

Bayesian Network to Predict Breast Cancer Risk of Mammographic Microcalcifications and Reduce Number of Benign Biopsy Results: Initial Experience¹

Elizabeth S. Burnside, MD, MPH, MS
 Daniel L. Rubin, MD, MS
 Jason P. Fine, PhD
 Ross D. Shachter, PhD
 Gale A. Sisney, MD
 Winifred K. Leung, MD

Purpose:

To retrospectively determine whether a Bayesian network (BN) computer model can accurately predict the probability of breast cancer on the basis of risk factors and mammographic appearance of microcalcifications, to improve the positive predictive value (PPV) of biopsy, with pathologic examination and follow-up as reference standards.

Materials and Methods:

The institutional review board approved this HIPAA-compliant study; informed consent was not required. Results of 111 consecutive image-guided breast biopsies performed for microcalcifications deemed suspicious by radiologists were analyzed. Mammograms obtained before biopsy were analyzed in a blinded manner by a breast imager who recorded Breast Imaging Reporting and Data System (BI-RADS) descriptors and provided a probability of malignancy. The BN uses probabilistic relationships between breast disease and mammography findings to estimate the risk of malignancy. Probability estimates from the radiologist and the BN were used to create receiver operating characteristic (ROC) curves, and area under the ROC curve (A_z) values were compared. PPV of biopsy was also evaluated on the basis of these probability estimates.

Results:

The BN and the radiologist achieved A_z values of 0.919 and 0.916, respectively, which were not significantly different. If the 34 patients estimated by the BN to have less than a 10% probability of malignancy had not undergone biopsy, the PPV of biopsy would have increased from 21.6% to 31.2% without missing a breast cancer ($P < .001$). At this level, the radiologist's probability estimation improved the PPV to 30.0% ($P < .001$).

Conclusion:

A probabilistic model that includes BI-RADS descriptors for microcalcifications can distinguish between benign and malignant abnormalities at mammography as well as a breast imaging specialist can and may be able to improve the PPV of image-guided breast biopsy.

© RSNA, 2006

¹ From the Department of Radiology, University of Wisconsin Medical School, E3/311 Clinical Science Center, 600 Highland Ave, Madison, WI 53792-3252 (E.S.B., G.A.S., W.K.L.); Section on Medical Informatics, Stanford University, Stanford, Calif (D.L.R.); Department of Statistics and Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Madison, Wis (J.P.F.); and Management Science and Engineering, Stanford University, Terman Engineering Center, Stanford, Calif (R.D.S.). Received June 29, 2005; revision requested August 31; revision received September 15; accepted October 4; final version accepted November 23. E.S.B. supported by the GE Research in Radiology Academic Fellowship. **Address correspondence to** E.S.B. (e-mail: es.burnside@hosp.wisc.edu).

Improved detection of microcalcifications at mammography has resulted in an increased number of breast biopsies with benign results in parallel with an increased number of ductal carcinoma in situ diagnoses. Although current practice standards recommend that the positive predictive value (PPV) of breast biopsy should be between 25%–40%, not everyone practices within this range (1,2). It has been established that subspecialty breast imagers have a higher PPV for biopsy recommendations than do general radiologists (3). Unfortunately, there is a nationwide shortage of subspecialists (4). In addition, results of a recent study from one practice (5) suggest the trend that an increasing number of biopsies in a breast imaging practice may increase only the number of benign biopsy results, not the number of cancer diagnoses. The practice that reported this result attributed these findings, in large measure, to increasing detection and biopsy of microcalcifications.

Our computer model, a Bayesian network, was designed to calculate the risk of malignancy for mammography findings on the basis of personal risk factors (age, family history, and hormone replacement therapy) and imaging features. Our system uses predictive imaging features catalogued in the Breast Imaging Reporting and Data System (BI-RADS) lexicon to assess disease risk (6). BI-RADS uses terms that were studied and determined to be most predictive of breast malignancy (7). It has been documented in the literature (8,9) that the mammographic appearance described by the radiologist can help predict the histologic features and prognosis of breast cancers. In addition, computerized classification algorithms have also succeeded in predicting the histo-

logic features of abnormalities identified as microcalcifications at mammography by using image-processing features such as shape and distribution (10).

