Breast US Computer-aided Diagnosis Workstation: Performance with a Large Clinical Diagnostic Population[1]

Karen Drukker, PhD
Nicholas P. Gruszauskas, MS
Charlene A. Sennett, MD
Maryellen L. Giger, PhD

Purpose: To evaluate the performance of a computer-aided diagnosis (CAD) workstation in classifying cancer in a realistic data set representative of a clinical diagnostic breast ultrasonography (US) practice.

Materials and Methods: The database consisted of consecutive diagnostic breast US scans collected with informed consent with a protocol approved by the institutional review board and compliant with the HIPAA. Images from 508 patients with a total of 1046 distinct abnormalities were used. One hundred one patients had breast cancer. Results both for patients in whom the lesion abnormality was proved with either biopsy or aspiration ($n = 183$) and for all patients irrespective of biopsy status ($n = 508$) are presented. The ability of the CAD workstation to help differentiate malignancies from benign lesions was evaluated with a leave-one-out-by-case analysis. The clinical specificity of the radiologists for this dataset was determined according to the biopsy rate and outcome.

Results: In the task of differentiating cancer from all other lesions sent to biopsy, the CAD workstation obtained an area under the receiver operating characteristic curve (AUC) value of 0.88, with 100% sensitivity at 26% specificity (157 cancers and 362 lesions total). The radiologists' specificity at 100% sensitivity for this set was zero. When analyzing all lesions irrespective of biopsy status, which is more representative of actual clinical practice, the CAD scheme obtained an AUC of 0.90 and 100% sensitivity at 30% specificity (157 cancers and 1046 lesions total). The radiologists' specificity at 100% sensitivity for this set was 77%.

Conclusion: Current levels of computer performance warrant a clinical evaluation of the potential of US CAD to aid radiologists in lesion work-up recommendations.

[1] From the Department of Radiology, University of Chicago, MC 2026, 5841 S Maryland Ave, Chicago, IL 60637. From the 2006 RSNA Annual Meeting. Received October 9, 2007; revision requested November 27; final revision received January 21, 2008; accepted January 30. Supported in part by NIH grants R01-CA89452, R21-113800, and P50-CA125183. M.L.G. is a stockholder of and supported by a grant from R2 Technology/Hologic. **Address correspondence to** K.D. (e-mail: *kdrukker@uchicago.edu*).

To date, breast ultrasonography (US) largely functions as a diagnostic—rather than a screening—method and is used to improve specificity in the assessment of abnormalities seen at mammography or of palpable masses found during clinical breast examinations. Patients with a suspicious abnormality seen at mammography frequently undergo a subsequent breast US examination to avoid unnecessary biopsies. We believe that computerized analysis has the potential to help radiologists make correct diagnoses. Previously, a computer-aided diagnosis (CAD) scheme (1,2) and a combined computer-aided detection and diagnosis scheme were developed (3–7). Robustness of those schemes was demonstrated across different users and institutions and with scanner models of two manufacturers (5). Moreover, a clinical CAD workstation prototype was developed and tested in daily clinical practice with regard to workflow while blinding the radiologists to the computer output in order to not influence patient care (8). The purpose of our study was to evaluate the performance of a CAD workstation in the task of classifying cancer in a realistic dataset representative of a clinical diagnostic breast US practice.

Advances in Knowledge

- In the task of differentiating cancer from all other lesions sent to biopsy, the computer-aided diagnosis (CAD) workstation obtained an area under the receiver operating characteristic curve (AUC) value of 0.88, with 100% sensitivity at 26% specificity (157 cancers and 362 lesions total).
- When analyzing all lesions irrespective of biopsy status, which is more representative of actual clinical practice, the CAD scheme obtained an AUC of 0.90 and 100% sensitivity at 30% specificity (157 cancers and 1046 lesions total).

