AD_____

Award Number: DAMD17-96-1-6145

TITLE: A Pilot Study to Explore Linkages Among Isomers of Organochlorines, Promutagenic DNA Lesions and Breast Cancer Using Sensitive Techniques

PRINCIPAL INVESTIGATOR: Joseph Lo, Ph.D.

CONTRACTING ORGANIZATION: Duke University Medical Center
Durham, North Carolina 27710

REPORT DATE: January 2000

TYPE OF REPORT: Final

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

REPORT DOCUMENTATION PAGE

OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 2000	Final (1 Jul 96 - 31 Dec 99)

4. TITLE AND SUBTITLE
A Pilot Study to Explore Linkages Among Isomers of Organochlorines, Promutagenic DNA Lesions and Breast Cancer Using Sensitive Techniques

5. FUNDING NUMBERS
DAMD17-96-1-6145

6. AUTHOR(S)
Joseph Lo, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Duke University Medical Center
Durham, North Carolina 27710

E-MAIL:
joseph.lo@duke.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The purpose of this grant was to construct an artificial neural network (ANN) to assist radiologists in differentiating benign from malignant solid lesions in ultrasound (US) breast imaging. A data set of patient cases was collected, consisting of 192 biopsy-proven breast lesions for which radiologists provided descriptive terms to characterize the US appearance of the lesions. An ANN model was developed to predict probably benign lesions based upon those descriptors and the patient age. The model was potentially able to maintain 100% sensitivity of cancer detection, while improving the radiologists' specificity from 0% to 35% (42 out of 121 benign biopsies obviated). This corresponded to improving the PPV of the radiologists from 37% to 47%. Moreover, we also identified that the mass margin and patient age were the two most important input features for this model, and that highly simplified models based on those two features alone could still perform as well as the more complicated models using all available information. Predictive models such as these can provide physicians and patients with accurate information for managing suspicious breast lesions without the invasiveness of biopsy procedures.

14. SUBJECT TERMS
ultrasound, breast imaging, computer-aided diagnosis

15. NUMBER OF PAGES
25

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

SIL X Where copyrighted material is quoted, permission has been obtained to use such material.

___ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

___ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

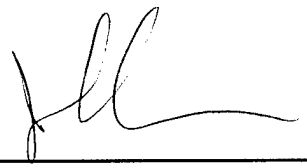
N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

SIL X For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.



7/20/2000

PI - Signature Date

Table of Contents

5. Introduction

Currently the accepted role of ultrasound (US) in diagnostic breast imaging is the differentiation of simple cysts from solid breast masses [3]. Stavros et al suggested it is possible to differentiate benign vs. malignant masses based upon US findings [5]. During the course of this project, newer studies provided further supporting evidence [10, 11], but there is still no widely accepted diagnostic criteria or system. The purpose of this study was address this need by developing an artificial neural network (ANN) model to assist radiologists in differentiating between benign vs. malignant masses based upon US findings. In particular, the goal is to be able to identify probably benign breast masses, for which follow-up may be recommended in lieu of biopsy, thus reducing the cost and trauma associated with unnecessary biopsies of benign lesions.

6. Body

Revised Statement of Work

This is the third and final report for this project, which was originally a two-year project scheduled for completion by Jan 31, 1999. During the second year, the USAMRMC approved a change in PI as well as a no-cost extension into a third year (2/1/1999 to 1/31/2000) to accomplish the following specific aims:

A. Resume collection of retrospective cases. We will attempt to double the current database of approximately 100 patient cases to 200 overall. For each patient, we will record ultrasound (US) and mammography findings and patient history data.

B. Given the larger database of patient cases, optimize the performance of an artificial neural network (ANN) to predict malignancy among breast masses. The ability of the ANN to generalize from training cases will be evaluated using retrospective data sampling rather than prospective clinical evaluation.

C. Evaluate the contribution of different input features in order to develop a simplified ANN that maintains diagnostic performance while requiring fewer features.

D. Evaluate the usefulness of the ANN in improving observer variability in US examination of breast masses. Specifically, compare the consistency and accuracy of the radiologists' assessments with that of the predictions of the ANN using the radiologists' findings as inputs.

The accomplishments of the entire effort will be summarized based upon these aims, since they extend or supercede all of the original aims. (The progress report for year two describes how the original aims were converted into the above new aims.)

Overview of Progress for Each Aim

Task A. Resume collection of retrospective cases. We will attempt to double the current database of approximately 100 patient cases to 200 overall. For each patient, we will record ultrasound (US) and mammography findings and patient history data.

The bulk of the effort in this project has been to collect the data set of US findings and biopsy outcomes which were used to train and test the ANN models. This has been a very time-consuming process. The collection of mammography findings data at this institution culminated from several different projects spanning approximately seven years [1, 2, 8]. We now record mammography findings prospectively as part of the standard operating procedure for all breast biopsy cases. The collection of US findings only began with the current project, however, and the procedure continues to evolve.

The original PI, Dr. Jay Baker, collected 65 cases during the first year and 35 cases during the first half of the second year prior to his departure. In the past (third) year, the new PI, Dr. Joseph Lo, supervised the collection of an additional 92 new cases. Each case consisted of a woman who had an abnormal US examination and then underwent underwent biopsy to yield definitive histopathologic diagnosis. The 7 US findings and biopsy outcome were recorded retrospectively. All studies were performed in accordance with standard clinical indications, with adequate safeguards for patient anonymity. The research conducted had no effect on the management of the patients.

The 192 cases consisted of 121 benigns and 71 malignancies, corresponding to a positive predictive value (PPV) of 37%. The women had an age range from 18 to 82, with mean age of 50.2 years. The increase in patient yield in the last year was possible due to both the increase in number of US-guided biopsies performed at this institution, as well as considerable non-salaried support from several personnel (notably John Zhang, medical student, and Dr. Patricia Walsh, attending radiologist) to augment the efforts of the PI and Dr. Mary Scott Soo. With the total of 192 cases, we reached our goal of collecting approximately 200 cases for model development. We are now in the process of identifying weaknesses in the current data collection schemes. The eventual goal will be to collect prospectively US findings for all cases which undergo biopsy at this institution, in a similar manner as the mammography data collection procedure.

Task B. Given the larger database of patient cases, optimize the performance of an artificial neural network (ANN) to predict malignancy among breast masses. The ability of the ANN to generalize from training cases will be evaluated using retrospective data sampling rather than prospective clinical evaluation.

ANN models were developed using just the seven US findings to predict if the mass described was benign vs. malignant. New models were designed at several different points during the course of this project using all data available at the time. In year two, results from the first 65 patients were presented at the First International Workshop on Computer-Aided Diagnosis sponsored by the University of Chicago department of radiology in Chicago, IL [6], see Appendix A. In year three, final results with all 192 cases were presented at the annual meeting of the Radiological Society of

North America, RSNA 1999 [9], see Appendeix B. The latter presentation was well received, resulting in unsolicited write-ups in WebMD (Nov. 29, 1999, http://my.webmd.com/content/article/1728.52643), the RSNA Daily Bulletin (Nov. 30, 1999) and Physician's Weekly (Feb. 21, 2000).

This final model based upon the seven US findings and patient age over 192 cases resulted in receiver operating characteristic (ROC) A_z of 0.92 ± 0.02, which was somewhat lower than reported before with the smaller data set, but nevertheless indicative of good performance. The continuous output values of the model could be thresholded to achieve a desired tradeoff between sensitivity on the one hand and specificity and positive predictive value (PPV) on the other. At 100% sensitivity, the model performed with 35% specificity and 47% PPV. In terms of the cases in this data set, it would have correctly referred all 71 actual cancers to biopsy, while obviating 42 out of 121 benign biopsies. At 96% sensitivity, the model performed with 63% specificity and 60% PPV. At the cost of delaying the diagnosis for only 3 of the 71 cancers, it could have obviated the majority (76 out of 121) of benign biopsies.

Task C. Evaluate the contribution of different input features in order to develop a simplified ANN that maintains diagnostic performance while requiring fewer features.

