Published in final edited form as:

Acad Radiol. 2009 July ; 16(7): 810–818. doi:10.1016/j.acra.2009.01.011.

Multi-modality CADx: ROC study of the effect on radiologists' accuracy in characterizing breast masses on mammograms and 3D ultrasound images

Berkman Sahiner, PhD, Heang-Ping Chan, PhD, Lubomir M. Hadjiiski, PhD, Marilyn A. Roubidoux, MD, Chintana Paramagul, MD, Janet E. Bailey, MD, Alexis V. Nees, MD, Caroline E. Blane, MD, Dorit D. Adler, MD, Stephanie K. Patterson, MD, Katherine A. Klein, MD, Renee W. Pinsky, MD, and Mark A. Helvie, MD

Department of Radiology, The University of Michigan

Abstract

Rationale and Objectives—To investigate the effect of a computer-aided diagnosis (CADx) system on radiologists' performance in discriminating malignant and benign masses on mammograms and 3D ultrasound (US) images.

Materials and Methods—Our data set contained mammograms and 3D US volumes from 67 women (median age: 51, range: 27–86) with 67 biopsy-proven breast masses (32 benign and 35 malignant). A CADx system was designed to automatically delineate the mass boundaries on mammograms and the US volumes, extract features, and merge the extracted features into a multi-modality malignancy score. Ten experienced readers (subspecialty academic breast imaging radiologists) first viewed the mammograms alone, and provided likelihood of malignancy (LM) ratings and Breast Imaging and Reporting System (BI-RADS) assessments. Subsequently, the reader viewed the US images with the mammograms, and provided LM and action category ratings. Finally, the CADx score was shown and the reader had the opportunity to revise the ratings. The LM ratings were analyzed using receiver operating characteristic (ROC) methodology, and the action category ratings were used to determine the sensitivity and specificity of cancer diagnosis.

Results—Without CADx, readers' average area under the ROC curve, A_z , was 0.93 (range: 0.86 to 0.96) for combined assessment of the mass on both the US volume and mammograms. With CADx, their average A_z increased to 0.95 (range: 0.91 to 0.98), which was borderline significant ($p=0.05$). The average sensitivity of the readers increased from 98% to 99% with CADx, while the average specificity increased from 27% to 29%. The change in sensitivity with CADx did not achieve statistical significance for the individual radiologists, and the change in specificity was statistically significant for one of the radiologists.

Conclusion—A well-trained CADx system that combines features extracted from mammograms and US images may have the potential to improve radiologists' performance in distinguishing malignant from benign breast masses and making decisions about biopsies.

© 2009 The Association of University Radiologists. Published by Elsevier Inc. All rights reserved.

Corresponding Author: Berkman Sahiner, Ph.D., MIB C480A, 1500 East Medical Center Drive, Ann Arbor, MI 48109-5842, Phone: 734-647-7429, Fax: 734-615-5513, Email: E-mail: berki@umich.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

Computer-Aided Diagnosis; ROC Observer Study; Mammography; Ultrasound; Malignancy

INTRODUCTION

Breast cancer is the second leading cause of cancer death and the most prevalent non-cutaneous cancer among American women (1). Because early detection of breast cancer may improve the chance of survival, Breast Imaging Reporting and Data System (BI-RADS) categories 4 and 5 findings are typically referred for biopsy. Although the optimal positive biopsy rate for abnormalities detected by mammography is still being debated, positive breast biopsy rates of 25–40 percent have been recommended as appropriate (2). Studies indicate that academic and community practices perform close to the lower end of this recommendation (3,4).

Increasing the positive predictive value of biopsy would reduce the cost of health care, spare patients the anxiety and discomfort of biopsy, and avoid the possible scarring of the breast tissue, which might complicate future exams. However, this increased positive predictive value should not come at the cost of missed cancers, but rather as a result of an overall improvement in the accuracy of breast cancer detection and characterization. Computer-aided diagnosis (CADx) is one of the techniques that strive to improve radiologists' differentiation between malignant and benign lesions, thus improving the positive biopsy rate without missing malignancies.

Breast masses are often evaluated using both mammography and ultrasound (US). CADx methods have been developed for both mammographic (5–12) and sonographic (13–20) masses. Several research groups have investigated how the mass characteristics in sonographic and mammographic images can be combined to improve the performance of a CADx system (21–23). The effects of CADx systems on radiologists' performance in differentiating between malignant and benign breast masses on mammograms (7,24–26) and on US images (16,27) have also been reported. However, to our knowledge, very few studies (28) have investigated the effect of CADx on radiologist's evaluation of breast masses when both modalities are simultaneously available, similar to routine clinical practice.

The purpose of our study was to investigate the effect of a classifier that uses computer-extracted features from both mammograms and US images on radiologists' performance in characterizing malignant and benign masses using both modalities.

