Published in final edited form as:
Radiology. 2007 March ; 242(3): 716–724. doi:10.1148/radiol.2423051464.

Malignant and Benign Breast Masses on 3D US Volumetric Images: Effect of Computer-aided Diagnosis on Radiologist Accuracy

Berkman Sahiner, PhD, **Heang-Ping Chan, PhD**, **Marilyn A. Roubidoux, MD**, **Lubomir M. Hadjiiski, PhD**, **Mark A. Helvie, MD**, **Chintana Paramagul, MD**, **Janet Bailey, MD**, **Alexis V. Nees, MD**, and **Caroline Blane, MD**
Department of Radiology, University of Michigan Medical Center, CGC B2102, 1500 E Medical Center Dr, Ann Arbor, MI 48109-0904.

Abstract

Purpose—To retrospectively investigate the effect of using a custom-designed computer classifier on radiologists' sensitivity and specificity for discriminating malignant masses from benign masses on three-dimensional (3D) volumetric ultrasonographic (US) images, with histologic analysis serving as the reference standard.

Materials and Methods—Informed consent and institutional review board approval were obtained. Our data set contained 3D US volumetric images obtained in 101 women (average age, 51 years; age range, 25–86 years) with 101 biopsy-proved breast masses (45 benign, 56 malignant). A computer algorithm was designed to automatically delineate mass boundaries and extract features on the basis of segmented mass shapes and margins. A computer classifier was used to merge features into a malignancy score. Five experienced radiologists participated as readers. Each radiologist read cases first without computer-aided diagnosis (CAD) and immediately thereafter with CAD. Observers' malignancy rating data were analyzed with the receiver operating characteristic (ROC) curve.

Results—Without CAD, the five radiologists had an average area under the ROC curve (A_z) of 0.83 (range, 0.81–0.87). With CAD, the average A_z increased significantly ($P = .006$) to 0.90 (range, 0.86–0.93). When a 2% likelihood of malignancy was used as the threshold for biopsy recommendation, the average sensitivity of radiologists increased from 96% to 98% with CAD, while the average specificity for this data set decreased from 22% to 19%. If a biopsy recommendation threshold could be chosen such that sensitivity would be maintained at 96%, specificity would increase to 45% with CAD.

Conclusion—Use of a computer algorithm may improve radiologists' accuracy in distinguishing malignant from benign breast masses on 3D US volumetric images.

Address correspondence to B.S. (berki@umich.edu)..

Author contributions:
Guarantors of integrity of entire study, B.S., H.P.C.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; manuscript final version approval, all authors; literature research, B.S., H.P.C., M.A.R., L.M.H.; clinical studies, M.A.R., C.P.; experimental studies, B.S., M.A.R., M.A.H., C.P., J.B., A.V.N., C.B.; statistical analysis, B.S., H.P.C., L.M.H.; and manuscript editing, B.S., H.P.C., M.A.R., M.A.H.

Supplemental material: http://radiology.rsnajnls.org/cgi/content/full/2423051464/DC1

Authors stated no financial relationship to disclose.

In current clinical practice, the rate of positive biopsy results for breast cancer is about 15%–30% (1–3). To reduce patient anxiety and morbidity and to decrease health care costs, it is desirable to reduce the number of benign biopsy results without missing malignant lesions. Results of previous studies of mammography have shown that radiologists' accuracy in distinguishing malignant from benign masses can significantly improve when they use a computer-aided diagnosis (CAD) system as a second opinion (4–6).

Ultrasonography (US) is an important imaging modality in the characterization of breast masses. For differentiation of simple cysts from other lesions, interpretation of US images by experienced breast radiologists results in an accuracy close to 100% (7). In current clinical practice, if a palpable or mammographically suspicious mass cannot be confidently categorized as a cyst on US images, it is often recommended for biopsy. Several reports (8–10) have indicated that the improvement in US imaging technology and the interpretation of US images by experienced radiologists may make it possible to characterize solid breast masses as malignant or benign with a high level of accuracy.

Advances in Knowledge

- The computer-aided diagnosis (CAD) algorithm used in this study assisted experienced breast imaging radiologists in accurately characterizing masses as benign or malignant on 3D US volumetric images. The average area under the receiver operating characteristic (ROC) curve improved significantly ($P = .005$) from 0.83 to 0.90, and the average partial area index above a sensitivity of 0.9 value improved significantly ($P = .015$) from 0.30 to 0.44.
- When we confined our ROC analysis to the subset of 95 solid masses, the results were virtually unchanged from the entire set of 101 masses.
- With CAD, the average likelihood of malignancy rating decreased for benign masses ($P = .51$) and increased for malignant masses ($P < .001$).

Several groups of researchers have been developing methods for computerized characterization of masses on two-dimensional (2D) US images (11–14). We have developed an automated computer classifier for differentiation of malignant and benign breast masses on three-dimensional (3D) US volumetric images (15). Thus, the purpose of our study was to retrospectively investigate the effect of using the computer classifier we developed on radiologists' sensitivity and specificity for discriminating malignant masses from benign masses on 3D volumetric US images, with histologic analysis serving as the reference standard.

Materials and Methods

Patients and Diagnoses

We received institutional review board approval prior to the commencement of our study; informed patient consent, including consent for future retrospective data analysis, was obtained. Use of the data set in this study was Health Insurance Portability and Accountability Act compliant. The study group consisted of 130 consecutive women who underwent 3D breast US between 1998 and 2002. All patients had a US mass that was classified as suspicious or highly suggestive of malignancy, and they were scheduled to undergo core-needle biopsy, surgical biopsy, or fine-needle aspiration biopsy. A total of 29 patients from this study group were excluded from analysis for the following reasons: *(a)* They had undergone prior biopsy in the same region of the breast, *(b)* they had sonographically simple cysts, *(c)* US scanning was deemed technically unsuccessful because of motion or other artifacts, *(d)* they did not undergo biopsy, and/or *(e)* masses were incompletely scanned in any dimension because of

large size or eccentric position on the image. Thus, our study group consisted of 101 patients (average age, 51 years; age range, 25–86 years). On the basis of core-needle biopsy, surgical biopsy, or fine-needle aspiration biopsy findings, 56 masses were classified as malignant and 45 were classified as benign (Fig 1). The mean mass diameter was 1.29 cm ± 0.77 (standard deviation).