Using a technique we previously established [4], a simplified ANN was developed which utilized an optimized subset of the input findings while maintaining diagnosistic performance. This involved a two step process. First, the input features were rank ordered in order to determine which contributed more to the overall prediction. Separate ANN models were developed, each excluding one of the input features, and their performances were compared. The hypothesis was that the exclusion of a more important feature would reduce performance more, as measured by the ROC area index (A_z) and the partial area index A_z for sensitivity ≥ 0.90 (partial A_z). The results are summarized below in Table 1.

Table 1. Effect of excluding individual findings.

Finding	A_z	partial A_z
US mass shape	0.91 ± 0.02	0.60 ± 0.08
US mass margin	0.88 ± 0.02	0.55 ± 0.08
acoustic transmission	0.92 ± 0.02	0.64 ± 0.08
mass echogenicity	0.92 ± 0.02	0.62 ± 0.08
echotexture	0.92 ± 0.02	0.59 ± 0.08
thin, echogenic pseudocapsule	0.92 ± 0.02	0.60 ± 0.08
calcifications within nodule on US	0.92 ± 0.02	0.62 ± 0.08
patient age	0.90 ± 0.02	0.57 ± 0.08
ALL FINDINGS INCLUDED	**0.92 ± 0.02**	**0.62 ± 0.08**

The above results indicated that the model's performance depended only upon very few findings. In particular, performance was only noticeably reduced with the exclusion of the mass margin, patient age, and mass shape.

The second step of the process was to reduce the number of inputs. Using the rank ordered findings from Table 1, the ANN was simplified by eliminating the findings one at a time, starting from those which contributed least. This process is illustrated in Figure 1. Starting from left to right, the number of findings was successively reduced by one at a time. The performance was surprisingly not affected even when the ANN was reduced down to only its two most important input findings (mass margin and patient age). The performance of this drastically simplified two-input ANN was not statistically significantly different from that of the full eight-input model ($p=0.9$ for A_z, 0.7 for partial A_z). Only when the model was reduced to a single-input perceptron using only the mass margin feature did performance drop, although even in this extreme case, the difference was significant only for A_z ($p<0.001$) but not partial A_z ($p=0.14$). It should be noted that although the partial A_z is the more clinically relevant measure of performance, over these cases the standard deviations for partial A_z were quite large (0.08 for all trials in Table 1).

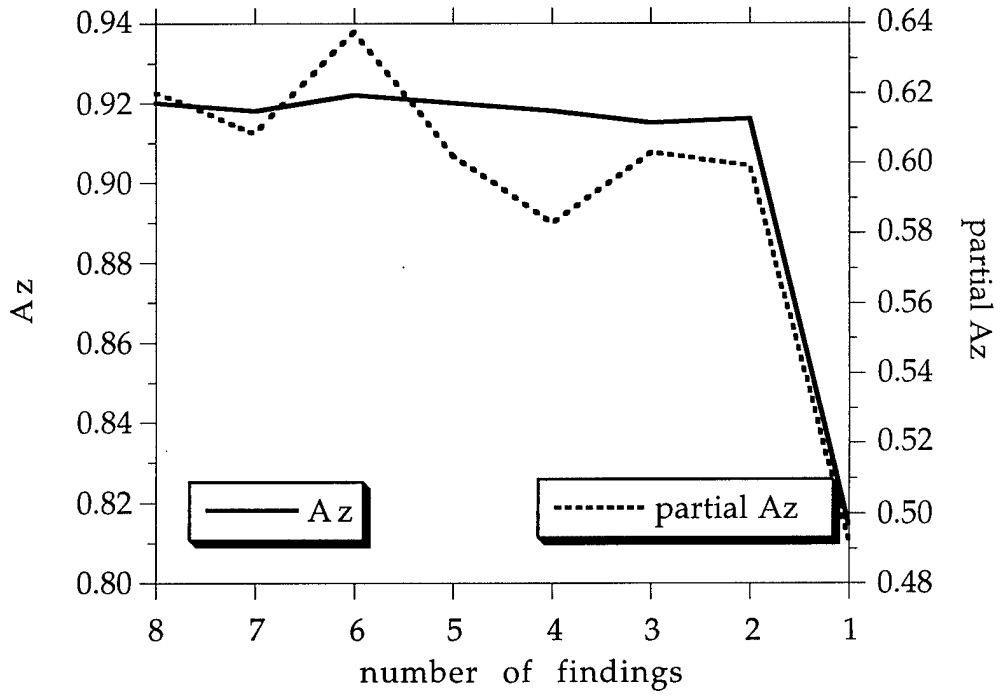


Figure 1. Effect of eliminating less important findings

These results were intriguing, especially since they were similar to those from a similar study for ANN predictors of probably benign lesions using mammographic findings, which identified the mammographic mass margin and patient age as the two most important findings as well [4]. As was the case with that previous study, we anticipate that although the exact performance values may not generalize to a larger data set, but that the general trends will hold true.

Task D. Evaluate the usefulness of the ANN in improving observer variability in US examination of breast masses. Specifically, compare the consistency and accuracy of the radiologists' assessments with that of the predictions of the ANN using the radiologists' findings as inputs.

As reported in the first and second annual reports, we assessed the usefulness of the ANN in reducing observer variability. In brief, 60 cases were read independently by 5 radiologists, and the consistency of their US findings as well as diagnostic assessment of likelihood of malignancy were measured. It was found that "considerable" interobserver variability existed for choosing terms for describing US findings (kappa=0.09 to 0.80) as well as assessing the likelihood of malignancy (kappa=0.51). This work was published in *AJR. American Journal of Roentgenology* after peer review [7], see Appendix C.

The BI-RADS (Breast Imaging Reporting and Data System, American College of Radiology) lexicon employs a five point rating scale for the radiologist's assessment, with four recommendations for cases with findings. We initially used this same four point rating scale. Unfortunately, the rating of one was never selected for any of these 192 cases (consistent with the retrospective knowledge that all of these cases did go to biopsy originally), and it was not possible to perform ROC analysis on the remaining three-category assessments. More importantly, there is a growing consensus that it is incorrect to use the radiologist's clinical recommendations (whether to follow-up vs. biopsy) as an assessment of the likelihood of malignancy.

For the above reasons, we compared the performance of the ANN model against the radiologists only at their actual clinical operating point, namely the fact that they did originally recommend biopsy for all of these 192 cases. By definition their sensitivity was 100% and their specificity 0% over these cases. Their PPV would correspond to that of the data set, which was 37% as noted previously. In comparison, the ANN model had the potential to maintain 100% sensitivity, while improving specificity to 35% and PPV to 47%. Also as noted previously, with small tradeoffs in sensitivity to 96%, the PPV could be improved to 60%. There may be an optimistic bias due to the relatively small number of cases in this data set, but it is evident that the model has the potential to improve the performance of the radiologists.

7. Key Research Accomplishments

This research resulted in the following major accomplishments:

(a) A data set of ultrasound (US) findings interpreted by expert radiologists was collected for 192 biopsy-proven breast masses from this institution.

(b) Using US findings and the patient age as inputs, artificial neural network (ANN) models were developed to predict whether the mass described was benign vs. malignant. This was the first successful model for this purpose, and represented several important extensions over the current practice of breast US imaging.

(c) The model had the potential to improve upon the diagnostic accuracy of the radiologists who extracted the US findings in the first place. While maintaining 100% sensitivity for cancers, the model could have obviated 35% (42 out of 121) of the benign biopsies, improving the radiologists' PPV from 37% to 47%.

(d) The model was simplified dramatically to reveal the important diagnostic contribution of just two findings, the US mass margin and patient age, over these cases. A performance of a new model based upon only those two findings was not statisticially significantly different from that of the more complicated models described above.

8. Reportable Outcomes

Publications:

The following publications resulted directly from this work. They consist of a conference proceeding published in hardcopy book form, a peer reviewed journal article, and a conference abstract published as a supplement to a journal, respectively. Copies are attached as appendices A–C.

A. Lo JY, and Floyd CE, Jr, "Computer-aided diagnosis of breast cancer," Doi K et al., Ed., First International Workshop on Computer-Aided Diagnosis, Elsevier Science, Univ. of Chicago, Chicago, IL, 1182 (ICS 1182): 221-5 (1998).

B. Lo JY, and Floyd CE, Jr, "Predicting malignancy of breast masses with ultrasound findings," *Radiology* 213(P), 198 (1999).