Materials and Methods

Patients and Lesions

The database consisted of consecutive diagnostic breast US scans collected under protocols approved by the institutional review board and in compliance with the Health Insurance Portability and Accountability Act. Informed consent was obtained from 695 patients for participation in this study. Of these patients, 187 had no US abnormality and the remaining 508 patients had at least one lesion, with a total of 1046 distinct abnormalities on 2266 images (Fig 1).

When biopsy data were not available, imaging characteristics on US scans, magnetic resonance (MR) images, and mammograms were used to determine whether a lesion was benign or malignant.

In our patient population, the reasons why biopsies were not performed after a diagnostic breast US examination included a benign US appearance of lesion(s), a patient history of benign breast disease combined with a stable appearance of the lesion(s), and lesion(s) with questionable character at US but with a benign appearance at subsequent follow-up with breast MR imaging. Conversely, patients with lesions appearing to be benign at US were occasionally referred to biopsy for various reasons, including patient discomfort or anxiety, patient age, a highly suspicious appearance with another imaging modality (MR imaging), or a family history of breast cancer. In this report, the term *lesion* includes all findings observed at US, including, for example, cancers, benign solid lesions, cystic lesions, hematomas, surgical scars, and lymph nodes. One hundred eighty-three of the 508 patients (36%) with a US abnormality were referred to biopsy, and 362 lesions underwent biopsy. The clinical positive predictive value for biopsy was 43% (157 of 362 lesions). There were 101 patients with breast cancer and a total of 157 cancerous lesions (including 26 metastatic lymph nodes), bringing the cancer prevalence in this study population to 20% by patient (101 of 508 patients) and to 15% by lesion (157 of 1046 lesions). The most prevalent lesion type was cystic, with most being small subcentimeter cysts. Patients with lesions seen at US in whom the disease was not confirmed with biopsy were followed up for an average of 3 years (range, 2–4 years) to minimize the risk of including missed cancers as benign lesions in our analysis. It is important to note that herein the word "case" refers to a physical lesion, not a patient. An HDI5000 scanner with an HDI L12-5 scan head (Philips Medical Systems, Bothell, Wash) was used for image acquisition.

Implication for Patient Care

- Computerized lesion characterization with breast US in a population representative of clinical practice is accurate and may aid in the prospective diagnosis of breast malignancy.

CAD Lesion Characterization

The CAD methodology for breast US images used in this work has been described extensively elsewhere (1,2,5,6) and will only be briefly summarized here. All imaged lesions were outlined by an experienced breast radiologist (with more than 10 years of experience and who was certified in accordance with the Mammography Quality Stan-

Published online before print
10.1148/radiol.2482071778

Radiology 2008; 248:392–397

Abbreviations:
AUC = area under the ROC curve
CAD = computer-aided diagnosis
ROC = receiver operating characteristic

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

Funding:
This work was supported by the National Institutes of Health (grants R01-CA89452, R21-113800, P50-CA125183).

dards Act). For each lesion, a seed point was calculated as the center of mass of its manually delineated outline that, in combination with the image data, formed the input to the computerized analysis. The manual lesion outlining was performed in several sessions for another ongoing study and took about 25 hours (508 patients with 2409 images depicting at least one US lesion and 2966 corresponding outlines). It is important to note that, in this work, the radiologist-drawn lesion outlines were only used to determine the center points for input to the computerized analysis and that, hence, our analysis did not depend on details of the manually drawn contours. The first step in the analysis was a preprocessing stage, which consisted of gray-scale inversion and median filtering (filter size, 1.75 mm^2). Each lesion was subsequently segmented automatically (Fig 2) by using contour optimization based on the average radial derivative (1). Four image features, that is, mathematical lesion descriptors, were extracted for each computer-determined lesion contour: the depth-to-width ratio, radial gradient index (9), posterior acoustic signature, and autocorrelation texture feature (2). Each physical lesion was imaged in at least two views, and the feature values were averaged over all applicable views. These lesion-averaged features formed the input to a Bayesian neural network classifier with five hidden units that we used for the task of cancer classification, with the Bayesian neural network output being the computer-estimated probability of malignancy. Note that the size of the preprocessing median filter was the only parameter that differed with respect to values reported in previously published work in which older scanner models were used (2). It was adjusted empirically by using the image data of a random subset of 200 patients.