3D US Imaging

The 3D US data were acquired with an experimental system that was previously developed and tested at our institution (16,17). The 3D system consisted of a commercially available Logiq 700 (GE Medical Systems, Milwaukee, Wis) US scanner with an M12 linear-array transducer, a mechanical transducer-guiding system, and a computer workstation. The linear-array transducer was operated at a frequency of 11 MHz. All 3D US images were acquired by the same US technologist, who had 15 years of breast US experience. Acquisition of 3D US images is different from acquisition of clinical US images; the latter is performed by radiologists at our institution. The technologist was free to set the focal distance and overall gain adjustment to obtain the best possible image. Before 3D image acquisition, the technologist used clinical US images and mammographic images and reports to identify the suspicious mass. During 3D image acquisition, the technologist manually moved the transducer linearly in the cross-plane or the z direction, while the image acquisition system recorded 2D B-mode images in the scanning plane (x-y plane). The 2D images were obtained at approximately 0.5-mm incremental translations, which were measured and recorded with a translation sensor. The scanned breast region typically measured 4.5 cm long by 4.0 cm wide by 4.0 cm deep. The typical in-section pixel size was approximately 0.11 × 0.11 mm.

The B-mode images were recorded in the memory buffer of the US scanner. After data acquisition, US images and position data were transferred digitally to the workstation, where individual planes were cropped and stacked to form a 3D volumetric image. The biopsy-proved mass on each image was identified by a Mammography Quality Standards Act–qualified radiologist (M.A.R., 8 years of experience in breast US imaging, referred to as radiologist 0 hereafter), who used clinical US and mammographic images to confirm that the 3D images contained the mass of interest and showed the mass in its entirety.

Computerized Classification of Masses in US Volumetric Data Sets

The first step of computerized analysis (15) involved extraction of the mass boundaries in the 3D volumetric data set (ie, mass segmentation). Automated segmentation of breast masses on US images is a difficult task because of image speckles, posterior shadowing, and variations of the gray level both within the mass and within the normal breast tissue. We developed a 3D active contour model for segmentation (Fig 2). The active contour model combined prior knowledge about the relative smoothness of the 3D mass shape on the US volumetric image with information in the image data.

After mass segmentation, image features were extracted from the mass and its margins for classification. Our feature space consisted of width-to-height ratio, posterior shadowing, and texture descriptors. The mass shape in terms of relative width to height was described by the ratio of the widest cross section of the automatically segmented lesion shape to the tallest cross section. Posterior shadowing features were defined in terms of the normalized average gray-level values in strips posterior to the mass. Texture features were extracted from two disk-shaped regions containing the boundary of each mass and, presumably, the mass and the normal tissue adjacent to the boundary of the mass. These regions followed the contour determined with the active contour model (Fig 3, Appendix E1 [http://radiology.rsnajnls.org/cgi/content/full/2423051464/DC1]).

The features described previously were extracted from each section of the US volumetric image that contained a mass to define section-based features. For a given mass, features extracted from different sections were combined to define case-based features. Linear discriminant analysis with stepwise feature selection (18) was applied to the case-based feature vectors to obtain computer-estimated malignancy scores. A leave-one-case-out resampling method (19) was used to train the classification system and to obtain test scores. The test scores obtained with the leave-one-case-out method were used as the malignancy scores in the observer performance study. Two Gaussian functions were fitted to the distributions of the malignancy scores of the benign and malignant classes separately and were used in the observer performance study.

Observer Performance Study

Five radiologists (M.A.H., C.P., J.B., A.V.N., C.B.), who were referred to as radiologists 1–5, had 3–26 years of experience in mammographic and breast US image interpretation. They were all Mammography Quality Standards Act qualified, and four were fellowship trained in breast imaging. In our department, about 4300 breast US examinations are performed annually.

An interactive graphical user interface facilitated navigation through the scanned 3D volumetric images of interest that contained the mass and allowed adjustment of the window and level settings of the displayed images. The location of the mass of interest, as determined by radiologist 0 with all available imaging and histologic findings, was marked on each section so that all radiologists would rank the same mass and ignore others if more than one mass could be seen on the volumetric image.

Observers first interpreted studies without CAD. This involved assessing the mass for shape, margins, echogenicity, cystic versus solid appearance, and through transmission, as well as estimating the likelihood of malignancy (LM) on a scale of 0% to 100%. For assessment of mass characteristics, the radiologists chose terms from a list of descriptors that were similar to but not exactly the same as those in the US Breast Imaging Reporting and Data System lexicon of the American College of Radiology, as this observer study was performed before the lexicon was published. For example, the descriptors for shape were “oval,” “round,” “lobulated,” and “irregular,” whereas the descriptors for margins were “circumscribed,” “spiculated,” “microlobulated,” and “ill defined.”

A button corresponding to an LM rating of 0% was provided for benign masses. Another button corresponding to LM ratings of less than 2% was provided for probably benign masses. This second button was set to correspond to Breast Imaging Reporting and Data System category 3 (ie, probably benign) lesions, for which short-interval follow-up is recommended (20). The radiologists used a slide bar to enter ratings between 2% and 100%. The discrete buttons facilitate the selection of the LM ratings more precisely for the benign and probably benign masses because our previous experience indicates that the uncertainty of observers when selecting ratings on a slide bar can be much greater than 2%. The observers were reminded at the beginning of the study that if they rated a mass as having an LM of more than 2% it would indicate that they would recommend the mass for biopsy (20,21). The assessment and the LM estimate were based on the findings in all of the volumetric images (stack of sections) that contained the mass.

