period, the baseline (prior to the radiologist's review of CAD results) recall rate of 13.5% in study subjects was higher than that of 10.3% in nonstudy patients. This was a 31% (3.2% of 10.3%) increase.

In 2,389 patients, 34 (1.42%) were recalled exclusively because of CAD results (Table 4). Five recalled patients underwent invasive procedures, and 29 were followed up. Three incremental biopsies were performed on the basis of CAD results; these data represented an 8% (three of 40) increase in the overall biopsy rate. In one (33%) of three cases, findings were malignant. Findings in two additional fine-needle aspiration biopsies performed on the basis of CAD results proved to be benign. Hence, five (0.21%) of 2,389 additional invasive studies were prompted on the basis of CAD results, and findings in one (20%) of five proved to be malignant. Of 29 additional patients who were called back because of CAD results and who were followed up, five were examined at short-term follow-up of 6 months, and the balance of the patients returned for follow-up at 1 year. The conditions of all patients who were examined at short-term follow-up were stable, and none required biopsy.

Follow-up

One-year follow-up data were available concerning 1,077 (65%) patients from institution A. Five (0.46%) of 1,077 subsequently developed cancer. Four (80%) of five had nonpalpable cancers found at yearly follow-up screening mammography. One (20%) of five developed a palpable cancer during the interval. At presentation, this cancer was mammographically occult on craniocaudal and mediolateral oblique views because of its geographic location. There were two cases of invasive ductal cancer, one of invasive lobular cancer, one of cancer with mixed lobular and ductal features, and one of ductal carcinoma in situ. Size range of invasive cancers was 10–22 mm (median, 14 mm). Retrospective review of the five cancers showed that three (60%) were not identified at the preceding year's mammographic examination with CAD. In two (40%) of five cancers, images had a CAD mark within 2 cm of the site of eventual cancer development, but in both cases, a mark was noted on only the craniocaudal view. All five cancers were deemed nonactionable findings by a study radiologist (M.A.H.) in retrospective review. In all five cases of cancer, results were considered negative at prospective review by study radiologists.

DISCUSSION

Our CAD program achieved very high sensitivity of 91% during this prospective clinical trial at two academic institutions. The single cancer that was not detected was obscured by overlying pectoral muscle and, because of its geographic location, was visualized only on the mediolateral oblique view. The sensitivity of our program exceeds that of 82% in previous clinical trials (7), although the CIs include this value.

Radiologist sensitivity for breast cancer was very high at 91%. This value also exceeds that pf 84% in a previous CAD clinical trial (7). One of the two institutions in our study had 100% radiologist sensitivity and no incremental cancer detection with CAD. The case of cancer that was not detected by a radiologist but was detected by CAD was an area of microcalcifications identified as ductal carcinoma in situ in a postmenopausal woman.

Conceptually, CAD systems can be programmed for high sensitivity or high specificity. For example, the performance of a CAD program can be limited to detection of obvious cancers with a moderate sensitivity and a relatively good specificity. Conversely, CAD programs can be oriented toward high sensitivity without regard for decreasing specificity. It is likely that each physician would hold a different opinion about how a CAD program should be structured within these extremes. Some physicians may choose a system that identifies only potential problems of obvious oversight, such as a large spiculated mass or a large group of calcifications. These physicians may dislike a plethora of false-positive marks, the majority of which will prove to be of no consequence, on a mammogram. Other physicians may choose a system that identifies any potential abnormality, because they know that the majority of marks on a mammogram will be benign. These physicians may not be bothered by having to characterize CAD marks as of no consequence. As a general rule, the more marks that are allowed with a CAD program, the higher the sensitivity yet the lower the specificity (16). Because of the subtle nature of early cancer, it is unlikely that higher sensitivity can be achieved without decrements in specificity. Experienced breast radiologists are well aware of this dilemma, and it is known that achievement of very high positive predictive values for biopsy is simply not possible without loss of sensitivity.

In our CAD program, we deliberately set a relatively low threshold for abnormality detection and allowed the CAD program to make up to four marks per image. This threshold is higher than that for some commercial applications but is in keeping with a philosophy that emphasizes sensitivity over specificity (7). Our high 91% sensitivity was achieved with some negative consequences. These consequences included a higher recall rate, which was increased by 1.4%, and a higher biopsy rate, which was increased by 8%. These negative risks were associated with a 9% improvement in cancer detection.