MATERIALS AND METHODS

Data Set

A set of 130 consecutive subjects who participated in a 3D breast US study between 1998 and 2002 was included in this study. The patients were recruited with Institutional Review Board (IRB) approved protocol and written informed consent. We received IRB approval for retrospective use of the data set prior to the commencement of our study. The study protocol was Health Insurance Portability and Accountability Act compliant. Following mammographic evaluation, all patients had a sonographic mass assessed as suspicious or highly suggestive of malignancy and were scheduled for biopsy or fine needle aspiration. Twenty-nine patients were excluded using first-stage exclusion criteria. These were patients who had prior biopsy in the same region of the breast, scans which were deemed technically unsuccessful because of motion or other artifacts, masses that did not undergo biopsy, and women in whom the mass was incompletely imaged in any dimension because of large size or eccentric position in the

scan. An additional set of 34 patients was excluded because their mammograms at the time of the 3D US examination were not available, or the mass was not mammographically visible.

Our final data set thus consisted of 3D US and mammogram images from 67 patients (average age: 51 years, range: 27–86), for whom the mass was visible in both modalities. Based on excisional or core biopsy, or fine needle aspiration results, 35 masses were malignant and 32 were benign. Thirty of the malignancies were invasive ductal carcinoma, 2 were ductal carcinoma in-situ, 1 was invasive lobular carcinoma, and 2 were other invasive carcinoma. Of the benign masses, 15 were fibroadenoma, 7 were fibrocystic disease, 5 were cysts, 2 were fat necrosis, and three were other benign breast diagnoses.

Mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of $50 \mu\text{m} \times 50 \mu\text{m}$ and 4096 gray levels. All mammograms were acquired with dedicated mammographic systems. The digitizer was calibrated so that gray level values were linearly proportional to the optical density (OD) within the range of 0.1 to greater than 3.0 OD units, with a slope of 0.001 OD/pixel value. The nominal O.D. range of the scanner is 0–4. The image matrix size was reduced by averaging every 2×2 adjacent pixels and down-sampled by a factor of 2, resulting in images with a pixel size of $100 \mu\text{m} \times 100 \mu\text{m}$ for further analysis. The total number of mammographic views was 163, with each case containing between one and three views (craniocaudal, mediolateral-oblique, and occasionally a lateral view).

The system for the acquisition of the 3D US data consisted of a commercially available GE Logiq 700 (Milwaukee, WI) US scanner with an M12 linear array transducer, a mechanical transducer guiding system, and a computer workstation. The transducer was manually linearly translated in the cross-plane direction, while 2D B-mode images were automatically recorded in the image scan plane at approximately 0.5 mm incremental translations. The 2D images were sent to a buffer in the US scanner and then transferred to a workstation where they were stacked to form a 3D volume. Additional details of US image acquisition can be found in the literature (27,29).

The biopsied mass on the mammograms and the US volumes was identified by an MQSA (Mammography Quality Standards Act) qualified radiologist using clinical images and accompanying patient medical records to confirm that the identified region contained the biopsied mass. The radiologist identified the region of interest (ROI) containing the mass on each mammogram, and the corresponding mass and the image slices containing the mass on the US volume. The radiologist also measured the mass size, defined as the largest dimension of the mass on the mammogram. The average mass size in the data set was estimated to be 1.58 ± 0.80 cm (range: 0.5–4.0 cm).

CADx System

Our CADx system included mass segmentation and extraction of relevant features from the mass and its margins on both modalities, and a multi-modality classifier that fused the information.

The mammographic segmentation method was based on an active contour (AC) model that followed an initial segmentation using K-means clustering (30). Morphological features, including a Fourier descriptor, convexity, rectangularity, perimeter, contrast, circularity, perimeter-to-area ratio, area, and normalized radial length features were extracted from the segmented mass shapes. Three spiculation features were extracted from a spiculation measure defined for the pixels along the boundary of the mass (31,32). Run-length statistics (RLS) texture features were extracted from the band of pixels surrounding the mass after the band was transformed into Cartesian coordinates using the rubber-band straightening transform (33).

Our 3D US segmentation method was based on a 3D AC model initialized with a radiologist-defined 3D ellipsoid (34). After mass segmentation, morphological and texture features were extracted from each slice containing the mass. The morphological features included the width-to-height ratio and posterior shadowing features. The texture features were extracted from the spatial gray level dependence (SGLD) matrices derived from the 2D slices of the 3D data set. Since the margins of the mass contain the richest information for characterization, these features were extracted from two disk-shaped regions containing the mass boundary on the upper and lower margins of the mass. Six texture feature measures that are invariant under linear, invertible gray scale transformations were extracted. We have previously described the extraction of these features in detail (34).

The extracted features were merged into a single malignancy score using a multi-modality classifier. Details of the classification method can be found in the literature (21). Briefly, corresponding feature vectors from each mammographic view were averaged to obtain a case-based mammographic feature vector, and corresponding feature vectors from each US slice were averaged to obtain a case-based US feature vector. A case-based combined feature space was defined by merging the case-based mammographic and US feature spaces. A leave-one-case-out method was used to train and test the classifier in the case-based combined feature space. The training included feature selection and the computation of linear discriminant analysis (LDA) coefficients for using all cases but one, which was used as the test case. The test case was changed in round-robin order so that a test score was obtained for each case. For the observer study, the computer malignancy score was linearly mapped and rounded to an integer between 1 and 10. In order to provide a reference of the computer rating scale to the radiologists, two normalized Gaussian distributions were fitted to the computer scores for the malignant and benign classes and displayed along with the score.