CAD Lesion Characterization Performance according to Benign Subtype

We used the computer-estimated probabilities of malignancy obtained in the round-robin-by-case analysis with all lesions and divided them into subgroups to assess the degree of difficulty each lesion type presented to the computerized analysis in the distinction of these different lesion types from cancer. Note that in this instance, a single round-robin analysis (by case) was performed and that the estimated probability of malignancy was calculated only once for each lesion. Area under the receiver operating characteristic (ROC) curve (AUC) values for the task of differentiating cancer from benign lesion subtypes were obtained by regrouping the classifier output data.

Statistical Analysis

Classifier training and testing was performed within a round-robin (leave-one-case-out) protocol, and classification performance was assessed by using ROC analysis (10–12). The ROC curves were calculated as parametric curves modeling the data (12), that is, by modeling the computer-estimated probabilities of malignancy for cancerous and noncancerous lesions. The figures of merit used to assess the computerized classification performance were the AUC and selected operating points on the ROC curve in terms of specificity and sensitivity. For the human clinical performance, only a single operating point in terms of specificity and sensitivity could be calculated because numeric estimated probabilities of malignancy were not part of the patient reports used herein. Sensitivity and specificity reported for the computerized lesion characterization were obtained from the modeled ROC curves. Sensitivity and specificity for clinical lesion work-up were obtained from the choices of clinical lesion work-up (biopsy vs follow-up) and their outcome. Hence, percentages for sensitivity and specificity are reported as "numerator/denominator" only for the human lesion assessment.

Two round-robin analyses were performed: one in which all lesions were included irrespective of biopsy status

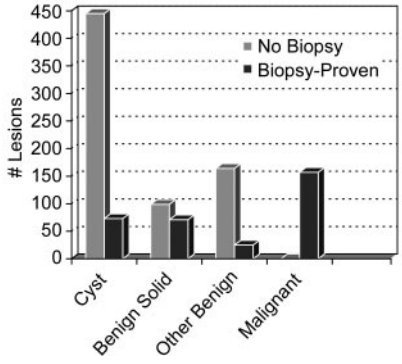
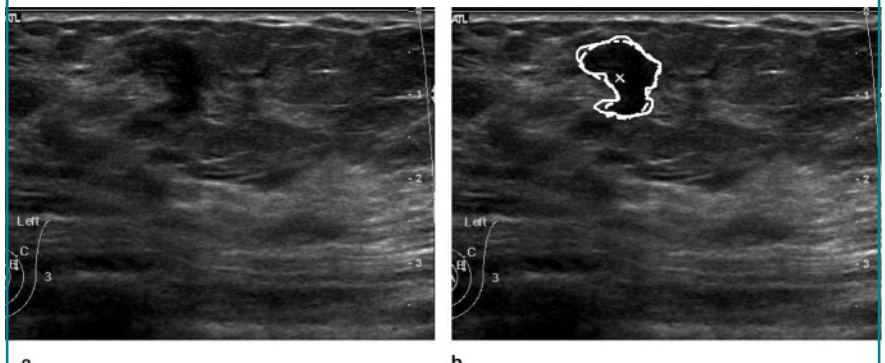


Figure 1: Bar chart shows the spectrum of lesions in the study. *Other Benign* includes all benign breast abnormalities not included under *Cystic* and *Benign Solid*, such as breast abscesses and lymph nodes.



a. b.