We used a two-step sequential reading design, which was found to be a sensitive technique for assessing the difference between the two conditions in previous studies (6,22). Immediately after reading without CAD, the computer-estimated malignancy score for the study was displayed on the screen and the radiologist estimated the LM with CAD. The estimate of LM without CAD was stored in a computer file, and the radiologist was unable to modify it after seeing the computer-estimated score. The computer-estimated malignancy score was linearly

mapped to an integer between 1 and 10 before the score was displayed on the graphical user interface. To provide radiologists with a reference of computer performance, the Gaussian distributions fitted to the computer scores for the malignant and benign lesions were also displayed on the interface. The radiologists could keep their original malignancy rating or change it by using the slide bar after they considered the computer-estimated score. The radiologists were not informed whether a mass was malignant or benign during or after the study, and the overall results of their assessment were not discussed with them before the study was completed.

There was no time limit for the radiologists to assign an LM rating. The radiologists were not informed of the proportion of malignant masses. The study reading order was randomized for each radiologist. To reduce fatigue, each radiologist read the data set in three separate sessions. The three sessions were separated by at least 2 days and at most 1 month. Before participating in the study, the radiologists were trained with five studies that were not part of the test set. They were familiarized with the study design, the functions on the graphical user interface, and the relative malignancy rating scale of the computer during the training session.

The data set used in this investigation was also used in an earlier study to develop the CAD technique (15). Three radiologists in the current investigation had already assigned an LM score for these masses without use of CAD in our earlier study, which had a different experimental design and involved use of a different graphical user interface. (Radiologists 1, 2, and 3 in the current study were referred to as radiologists 3, 4, and 2, respectively, in the earlier study.) The reading sessions in the past and current studies were separated by at least 6 months. The radiologists were not informed whether a mass was malignant or benign during or after the earlier study. The accuracies of these three radiologists in assigning LM scores without CAD in these two studies were compared.

Data and Statistical Analyses

There is no reference standard for mass characteristics since they are judged subjectively by radiologists. Thus, a majority assessment (ie, the mode) for each characteristic was determined according to majority rule by the six radiologists (radiologists 0–5). For example, if one radiologist described the echogenicity characteristics of a mass as hypoechoic, three described them as markedly hypoechoic, one described them as anechoic, and one described them as heterogeneous, the majority assessment for echogenicity of the mass would be markedly hypoechoic. When there was a tie between two descriptors, we used the descriptor chosen by radiologist 0—who was very familiar with the cases owing to her role in data collection—as the tie breaker. If there was a tie and the original descriptor provided by radiologist 0 was not one of the descriptors that were tied, radiologist 0 was asked to re-read the images and choose one of the tied descriptors. An alternative to the majority rule for summarizing the central tendencies is to use the mean of each descriptor. In this study, we chose to use the mode because we were interested in how each mass could be characterized and in the overall central tendency.

The LM ratings of the radiologists with and without CAD were evaluated with receiver operating characteristic (ROC) curve analysis (23,24). The area under the ROC curve *(A_z)* and the partial area index above a sensitivity of 0.9 ($A_z^{0.9}$) (25) were used as measures of accuracy. For an individual radiologist, the significance of the change in accuracy with CAD was also analyzed with the ROC method. For the group of five radiologists, the significance of the change in accuracy with CAD was tested with the Dorfman-Berbaum-Metz multireader multicase method (26) and the Student two-tailed paired *t* test (Microsoft Excel, version 2002; Microsoft, Redmond, Wash). The Dorfman-Berbaum-Metz method (http://xray.bsd.uchicago.edu/krlbp/KRL_ROC/) is normally the preferred method used to analyze the A_z values for multireader multicase data because it accounts for both reader and

case variances, whereas the *t* test does not account for case variance in calculation of the *P* value. Therefore, conclusions drawn from the *t* test can be generalized to the population of readers but not to the population of cases. The *t* test was applied to the evaluation of $A_z^{0.9}$. For this task, we are unaware of any available software that can account for both reader and case variances.

The sensitivity and specificity of each radiologist with and without CAD were compared by using an LM rating of 2% as the threshold above which biopsy would be recommended (20, 21). The radiologists in our study were familiar with Breast Imaging Reporting and Data System recommendations and were well aware that selecting an LM of more than 2% would be the equivalent of declaring that the mass was suspicious enough to warrant biopsy. If the radiologist intended to indicate an LM of less than 2%, he or she selected one of two graphical user interface buttons designated benign and less than 2% LM. The buttons were clearly labeled “benign” and “probably benign.”

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

To investigate whether the change in sensitivity with CAD was statistically significant for a given radiologist, we used the McNemar test (WinStat, version 2005.1; R. Fitch Software, Lehigh Valley, Pa) and considered the number of beneficial and detrimental changes in biopsy recommendation for malignant masses with CAD. If a malignant mass was not recommended for biopsy without CAD but was recommended for biopsy with CAD, this was defined as a beneficial change. If a malignant mass was recommended for biopsy without CAD but was not recommended for biopsy with CAD, this was defined as a detrimental change. We similarly applied the McNemar test to benign masses to investigate whether the change in specificity with CAD was statistically significant.

In addition to analyzing the change in the number of masses for which the LM rating increased above (or decreased below) the biopsy threshold of 2% with use of CAD, we also examined the number of masses for which CAD resulted in a substantial change in the LM rating. We defined a substantial change as an absolute value difference of greater than or equal to five between LM ratings with and without CAD. The substantial decreases and increases in the ratings of malignant and benign masses were examined. For each mass, we also averaged the changes in the LM ratings by the five radiologists and evaluated how CAD changes the average LM ratings for malignant and benign masses with one-sample *t* tests.