Published reports (11–13) have already demonstrated that a Bayesian network can predict the probability of malignancy from demographic risk factors and mammographic findings (11–12), as well as the probability of sampling error, from similar data and percutaneous breast biopsy results (13). The purpose of our study was to retrospectively determine whether a Bayesian network computer model can accurately predict a patient's probability of breast cancer on the basis of risk factors and the appearance of microcalcifications at mammography to improve the PPV of biopsy, by using pathologic examination and follow-up as the reference standards.

Materials and Methods

Bayesian Network

The details of our Bayesian network have been reported previously but are repeated here for the convenience of the reader (11,12). From the literature, we identified 25 diseases of the breast (Fig 1) that represent the most common diagnoses rendered at mammography. Eleven of these diseases are malignant and 14 are benign. In our Bayesian network, uncertain variables affecting the probability of disease are represented as “nodes” that can be understood by both human and computer. Nodes are data structures that contain an enumeration of possible values they can assume (“states”); they also store probabilities associated with each state. In our system the “disease” node has states that represent the possible diseases of the breast. This node stores the prior probabilities of disease (the prevalence of each disease of the breast) as conditions of age, hormone replacement therapy, and family history of breast cancer.

The remaining nodes in the network represent possible findings on a mammogram. These nodes are based on descriptors in the BI-RADS lexicon (6). The structure of the model is also com-

posed of directed arcs (Fig 2) that encode the conditional dependence relationships among the variables. The absence of an arc represents conditional independence. Each arc implies a state of conditional dependence (in most cases, a causal link) between the nodes joined by that arc. Parent-child relationships are defined by the direction of the arcs between related nodes. As is the case with a genealogy chart, a parent node points to a child node. Each of the finding nodes is associated with a probability table that quantifies the probability of each state of the node depending on the values of incoming nodes. In other words, the conditional probability table has a row for each possible combination of parent values. The deterministic (double-bordered) node maps all diseases in the “disease node” into the appropriate disease type—benign or malignant.

There are two approaches to building Bayesian networks: (a) Use preexisting knowledge about the probabilistic relationships among variables, and (b) learn all the probabilities from large existing data sets. Because the literature already contained much information about how often different radiologic features occur in association with breast disease, we chose the former approach; our model was not trained on clinical data and outcomes. To con-

Advances in Knowledge

- A Bayesian network can accurately predict the probability that microcalcifications at mammography are malignant.
- The Bayesian network performed at the level of a subspecialty-trained mammographer.

Published online

10.1148/radiol.2403051096

Radiology 2006; 240:666–673

Abbreviations:

A_2 = area under the ROC curve
 BI-RADS = Breast Imaging Reporting and Data System
 PPV = positive predictive value
 ROC = receiver operating characteristic

Author contributions:

Guarantor of integrity of entire study, E.S.B.; 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, E.S.B., R.D.S.; clinical studies, E.S.B., W.K.L.; experimental studies, E.S.B., D.L.R.; statistical analysis, E.S.B., J.P.F., R.D.S.; and manuscript editing, E.S.B., D.L.R., J.P.F., R.D.S., G.A.S.

Authors stated no financial relationship to disclose.

struct the Bayesian network, we made probability assessments on the basis of the medical literature or the opinion of breast imagers with years of experience or subspecialty training (11,12). Pretest probabilities and age-specific and risk factor-specific distributions of diseases were obtained from census data and results of large randomized trials (14–17). The Bayesian network is able to calculate a posttest probability of malignancy by using the structure of the model and the probabilities in the conditional probability tables. To implement our Bayesian network and perform inference, we used the Netica modeling environment (<http://www.norsys.com/netica.html>).