Observer Performance Study

Ten academic breast radiologists, referred to as RAD1-RAD10, who had 3–26 years (median=11) of experience in mammographic and breast US interpretation, participated as observers. They were all MQSA certified and nine were fellowship-trained in breast imaging. The observers used a graphical user interface (GUI) to view the US images, mammographic ROIs, and to provide ratings and assessments for the masses. The GUI allowed the radiologists to adjust the window and level of the displayed images, and to navigate through the 3D US volumes.

A three-step sequential reading study design was used. In the first step, (MAM mode) the radiologist examined the ROIs extracted from all the mammographic views containing the mass, and provided a mammographic BI-RADS assessment (categories 1 to 5, 0 was not allowed). The radiologist also provided an estimate of the likelihood of malignancy (LM) rating on a scale of 0–100, corresponding to the BI-RADS assessment. The observers were reminded at the beginning of the study that if they rated a mass as having a greater than 2% LM, i.e., BI-RADS category 4 or 5, a recommendation for biopsy of the mass would be implied (35,36). Masses considered to be “benign findings” and those with less than 2% LM are important sub-categories (BI-RADS categories 1–3). The GUI module for providing the LM therefore contained selection buttons to indicate a 0% LM for a benign finding and a <2% LM for a probably benign finding. For masses in BI-RADS 4 and 5 categories, the radiologists used a slide-bar to indicate an LM between 2% and 100%.

Immediately following the first step, the radiologists viewed the 3D US volume containing the mass in addition to the mammographic ROIs. This mode is referred to as the USM mode below, indicating that the radiologists used both the US and mammographic images to evaluate the mass. By default, at the beginning of this step, the first US slice containing the mass was shown, and the location of the mass center was marked. The radiologists used the GUI to navigate

through the 3D volume. At the end of the second step, the radiologist provided a second LM rating based on his/her combined assessment of the mass on both the US volume and mammograms, using the same GUI module as in the MAM mode. The radiologist also recommended one of three actions: 1-year follow-up, 6-months follow-up, or biopsy.

Immediately after reading in the USM mode, the CADx score for the case, and the reference class distributions for CADx scores of benign and malignant masses were displayed on the screen (CADx mode). The radiologists had the option to keep their original malignancy rating, or change it using the slide bar after considering the CADx score.

There was no time limit for the radiologists. The radiologists were not informed of the proportion of malignant masses, or whether a mass was malignant or benign after they finished rating of the mass. The reading order was randomized for each radiologist. In order to reduce fatigue, the data set was read in two separate sessions by each radiologist. The radiologists were familiarized with the study design, the functions on the GUI, and the computer's relative malignancy rating scale in a training session before the study. A data set of three patient masses which satisfied first-stage exclusion criteria (simple cysts and mass with prior biopsy) were used in the training session.

Data and Statistical Analysis

The LM ratings of the radiologists in MAM, USM, and CADx modes were analyzed using ROC methodology (37,38). The significance of the difference in the area A_z under the ROC curve between two conditions was tested using the Dorfman-Berbaum-Metz (DBM) multi-reader multi-case (MRMC) methodology (39). The partial area index above a sensitivity of 0.9, $A_z^{(0.9)}$ (40), was also computed for each radiologist. Student's two-tailed paired t-test was used to test the significance of the difference in $A_z^{(0.9)}$ between different modes. The DBM method accounts for both reader and case variances, while the t-test accounts only for reader variance in its significance estimation. In the Results Section, A_z and $A_z^{(0.9)}$ values are compared between two pairs of consecutive steps in the observer study, i.e., between MAM and USM modes, and between USM and CADx modes. Although a comparison between MAM and CADx modes was also performed, this comparison was not included below, because the study design did not allow the radiologists to move directly from the MAM mode to CADx mode.

The sensitivity and specificity of each radiologist were compared among the MAM, USM, and CADx modes. A two-tailed t-test was used to investigate the significance of the differences in overall sensitivities and specificities between different modes for the ten radiologists. McNemar's test (WinStat version 2005.1, Lehigh Valley, PA) was used to test the significance of the difference in biopsy recommendations between different modes for each radiologist.

In addition to an LM rating of 2%, we also tested a hypothetical biopsy threshold for LM with CADx. This hypothetical threshold was chosen to maintain the average sensitivity of the radiologists at the same level as that without CADx. We could then evaluate the change in specificity if the sensitivity was kept the same before and after use of CADx.

RESULTS

The A_z values of the radiologists in MAM, USM and CADx modes are listed in Table 1. The average A_z values, derived from the averages of the a and b parameters of the fitted ROC curves for the individual readers, were 0.87, 0.93, and 0.95 in the MAM, USM, and CADx modes, respectively. The A_z values of the individual reader ranged between 0.80 to 0.93 in MAM mode; 0.86 to 0.96 in USM mode; and 0.91 to 0.98 in CADx mode. The differences in the average A_z values between the MAM and USM, and between the USM and CADx modes were

statistically significant ($p=0.03$, and $p=0.05$, respectively) by MPMC ROC analysis. In step 2 (i.e., when advancing from MAM mode to USM mode) all radiologists showed an improvement in their A_z values. In step 3 (i.e., when advancing from USM mode to CADx mode), the A_z value of 8 radiologists improved, whereas that of 2 radiologists remained the same.