When an observer experiment is performed to investigate the effect of CAD on radiologists’ decisions in a laboratory environment, there may be a concern that the radiologists may rely too heavily on the CAD system without adequately merging the computer output with their own judgment. To investigate whether this is the case, we estimated the correlation between the radiologists’ readings with CAD and *(a)* their readings without CAD and *(b)* the computer scores. We then estimated the statistical significance of the difference between these two correlation coefficients by using the method described by Cohen and Cohen (27). If radiologists use the computer scores only when they believe that it makes a true contribution to their original assessment, then the correlation between the radiologists’ readings with CAD and their readings without CAD should be significantly higher than the correlation between the radiologists’ readings with CAD and the computer scores.

Results

Mass Categorization

A total of 95 masses were categorized as solid according to majority rule. Five masses were categorized as complex cysts, and one was categorized as a simple cyst by three or more radiologists. One mass that was categorized as a complex cyst was malignant, and the remaining five nonsolid masses were benign. The most common margin descriptor was “ill defined” (46%) for malignant masses and “circumscribed” (58%) for benign masses. Most of the malignant masses had an irregular shape (59%), and most of the benign masses had an oval shape (69%). Most of the masses (76% of benign masses and 64% of malignant masses) were categorized as hypoechoic. Calcifications were seen in 2% of benign masses and 25% of malignant masses.

ROC Analysis

The A_z values of the radiologists ranged from 0.81 to 0.87 without CAD and from 0.86 to 0.93 with CAD (Table 1). Radiologist 4 had the largest A_z value change when reading with CAD: The A_z value for this radiologist was 0.82 without CAD and 0.93 with CAD. The improvement in A_z values was statistically significant for four of five radiologists.

The average ROC curves (Fig 4) for the radiologists with and without CAD were derived from the average a and b parameters, which were defined as the means of each radiologist's a and b parameters for the fitted ROC curves. The test ROC curve of the computer classifier had an A_z value of 0.92 (Fig 4). With CAD, the average A_z value improved significantly ($P < .01$) from 0.83 to 0.90 and the average $A_z^{0.9}$ value improved significantly ($P = .017$) from 0.30 to 0.44 (Table 2). Improvement in the A_z and $A_z^{0.9}$ values was statistically significant ($P < .01$), even when radiologist 4—who showed the largest improvement with CAD—was excluded from the analysis.

The ROC curves for radiologists 1–3 showed average A_z values of 0.87 and 0.84 in the previous (15) and current studies, respectively. The difference between the current and previous studies was not statistically significant when ROC curves were analyzed as a group ($P = .19$) or when each radiologist's ROC curves were analyzed separately ($P = .86$, $P = 0.13$, and $P = 0.09$ for radiologists 1, 2, and 3, respectively).

Sensitivity and Specificity

On average, radiologist sensitivity increased from 96% to 98% with CAD; however, specificity decreased from 22% to 19% (Table 3). Sensitivity of three radiologists increased, while two radiologists maintained a sensitivity of 100%. The specificity of three radiologists decreased with CAD, the specificity of one radiologist increased, and the specificity of another did not change. Changes in sensitivity and specificity were not statistically significant for any radiologist (range of P values with the McNemar test, .157 to > .99 for sensitivity and .102 to > .99 for specificity). If the LM threshold was to be adjusted to 7% with CAD, the average sensitivity would remain at 96% (same as that without CAD) and the average specificity would increase to 45%. Under this condition, the improvement in specificity for four of five radiologists was statistically significant ($P < .003$, McNemar test), while the change in sensitivity for each radiologist was insignificant (Table 3).

LM Ratings

With 101 masses and five radiologists, we had a total of 505 pairs of LM ratings with and without CAD (Fig 5). The radiologists did not change their LM rating substantially (ie, more than five points) with CAD in 64% (321 of 505) of the readings. For malignant masses, the

ratings were substantially increased for 34% (95 of 280) and decreased for 7% (19 of 280) of the readings. For benign masses, the ratings were substantially increased for 14% (32 of 225) of the readings and decreased for 17% (38 of 225).

To determine the mean change in rating for a mass, the average change in LM rating after CAD for five radiologists was calculated (Fig 6). For benign masses, the decrease in the average LM rating was 0.79; this decrease was not statistically significant (Student two-tailed paired *t* test, $P = .51$). The increase in the average LM rating of malignant masses was 5.59, which was statistically significant (Student two-tailed paired *t* test, $P < .001$).

Correlation

Correlation between radiologists' readings with and their readings without CAD was higher than correlation between radiologists' readings with CAD and computer scores for all five radiologists (Table 4); the difference between the two correlations was statistically significant ($P < .001$) for four radiologists. This result indicates that when radiologists read images with CAD, they had a higher agreement with the diagnosis assigned without CAD as compared with the computer scores.

Solid Masses

To investigate how the radiologists performed with and without CAD for 95 solid masses, we applied ROC analysis to this subset by excluding masses that were categorized as cysts. The average A_z values without and with CAD for this subset were 0.83 and 0.90, respectively, and were unchanged from the entire set of 101 masses. The improvements in A_z values for the individual radiologists, as well as for all radiologists as a group, were significant ($P < .05$) for the subset of solid masses.

Discussion

Our results indicate that the CAD algorithm used in this study assisted even experienced breast imaging radiologists in the characterization of masses on 3D US volumetric images. At our institution, all clinical breast US examinations are performed by breast imaging radiologists and not sonographers; therefore, the readers in our ROC study were experienced in assessing whole volumetric images. Our CAD system improved the accuracy of these experienced radiologists in the interpretation of 3D US volumetric images in terms of the *A* and $A_z^{0.9}$ values.

During our observer experiment, 95 (94%) of the 101 masses were classified as solid according to majority rule. When analysis was limited to the subset of solid masses, the A_z values derived with and without CAD and the significance of the improvement with CAD were essentially unchanged when compared with the results for the entire data set of 101 masses. This indicates that CAD would be helpful even if we only considered the interpretation of the more difficult category of solid masses.