Study Design and Patients

Our study included results of consecutive image-guided biopsies that met our criteria (see below) and that were performed for diagnosis of microcalcifications detected and deemed suspicious by radiologists in our practice. Readers included two subspecialty-trained breast imagers (including E.S.B.) with 3–4 years of experience, two radiologists who practiced predominately in other specialties—each of whom spent approximately 10%–20% of his or her time in breast imaging (with 7 and 16 years of experience)—and three general radiologists who spent 10%–50% of their time in breast imaging (with 4–23 years of experience). The institutional review board of the University of Wisconsin Medical School approved our study and determined that it was exempt from the requirement for informed consent. The study was compliant with the Health Insurance Portability and Accountability Act.

Results of 11-gauge stereotactic biopsies and needle localizations performed for diagnosis were included in our study. Our exclusion criteria included the following: (a) Patient’s images were not available for review, (b) calcifications were not identified in the histologic specimen, and (c) mammographic or clinical follow-up of at least 12 months was not performed. These criteria ensured the accurate categorization of each abnormality as benign or malignant by recognizing possible sampling error at percutaneous bi-

opsy. Of 124 consecutive patients who underwent biopsy, 13 were excluded according to our protocol (images were unavailable for six patients, no calcifications were identified at biopsy in three patients, no follow-up was performed in two patients, and no images or follow-up results were available for two

patients). Breast density and pathologic diagnoses in the final study population were analyzed to reveal our case mix. The 111 study patients consisted of women between the ages of 26 and 82 years (mean age, 55.8 years ± 10.5 [standard deviation]) who underwent a biopsy procedure between November

Figure 1

Malignant	Benign
Ductal carcinoma (DC)	Cyst
Ductal carcinoma in situ (DCIS)	Fibroadenoma
DC/DCIS*	Papilloma
Lobular carcinoma	Hamartoma
LC/LCIS*	Lymph node
Tubular carcinoma	Focal fibrosis
Papillary carcinoma	Fat necrosis
Medullary carcinoma	Secretory disease
Colloid carcinoma	Postoperative change
Phyllodes tumor	Skin lesion
Metastasis	Radial scar
	Atypical ductal hyperplasia
	Lobular carcinoma in situ (LCIS)
	Normal

*Two individual diagnoses present simultaneously.

Figure 1: Diagnoses in the disease node of the Bayesian network. *LC* = lobular carcinoma.