In comparison to the radiologists' average A_z value of 0.93 in USM mode, the test A_z value of the designed computer classifier was 0.91 ± 0.04 . The A_z value of the computer was not significantly different from that of any of the ten radiologists in the USM mode (p -value range: 0.24 to 0.90). The radiologists' average ROC curves (derived from the average a and b parameters) in the three modes, and that of the computer classifier, are shown in Fig. 1.

The partial area indices above a sensitivity of 0.9, $A_z^{(0.9)}$, are shown in Table 1. The average $A_z^{(0.9)}$ value was 0.27 (range: 0.02 to 0.54) in MAM mode; 0.52 (range: 0.27 to 0.77) in USM mode; and 0.65 (range: 0.52 to 0.83) in CADx mode. The differences in the average $A_z^{(0.9)}$ values between the MAM and USM, and USM and CADx modes were statistically significant ($p=0.001$, and $p=0.008$, respectively) by the two-tailed paired t -test.

The average LM ratings for benign masses over ten radiologists were 25.6, 25.2, and 24.5 in MAM, USM, and CADx modes, respectively, while those for malignant masses were 64.9, 75.0, and 78.9, respectively. In step 2, the radiologists changed their LM rating for 60% (210/350) of the malignant and 67% (214/320) of the benign masses. In step 3, the radiologists further changed their LM rating for 31% (108/350) of the malignant and 41% (130/320) of the benign masses. The numbers and percentages of changes in beneficial and detrimental directions for malignant and benign masses are shown in Table 2.

The average sensitivity and specificity of the radiologists in MAM mode were 0.94 and 0.33, respectively. In step 2, the average sensitivity increased to 0.98 while the average specificity decreased to 0.27. In step 3, both the sensitivity and specificity improved with respect to USM mode, to 0.99 and 0.29, respectively. The sensitivities and specificities of each radiologist in each mode are listed in Table 3. The changes in average sensitivities and specificities in steps 2 and 3 did not achieve statistical significance (t -test, Table 3).

Table 4 shows the number of beneficial and detrimental changes in biopsy decisions in steps 2 and 3 for each radiologist. A beneficial change for a malignant mass is a decision to recommend biopsy for the mass that was not recommended for biopsy in the previous step and vice versa for a detrimental change. A beneficial change for a benign mass is a decision not to recommend biopsy for the mass that was recommended for biopsy in the previous step and vice versa for a detrimental change. The total number of correct biopsy decisions for both malignant and benign masses increased when the radiologists read in the CADx mode, as indicated in the last row of Table 4. In step 2, the number of correct biopsy decisions for malignant masses increased significantly (McNemar's test) for one radiologist (R8). However, the number of incorrect biopsy decisions for benign masses increased significantly for the same radiologist. In step 3, the number of correct biopsy decisions for benign masses increased significantly for one radiologist (R7), while the change in biopsy decision for malignant masses did not reach statistical significance for any radiologist.

We also investigated how the biopsy decisions would be affected if the LM threshold for CADx mode were to be adjusted to 10%, for which the average sensitivity would remain at 98% (same as that in USM mode). Under this condition, i.e., if biopsy would be recommended only if the LM were 10% or greater, the average specificity would increase to 39%, and the improvement would be significant for three of the radiologists (Table 5).

DISCUSSION

Our results indicate that the designed CADx system has the potential to aid radiologists in characterization of breast masses on mammograms and 3D US volumes. With CADx, the improvements in both A_z and $A_z^{(0.9)}$ values of ten MQSA radiologists were significant. This improvement was achieved despite comparatively lower A_z value of the stand-alone CADx system (0.91 ± 0.03) compared to the average of the observers without CADx (average $A_z = 0.93$). After reading with CADx, A_z of eight radiologists was higher than the CADx system, whereas two radiologists had the same A_z as the CADx system. This indicates that a CADx system with a performance similar to, or even slightly lower than that of the radiologists may still be useful. For achieving this gain, the radiologists need to be selective in choosing cases for which they change their LM rating based on the computer score. Table 2 indicates that this may indeed be the case. When they advanced from USM mode to CADx mode, the radiologist changed their LM rating for an average of 36% of the cases. In comparison, the corresponding number of changes when they advanced from MAM mode to USM mode was 63%.

Table 1 indicates that the radiologists had a larger gain in step 2 of the observer study compared to step 3 as measured by ROC analysis. This is not surprising; the addition of US images for assessment of masses seen on mammograms has been shown to be useful for differentiation of malignant and benign lesions and can be expected to be more beneficial than the addition of only computer scores for the modalities that the radiologist has already examined. However, the improvement in both steps were statistically significant, with the p values smaller in step 2 compared to step 3.

When the changes in sensitivity and specificity were considered, the effect of step 2 was mixed, with a relatively larger average increase in sensitivity but an average decrease in specificity. In step 3, the radiologists had small average increases in both sensitivity and specificity (Table 3). The changes in overall sensitivities and specificities among the ten radiologists in steps 2 and 3 did not achieve statistical significance.