C. Baker J, Kornguth P, Soo M, Walsh R, and Mengoni P, "Sonography of solid breast lesions: observer variability of lesion description and assessment," *AJR Am J Roentgenol* 172, 1621-5 (1999).

Personnel Receiving Salary:

1. Joseph Y. Lo, PhD, PI (current)
2. Jay A. Baker, MD, PI (original)
3. Mary Scott Soo, MD, clinical co-investigator
4. Phyllis J. Kornguth, MD, PhD, clinical co-investigator
5. Carey E. Floyd, Jr., PhD, scientific co-investigator

Funding:

The following applications for funding resulted directly from this research.

1. Translational Medicine Awards, Duke University School of Medicine. "Clinical implementation of artificial neural networks for breast cancer diagnosis," PI Baker J, co-PIs Lo JY and Floyd CE Jr., status: rejected.
2. Idea Award, US Army Breast Cancer Research Program, "Computer-aided diagnosis of breast masses: Combining ultrasound and mammographic findings to improve accuracy of breast cancer diagnosis," PI Lo JY, co-investigators Baker JA, Floyd CE Jr, et. al., total costs $449,923, 3/1/01–2/28/04, status: pending.

9. Conclusions

This research resulted in several major advancements in the fields of breast imaging and computer-aided diagnosis. At present, breast ultrasound imaging is used only to differentiate between cysts and solid masses. Although some criteria were recently suggested for distinguishing benign from malignant masses, there is still no consensus, and the criteria involve simple rules based upon individual findings.

In this project, we developed an artificial neural network model which was able to predict benign versus malignant breast masses based upon ultrasound findings extracted by radiologists and the patient age. Unlike the aforementioned relatively simple diagnostic criteria, this model provided quantitative predictions for all cases by taking into consideration nonlinear interactions between all available findings. This was the first such comprehensive, quantitative model. Given 192 cases of suspicious masses which underwent biopsy, this model had the potential to maintain the sensitivity of cancer detection at 100%, while improving the radiologists' specificity from 0% to 35% (42 out of 121 benign biopsies obviated). This corresponded to improving the PPV of the radiologists from 37% to 47%. Moreover, we also identified that the mass margin and patient age were the two most important input features for this model, and that highly simplified models based on those two features alone could still perform as well as the more complicated models using all available information.

In future work, it would be interesting to see if the inclusion of mammography findings would improve the accuracy or robustness of the current models which are based on ultrasound findings and patient age alone. The success of these models also depends on how well they generalize to larger data sets from multiple institutions.

Predictive models such as these can provide physicians and patients with accurate information for managing suspicious breast lesions without the invasiveness of biopsy procedures. These models have the potential to obviate many unnecessary biopsies of benign lesions and their associated cost to society and trauma to patients.

10. References

1 Floyd CE, Jr, Lo JY, Yun AJ, Sullivan DC, and Kornguth PJ, "Prediction of breast cancer malignancy using an artificial neural network," *Cancer* 74, 2944-2948 (1994).

2 Baker JA, Kornguth PJ, Lo JY, Williford ME, and Floyd CE, Jr, "Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon," *Radiology* 196, 817-822 (1995).

3 Jackson VP, "Management of solid breast nodules: what is the role of sonography?," *Radiology* 196, 14-15 (1995).

4 Lo JY, Baker JA, Kornguth PJ, and Floyd CE, Jr, "Computer-aided diagnosis of breast cancer: artificial neural network approach for optimized merging of mammographic features," *Acad Radiol* 2, 841-850 (1995).

5 Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker S, and Sisney G, "Solid breast nodules: use of sonography to distinguish between benign and malignant lesions," *Radiology* 196, 123-134 (1995).

6 Lo JY, and Floyd CE, Jr, "Computer-aided diagnosis of breast cancer," Doi K et al., Ed., First International Workshop on Computer-Aided Diagnosis, Elsevier Science, Univ. of Chicago, Chicago, IL, 1182 (ICS 1182): 221-5 (1998).

7 Baker J, Kornguth P, Soo M, Walsh R, and Mengoni P, "Sonography of solid breast lesions: observer variability of lesion description and assessment," *AJR Am J Roentgenol* 172, 1621-5 (1999).

8 Lo JY, Baker JA, Kornguth PJ, and Floyd CE, Jr, "Effect of patient history data on the prediction of breast cancer from mammographic findings with artificial neural networks," *Acad Radiol* 6, 10-15 (1999).

9 Lo JY, and Floyd CE, Jr, "Predicting malignancy of breast masses with ultrasound findings," *Radiology* 213(P), 198 (1999).

10 Rahbar G, Sie AC, Hansen GC, Prince JF, Melany ML, Reynolds HE, Jackson VP, Sayre JW, and Bassett LW, "Benign versus malignant solid breast masses: US differentiation," *Radiology* 213, 889-894 (1999).

11 Zonderland HM, Coerkamp EG, Hermans J, van de Vijver MJ, and van Voorthuisen AE, "Diagnosis of breast cancer: contribution of US as an adjunct to mammography," *Radiology* 213, 413-422 (1999).

11. Appendices

A. Lo JY, and Floyd CE, Jr, "Computer-aided diagnosis of breast cancer," Doi K et al., Ed., First International Workshop on Computer-Aided Diagnosis, Elsevier Science, Univ. of Chicago, Chicago, IL, 1182 (ICS 1182): 221-5 (1998).

B. Lo JY, and Floyd CE, Jr, "Predicting malignancy of breast masses with ultrasound findings," *Radiology* 213(P), 198 (1999).

C. Baker J, Kornguth P, Soo M, Walsh R, and Mengoni P, "Sonography of solid breast lesions: observer variability of lesion description and assessment," *AJR Am J Roentgenol* 172, 1621-5 (1999).

Computer-Aided Diagnosis in Medical Imaging
K. Doi, H. MacMahon, M.L. Giger and K.R. Hoffmann, editors

Computer-aided diagnosis of breast cancer

Joseph Y. Lo, Ph.D. and Carey E. Floyd, Jr., Ph.D.

Digital Imaging Research Division,
Dept. of Radiology, Duke University Medical Center, and
Dept. of Biomedical Engineering, Duke University,
Box 3302 DUMC, Durham, NC 27710, USA.

ABSTRACT

We will review two current projects pertaining to artificial neural network (ANN) computer models that merge radiologist-extracted findings to perform computer-aided diagnosis (CAD) of breast cancer. The goal of both projects is to obviate the diagnostic excisional biopsy, a costly and invasive surgical procedure, for certain groups of cases. These projects are (1) mammography-based model to predict both malignancy and invasion of all breast lesions, and (2) ultrasound-based model to predict malignancy of breast masses. By providing information which was previously available only through biopsy, these CAD models may help to reduce the number of unnecessary surgical procedures and their associated costs.

Keywords: computer-aided diagnosis, breast cancer, artificial neural network

1. INTRODUCTION

Mammography and ultrasonography are currently the main breast imaging modalities for the early detection of breast cancer, but they suffer from several important limitations. Mammography is very sensitive but has a low positive predictive value (PPV), resulting in benign biopsy rates of 65% or higher [1, 2]. Preventing these unnecessary surgical procedures is one of the most important ways to improve the efficacy of mammographic screening. Furthermore, although the remaining cases are malignant by biopsy, up to 80% are invasive cancers which require a second, therapeutic surgical procedure such as axillary dissection [3]. If these patients may be identified a priori, they may undergo a combination single-stage surgery, thus also obviating the separate biopsy surgery [4, 5].

Ultrasound (US) is used primarily in breast imaging to distinguish simple cysts from solid masses. Due to US's low cost, widespread availability, and use of nonionizing radiation, it has tremendous potential in helping to

assess masses identified both by screening mammography and physical exam. A previous report suggested it is further possible to differentiate benign vs. malignant breast masses based upon grayscale US features [6]. There is as yet no established model, however, to combine multiple features for consistent, accurate prediction of breast cancer.

To address these concerns, we have developed artificial neural network (ANN) computer models that merge radiologist-extracted findings to perform computer-aided diagnosis (CAD) of breast cancer [7-9]. These ANNs can provide consistent, accurate, and robust predictions, using readily available medical information. We will review here two studies: (1) predicting breast lesion malignancy and invasion using mammographic and patient history findings, and (2) predicting breast mass malignancy using ultrasound findings.