Figure 2: **(a)** US scan of a malignant breast lesion. **(b)** Same image with annotation. The solid line delineates the lesion as defined by the radiologist, and the dashed line delineates the lesion as segmented by the computer. *X* = the seed point derived from the radiologist outline as a starting point for computerized lesion segmentation.

and one in which only lesions of biopsy-proved abnormalities were included. The former analysis was also used to assess the degree of difficulty each lesion type presented to the computerized analysis in the task of differentiating them from cancer. For this purpose, the obtained probabilities of malignancy for different lesion types were divided into pathologic subgroups after completion of the round-robin analysis but before the performance of the ROC analysis. The obtained AUC values served as an indication of the level of difficulty different lesion types posed to the computerized analysis irrespective of human perception of their classification difficulty. The statistical significance of the perceived differences in CAD performance in terms of AUC value—for the task of differentiating benign lesion subtypes from breast cancers—was assessed by calculating the *P* value and 95% confidence interval for each difference in AUC (13). A *P* value of less than $\alpha = .05$ was considered to indicate a statistically significant difference for a single test. The sequential Bonferroni-Holm method (14) was used to assess statistical significance for multiple tests on the basis of the same data at an α level of .05.

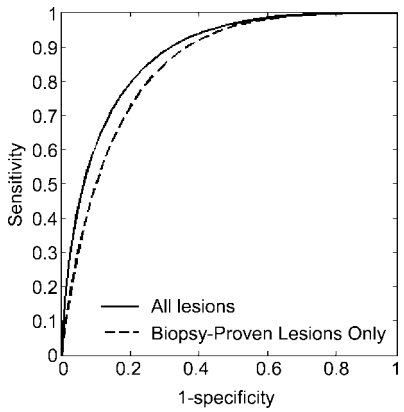


Figure 3: Graph shows the classification performance for the task of differentiating cancer from all benign lesions. The ROC curves were obtained from two separate round-robin (by-case) analyses, one in which all lesions were included (independent of biopsy status) and one that included only those lesions in which the pathologic condition was proved with biopsy. The AUC values were 0.90 and 0.88, respectively. We failed to find a statistically significant difference between these two AUC values (95% confidence interval: −0.02, 0.06).

Results

CAD Lesion Characterization Performance

The computerized classification of lesions with biopsy-proved abnormalities only (157 cancers and 362 lesions total) obtained an AUC value of 0.88 (Fig 3). One hundred percent sensitivity was achieved at 26% specificity. Note that because these cases all had gone to biopsy, the specificity for this set for the attending radiologists was zero at 100% sensitivity. When all lesions were included in a round-robin analysis (157 cancers and 1046 lesions total), the classification performance improved slightly at an AUC value of 0.90, achieving 100% sensitivity at 30% specificity. The specificity of the attending radiologists at 100% sensitivity was 77% (1 − false positive fraction = 1 − [362 − 157]/[1046 − 157]) for this set. With regard to the performance of the CAD scheme, we failed to find a statistically significant difference between the performance for lesions of biopsy-proved abnormalities only and that for all lesions (*P* = .09; 95% confidence interval: −0.02, 0.06 for the difference in AUC [13]).

Benign subtype	Solid	Solid	Cystic	Cystic	Other	Other	AUC	Standard Error
Biopsy proven	Yes	No	Yes	No	Yes	No		
Included in ROC	✓						0.84	0.025
		✓					0.84	0.028
			✓				0.91	0.022
				✓			0.96	0.011
					✓		0.83	0.048
						✓	0.78	0.031
	✓	✓					0.84	0.022
			✓	✓			0.95	0.011
					✓	✓	0.80	0.023
	✓		✓		✓		0.88	0.022
		✓		✓		✓	0.91	0.013
	✓	✓	✓	✓	✓	✓	0.90	0.014

Figure 4: Chart demonstrates CAD performance measured in terms of AUC and its standard error for the task of distinguishing cancer from the listed benign subtypes.