The average specificity in all three modes was low in our study. This is largely due to the fact that all lesions in our data set were clinically recommended for biopsy or fine needle aspiration after imaging. Therefore, the sensitivity and specificity in the clinical interpretation of these cases was 1 and 0, respectively. In clinical interpretation, the patient's previous exams, additional mammograms such as spot views, patient's age, and other information such as family and personal cancer history were available to radiologists, whereas they were not presented to the observers in the ROC study. In addition, a retrospective study cannot duplicate many factors that may affect radiologists' diagnostic decision in clinical practice. Due to these differences, the observers did not recommend all cases for biopsy in the ROC study. Nevertheless, the rate of biopsy recommendation remained quite high, which resulted in a low specificity although it was still substantially higher than 0.

For individual radiologists, the changes in biopsy decisions reached statistical significance for only one radiologists for both steps 2 and 3. This can be attributed to two reasons: our data set was small, and at each step, both the sensitivities and specificities changed, which makes it more difficult to detect a significant change in either of them. We investigated the effect of artificially changing the decision threshold for biopsy recommendation in CADx mode so that the sensitivity was kept the same for the CADx and USM modes. If the decision threshold were changed to 10% in CADx mode, the average sensitivity would match that of USM mode, and seven radiologists would obtain an improvement in specificity, of which three would be significant (Table 5). Similarly, if the decision threshold in USM mode were changed so that the average sensitivity matched that of MAM mode, eight radiologist would obtain an improvement in specificity, five of which would be significant (not shown in tables).

A problem common to many CAD studies is the lack of both a large training and a large independent test set. Our classifier was trained and tested using a leave-one-case-out resampling technique, because it was not possible to reserve an independent test set, due to the fact that the total number of available case samples was small. Although this technique is known as nearly unbiased (41), it is also known to be a method with a large standard deviation, especially for small data sets (42). A larger independent test data set is therefore necessary to assure the generalizability of the performance of our CADx system. Nevertheless, the observer ROC study design investigates whether a CAD system having a given performance level may improve the observers' relative performance in a certain task. Our observed improvement should not be affected by whether the malignancy ratings of the CADx system were independent test results.

A recent study by Horsch et al. (28) compared the performances of five breast radiologists and five breast imaging fellows with and without CADx in the task of distinguishing between malignant and benign breast masses on mammograms and US images. They found a statistically significant improvement in A_z , partial A_z , and sensitivity values of the radiologists with CADx. They designed their classifier and display such that the scores for mammography and US were determined and shown separately, on the same screen. An advantage of such a design is that radiologists may be able to weigh the computer scores from these two modalities differently. On the other hand, some radiologists may prefer the computer results to be summarized with a single score that has been jointly optimized for best computer performance on these two modalities. Additional studies are required to gain an understanding into the best methods for presenting the malignancy ratings by the computer classifier to radiologists.

Our data set had a higher prevalence of malignant masses (57%) than typical clinical care. The effect of prevalence in laboratory observer performance studies is an interesting topic that is not yet fully understood. In a previous study (43), Gur *et al.* found that, when the area under the ROC curve was used as the performance criterion, no significant effect could be observed when the prevalence was varied from 2 to 28%. However, when the same data was analyzed in terms of the difference in confidence ratings, a statistically significant trend was found (44). Interestingly, for both actually positive and actually negative cases, when the disease prevalence in the data set was *higher*, the confidence that the abnormality was present tended to be *smaller*. The opposite effect was previously observed in a smaller study, that found *larger* confidence ratings under the condition of *higher* prevalence (45). However, the prevalence rates in (45) were larger (20% and 60%) compared to those in (44). The main purpose of our study was to compare different reading conditions under the same prevalence. The effect of prevalence, if any, would be present under all three conditions. Consequently, the effect of the prevalence of malignant masses on the relative performances under the three reading conditions should be much weaker than that on the absolute performances. Nevertheless, a larger study that more closely emulates the reading conditions in the clinic would be needed to confirm this implication and to validate our results.

Our study had a number of limitations. Our data set was relatively small, which may explain why some of the changes failed to reach statistical significance. Our data set consisted only of masses that were recommended for tissue sampling, and it would be interesting to study the performance of our CADx system on masses that are not recommended for biopsy. All observers in our study were experienced breast radiologists from an academic center. Our results may therefore not generalize to radiologists with less experience or in different practice settings. At our institution, all clinical breast US examinations are performed by breast imaging radiologists, not sonographers, and therefore the readers in our ROC study are experienced in assessing whole volume images. However, the 3D US acquisition method in this study may not emulate hand-held imaging and real-time assessment by radiologists. Finally, retrospective ROC studies cannot duplicate many factors that may affect a diagnostic decision in clinical

practice. For example, in an ROC study, there is no concern for medicolegal liability of misdiagnosing a malignancy, while there may be some pressure for achieving the best performance among one's peers. As a result, radiologist decisions in an experimental situation will not completely reflect clinical reality. As discussed in a recent study, radiologists may perform differently in the clinic than in the laboratory when interpreting the same mammograms (46). Due to these limitations, our results cannot be directly transferred to daily practice. Nevertheless, laboratory ROC studies are commonly accepted as the best method to date to compare the relative performance of different imaging modalities before more costly realistic clinical trials are conducted. Our study demonstrates that CADx is promising for improving breast mass diagnosis even after the highly accurate combined mammography and ultrasound evaluation. Additional investigations that address the limitations above are warranted in future studies.

We conclude that a well-trained CADx system that combines features extracted from mammograms and US images has the potential to improve radiologists' performance in distinguishing malignant from benign breast masses.