These projects share in common the use of feedforward, error-backpropagation ANNs with one hidden layer. Inputs to the ANNs were quantitatively encoded medical findings, including mammographic descriptors of lesion morphology according to the Breast Imaging Reporting and Data System (BI-RADS) [10], ultrasound lesion descriptors according to Stavros et al [6], and patient history data. The output to each ANN was a number between zero and one corresponding to the biopsy outcome which was being predicted, such as benign vs. malignant or in situ vs. invasive cancer. Each ANN underwent supervised training and independent testing with actual patient data. Performance was evaluated by several clinically relevant metrics, including ROC area, specificity for a given near-perfect sensitivity, and/or positive predictive value (PPV).

2. MAMMOGRAPHY-BASED MODEL

2.1. Methods

We developed a cascaded, multi-stage system consisting of two ANNs to predict first malignancy and then invasion. The goal was to identify as many benign lesions and invasive cancers, respectively. Together these two categories comprise approximately 90% of all currently biopsied cases, yet as explained before many of these cases are candidates for obviating the diagnostic excisional biopsy surgery.

The data set consisted of 500 consecutive cases of mammographically suspect, nonpalpable lesions which underwent excisional biopsy and resulted in definitive histopathologic diagnoses. The first ANN predicted whether each of the 500 cases was benign vs. malignant. A threshold was set over the ANN output values such that almost all malignancies had outputs which were *above* the threshold and were thus correctly classified as true positives. Many benign cases *below* the threshold were also correctly classified as true negatives which may be spared the unnecessary biopsy. The goal was to achieve fair specificity at near-perfect sensitivity, i.e. to obviate as many

benign biopsies as possible while missing very few cancers, since the cost of the latter mistake far exceeds the former.

All cases above the threshold (consisting of almost all malignancies and some false-positive benign cases) were then referred to the second ANN, which predicted whether these cases were in situ vs. invasive cancer. The goal was to find a threshold such that the outputs for almost all cases which were not invasive cancers (benign lesions and in situ carcinomas) were *below* the threshold. Many invasive cancers would lie *above* the threshold and be correctly identified as true positives and thus candidates for the single-stage surgery, thus obviating the excisional biopsy. Unlike the previous stage, the goal here was to achieve fair sensitivity at near-perfect specificity, i.e. to identify as many invasive cancers as possible while avoiding almost all benign lesions and in situ carcinomas as candidates for single-stage surgery.

2.2. Results

The first-stage ANN was able to identify many probably benign cases. At an arbitrary threshold over the output values which corresponded to 98% sensitivity, the ANN performed with 41% specificity. In other words, it missed only 3 of 174 malignancies (false negatives), while correctly identifying 134 out of 326 benign biopsies (true negatives) which may have been obviated.

The remaining 363 cases above the threshold consisted of 120 invasive cancers and 243 other cases (benign lesions and in situ carcinomas). These 363 cases were referred to the second-stage ANN to identify as many probably invasive cancers as possible. Again at an arbitrary threshold corresponding to 90% specificity, the ANN performed with 54% sensitivity. In other words, it correctly ruled out 218 of the 243 other cases, while identifying 65 of 120 invasive cancers as candidates for single-stage surgery.

3. ULTRASOUND-BASED MODEL

3.1. Methods

For the US-based ANN, 175 consecutive patients at this institution had an abnormal US examination. Of those with a solid lesion; definitive histologic diagnosis was available for 65 who underwent needle core biopsy, fine needle aspiration, or open excisional biopsy, yielding 34 benign lesions and 31 malignancies. For each of these 65 lesions, a radiologist recorded 7 morphologic findings as previously suggested by Stavros, et al: mass shape, mass margin, presence of an echogenic pseudocapsule, presence of calcification within the lesion visible by US, acoustic transmission, lesion echogenecity, and lesion echotexture. The ANN was developed to merge the 7 US findings and patient age in order to predict whether each case was benign or malignant.

3.2. Results

For the task of distinguishing benign vs. malignant masses using US features and patient age, the ANN performed with ROC area of 0.96 ± 0.02, indicating nearly perfect performance. At an arbitrary threshold, the ANN provided a PPV of 81%, compared to the original radiologist's PPV of 48%. At that same threshold the ANN had a sensitivity of 97% (missing only 1 of 31 malignancies) and specificity of 79% (correctly sparing 27 of 34 benign lesions).

We have developed ANNs which have the ability to predict the outcome of breast biopsy at a level comparable or better than expert radiologists. For example, using only 10 BI-RADS mammographic findings and the patient age, the ANN predicted malignancy with ROC area of 0.86 ± 0.02, a specificity of 42% at a given sensitivity of 98%, and a 43% PPV.

4. CONCLUSION

We described here two separate studies indicating the potential of using ANNs for CAD of breast cancer. With the mammography-based model, we were able to predict first malignancy and then invasion over a relatively large data set of 500 indeterminate cases which previously all underwent biopsy. At each stage the ANN correctly classified approximately half of the target category while ruling out the vast majority of the other category. We identified 134 benign lesions and 65 invasive cancers, thus potentially obviating 199 or 40% of the biopsies.

Likewise the results from the US-based model were very encouraging and compared favorably with previous models based upon mammographic findings. The important distinction is that US is low cost, widely available, and uses nonionizing radiation. The ANN performed nearly perfectly, potentially sparing 79% of benign biopsies. This work was preliminary, however, due to the small number of patient cases and other factors. On-going studies will evaluate the ANN's performance with more cases and with the inclusion of other patient information, such as mammographic findings and history data.

The two studies shared in common the use of readily available input data such as radiologist-extracted image findings and patient history. The first study uses the standardized BI-RADS lexicon of mammographic descriptors, so the results reported herein should generalize to any other institution which has adopted this standard. The US lexicon proposed by Stavros et al is not considered a standard yet, but we believe it is a thorough and consistent scheme for codifying the US data.

With further development, these CAD models have the potential to provide important knowledge which may assist in surgical planning for patients with breast lesions. This may help reduce the number of unnecessary biopsies and the considerable cost associated with them.

REFERENCES

1. Kopans DB. The positive predictive value of mammography. *AJR Am J Roentgenol* **1992**; 158:521-526.
2. Knutzen AM, Gisvold JJ. Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions. *Mayo Clin Proc* **1993**; 68:454-460.
3. Ciatto S, Cataliotti L, Distante V. Nonpalpable lesions detected with mammography: review of 512 consecutive cases. *Radiology* **1987**; 165:99-102.
4. Jackman RJ, Nowels KW, Shepard MJ, Finkelstein SI, Marzoni FA, Jr. Stereotaxic large-core needle biopsy of 450 nonpalpable breast lesions with surgical correlation in lesions with cancer or atypical hyperplasia. *Radiology* **1994**; 193(1):91-5.
5. Liberman L, Dershaw DD, Rosen PP, Cohen MA, Hann LE, Abramson AF. Stereotaxic core biopsy of impalpable spiculated breast masses. *AJR Am J Roentgenol* **1995**; 165(3):551-4.
6. Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker S, Sisney G. Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. *Radiology* **1995**; 196:123-134.
7. Lo JY, Baker JA, Kornguth PJ, Iglehart JD, Floyd CE, Jr. Predicting breast cancer invasion with artificial neural networks on the basis of mammographic features. *Radiology* **1997**; 203(1):159-163.
8. Lo JY, Baker JA, Kornguth PJ, Floyd CE, Jr. Computer-aided diagnosis of breast cancer: artificial neural network approach for optimized merging of mammographic features. *Acad Radiol* **1995**; 2(10):841-850.
9. Baker JA, Kornguth PJ, Lo JY, Williford ME, Floyd CE, Jr. Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. *Radiology* **1995**; 196(3):817-822.
10. BIRADS. Breast Imaging - Reporting and Data System (BI-RADS). Reston, VA: American College of Radiology, **1993**.

analysis of bronchial diseases, become now available in an automatic or interactive way.
METHOD AND MATERIALS: The study uses volumetric helical CT data sets acquired with a pitch of 1.5 and a collimation varying between 1 mm and 3 mm, without any contrast agent. Axial CT scans were reconstructed at 0.6 mm intervals on a 512x512 pixel matrix. The 3D reconstruction of the bronchial tree is achieved by applying a 3D topology-based propagation of the segmented 2D bronchial lumen. The automatic 2D segmentation method relies on the mathematical morphology theory and involves a morphological marking exploiting the connection cost concept, together with a contour extraction by using a conditional watershed. Stacking the result of the 2D segmentation step provides a primary incomplete and artifacted 3D reconstruction. We then developed a specific 3D propagation procedure exploiting the oriented, multivalued and evolutive 3D graph describing the 3D topology of the stacked volume. The resulting reconstructed bronchial tree recovers the branch discontinuities while pointing out the airway pathologies (bronchial stenosis, mucoid impactions). Finally, the bronchial tree is visualized by using a semi-transparent volume rendering technique. All the above-mentioned functionalities are integrated within a user-friendly software package.
RESULTS: Tests performed on 10 patients with chronic airway diseases showed an accurate and robust 3D reconstruction up to 6-7th order divisions.The procedure proved to be stable with respect to bronchial stenosis, bronchiectasis and mucoid impaction.
CONCLUSIONS: Following this preliminary assessment stage, we are now conducting an extensive validation of this 3D CT bronchography package within a clinical routine application framework. This work was supported by a Grant from Ministry of Industry of France (CIFRE No 525/96) and Picker International.