CAD Lesion Characterization Performance according to Benign Subtype

We observed a wide range in AUC values for the distinction of different subgroups from cancerous lesions (Fig 4). The cystic lesions presented the least difficulty to the CAD scheme, with an AUC value of 0.95, and the other benign-type lesions were the most difficult to analyze, with an AUC value of 0.80. The latter group of abnormalities included those with a wide range of imaging characteristics, such as lymph nodes, hematomas, scars, and abscesses. The AUC value for the classification performance for lesions in which the pathologic condition was not proved with biopsy was higher than that for lesions in which the pathologic condition was proved with biopsy (0.91 vs 0.88) due to the large number of cystic lesions in the former. Note that in these AUC calculations, the cancers were the same for each subgroup and, thus, the subgroups were (partially) correlated. By using the AUC values as indicators, lesions of different benign subtypes posed varying degrees of difficulty to our CAD scheme in the differentiation from cancer (Fig 4). Cystic lesions

Benign Subtype 1		Benign Subtype 2		*P* Value	Level Required	Statistically Significant
Biopsy proven	Solid	Nonbiopsy	Solid	.9987	0.0500	–
Biopsy proven	Cystic	Nonbiopsy	Cystic	.0478	0.0125	–
Biopsy proven	Other	Nonbiopsy	Other	.4761	0.0250	–
Biopsy proven	Solid + cystic + other	Nonbiopsy	Solid + cystic + other	.0436	0.0100	–
All*	Solid	All*	Cystic	.0000	0.0083	Yes
All*	Solid	All*	Other	.1813	0.0167	–

Figure 5: Chart demonstrates *P* values for the differences in AUC for distinguishing the listed benign subtypes from cancer. *All** = both biopsy-proven lesions and those without biopsy.

proved to be more easily distinguishable from cancers than solid lesions (Fig 5). It is worth noting that some *P* values (.0436 vs .0478) would have indicated statistical significance if the hypothesis was stated differently, that is, outside of a multiple comparison setting (14).

Discussion

In work using images acquired with an older scanner type (Philips ATL 3000), an AUC value of 0.87 was obtained in the classification of automatically segmented lesions of biopsy-proved abnormalities (2). Our results indicate that computerized diagnosis methods can be recalibrated as the image acquisition systems improve. The main difference in the CAD scheme with respect to other work was in the image preprocessing stage. Here, the size of the median filter was adjusted to a value smaller than that used in previously published work (2) because of the improved image quality of the newer-generation US scanner used here. Others (15) obtained an AUC of 0.95 in a cross-validation scheme for the task of cancer classification on images obtained with an ATL 3000 scanner. It is important to note, however, that the lesion segmentation in that work was based on selection by the user of a rectangular region of interest only slightly larger than the imaged lesion, which is known to result in a higher classification performance (2) than for automatically segmented lesions based only on a user-identified seed point (1), as used in this work.

The computer performance was largely unaffected by the inclusion of large numbers of lesions that did not undergo biopsy in the analysis, achieving overall good lesion characterization performance at an AUC value of 0.90. Although the performance of the computerized lesion characterization for the lesions with biopsy-proved abnormalities outperformed that of the radiologists in terms of specificity at 100% sensitivity (26% vs 0%), the radiologists outperformed the computer analysis when the additional nonbiopsied lesions were included. In that instance, the specificity at 100% sensitivity was 30% for the computer analysis and 77% for the radiologists, respectively. When actually deciding whether to recommend a biopsy, radiologists were at a great advantage over our current CAD implementation because the radiologists had access to all clinically available patient information (eg, previous mammograms and patient history), whereas the computer analysis was purely based on the appearance of the lesions at the current US examination. Our analysis also included a fairly large number of "other" lesions—lesions other than typical cystic, benign solid, and malignant ones—such as lymph nodes. Although our CAD system was initially not specifically designed to analyze these types of findings, it was retrained (in the round-robin analyses) to characterize the wide range of lesion types seen in clinical practice. The heterogeneous nature of the lesion population likely complicated the computer analysis, as illustrated in part by the lower classification performance for "other" lesions, although we failed to find the difference in performance for solid lesions and "other" ones statistically significant.