Acknowledgments

Grant Supporting the Research:

This work was supported by USPHS grants CA118305, CA095153, and CA091713.

REFERENCES

1. Jemal A, Siegel J, Ward E, Murray T, Xu J, Thun MJ. Cancer Statistics, 2007. *CA - A Cancer Journal for Clinicians* 2007;57:43–66. [PubMed: 17237035]
2. Bassett, LW.; Hendrick, RE.; Bassford, TL., et al. Quality Determinants of Mammography. Clinical Practice Guideline No. 13. AHCPR Publication No. 95-0632. Rockville, MD: Agency for Health Care Policy and Research, Public Health Service, U.S. Department of Health and Human Services; 1994.
3. Brown ML, Houn F, Sickles EA, Kessler LG. Screening mammography in community practice: positive predictive value of abnormal findings and yield of follow-up diagnostic procedures. *American Journal of Roentgenology* 1995;165:1373–1377. [PubMed: 7484568]
4. Gur D, Wallace LP, Klym AH, et al. Trends in recall, biopsy, and positive biopsy rates for screening mammography in an academic practice. *Radiology* 2005;235:396–401. [PubMed: 15770039]
5. Huo ZM, Giger ML, Vyborny CJ, Wolverton DE, Metz CE. Computerized classification of benign and malignant masses on digitized mammograms: A study of robustness. *Academic Radiology* 2000;7:1077–1084. [PubMed: 11131052]
6. Mudigonda NR, Rangayyan RM, Desautels JEL. Gradient and texture analysis for the classification of mammographic masses. *IEEE Transactions on Medical Imaging* 2000;19:1032–1043. [PubMed: 11131493]
7. Leichter I, Fields S, Nirel R, et al. Improved mammographic interpretation of masses using computer-aided diagnosis. *European Radiology* 2000;10:377–383. [PubMed: 10663772]
8. Bilska-Wolak AO, Floyd CE, Nolte LW, Lo JY. Application of likelihood ratio to classification of mammographic masses. performance comparison to case-based reasoning *Medical Physics* 2003;30:949–958.
9. Varela C, Timp S, Karssemeijer N. Use of border information in the classification of mammographic masses *Physics in Medicine and Biology*. 2006;51:425–441.
10. Cheng HD, Shi XJ, Min R, Hu LM, Cai XP, Du HN. Approaches for automated detection and classification of masses in mammograms. *Pattern Recognition* 2006;39:646–668.
11. Delogu P, Fantacci ME, Kasae P, Retico A. Characterization of mammographic masses using a gradient-based segmentation algorithm and a neural classifier. *Computers in Biology and Medicine* 2007;37:1479–1491. [PubMed: 17383623]

12. Shi J, Sahiner B, Chan HP, et al. Characterization of mammographic masses based on level set segmentation with new image features and patient information. *Medical Physics* 2008;35:280–290. [PubMed: 18293583]
13. Chen DR, Chang RF, Huang YL. Computer-aided diagnosis applied to US of solid breast nodules by using neural networks. *Radiology* 1999;213:407–412. [PubMed: 10551220]
14. Dumane VA, Shankar PM, Piccoli CW, Reid JM, Forsberg F, Goldberg BB. Computer aided classification of masses in ultrasonic mammography *Medical Physics*. 2002;29:1968–1973.
15. Chen CM, Chou YH, Han KC, et al. Breast lesions on sonograms: computer-aided diagnosis with nearly setting-independent features and artificial neural networks. *Radiology* 2003;226:504–514. [PubMed: 12563146]
16. Horsch K, Giger ML, Vyborny CJ, Venta LA. Performance of computer-aided diagnosis in the interpretation of lesions on breast sonography. *Academic Radiology* 2004;11:272–280. [PubMed: 15035517]
17. Joo S, Yang YS, Moon WK, Kim HC. Computer-Aided Diagnosis of Solid Breast Nodules: Use of an Artificial Neural Network Based on Multiple Sonographic Features. *IEEE Trans Medical Imaging* 2004;23:1292–1300.
18. Sahiner B, Chan HP, Roubidoux MA, et al. Computerized characterization of breast masses on 3-D ultrasound volumes. *Medical Physics* 2004;31:744–754. [PubMed: 15124991]
19. Sehgal CM, Cary TW, Kangas SA, et al. Computer-based margin analysis of breast sonography for differentiating malignant and benign masses. *Journal of Ultrasound in Medicine* 2004;23:1201–1209. [PubMed: 15328435]
20. Shen WC, Chang RF, Moon WK, Chou YH, Huang CS. Breast ultrasound computer-aided diagnosis using BI-RADS features *Academic Radiology*. 2007;14:928–939.
21. Sahiner B, Chan HP, Hadjiiski LM, et al. Fusion of mammographic and sonographic computer-extracted features for improved characterization of breast masses. *Proc. 7th International Workshop on Digital Mammography* 2004:199–202.
22. Drukker K, Horsch K, Giger ML. Multimodality computerized diagnosis of breast lesions using mammography and sonography. *Academic Radiology* 2005;12:970–979. [PubMed: 16087091]
23. Jesneck JL, Lo JY, Bake JA. Breast mass lesions: Computer-aided diagnosis models with mammographic and sonographic descriptors. *Radiology* 2007;244:390–398. [PubMed: 17562812]
24. Chan HP, Sahiner B, Helvie MA, et al. Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC study. *Radiology* 1999;212:817–827. [PubMed: 10478252]
25. 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]
26. Huo ZM, Giger ML, Vyborny CJ, Metz CE. Breast Cancer: Effectiveness of Computer-aided Diagnosis - Observer Study with Independent Database of Mammograms. *Radiology* 2002;224:560–568. [PubMed: 12147857]
27. Sahiner B, Chan HP, Roubidoux MA, et al. Computer-Aided Diagnosis of Malignant and Benign Breast Masses in 3D Ultrasound Volumes: Effect on Radiologists' Accuracy. *Radiology* 2007;242:716–724. [PubMed: 17244717]
28. Horsch K, Giger ML, Vyborny CJ, Lan L, Mendelson EB, Hendrick RE. Classification of breast lesions with multimodality computer-aided diagnosis: Observer study results on an independent clinical data set *Radiology*. 2006;240:357–368.
29. Bhatti PT, LeCarpentier GL, Roubidoux MA, Fowlkes JB, Helvie MA, Carson PL. Discrimination of sonographically detected breast masses using frequency shift color Doppler imaging in combination with age and gray scale criteria. *Journal of Ultrasound in Medicine* 2001;20:343–350. [PubMed: 11316312]
30. Sahiner B, Petrick N, Chan HP, et al. Computer-Aided Characterization of Mammographic Masses: Accuracy of Mass Segmentation and its Effects on Characterization. *IEEE Transactions on Medical Imaging* 2001;20:1275–1284. [PubMed: 11811827]
31. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA, Gurcan MN. Analysis of Temporal Change of Mammographic Features: Computer-Aided Classification of Malignant and Benign Breast Masses. *Medical Physics* 2001;28:2309–2317. [PubMed: 11764038]