430 • 3:06 PM

Computerized Detection of Pulmonary Embolism in Spiral CT Angiography: Segmentation and 3D Image Feature Analysis of Thrombi

Y. Masutani, PhD, Chicago, IL • K.R. Hoffmann, PhD • H. MacMahon, MD • K. Doi, PhD

PURPOSE: Spiral CT-Angiography (CTA) has been recently reported as a superior modality for diagnosis of pulmonary embolism (PE). However, radiologists must view more than 50 images per case and the manifestations of PE in small vessels can be difficult to detect. Our purpose is to develop a computerized scheme for automated detection of pulmonary embolism in CTA images as an aid to radiologists. In this study, we present new methods for segmentation and detection of thrombus candidates and for distinction of thrombi from false positives.
METHOD AND MATERIALS: We used clinical CTA data acquired with 3.0 mm collimation and a pitch of 1.7, reconstructed at 1.5 mm intervals. The data were interpolated in an axial direction to yield isotropic data for 3D analysis. Segmentation proceeds automatically and uses a combination of thresholding, morphological operations, connectivity analysis, and region-growing. Local diameters of pulmonary vessels were determined for feature analysis using morphological operations. The distances of the thrombus candidates from the vessel wall were examined. A 3D line enhancement filter was also employed for determination of a feature value which would be related to line-like structures of relatively large thrombi. In addition, the average CT value, contrast and volume of candidate regions were determined and analyzed.
RESULTS: Automated segmentation was successfully performed on several clinical cases with the adjustment of a few parameters. The segmented volumes of pulmonary vessels occupied about 3-4 % of the total data volume with thrombus candidates being less than 1 % of the segmented vessel volume. For thrombus candidates, false positives resulted mainly from artifacts due to partial-volume effects and breathing motion. The distance criteria were effective for elimination of false positives due to partial-volume artifact, whereas the shape feature by the line filter was useful for detection of line-like thrombi.
CONCLUSIONS: Pulmonary vessels, and thrombi were effectively segmented in spiral CTA data. Analysis of 3D image features shows promises for detection of thrombi.

431 • 3:15 PM

Quantitative in Vivo Analysis of the Kinematics of Carpal Bones Using a Deformable Surface Model and a 3D Matching Technique

J.G. Snel, MS, Amsterdam, Netherlands • H.W. Venema, PhD • C.A. Grimbergen, PhD • T.M. Mooijen, MD • M.J. Ritt, PhD • G.J. Den Heeten, MD, PhD

PURPOSE: To obtain quantitative information of the relative displacements and rotations of the carpal bones during movement of the wrist of both normal volunteers and of patients before and after operative intervention.
METHOD AND MATERIALS: Axial helical CT-scans were made with a CT-scanner with a double detector array (Elscint CT-Twin/Flash). The wrists were imaged in the neutral position with a conventional CT-technique, and in 10-15 other postures (volair and palmar flexion, radial and ulnar abduction) with a low dose technique. The imaging protocol was as follows: collimation: 2 ' 0.5 mm, scan time: 1 s (360°), pitch 0.5 (conventional) or 2.0 (low dose), 120 kV, 135 mAs (conventional) or 13 mAs (low dose). The ultra high resolution (UHR) mode of the CT-Twin was used. A segmentation of the carpal bones, radius and ulna was obtained by applying a deformable surface model (DSM) to the high dose scan. Next each bone of the high dose scan was registered with the corresponding bone in each low dose scan using a 3-D matching technique.
RESULTS: A very detailed definition of the surfaces of the carpal bones was obtained from the high dose scans. The low dose scans provided sufficient information to obtain an accurate match of each carpal bone with the corresponding carpal bone in the high dose scan. Accurate estimates of the relative positions and orientations of the carpal bones during flexion and deviation were obtained.
CONCLUSIONS: The movement of the carpal bones can be quantified accurately by matching a single high dose CT-scan with a number of low dose CT-scans. This quantification is especially useful when monitoring changes in kinematics before and after operative interventions, like mini-arthrodeses. This technique can also be applied in the quantification of the movement of other bones in the body (e.g. ankle and cortical spine).
(This presentation was supported in part by a grant from Elscint Ltd.)

Monday Afternoon • Room S402AB

■ Breast (Ultrasonography)

PRESIDING: **Ellen B. Mendelson, MD,** Pittsburgh, PA
Computer Code: G04 • 1½ hours
To receive credit, relinquish attendance voucher at end of session.

432 • 2:30 PM

Predicting Malignancy of Breast Masses with Ultrasound Findings

J.Y. Lo, PhD, Durham, NC • C.E. Floyd, Jr, PhD

PURPOSE: An artificial neural network (ANN) model was developed to use only ultrasound (US) findings to predict whether solid breast masses were benign vs. malignant.
METHOD AND MATERIALS: Among women who had an abnormal US examination at this institution, 102 cases of solid lesions which underwent biopsy to yield definitive histopathologic diagnosis were selected. For each of these 102 lesions (53 benign, 49 malignant), a radiologist was blinded to the biopsy outcome and recorded 7 morphologic findings as previously suggested by Stavros, et al.: mass shape, mass margin, presence of an echogenic pseudocapsule, presence of calcification within the lesion visible by US, acoustic transmission, lesion echogenecity, and lesion echotexture. A backpropagation ANN was developed to merge these US findings in order to predict whether each case was benign or malignant. Round robin data sampling was employed to ensure independence between training and testing cases.
RESULTS: The ANN model performed with a positive predictive value (PPV) of 55%, which was better than the 48% PPV of the original radiologists' decision to recommend biopsy. The ROC area index of the ANN was 0.92 ±0.03. Note that the ANN based its decision on the 7 US findings only, while the radiologists took into consideration all available information, including not only the US films but also mammograms, prior films, and patient history.
CONCLUSIONS: Using only US findings, the ANN model accurately predicted malignancy of breast masses, improving the PPV for the radiologists' biopsy recommendations. Since US is cheap, uses nonionizing radiation, and widely available, this ANN approach has considerable potential in helping to assess masses identified by screening mammography or physical exam.

433 • 2:39 PM

Tissue Harmonic Imaging Sonography of Breast Lesions: Improved Margin Analysis, Conspicuity, and Image Quality Compared to Standard Ultrasound

E.L. Rosen, MD, Durham, NC • M.S. Soo, MD

PURPOSE: To determine if tissue harmonic imaging (THI) afforded a qualitative advantage compared to conventional sonography in the evaluation of breast masses. MATERIALS AND METHODS: A prospective evaluation 103 image pairs (each consisting of two identical images, one obtained with conventional sonography, and the other with THI sonography) were obtained. Each image set was masked and then independently evaluated by two experienced breast imagers who determined whether the lesion was solid, cystic or indeterminate and then contrasted lesion conspicuity, margins, and overall quality between the two images. Statistical analysis was performed with the sign test (modified t-test). RESULTS:

Sonography of Solid Breast Lesions: Observer Variability of Lesion Description and Assessment

Jay A. Baker[1]
Phyllis J. Kornguth[2]
Mary Scott Soo[2]
Ruth Walsh[2]
Patricia Mengoni[3]

OBJECTIVE. The purpose of this study was to measure the level of inter- and intraobserver agreement and to evaluate the causes of variability in radiologists' descriptions and assessments of sonograms of solid breast masses.