The effect of CAD was mixed when measured in terms of the radiologists' sensitivity and specificity at the current threshold of biopsy recommendation (LM of 2%). With CAD, the average sensitivity of the five radiologists increased from 96% to 98%, while their average specificity for this data set decreased from 22% to 19%. The significant improvement in the ROC curves strongly suggests that these changes do not reflect only a shift in decision threshold along the same ROC curve. Although the changes in sensitivity and specificity were not statistically significant because of the relatively small data set available in this study, these observations indicate a promising trend that may be achieved with CAD.

Since the cost of failing to perform a biopsy for a malignant lesion is much greater than the cost of performing a biopsy for a benign lesion, it can logically be expected that radiologists may use the CAD system to confirm and increase their LM estimate for malignant lesions but not to decrease their LM estimate for low-suspicion lesions. This will result in an overall increase in radiologists' LM ratings, as observed in our study. While the ratings for malignant masses demonstrated a strong tendency to increase with CAD, the ratings for benign masses did not show a strong trend either way. These results led to an increase in sensitivity and a decrease in specificity. However, since the ROC curves of all radiologists improved with CAD, there is a chance that radiologists can adjust their decision thresholds along the higher ROC curves and thus increase both their sensitivity and their specificity. Alternatively, it may be possible to convince them to reduce the LM ratings of very-low-suspicion masses, as indicated by the CAD system, and thus improve the specificity. These improvements may be realized after radiologists accumulate experience and increase their confidence in the use of CAD.

Horsch et al (28) found that the accuracy of both expert mammographers and community radiologists improved significantly when they read 2D US images with CAD. Our study design differs from that used by Horsch et al (28) in that 3D US images were used, but our results reinforce their finding that experienced radiologists can benefit from reading US images with CAD.

The radiologists were not informed of the prevalence of cancer in the data set. However, they probably assumed that the prevalence of the disease was higher than that in the diagnostic population in clinical practice because most laboratory ROC studies are designed to have an approximately equal number of positive and negative cases in order to increase the statistical power for the same total number of cases read (23). Gur et al (29) found that no significant effects could be measured for prevalence in the range of 2%–28% in laboratory ROC experiments. It is not known if their findings could be extended to a prevalence of nearly 50%. On the other hand, since ROC studies are usually performed to measure the relative performances of two modalities instead of their absolute performances in the patient population at large, the prevalence effects should be comparable for both modalities and would be unlikely to change the relative performances, as assumed in most laboratory ROC studies.

Our observations indicate that the radiologists were not overly reliant on computer ratings in this study. First, they did not change their LM rating substantially (ie, a change of five or more points on the 100-point scale) with CAD in 64% of the readings. Second, correlation analysis revealed that the LM ratings assigned by a radiologist with and without CAD were highly correlated, whereas the correlation between the computer scores and the radiologists' LM ratings with CAD was significantly lower for four readers. Third, before all the readings were completed, the radiologists did not receive any feedback regarding whether the computer rating was more accurate than their rating. Thus, they had no way to know that their accuracy would improve by simply following the computer rating.

Our study had a number of limitations. Our data set consisted of only masses that were recommended for core-needle biopsy, surgical biopsy, or fine-needle aspiration biopsy. It is therefore important to investigate the performance of the CAD system in the evaluation of masses that are not recommended for biopsy. Second, all studies in our data set were obtained with the same US machine; the CAD system needs to be evaluated with images acquired with different US imaging systems. Third, all the observers in our study were experienced in the interpretation of mammograms and US images; thus, the effects of CAD on less experienced radiologists were not studied. Fourth, the classifier in our CAD system was trained and tested by using a leave-one-case-out method, and the segmentation method was optimized by using a small subset of the data set. Although the leave-one-case-out resampling method is known to be a nearly unbiased classifier design method (19), the performance of our CAD system

needs to be evaluated by using independent test sets to ensure the generalizability of our approach. Fifth, radiologists generally combine information from US images with information from mammograms to reach a diagnosis; however, we used only information from US images. Sixth, the components of retrospective ROC studies cannot emulate many factors that exist in clinical practice, such as the psychologic effects of the liability of misdiagnosing a malignant lesion.

We conclude that use of a well-trained computer algorithm may improve radiologists' accuracy in distinguishing malignant from benign breast masses on 3D US volumetric images.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to Charles E. Metz, PhD, for the LABMRMC program.

Supported in part by U.S. Army Medical Research Materiel Command grant DAMD17-01-1-0328 and by U.S. Public Health Service grants CA095153 and CA091713.

Abbreviations

A_z	area under the ROC curve
CAD	computer-aided diagnosis
LM	likelihood of malignancy
ROC	receiver operating characteristic
3D	three-dimensional
2D	two-dimensional

References

1. Kopans DB. The positive predictive value of mammography. AJR Am J Roentgenol 1992;158:521–526. [PubMed: 1310825]
2. Adler DD, Helvie MA. Mammographic biopsy recommendations. Curr Opin Radiol 1992;4:123–129. [PubMed: 1524971]
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. AJR Am J Roentgenol 1995;165:1373–1377. [PubMed: 7484568]
4. 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]
5. Huo Z, 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]
6. Hadjiiski L, Chan HP, Sahiner B, et al. Improvement of radiologists' characterization of malignant and benign breast masses in serial mammograms by computer-aided diagnosis: an ROC study. Radiology 2004;233:255–265. [PubMed: 15317954]
7. Jackson VP. The role of US in breast imaging. Radiology 1990;177:305–311. [PubMed: 2217759]
8. Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker SH, Sisney GA. Solid breast nodules: use of sonography to distinguish between malignant and benign lesions. Radiology 1995;196:123–134. [PubMed: 7784555]