32. Sahiner B, Chan HP, Petrick N, Helvie MA, Hadjiiski LM. Improvement of mammographic mass characterization using spiculation measures and morphological features. *Medical Physics* 2001;28:1455–1465. [PubMed: 11488579]
33. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis. *Medical Physics* 1998;25:516–526. [PubMed: 9571620]
34. Sahiner B, Chan HP, Roubidoux MA, et al. Computerized characterization of breast masses on 3-D ultrasound volumes. *Medical Physics* 2004;31:744–754. [PubMed: 15124991]
35. American College of Radiology Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas). Reston, VA: American College of Radiology; 2003.
36. Sickles EA. Nonpalpable, circumscribed, noncalcified solid breast masses: likelihood of malignancy based on lesion size and age of patient. *Radiology* 1994;192:439–442. [PubMed: 8029411]
37. Metz CE. ROC methodology in radiologic imaging. *Investigative Radiology* 1986;21:720–733. [PubMed: 3095258]
38. McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Med Decis Making* 1984;4:137–150. [PubMed: 6472062]
39. Dorfman DD, Berbaum KS, Metz CE. ROC rating analysis: Generalization to the population of readers and cases with the jackknife method. *Investigative Radiology* 1992;27:723–731. [PubMed: 1399456]
40. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* 1996;201:745–750. [PubMed: 8939225]
41. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*. New York: Springer-Verlag; 2001.
42. Sahiner B, Chan HP, Hadjiiski L. Classifier performance prediction for computer-aided diagnosis using a limited data set. *Medical Physics* 2008;35:1559–1570. [PubMed: 18491550]
43. Gur D, Rockette HE, Armfield DR, et al. Prevalence effect in a laboratory environment. *Radiology* 2003;228:10–14. [PubMed: 12832568]
44. Gur D, Bandos AI, Fuhrman CR, Klym AH, King JL, Rockette HE. The prevalence effect in a laboratory environment: Changing the confidence ratings. *Academic Radiology* 2007;14:49–53. [PubMed: 17178365]
45. Egglin TKP, Feinstein AR. Context bias - A problem in diagnostic radiology. *Jama-Journal of the American Medical Association* 1996;276:1752–1755.
46. Gur D, Bandos AI, Cohen CS, et al. The "Laboratory" effect: Comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 2008;249:47–53. [PubMed: 18682584]

Figure 1.

The average ROC curves of the radiologists in the MAM, USM, and CADx modes, and the ROC curve of the combined mammography-US CADx system.

The area A_z under ROC curve, and the partial area index above a sensitivity of 0.9, $A_z^{(0.9)}$, for characterization of the masses in our data set in MAM, USM, and CADx reading modes. The p-values of the difference in average A_z value observed in USM mode compared to MAM mode, and in CADx mode compared to USM mode were estimated using MRMC ROC analysis. The corresponding p-values for $A_z^{(0.9)}$ were estimated using two-tailed paired t-test.