The inclusion of nonbiopsied lesions in the analysis presented herein was important to more fully assess potential CAD performance in clinical practice. However, the consequence was that the "truth" was based on radiologists' opinion regarding the probability of malignancy rather than disease for many lesions. Even though patients were followed up for an average of 3 years after the US examination, false-negative lesions—that is, cancers misdiagnosed by the attending radiologists—could not be entirely excluded. Another assumption in this work (which is commonly made) was the validity of normality hypotheses underlying parametrized ROC analysis (10–12). More extensive performance analysis with bootstrapping (16,17) could potentially yield additional insights.

In our study, the only interaction between humans and the CAD scheme was the determination of lesion seed points on the basis of lesion outlines drawn by a radiologist on full-film images. That is, the algorithm used as input an image with a seed point per visible lesion, and the entire analysis, including lesion segmentation, was automatic. The stand-alone lesion classification performance of a CAD scheme does not necessarily predict its influence on the performance of radiologists when using the CAD scheme in clinical practice. Moreover, the failure to find a statistically significant difference in CAD performance between the classification of lesions of biopsy-proved abnormalities only and the classification of all lesions irrespective of biopsy status does not necessarily

imply equivalent performance from a clinical perspective (18).

It should be noted that our study was limited by not performing an a priori power analysis. This limitation does not invalidate our results since we found no evidence to imply a significant difference in ROC performance for biopsy-proved versus all lesions. We believe that the current levels of computer performance for breast US are promising and warrant further testing.

References

1. Horsch K, Giger ML, Venta LA, Vyborny CJ. Automatic segmentation of breast lesions on ultrasound. Med Phys 2001;28:1652–1659.
2. Horsch K, Giger ML, Venta LA, Vyborny CJ. Computerized diagnosis of breast lesions on ultrasound. Med Phys 2002;29:157–164.
3. Drukker K, Giger ML, Horsch K, Kupinski MA, Vyborny CJ, Mendelson EB. Computerized lesion detection on breast ultrasound. Med Phys 2002;29:1438–1446.
4. Drukker K, Giger ML, Mendelson EB. Computerized analysis of shadowing on breast ultrasound for improved lesion detection. Med Phys 2003;30:1833–1842.
5. Drukker K, Giger ML, Metz CE. Robustness of computerized lesion detection and classification scheme across different breast US platforms. Radiology 2005;237:834–840.
6. Drukker K, Giger ML, Vyborny CJ, Mendelson EB. Computerized detection and classification of cancer on breast ultrasound. Acad Radiol 2004; 11:526–535.
7. Drukker K, Horsch K, Giger ML. Multimodality computerized diagnosis of breast lesions using mammography and sonography. Acad Radiol 2005;12:970–979.
8. Drukker K, Gruszauskas N, Sennett CA, Giger ML. CADx in diagnostic breast ultrasound: study of a large sample of patients [abstr]. In: Radiological Society of North America Scientific Assembly and Annual Meeting Program. Oak Brook, Ill: Radiological Society of North America, 2006; 567.
9. Kupinski MA, Giger ML. Automated seeded lesion segmentation on digital mammograms. IEEE Trans Med Imaging 1998;17: 510–517.
10. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283–298.
11. Metz CE. Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. J Am Coll Radiol 2006;3:413–422.
12. Pan X, Metz CE. The "proper" binormal model: parametric receiver operating characteristic curve estimation with degenerate data. Acad Radiol 1997;4:380–389.
13. Metz CE, Herman BA, Roe CA. Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. Med Decis Making 1998;18:110–121.
14. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6: 65–70.
15. Joo S, Moon WK, Kim HC. Computer-aided diagnosis of solid breast nodules on ultrasound with digital image processing and artificial neural network. Conf Proc IEEE Eng Med Biol Soc 2004;2:1397–1400.
16. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis. Acad Radiol 2000;7:341–349.
17. Mossman D. Resampling techniques in the analysis of non-binormal ROC data. Med Decis Making 1995;15:358–366.
18. Obuchowski NA. Testing for equivalence of diagnostic tests. AJR Am J Roentgenol 1997; 168:13–17.