9. Skaane P, Engedal K. Analysis of sonographic features in differentiation of fibroadenoma and invasive ductal carcinoma. AJR Am J Roentgenol 1998;170:109–114. [PubMed: 9423610]
10. Taylor KJ, Merritt C, Piccoli C, et al. Ultrasound as a complement to mammography and breast examination to characterize breast masses. Ultrasound Med Biol 2002;28:19–26. [PubMed: 11879948]
11. Garra BS, Krasner BH, Horii SC, Ascher S, Mun SK, Zeman RK. Improving the distinction between benign and malignant breast lesions: the value of sonographic texture analysis. Ultrason Imaging 1993;15:267–285. [PubMed: 8171752]
12. 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]
13. 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]
14. Horsch K, Giger ML, Venta LA, Vyborny CJ. Computerized diagnosis of breast lesions on ultrasound. Med Phys 2002;29:157–164. [PubMed: 11865987]
15. Sahiner B, Chan HP, Roubidoux MA, et al. Computerized characterization of breast masses on 3-D ultrasound volumes. Med Phys 2004;31:744–754. [PubMed: 15124991]
16. 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. J Ultrasound Med 2001;20:343–350. [PubMed: 11316312]
17. Carson PL, Fowlkes JB, Roubidoux MA, et al. 3-D color Doppler image quantification of breast masses. Ultrasound Med Biol 1998;24:945–952. [PubMed: 9809628]
18. Draper, NR. Applied regression analysis. Wiley; New York, NY: 1998.
19. Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning. Springer-Verlag; New York, NY: 2001.
20. American College of Radiology. Breast imaging reporting and data system atlas (BIRADS atlas). Vol. 4th ed. American College of Radiology; Reston, Va: 2003.
21. 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]
22. Beiden SV, Wagner RF, Doi K, et al. Independent versus sequential reading in ROC studies of computer-assist modalities: analysis of component of variance. Acad Radiol 2002;9:1036–1043. [PubMed: 12238545]
23. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986;21:720–733. [PubMed: 3095258]
24. McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis Making 1984;4:137–150. [PubMed: 6472062]
25. 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]
26. Dorfman DD, Berbaum KS, Metz CE. ROC rating analysis: generalization to the population of readers and cases with the jackknife method. Invest Radiol 1992;27:723–731. [PubMed: 1399456]
27. Cohen, J.; Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. Lawrence Erlbaum; Hillside, NJ: 1983.
28. Horsch K, Giger ML, Vyborny CJ, Venta LA. Performance of computer-aided diagnosis in the interpretation of lesions on breast sonography. Acad Radiol 2004;11:272–280. [PubMed: 15035517]
29. Gur D, Rockette HE, Armfield DR, et al. Prevalence effect in a laboratory environment. Radiology 2003;228:10–14. [PubMed: 12832568]

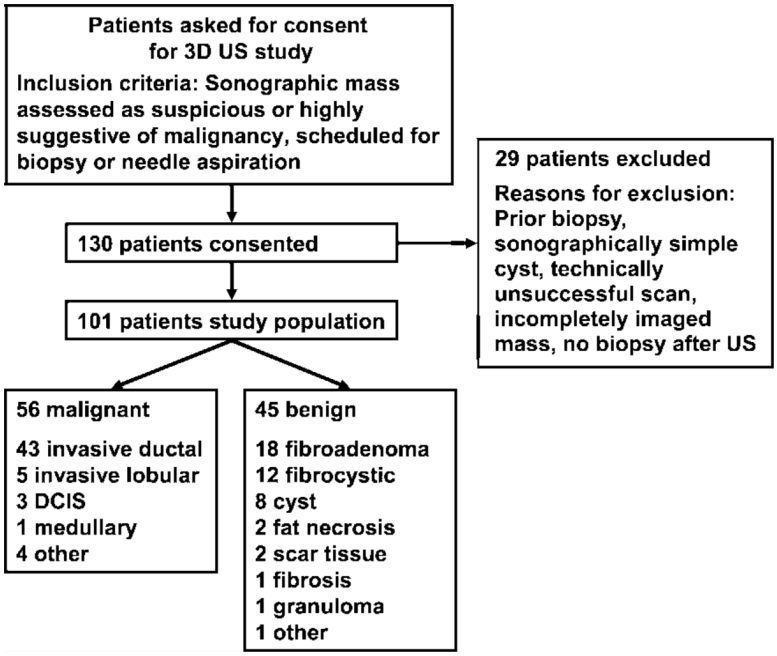

Figure 1.
Flow chart shows the study population, inclusion and exclusion criteria, and findings at core-needle biopsy, surgical biopsy, or fine-needle aspiration biopsy. *DCIS* = ductal carcinoma in situ.

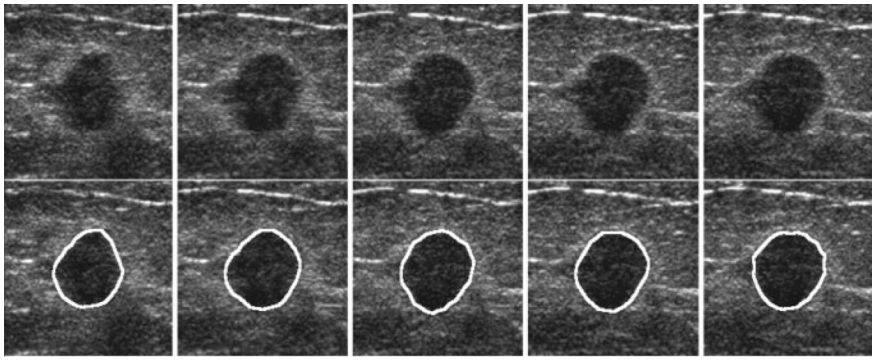

Figure 2.
Five US sections containing a malignant mass (upper row) and results of computer segmentation (lower row). The outlined area indicates the boundary of the mass extracted with the computer segmentation algorithm.