Table 1

Rad.	Years of experience	Area A_z under ROC curve			The partial area index above a sensitivity of 0.9, $A_z^{(0.9)}$		
		MAM mode	USM mode	CADx mode	MAM mode	USM mode	CADx mode
1	3	0.88±0.04	0.96±0.02	0.98±0.01	0.23±0.15	0.43±0.17	0.52±0.15
2	10	0.83±0.05	0.86±0.04	0.91±0.04	0.30±0.14	0.42±0.13	0.63±0.21
3	26	0.87±0.04	0.92±0.04	0.96±0.02	0.54±0.16	0.77±0.19	0.72±0.13
4	20	0.84±0.05	0.92±0.03	0.95±0.03	0.38±0.12	0.66±0.16	0.83±0.11
5	6	0.93±0.03	0.96±0.02	0.96±0.02	0.04±0.18	0.49±0.10	0.60±0.18
6	5	0.87±0.04	0.95±0.03	0.98±0.02	0.41±0.15	0.64±0.19	0.55±0.20
7	17	0.90±0.04	0.95±0.03	0.95±0.03	0.35±0.16	0.40±0.17	0.53±0.14
8	14	0.84±0.07	0.93±0.03	0.95±0.03	0.36±0.15	0.62±0.19	0.81±0.13
9	12	0.80±0.06	0.87±0.05	0.91±0.03	0.07±0.06	0.50±0.15	0.71±0.10
10	10	0.85±0.05	0.90±0.04	0.93±0.03	0.02±0.14	0.27±0.17	0.55±0.14
Avg.	12	0.87	0.93	0.95	0.27	0.52	0.65
p-value			0.03	0.05		0.001	0.008

Table 2

Numbers of benign and malignant masses for which the 10 radiologists' LM rating changed at step 2 (MAM→USM) and step 3 (USM→CADx) in the observer study. Numbers in parentheses are percentages. The total number of readings for malignant masses was 350 (35 masses × 10 radiologists) and the total number of readings for benign masses was 320 (32 masses × 10 radiologists). A beneficial change for a malignant mass is an increase in the LM rating, and that for a benign mass is a decrease in the LM rating.

		MAM→USM	USM→CADx
LM changes for malignant masses	Beneficial	190(54)	97(28)
	Detrimental	20 (6)	11 (3)
	Total	210(60)	108(31)
LM changes for benign masses	Beneficial	104(33)	74(23)
	Detrimental	110(34)	56(18)
	Total	214(67)	130(41)
Total changes	Beneficial	294(44)	171(26)
	Detrimental	130(19)	67(10)
	Total	424(63)	238(36)

Table 3

The sensitivity and specificity for each radiologist. The significances of the difference in the averages between USM mode vs. MAM mode, and CADx mode vs, USM mode was estimated using Student's paired t-test.

Radiologist	MAM mode		USM Mode		CADx Mode	
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
1	0.97	0.34	1.00	0.16	1.00	0.16
2	0.83	0.66	0.89	0.66	0.94	0.63
3	1.00	0.13	1.00	0.13	1.00	0.19
4	0.91	0.41	0.97	0.34	1.00	0.34
5	1.00	0.16	1.00	0.16	1.00	0.28
6	0.91	0.44	0.97	0.44	1.00	0.34
7	1.00	0.13	1.00	0.09	0.97	0.28
8	0.77	0.84	0.94	0.53	0.97	0.50
9	1.00	0.00	1.00	0.09	1.00	0.09
10	1.00	0.25	1.00	0.13	1.00	0.09
Average	0.94	0.33	0.98	0.27	0.99	0.29
p-value			0.06	0.13	0.17	0.50

Table 4

Changes in biopsy decisions for each radiologist in step 2 (MAM→USM,) and step 3 (USM→CADx) in the observer study. A beneficial change for a malignant mass is a decision to recommend biopsy for the mass that was not recommended for biopsy in the previous step. A beneficial change for a benign mass is a decision not to recommend biopsy for the mass that was recommended for biopsy in the previous step.

Rad.	Change in biopsy decision, step 2 (MAM→USM)				Change in biopsy decision, step 3(USM→CADx)			
	Malignant masses		Benign masses		Malignant masses		Benign masses	
	Beneficial	Detrimental	Beneficial	Detrimental	Beneficial	Detrimental	Beneficial	Detrimental
1	1	0	1	7	0	0	0	0
2	4	2	4	4	2	0	1	2
3	0	0	2	2	0	0	2	0
4	3	1	4	6	1	0	1	1
5	0	0	3	3	0	0	5	1
6	3	1	6	6	1	0	1	4
7	0	0	2	3	0	1	6 [†]	0
8	6 [*]	0	1 [†]	11	1	0	3	4
9	0	0	3	0	0	0	1	1
10	0	0	2	6	0	0	0	1
Total	17	4	28	48	5	1	20	14

* A statistically significant change for malignant masses (McNemar's test)

† A statistically significant change for benign masses (McNemar's test)

Table 5

Changes in biopsy decisions for each radiologist in step 3 (USM→CADx) if the decision threshold for biopsy recommendation were chosen to be 10% in CADx mode, and maintained at 2% in USM mode, with the sensitivity fixed at 98% for both modes

Rad.	Change in biopsy decision for malignant mass		Change in biopsy decision for benign mass	
	Beneficial	Detrimental	Beneficial	Detrimental
1 [†]	0	0	7	0
2	2	1	1	1
3	0	0	4	0
4	1	0	2	1
5 [†]	0	0	9	1
6	1	0	1	3
7 [†]	0	1	12	0
8	1	1	3	3
9	0	0	4	1
10	0	0	4	1
Total	5	3	47	11

[†] A statistically significant change for benign masses (McNemar's test)