The increase in the biopsy rate was associated with an increased detection rate so that the positive predictive value was similar with (27.5%) and without (27.0%) CAD. Similar results were found by Freer and Ulissey (7), who noted a stable positive predictive value. This finding suggests that the primary use of CAD by the radiologist is for detection and not for characterization in clinical practice. Marked oversights in detection by the radiologists are recognized as such. Once a lesion is noted with CAD, radiologists will assess it as true or not on the basis of existing radiologic knowledge and experience. This makes sense when one realizes the poor specificity of all current CAD systems. If a CAD system allows one mark per mammographic view (eg, four marks for a typical four-view bilateral mammogram), the radiologist is forced to deal with false-positive marks for nearly every case. In fact, at one mark per view, up to 99.9% of all marks will be false-positive. He or she must rely on existing characterization knowledge, since most CAD systems do not provide an assessment of probability of malignancy for detected lesions. This activity results in similar clinical characterizations, with the potential for improvement in detection sensitivity. However, a substantial burden of characterization is placed on the radiologist because of CAD results.

The clinical CAD study of Freer and Ulissey (7) showed that most incremental cancer detection involved calcifications, and many of the calcifications represented ductal carcinoma in situ. The 19.5% incremental cancer detection rate found by Freer and Ulissey (7) was explained by seven (87%) of eight calcification cases. Our only case of incremental detection involved calcifications. This suggests that, prospectively, radiologists overlooked calcifications more frequently than they overlooked masses and/or that detection of masses is more problematic for CAD. Most CAD reports note better sensitivity for detection of calcifications than that of masses. Bird-well et al (17) reported 86% sensitivity for calcifications and 73% for masses in a retrospective review of missed cancers. Likewise, Warren Burhenne et al (15) noted 99% sensitivity for calcifications and 75% for masses in a retrospective review. Markey et al (11)

found that a radiologist description–based CAD program performed better for detection of masses than for that of calcifications.

Also, just as Freer and Ulissey (7) observed, ductal carcinoma in situ was the most frequent incremental cancer detected in our study. The biologic and survival importance in the detection of ductal carcinoma in situ at a screening interval prior to future mammographic detection is uncertain. Some of this incremental detection of ductal carcinoma in situ no doubt is biologically unimportant and would not be associated with changes in mortality. Improved sensitivity for invasive cancer detection may be a more important criterion with which to judge a CAD system. With this criterion, we had no incremental gain, and the incremental gain of Freer and Ulissey (7) would be markedly reduced from 19.5% to 4.9%.

Recall rates were increased by an absolute value of 1.4%, which was similar to the 1.2% increase in the study of Freer and Ulissey (7). These higher recall rates were caused by the low specificity of CAD and the burden placed on the radiologist to characterize many marked areas. These findings strongly suggest that radiologists are not comfortable with discounting some marked lesions without obtaining additional diagnostic views.

The academic breast radiologists in our study were subspecialty trained and had subspecialty interests in breast imaging. They were frequent readers, and they generally had image reading volumes that were several times those in the minimal recommended standards of the Food and Drug Administration. Nonetheless, their behavior was influenced by the clinical experiment. For example, at institution A, where detailed information was available, in the study the recall rate for mammographic interpretation prior to the radiologist's review of the CAD results was different from the recall rate for mammographic interpretation of screening mammograms read on the same alternator at the same time for non-study patients. A recall rate of 13.5% was initially noted for the subjects in whom CAD was not used, and this rate was substantially higher than the 10.3% rate for the nonstudy population. This was a 31% increase in recall rate. This increase reflects the phenomenon that behavior will change when that behavior is being evaluated. Radiologists were aware that in some sense they were competing against the CAD study and erred on the side of recalling more patients for any questionable abnormality. Yankaskas et al (22) demonstrated that sensitivity increases with increasing recall rates to certain limits (50% sensitivity at a recall rate of 2.1% and 80% sensitivity at a recall rate of 8.9%–13.4%), and it is likely that the physician's psychologic behavior changed to maximize sensitivity so that no cancer would be missed. This occurred even though we made every attempt to make this experiment a true comparison between routine reading and reading plus CAD. Hence, increased recall rates reported with CAD may actually lead to underestimation of the change caused by CAD results. With these criteria, institution A had a 49% increase in recalled patients (10.3% vs 15.3%). The incremental benefit of CAD may also be underestimated because of the elevated sensitivity of the radiologists alone for the study cases.