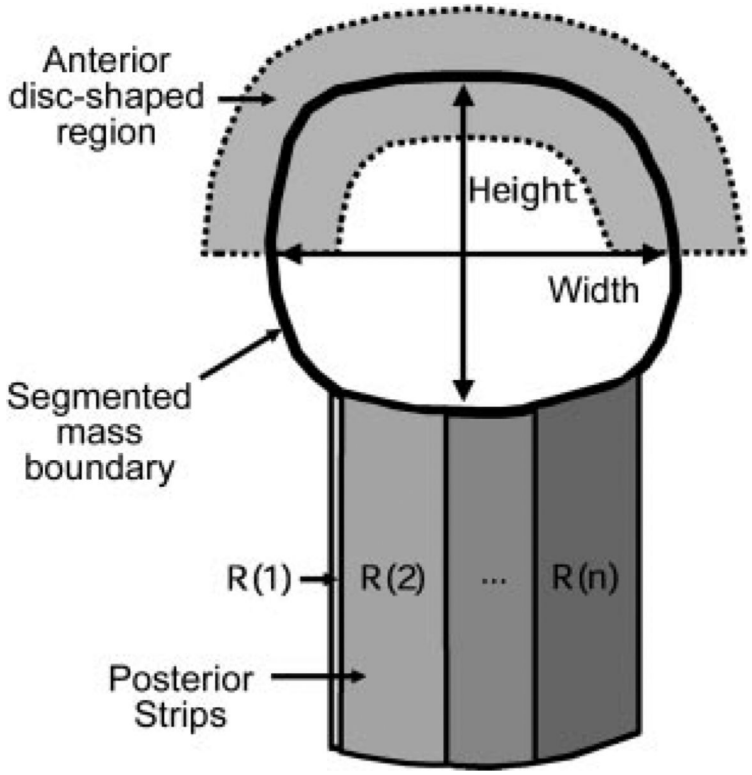

Figure 3.
Schematic drawing shows that for feature extraction, the width and height of the mass on a US section were defined as the widest and tallest cross-sections of the mass on that image. The mean gray level values within the overlapping posterior strips *(R[n], R[1], and R[2])* and the segmented mass were used to define the posterior shadowing features. The disk-shaped regions for texture feature extraction followed the shape of the mass and contained part of the segmented mass and part of its margins. An example of the anterior disk-shaped region is shown as the gray area above the segmented mass.

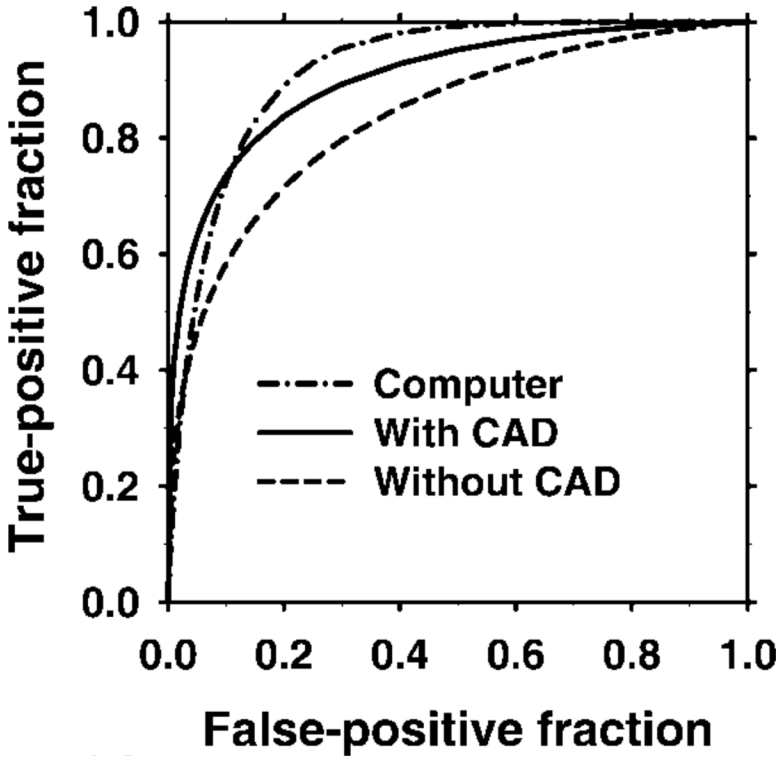

Figure 4.
Graph shows the average ROC curves of the computer classifier and of the radiologists working with and without CAD. Average ROC curves were constructed by using the mean a and b parameters of the individual observers' ROC curves.

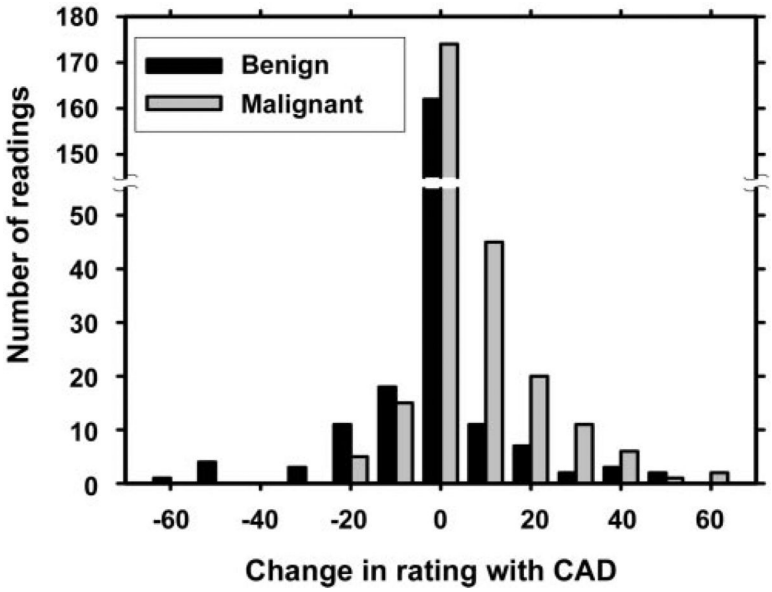

Figure 5.

Histogram shows the change in radiologists' LM ratings with use of CAD. For the majority of masses (59% of malignant masses and 69% of benign masses), the change ranged from −4 to 4. When the change was less than −4 or more than 4, it was considered substantial. For malignant masses, the ratings were substantially increased for an average of 34% (95 of 280) of the readings and substantially decreased for an average of 7% (19 of 280) of the readings. For benign masses, the ratings were substantially increased for 14% (32 of 225) of the readings and substantially decreased for 17% (38 of 225) of the readings.