MATERIALS AND METHODS. Sixty sonograms of solid masses were evaluated independently by five radiologists. Observers used the lexicon of a recently published benchmark report on sonographic appearances of breast masses to determine mass shape, margin, echogenicity, echotexture, presence of echogenic pseudocapsule, and acoustic transmission. Final diagnostic assessments were determined by applying the rule-based model of the same benchmark report to the radiologists' descriptions. In addition, one observer interpreted each case twice to evaluate intraobserver variability. Inter- and intraobserver variability were measured using Cohen's kappa statistic. We also investigated causes of variability in radiologists' descriptions.

RESULTS. Interobserver agreement ranged from lowest for determining the presence of an echogenic pseudocapsule ($\kappa = .09$) to highest for determining mass shape ($\kappa = .8$). Intraobserver agreement was lowest for mass echotexture ($\kappa = .24$) and greatest for mass shape ($\kappa = .79$). Variability in descriptions of lesions contributed to interobserver ($\kappa = .51$) and some intraobserver ($\kappa = .66$) inconsistency in assessing the likelihood of malignancy.

CONCLUSION. Lack of uniformity among observers' use of descriptive terms for solid breast masses resulted in inconsistent diagnoses. The need for improved definitions and additional illustrative examples could be addressed by developing a standardized lexicon similar to that of the *Breast Imaging Reporting and Data System.*

Sonographic imaging of the breast is a well-established adjunct to film-screen mammography. However, sonography has not been widely accepted in the United States for characterization of solid breast masses because numerous attempts to accurately classify and differentiate benign from malignant solid breast nodules have been unsuccessful [1–7].

In a recent benchmark study, Stavros et al. [8] described a classification model with a reported 99.5% negative predictive value and 98.4% sensitivity. The model is based on 20 specific sonographic features of breast masses, including morphologic descriptors of the shape, margin, and texture of a mass, and acoustic properties such as sonographic sound transmission and mass echogenicity.

Implicit in models for classifying solid breast masses is the assumption that morphologic and acoustic features of breast masses can be identified reliably and reproducibly from observer to observer. Substantial variability in identification of the specific sonographic features could yield varying conclusions and result in inconsistent treatment practices. Lack of consistency and reproducibility is a recognized focus of concern in breast imaging, and considerable inter- and intraobserver variability has been shown using film-screen mammography [9–14]. This study proposes to evaluate the inter- and intraobserver variability of radiologists' characterization of sonographic features of solid breast masses based on the imaging features defined by Stavros et al. [8].

Materials and Methods

Case Selection and Imaging

Sixty consecutive sonographic studies of solid breast lesions obtained between August and October 1997 were selected. To be included in this investigation, patients were required to be female with a solid breast mass visible on sonographic imaging. All of the lesions were identified on screening mammography, physical examination, or both. No masses inci-

Received September 21, 1998; accepted after revision December 10, 1998.

[1]Florida Radiology Associates P.A., P. O. Box 150505, Altamonte Springs, FL 32715. Address correspondence to J. A. Baker.

[2]Division of Breast Imaging, Department of Radiology, Box 3808, Duke University Medical Center, Durham, NC 27710.

[3]Northwestern Memorial Hospital, Lynn Sage Breast Center, 333 E. Superior St., Ste. 260, Chicago, IL 60611.

AJR 1999;172:1621–1625

0361–803X/99/1726–1621

dentally noted during sonography of other lesions were included.

Static sonographic images of each solid breast lesion were acquired and reviewed by five radiologists with experience in breast imaging. All images were obtained with high-resolution, state-of-the-art sonography equipment (Sonoline Elegra; Siemens, Issaquah, WA) using a variable-frequency linear transducer set at 9 MHz. In each case, at least four static images including radial and antiradial images with and without caliper measurements were acquired. The radial and antiradial planes are defined with the breast viewed as if it were a clock face with the nipple at the center. The radial plane is obtained by rotating the transducer around the clock face in the plane of a clock hand. The antiradial plane is perpendicular to the radial plane. Additional representative gray-scale images were available in almost all cases. Other sonographic images including Doppler, color Doppler, and power Doppler images were also available for review when obtained during the examination. Mammographic images and medical history were not provided for correlation to eliminate bias in description and assessment of the sonographic images.

Evaluation of Sonographic Images

The sonographic features chosen for investigation were those reported by Stavros et al. [8] in a study of 750 solid breast lesions. These features were chosen because of the reported accuracy of the classification scheme and the availability of definitions and representative images illustrating the lexicon used in that report. That study defined 20 morphologic and acoustic features for describing solid breast masses. For the purposes of this observer variability study, those features were grouped into seven broad categories: mass shape, mass margin, echogenic pseudocapsule, acoustic transmission, mass echogenicity, mass echotexture, and sonographic evidence of calcification.

Each of five radiologists independently evaluated all cases and selected the term from each category of the lexicon that best described each mass. All five carefully reviewed the report by Stavros et al. [8] before this study, and the definitions and example images depicted in that report were available to the radiologists at the time they evaluated the cases. Observers were limited to selecting a single term from each of six of the seven categories listed above. The presence of sonographically identifiable calcification within a mass was not evaluated. Because mammograms were not provided, observers could not correlate the appearance of any particular echogenic focus with the appearance of calcification on a radiograph.

Assessment of the likelihood of malignancy of the lesions was determined by applying the decision model proposed by Stavros et al. [8] to the descriptive terms chosen by the observers. Following the rules described in that model, each observer classified the lesions as either benign or malignant.

To assess intraobserver variability in evaluating breast sonography examinations, one of the five radiologists reevaluated all 60 cases 6 months after the initial exercise. The reevaluation consisted of selecting terms for describing the morphologic features of the lesions and assessing the likelihood of malignancy by applying those sonographic descriptions to the same rule-based decision model.

To determine the source of any variability in radiologists' descriptions or assessments, observers' comments regarding difficulty in selecting lesion descriptors were elicited. Each case was subsequently reviewed with these comments and with the statistical analysis of variability described later so that explanations for concordance or variability could be discerned.

Statistical Analysis

Inter- and intraobserver variability in choosing sonographic descriptors in each category was determined using Cohen's kappa statistic. Variability in observers' diagnostic assessments based on the criteria reported by Stavros et al. [8] was also calculated. Cohen's kappa measures the proportion of decisions in which observers agree while accounting for the possibility of agreements based on chance alone. Perfect agreement results in a kappa value of 1.0, and a kappa value of 0 indicates the level of agreement expected based on chance alone. Less agreement than that expected by chance results in a negative kappa value. Although no absolute scale exists, prior reports have suggested that kappa values of .2 or less indicate slight agreement, .21–.40 fair, .41–.60 moderate, .61–.80 substantial, and .81–1.00 indicates almost perfect agreement between observers [15]. This scale will be used throughout this study. Other researchers have advocated that kappa values of .5 or less be considered poor and values of .75 or more be considered excellent reproducibility [16].

Results

Interobserver Variability

The cases included in this study were typical of those routinely encountered. The distribution of lesion descriptors chosen by the five observers illustrates the range of appearance of the lesions included (Table 1). The relatively small number of cases described as "duct extension," "branch pattern," or "spiculation" is expected, because lesions obviously malignant on mammography did not require further sonographic imaging and were therefore not included in this study.

Statistical analysis of agreement among observers for choosing lesion descriptions showed that levels of agreement ranged from slight to substantial concordance (Table 2). The greatest reproducibility was found among observers determining the shape of a mass. However, only moderate levels of interobserver agreement were found for three of the six descriptive categories (mass margin, posterior acoustic transmission, and lesion echotexture), whereas only fair reproducibility was found for lesion echogenicity. The least concordance—slight agreement—was measured for observers determining the presence of an echogenic pseudocapsule.