Even with extremely high CAD sensitivity, the incremental detection rate of a CAD program at two academic centers was less than that in other clinical trials (9% in our study compared with 19.5% in the study of Freer and Ulissey [7]). Although CIs are large because of the small number of patients in the study, the result is not unexpected. Subspecialty expertise and volume have been positively associated with physician performance (23,24). The value of a second reader in mammography has been shown to vary in benefit, depending on the ability of the pair of readers (19). It is likely that CAD may benefit a certain reader more than it may benefit another reader. Further, it is possible that this benefit may fade, as a reader is essentially trained by a CAD program, over time, to detect breast cancer. Follow-up results both in patients with cancers detected by the radiologist with use of CAD over time and in those with cancers missed by the CAD program in clinical trials are necessary for long-term validation of the effects of these programs.

Overreliance on CAD could paradoxically result in diminished cancer detection if the radiologist's attention is focused on only marked areas. Not all cancer locations are marked by any CAD system.

In addition, findings in retrospective studies indicate that some of the places where cancer subsequently developed had been marked by the CAD system 1 year earlier. In our study, 40% of future cancer locations were marked by CAD in the preceding year, but the findings were deemed not actionable by the study radiologist. The ability to detect and characterize cancer differs between prospective and retrospective review and is, in part, caused by the nonspecific appearance of early cancer (17,25–27). Birdwell et al (17) noted a cancer detection rate of 73% for masses and 86% for calcification clusters with prior readings with CAD in a selected group of cases. Review of our cases showed normal appearing areas that were marked by the CAD system but were not suspected of being malignant by the radiologists. Until CAD offers better specificity, these problems of interpretation may be expected to continue, since more than 99.9% of marks made by CAD are false-positive.

There were several limitations of the current study. As a pilot trial, the number of cases was relatively small, and the small number results in rather large CIs. The selection of breast specialists at academic medical centers was not representative of the U.S. radiologist population in general. Sensitivity measures in our study are relative because cancer detection was based on clinical assessment and not on whole-breast histologic review. The lack of 1-year follow-up findings in all patients limits sensitivity measures. In addition, the patients who underwent screening at these sites and those who volunteered for the study may or may not be representative of the screening population in the United States. Because the radiologists were aware that they were participating in the study, their behavior was potentially affected.

In summary, we evaluated a noncommercial CAD program with a very high sensitivity of 91% for cancer detection in a pilot clinical trial. An incremental yield of 9% was observed. These effects were balanced by a higher recall rate and a higher subsequent biopsy rate.

Acknowledgments

The authors are grateful to the mammographic technologists who assisted in patient recruitment.

Supported by U.S. Army grant DAMD 17-96-1-6254 and U.S. Public Health Service grant CA48129.

Abbreviations

BI-RADS
: Breast Imaging Reporting and Data System

CAD
: computer-aided detection

References

1. American Cancer Society. Statistics for 2003. [Accessed August 7, 2003]. Available at: www.cancer.org
2. Feig S, D'Orsi C, Hendrick R, et al. American College of Radiology guidelines for breast cancer screening. AJR Am J Roentgenol 1998;171:29–33. [PubMed: 9648758]
3. Cady B, Michaelson JS. The life-sparing potential of mammographic screening. Cancer 2001;91:1699–1703. [PubMed: 11335893]
4. Tabar L, Vitak B, Chen HC, Yen MF, Duffy SW, Smith RA. Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. Cancer 2001;91:1724–1731. [PubMed: 11335897]