Figure 2

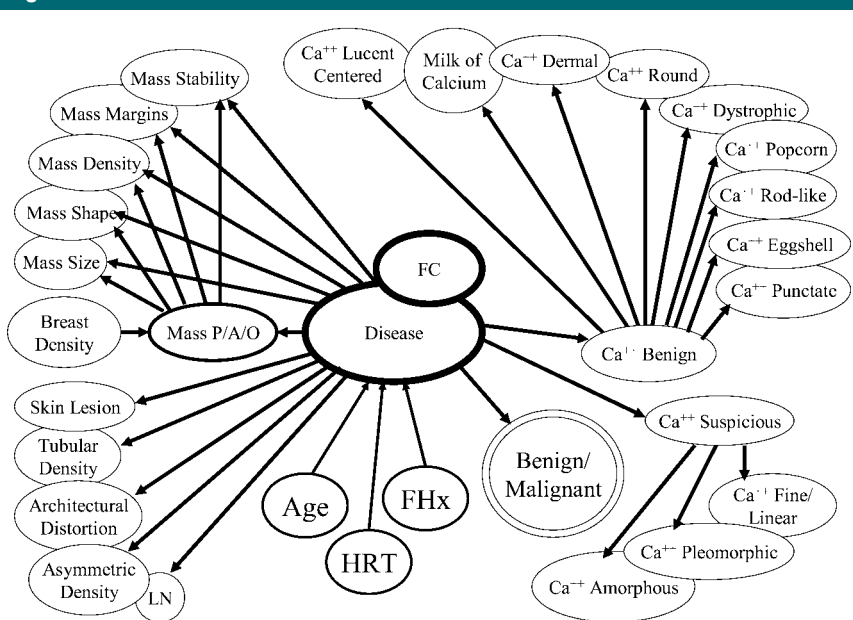


Figure 2: Structure of Bayesian network. Labeled ovals represent nodes; arrows (arcs) represent conditional dependence relationships. Each node is a data structure that contains conditional probability tables to quantify probabilistic relationships between variables. *Ca++* = calcifications, *FC* = fibrocystic change, *FHx* = family history of breast cancer, *HRT* = hormone replacement therapy, *LN* = lymph node, *P/A/O* = present, absent, or obscured.

2001 and September 2002 at the University of Wisconsin Medical School.

Screening and diagnostic mammograms obtained before the biopsy were analyzed in a blinded manner by a subspecialty-trained breast imager (E.S.B., with 3 years of experience) who recorded BI-RADS descriptors on a checklist and provided an estimate of the probability of malignancy. Age, family history, and history of hormone replacement therapy were collected for each patient (by W.K.L.) and were made available both to the radiologist at the time of the reader study and to the Bayesian network. The rate of malignant and high-risk lesions associated with morphologic and distribution descriptors were calculated to characterize the case mix and enable comparison of our results with those of previous studies of similar populations. The variables for each case, including the radiologist's descriptors, were entered into a research database by using a Web-

based interface that allowed the Bayesian inference engine to calculate the probability of malignancy.

Study Endpoints and Statistical Analysis

For malignant cases, surgical pathology findings at the time of the patient's definitive surgical intervention served as the reference standard in our study. To avoid the possibility of undersampling at percutaneous biopsy, we considered definitive surgical intervention to be either lumpectomy with established negative margins or mastectomy. For benign cases, clinical follow-up was the reference standard and was assessed by using either imaging or clinical records to determine whether a patient developed breast cancer at or adjacent to the biopsy site within 1 year of a benign biopsy result. We used 12-month follow-up as our reference standard because it has been recommended in mammography practice audits (18) as a sufficient interval to reveal false-negative results.

For evaluation of the performance of both the radiologist and the Bayesian network in the task of predicting the probability of malignancy, we created receiver operating characteristic (ROC) curves and calculated the area under the ROC curve (A_z) for comparison. In an attempt to determine the benefit of the radiologist and the Bayesian network working together to estimate probabilities, we averaged the two estimates to obtain an integrated probability of malignancy. For example, if, for a given case, the radiologist estimates a 30% probability of malignancy and the Bayesian network estimates a 50% probability of malignancy, the average probability is 40%. This procedure was performed for each patient, and an ROC curve was constructed and compared in a pairwise fashion with the ROC curves for the radiologist alone and for the Bayesian network alone. The PPV of biopsy was also evaluated on the basis of the probability estimates for the radiologist and the Bayesian network (each at thresholds for biopsy of 2% and 10%); these results were compared with the baseline PPV of the recommendation for biopsy in this population.

We used programming software for data calculations and statistical analysis (S-PLUS; Insightful, Seattle, Wash). Standard binormal ROC curve analysis was performed for the individual probability assessments made by the Bayesian network and the radiologist and the average of the two probability assessments (19). Point estimates and 95% confidence intervals for A_z values were calculated, and the different methods of assessment were compared by using tests for paired data. These analyses were implemented by using ROCKIT (http://www-radiology.uchicago.edu/krl/roc_soft.htm). The Fisher exact test was used to calculate the difference between the PPV of our practice and the PPV of the radiologist and the Bayesian network separately. Our specific application of the Fisher exact test is described in the Appendix. A *P* value of less than .05 was considered to indicate a statistically significant difference.