Applying the rules of the model of Stavros et al. [8] to the morphologic descriptions selected by the observers, each of the 60 lesions was

TABLE 1 Distribution of Descriptors Used in Interobserver Study Cases

Category	Sonographic Descriptors	Combined Responses of Observers for All Cases
Mass shape	Ellipsoid (wider-than-tall)	228 (76)
	Taller-than-wide	72 (24)
Mass margin	Well-circumscribed lobulation	170 (57)
	Microlobulation	45 (15)
	Angular margins	67 (22)
	Duct extension	4 (1)
	Branch pattern	3 (1)
	Spiculation	11 (4)
Mass echogenicity	Intensely hyperechoic	19 (6)
	Isoechoic	45 (15)
	Mildly hypoechoic	162 (54)
	Markedly hypoechoic (solid)	74 (25)
Echogenic pseudocapsule	Absent	278 (93)
	Present	22 (7)
Acoustic transmission	Enhanced through-transmission	65 (22)
	Normal sound transmission	143 (48)
	Shadowing/decreased transmission	92 (31)
Mass echotexture	Homogeneous texture	135 (45)
	Heterogeneous texture	165 (55)

Note.—Figures are numbers of times observers selected each descriptor. Five observers each interpreted sixty cases (300 total observations). Numbers in parentheses are percentages within each subgroup.

TABLE 2 Interobserver Agreement in Evaluation of Sonography of Solid Breast Masses

Sonographic Feature	Kappa Value	Level of Reproducibility
Echogenic pseudocapsule	.09	Slight
Mass echogenicity	.40	Fair
Mass margin	.43	Moderate
Mass echotexture	.44	Moderate
Acoustic transmission	.55	Moderate
Mass shape	.8	Substantial
Final diagnostic assessment	.51	Moderate

Note.— Level of reproducibility is calculated as described by Landis and Kock [15].

TABLE 3 Intraobserver Agreement in Evaluation of Sonography of Solid Breast Masses

Sonographic Feature	Kappa Value	Level of Reproducibility
Echogenic pseudocapsule	.63	Substantial
Mass echogenicity	.69	Substantial
Mass margin	.62	Substantial
Mass echotexture	.24	Fair
Acoustic transmission	.63	Substantial
Mass shape	.79	Substantial
Final diagnostic assessment	.66	Substantial

Note.—Level of reproducibility is calculated as described by Landis and Kock [15].

classified as benign or malignant for each of the five observers. Consistency in observers' assessments was only moderate using the rules of this model ($\kappa = .51$).

Intraobserver Variability

Substantial intraobserver agreement was found for selecting all morphologic features except mass echotexture (Table 3). Substantial reproducibility ($\kappa = .66$) was also found for the assessment of one observer for diagnosing lesions as benign or malignant on the basis of the model of Stavros et al. [8].

Discussion

Six studies have found significant observer variability in radiologists' description and assessment of breast lesions on film-screen mammography [9–14]. This study shows a similar level of inconsistency between observers using sonographic images for lesion evaluation.

The greatest degree of interobserver agreement was found in determining the shape of a mass. To determine the shape, observers simply judge whether the lesion is ellipsoid (i.e., wider than tall), which is reportedly characteristic of benign masses, or taller than wide, which is characteristic of malignant lesions [8]. Such a determination is generally easily measured, explaining the relatively high level of observer agreement. However, the margins of a lesion may be poorly defined, making accurate measurement of the width or height difficult. Furthermore, edge shadowing can obscure the lateral margins (Fig. 1) and acoustic shadowing can completely conceal the posterior margin of a mass, making measurement of the height of the lesion guesswork. Observers also reported difficulty categorizing lesions that measure nearly the same in maximum height and depth, a circumstance not addressed in the model proposed by Stavros et al. [8]. Our study found considerable interobserver variation in determining the shape of such a lesion (Fig. 2).

We found moderate agreement for choosing one of three descriptors for posterior acoustic transmission of the ultrasound beam. According to the model of Stavros et al. [8], decreased through-transmission identified from any portion of a lesion raises suspicion of malignancy, whereas normal acoustic transmission and increased through-transmission are indeterminate features with no prognostic value. Much of the variability in evaluating this feature was due to observer differentiation between normal through-transmission and decreased transmission of the ultrasound beam (Fig. 3). Because any acoustic shadowing in the model identifies a lesion as worrisome, the inconsistency that observers showed in determining this feature could lead directly to inconsistency in the final interpretation of a lesion as benign or malignant.

Only moderate agreement was found among observers in characterizing lesion echotexture, which is the uniformity of echogenicity throughout solid breast masses. However, given that both heterogeneous and homogeneous echotexture have been categorized by Stavros et al. [8] as indeterminate features of solid breast masses, evaluating this feature has little clinical usefulness. Therefore, although we found considerable variation in characterization of mass echotexture, such characterization is not

Fig. 1.—18-year-old woman with palpable fibroadenoma in left breast. Sonogram shows well-circumscribed anterior and posterior margins (*arrowheads*), with lateral margins obscured by edge shadowing (*arrows*).

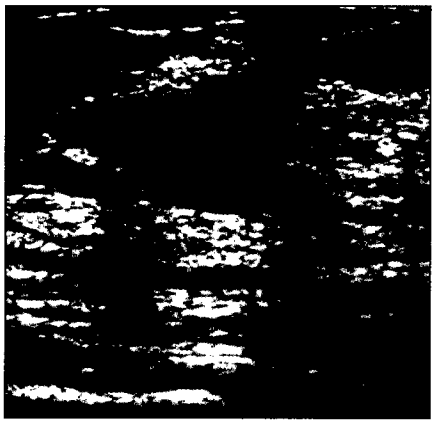
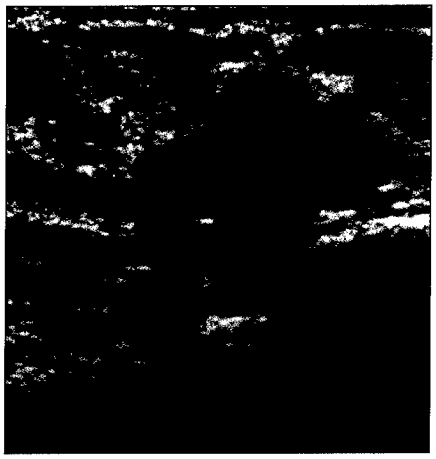


Fig. 2.—60-year-old woman with impalpable fibroadenoma identified at screening mammography. Sonogram shows variability in determining lesion shape due to indistinct margins and width and height that are nearly identical. Each of three observers described this mass as ellipsoid, whereas two other observers described it as taller than wide.

Fig. 3.—48-year-old woman with benign fibrosis in left breast (*arrowhead*). Sonogram illustrates interobserver variability in determining sonographic characteristics of lesion. Three observers said sound transmission was normal through-transmission, whereas two observers characterized it as decreased sound transmission.

A



B

Fig. 4.—68-year-old woman with infiltrating carcinoma in left breast.
A, Antiradial sonogram shows lesion.
B, Radial sonogram of mass illustrates variability in determining mass margin. Five observers used three different descriptors to characterize mass margin: "well-circumscribed gentle lobulation," "microlobulated," and "angular" margins.

Fig. 5.—41-year-old woman with lesion in right breast identified as indeterminate nodule on screening mammogram. On sonogram, nodule (*arrowhead*) has target or bull's-eye appearance with hyperechoic central portion and hypoechoic outer rim, often observed with intramammary lymph nodes.

useful in differentiating benign from malignant masses, so the variability is not relevant.

Mass margin is a critical feature for determining whether a lesion is benign or malignant according to the model of Stavros et al. [8]. We found only moderate agreement between observers in characterizing the margins of masses on breast sonograms. Observers reported that the seven terms available to describe a margin did not adequately characterize all possible margins for solid masses, so they had to select the term they deemed least wrong. For example, observers reported that the margins of many lesions were ill-defined. Although a solid mass was clearly present, the interface between the mass and the surrounding parenchyma was not sharp (Fig. 2). This appearance has been described elsewhere as "indistinct margins" [17]. Observers varied in how they ultimately described such margins, ranging from well-defined (a benign characteristic) to microlobulated or angular margins (malignant characteristics) (Fig. 4).