5. Duffy SW, Tabar L, Chen HH, et al. The impact of organized mammography service screening on breast carcinoma mortality in seven Swedish counties. Cancer 2002;95:458–469. [PubMed: 12209737]
6. Vyborny CJ. Can computers help radiologists read mammograms? Radiology 1994;191:315–317. [PubMed: 8153298]
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. [PubMed: 11526282]
8. 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;220:787–794. [PubMed: 11526283]
9. Chan HP, Sahiner B, Lam KL, et al. Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces. Med Phys 1998;25:2007–2019. [PubMed: 9800710]
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. [PubMed: 9891149]
11. Markey MK, Lo JY, Floyd CE. Differences between computer-aided diagnosis of breast masses and that of calcifications. Radiology 2002;223:489–493. [PubMed: 11997558]
12. Petrick N, Chan HP, Sahiner B, Helvie MA. Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms. Med Phys 1999;26:1642–1654. [PubMed: 10501064]
13. Chan HP, Sahiner B, Helvie MA, et al. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. Radiology 1999;212:817–827. [PubMed: 10478252]
14. Nishikawa RM, Doi K, Giger ML, et al. Computerized detection of clustered microcalcifications: evaluation of performance on mammograms from multiple centers. RadioGraphics 1995;15:443–452. [PubMed: 7761647]
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. [PubMed: 10796939]
16. Te Brake GM, Karssemeijer N, Hendriks JH. Automated detection of breast carcinomas not detected in a screening program. Radiology 1998;207:465–471. [PubMed: 9577496]
17. 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. [PubMed: 11274556]
18. Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994;191:241–244. [PubMed: 8134580]
19. Beam CA, Sullivan DC, Layde PM. Effect of human variability on independent double reading in screening mammography. Acad Radiol 1996;3:891–897. [PubMed: 8959178]
20. Petrick N, Sahiner B, Chan HP, Helvie MA, Paquerault S, Hadjiiski LM. Breast cancer detection: evaluation of a mass detection algorithm for computer-aided diagnosis—experience in 263 patients. Radiology 2002;224:217–224. [PubMed: 12091686]
21. Chan HP, Lo SC, Sahiner B, Lam KL, Helvie MA. Computer-aided detection of mammographic microcalcifications: pattern recognition with an artificial neural network. Med Phys 1995;22:1555–1567. [PubMed: 8551980]
22. 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. [PubMed: 11517044]
23. Esserman L, Cowley H, Eberle C, et al. Improving the accuracy of mammography: volume and outcome relationships. J Natl Cancer Inst 2002;94:369–376. [PubMed: 11880475]
24. Sickles EA, Wolverton PE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology 2002;224:861–869. [PubMed: 12202726]
25. Goergen SK, Evans J, Cohen GP, MacMillan JH. Characteristics of breast carcinomas missed by screening radiologists. Radiology 1997;204:131–135. [PubMed: 9205234]
26. Harvey JA, Fajardo LL, Innis CA. Previous mammograms in patients with impalpable breast carcinoma: retrospective vs blinded interpretation. AJR Am J Roentgenol 1993;161:1167–1172. [PubMed: 8249720]

27. Ikeda DM, Birdwell RL, O'Shaughnessy KF, Sickles EA. Analysis of 127 subtle, undetected or non-recalled findings on prior "negative" mammograms in women with screening-detected breast cancers (abstr). Radiology 1999;213(P):240.

TABLE 1

Cancer Type versus Detection Method

Cancer	Radiologist	CAD	Total
Invasive ductal	5 (100)	4 (80)	5
Invasive (not otherwise specified)	1 (100)	1 (100)	1
Ductal carcinoma in situ	4 (80)	5 (100)	5
Total	10 (91)	10 (91)	11

Note.—Data are numbers of cancers. Numbers in parentheses are percentages.

TABLE 2

Mammographic Appearance for Detected Cancers versus Detection Method

Appearance	Radiologist	CAD	Total
Mass	6 (100)	5 (83)	6
Calcification	3 (75)	4 (100)	4
Architectural distortion	1 (100)	1 (100)	1
Total	10 (91)*	10 (91)*	11

Note.—Data are numbers of cancers. Numbers in parentheses are percentages. The 95% CI was 74%, 100%.

*Both radiologist and CAD missed one.

TABLE 3

Performance of Radiologist, Radiologist with CAD, and CAD in 378 Recalled Patients

	Recommended Action					
Detection and Statistic	**Biopsy**	**FNAB***	**6-month Follow-up**	**12-month Follow-up**	**Overall**	**Cancer**
Radiologist alone	37	12	78	217	344	10†
Radiologist with CAD	40	14	83	241	378	11
CAD alone	36	12	68	175	291	10†
Sensitivity of CAD (%)‡	90	86	82	73	77	91
95% CIs (%)	81, 99	67, 100	73, 90	67, 78	73, 81	74, 100

*FNAB = fine-needle aspiration biopsy.

†The total number of detected cancers was 11, but both CAD and radiologist missed one.

‡Sensitivity was calculated by dividing numbers for CAD alone by those for radiologist with CAD.

TABLE 4

Incremental Recall of Patients Caused by CAD Results versus Outcome

	Recommended Action					
Appearance and Statistic	**Biopsy**	**FNAB***	**6-month Follow-up**	**12-month Follow-up**	**Overall**	**Cancer**
Mass	1	1	3	22	27	0
Calcification	2	1	2	2	7	1
Overall†	3/40 (7.5)	2/14 (14.3)	5/83 (6.0)	24/241 (10.0)	34/378 (9.0)	1/11 (9.1)
95% CIs (%)	0, 15.8	0, 33.0	0.8, 11.3	6.1, 13.8	6.1, 11.9	0, 26.4

*FNAB = fine-needle aspiration biopsy.

†Data are numbers of patients recalled. Numbers in parentheses are percentages. Denominators are from “Radiologist with CAD” category in Table 3.