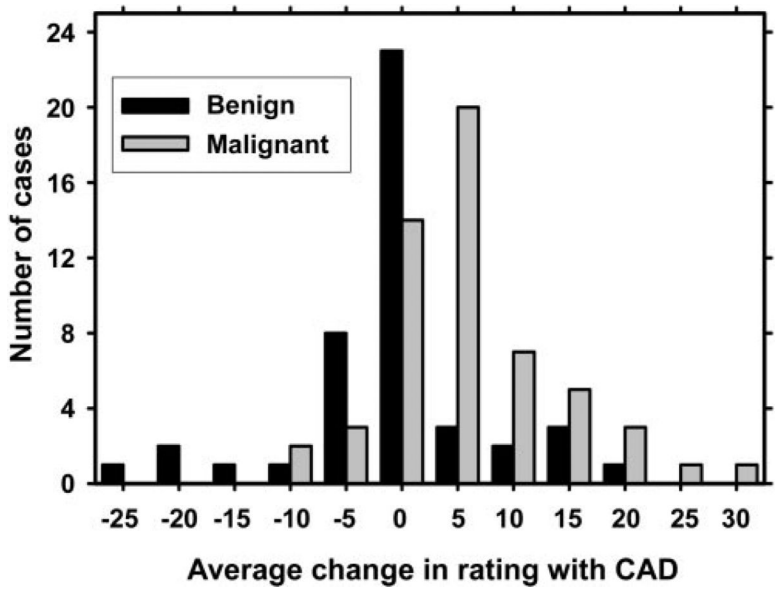

Figure 6.

Histogram shows the mean change in radiologists' LM ratings with use of CAD. The mean change for a mass was computed by averaging the changes in the LM ratings for the mass over the five radiologists who participated in the study. For benign masses, the overall average LM rating decrease was 0.79; this difference was not statistically significant ($P = .51$). For malignant masses, the overall average LM rating increase was 5.59; this difference was statistically significant ($P < .001$).

Table 1

A_z and $A_z^{0.9}$ Values for Characterization of the Masses in the Data Set without and with CAD

	A_z			$A_z^{0.9}$		
Radiologist No.	**Without CAD**	**With CAD**	***P* Value**	**Without CAD**	**With CAD**	***P* Value**
1	0.83 ± 0.04	0.89 ± 0.03	.002	0.26 ± 0.10	0.33 ± 0.13	.32
2	0.81 ± 0.04	0.86 ± 0.04	<.001	0.14 ± 0.08	0.26 ± 0.12	.07
3	0.87 ± 0.03	0.91 ± 0.03	.092	0.39 ± 0.12	0.52 ± 0.12	.08
4	0.82 ± 0.04	0.93 ± 0.03	<.001	0.40 ± 0.10	0.68 ± 0.09	.001
5	0.83 ± 0.04	0.90 ± 0.03	<.001	0.30 ± 0.10	0.41 ± 0.12	.06

Note.—The A_Z and $A_z^{0.9}$ values are means ± standard deviations. The statistical significance for each radiologist was estimated, as described in the literature (24,25).

Table 2

Average A_z and $A_z^{0.9}$ Values without and with CAD for the Five Radiologists

Accuracy Measure	Without CAD	With CAD	*P* Value*	*P* Value†
A_z	0.83	0.90	.006	.005
$A_z^{0.9}$	0.30	0.44	. . .	.017

Note.—Data were obtained by using the average a and b parameters from the fitted ROC curves. The significance of the change in the A_Z value with CAD for the group of five radiologists was estimated by using both the Dorfman-Berbaum-Metz method and the Student two-tailed paired *t* test. The significance of the change in the $A_z^{0.9}$ value was estimated by using the Student two-tailed paired *t* test.

*Calculated with the Dorfman-Berbaum-Metz method.

†Calculated with the Student two-tailed paired *t* test.

Table 3

Sensitivity and Specificity for Each Radiologist at Decision Thresholds of 2% and 7% LM

Radiologist No.	Sensitivity			Specificity		
	Without CAD*	With CAD*	With CAD†	Without CAD*	With CAD*	With CAD†
1	56 (100)	56 (100)	56 (100)	4 (9)	5 (11)	14 (31)
2	51 (91)	53 (95)	49 (88)	12 (27)	11 (24)	27 (60)
3	52 (93)	54 (96)	53 (95)	23 (51)	21 (47)	28 (62)
4	55 (98)	56 (100)	56 (100)	9 (20)	5 (11)	22 (49)
5	56 (100)	56 (100)	56 (100)	1 (2)	1 (2)	11 (24)
Average	54 (96)	55 (98)	54 (97)	10 (22)	9 (19)	20 (45)

Note.—Data are the number of correctly classified lesions. Data in parentheses are percentages. The total numbers of malignant and benign lesions were 56 and 45, respectively.

*Data show the sensitivity and specificity at the decision threshold of 2% LM without and with CAD.

†Data show the sensitivity and specificity with CAD at a hypothetical decision threshold of 7% LM, for which the average sensitivity would be the same as that without CAD (96%); however, the average specificity would increase to 45%.

Table 4

Correlation between Radiologists' LM Ratings with and without CAD and between Radiologists' LM Ratings with CAD and Computer Scores

Radiologist No.	Correlation between LM Rating with CAD and LM Rating without CAD*	Correlation between LM Rating with CAD and Computer Scores*	*P* Value
1	0.94	0.70	<.001
2	0.96	0.61	<.001
3	0.96	0.71	<.001
4	0.86	0.82	.27
5	0.94	0.70	<.001

Note.—The statistical significance in the difference between the two correlation coefficients of each radiologist was estimated by using the Cohen and Cohen method (27).

* Data are correlation coefficient values.