Determining the echogenicity of a mass was difficult for many observers, resulting in only fair levels of consistency. Echogenicity is the shade of gray constituting the lesion, ranging from markedly hypoechoic, which is essentially black, to intensely hyperechoic, which is primarily white. Many lesions had several different echogenic components. Several lesions had a hyperechoic inner portion and hypoechoic outer rim (Fig. 5). This description is often considered typical of an intramammary lymph node, but this type of target lesion is not addressed in the model of Stavros et al. [8]. For hypoechoic lesions, parts of the lesion may be slightly hypoechoic whereas other parts are markedly hypoechoic. It is unclear from the definitions of these terms whether the presence of any markedly hypoechoic tissue in a nodule is sufficient to declare the entire mass possibly malignant. Observers reported similar difficulties when evaluating hyperechoic lesions. Even for lesions that all observers agreed were hyperechoic relative to adjacent adipose tissue, observers disagreed about the degree of hyperechogenicity necessary to declare the lesion markedly or intensely hyperechoic and therefore benign (Fig. 6).

The category of mass echogenicity could be simplified without loss of diagnostic accuracy. According to the model of Stavros et al. [8], differentiating mildly hypoechoic lesions from isoechoic lesions offers no additional information in assessing breast lesions. Rather than choosing among four sometimes subtly different descriptors (markedly hyperechoic, isoechoic, mildly hypoechoic, markedly hypoechoic), the model could be simplified by requiring the observer to determine whether well-circumscribed or gently lobulated masses are markedly hypoechoic (and therefore likely malignant), markedly hyperechoic (and therefore benign), or neither.

The greatest variation in observer responses was found in determining whether an echogenic pseudocapsule was present for well-circumscribed or gently lobulated lesions (Fig. 7). This level of variability may in part be ascribed to our use of static images for evaluation. This scenario is doubtless common at busy breast imaging centers where the examination is performed by sonographers with only representative static images presented to the radiologist for interpretation. On the other hand, Stavros et al. [8] scanned the masses in real time, rocking the transducer beam to identify a pseudocapsule around all portions of a mass. Two of the five radiologists in our study insisted that only the most exhaustive set of static images could adequately depict a pseudocapsule. The other three radiologists determined that, although they would have preferred the information available with real-time imaging, representative static images were sufficient to judge the presence of a pseudocapsule. This difference in preference likely explains why all five observers did not agree that an echogenic pseudocapsule was present in any of the 60 cases in this study.

Variability in observers' descriptions of breast masses in this study is a concern because it resulted in inconsistent final interpretations of the masses using the rule-based model of Stavros et al. [8] (Figs. 6 and 7). Only a moderate level of agreement was found for this final assessment. Presumably, this inconsistency could be reflected in inconsistent recommendations to biopsy rather than closely monitor some solid breast lesions.

In contrast to the considerable variability between observers, we found substantial intraobserver agreement in characterizing each sonographic feature except lesion echotexture, for which variability has been shown to be unimportant. Although generalization based on the interpretations of one observer is necessarily limited, these results suggest that a single observer may be consistent in applying a defined lexicon. Such consistency in the use of this lexi-

Fig. 6.—60-year-old woman with nodule in left breast identified on screening mammogram. Three observers labeled lesion on sonogram (*arrowhead*) markedly hyperechoic, whereas two observers classified it isoechoic.

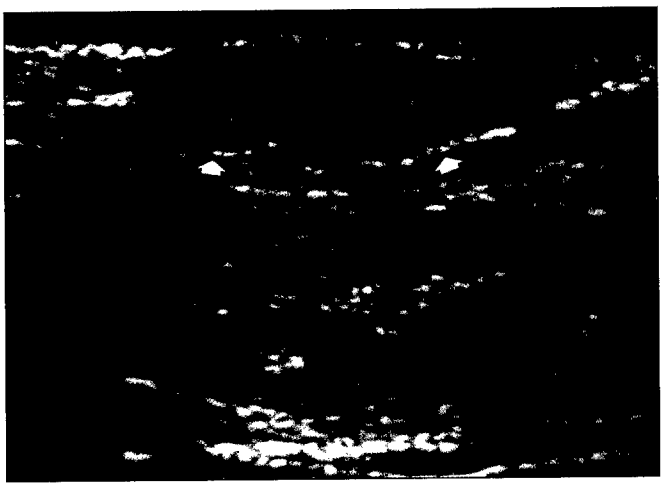


Fig. 7.—37-year-old woman with benign adenosis and fibroadenomatous change in superficial right breast. Sonogram of mass shows variability in determining presence of echogenic pseudocapsule. Two observers said pseudocapsule (*arrows*) was present, resulting in final assessment of benign lesion based on rule-based model of Stavros et al. [8]. Three other observers characterized lesion as malignant because they did not definitively identify pseudocapsule.

con resulted in substantial intraobserver consistency in determining the need for biopsy using the assessment model of Stavros et al. [8]. In contrast, the interobserver variability we found suggests that whether a lesion is interpreted as benign or malignant may depend in large part on which radiologist reviews the images.

The lexicon described and defined by Stavros et al. [8] was chosen for this study because the terms are defined and explained, with one or more examples of each descriptor provided. Nevertheless, the lack of consistency in applying these terms suggests further definition is needed. In an attempt to improve consistency, descriptive terms and their definitions could be agreed upon by a multiinstitutional panel in a document similar to the *Breast Imaging Reporting and Data System* [18], which was devised to improve the consistency of film-screen mammogram interpretations. Given the results of the present study, such a breast sonography lexicon should incorporate descriptors for commonly encountered findings such as "ill-defined" or "target" lesions. Furthermore, observers in this study desired additional example images to complement written definitions. Like the most recent edition of the *Breast Imaging Reporting and Data System*, illustrative images could be included in a consensus document defining a breast sonography lexicon. The development of such a standardized sonography lexicon may increase the consistency and reproducibility of sonographic imaging of solid breast lesions.

References

1. Venta LA, Dudiak CM, Salomon CG, Flisak ME. Sonographic evaluation of the breast. *RadioGraphics* **1994**;14:29–50
2. Jackson VP. The role of US in breast imaging. *Radiology* **1990**;177:305–311
3. Jokich PM, Monticciolo DL, Adler YT. Breast ultrasonography. *Radiol Clin North Am* **1992**;30: 993–1009
4. Jackson VP, Rothschild PA, Kreipke DL, et al. The spectrum of sonographic findings of fibroadenoma of the breast. *Invest Radiol* **1986**;21:34–40
5. Hilton SW, Leopold GR, Olson LK. Real-time breast sonography: application in 300 consecutive patients. *AJR* **1986**;147:479–486
6. Cole-Beuglet C, Soriano RZ, Kurtz B. Fibroadenoma of the breast: sonomammography correlated with pathology in 122 patients. *AJR* **1983**;140: 369–375
7. Cole-Beuglet C, Soriano RZ, Kurtz B. Ultrasound analysis of 104 primary breast carcinomas classified according to histopathologic type. *Radiology* **1983**; 147:191–196
8. Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker SH, Sisney GA. Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. *Radiology* **1995**;196:123–134
9. Skaane P, Engedal K, Skjennald A. Interobserver variation in the interpretation of breast imaging. *Acta Radiol* **1997**;38:497–502
10. Baker JA, Kornguth PJ, Floyd CE. Breast Imaging Reporting and Data System standardized mammography lexicon: observer variability in lesion description. *AJR* **1996**;166:773–778
11. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretation of mammograms. *N Engl J Med* **1994**;331:1493–1499
12. Ciccone G, Vineis P, Frigerio A, Segnan N. Inter-observer and intra-observer variability of mammogram interpretation: a field study. *Eur J Cancer* **1992**; 28A:1054–1058
13. Baines CJ, McFarlane DV, Miller AB. Estimates of inter-observer agreement and potential delay in cancer detection in the National Breast Screening Study. *Invest Radiol* **1990**;25:971–976
14. Vineis P, Sinistrero G, Temporelli A, et al. Inter-observer variability in the interpretation of mammograms. *Tumori* **1988**;74:275–279
15. Landis JR, Kock GG. The measurement of observer agreement for categorical data. *Biometrics* **1977**; 33:159–174
16. Svanholm H, Starklint H, Gundersen H, Fabricius J, Barlebo H, Olsen S. Reproducibility of histomorphologic diagnoses with special reference to the kappa statistic. *APMIS* **1989**;97:689–698
17. Leucht W. *Teaching atlas of breast ultrasound*. New York: Thieme, **1992**:21
18. American College of Radiology. *Breast imaging reporting and data system*, 3rd ed. Reston, VA: American College of Radiology, **1998**