Results

Breast Density and Pathologic Features

Analysis of the study population revealed that almost half of the patients had heterogeneously dense fibroglandular tissue (Table 1). The remaining patients had either scattered fibroglandular densities or extremely dense tissue. No patients in our series had predominantly fatty breasts. The pathologic diagnoses included infiltrating, as well as in situ, carcinomas. An array of benign conditions, all of which contained microcalcifications at pathologic examination, were also diagnosed.

Microcalcifications

Microcalcification shape (Table 2) and distribution (Table 3) descriptors reflected rates of malignancy that corresponded to expected levels of suspicion described in BI-RADS. Analysis of morphologic descriptors, for example, revealed that the risk for typically benign microcalcifications was less than that for amorphous calcifications (which are categorized as of intermediate concern in BI-RADS), which was less than that for pleomorphic or amorphous calcifica-

Table 1

Characteristics of 111 Study Patients

Characteristic	No. of Patients [†]
Breast density	
Fatty	0 (0)
Scattered	34 (30.6)
Heterogeneously dense	54 (48.6)
Extremely dense	23 (20.7)
Pathologic diagnosis	
Invasive ductal carcinoma not otherwise specified	
Colloid carcinoma	3 (2.7)
Colloid carcinoma	1 (0.9)
Ductal carcinoma and ductal carcinoma in situ	
Ductal carcinoma in situ	9 (8.1)
Ductal carcinoma in situ	11 (9.9)
Atypical ductal hyperplasia	
Lobular carcinoma in situ	4 (3.6)
Lobular carcinoma in situ	2 (1.8)
Fibrocystic changes	52 (46.8)
Fibroadenoma	7 (6.3)
Papilloma	3 (2.7)
Fat necrosis	2 (1.8)
Normal*	17 (15.3)

* Microcalcifications contained in otherwise normal benign ducts or stroma.

[†] Number in parentheses is the percentage.

tions (which are categorized as highly suspicious in BI-RADS) (6).

Prediction of Malignancy

The Bayesian network and the radiologist achieved A_z values of 0.919 (95% confidence interval: 0.852, 0.960) and 0.916 (95% confidence interval: 0.835, 0.963), respectively. Using the average probability of malignancy (for the radiologist and the Bayesian network combined), we calculated an A_z of 0.948 (95% confidence interval: 0.892, 0.978) (Fig 3). Comparison of each of these A_z values in a pairwise fashion revealed that the radiologist alone and the Bayesian network alone showed no statistically significant difference in their ability to predict the probability that microcalcifications were malignant in this population. The A_z value of 0.948 calculated by using the average probability was significantly better than the A_z value of 0.916 calculated by using the radiologist's probabilities ($P = .03$). The A_z value of 0.948 calculated by using the average probability was suggestively but not significantly different from the A_z value of 0.919 calculated by using the Bayesian network probabilities ($P = .06$).

With the use of probabilities as risk estimates, patients and physicians have the opportunity to forego biopsy if the chance of malignancy is low. Twenty-one of the Bayesian network's estimates fell below a 2% probability of malignancy (Table 4). Foregoing biopsy in these patients would have improved the PPV of biopsy in this group from 21.6% to 26.7% ($P < .01$) without missing a cancer. None of the radiologist's estimates were below the threshold of 2%. Therefore, no biopsies would have been avoided at this level according to the radiologist's probability estimates. With a 10% threshold for biopsy, 34 patients would not have undergone biopsy according to the Bayesian network probability estimates, improving the PPV of biopsy to 31.2% without missing a breast cancer ($P < .001$). At this level, the radiologist's probability estimation improved the PPV to 30.0%, also without missing a breast cancer ($P < .001$).

Table 2

Rates of Malignant and High-Risk Lesions according to Shape Descriptors

Shape	No. of Lesions	No. of Malignant Lesions	No. of High-Risk Lesions*
Typically benign [†]	19	0 (0)	1 (5)
Amorphous	29	4 (14)	4 (14)
Pleomorphic	44	10 (23)	0 (0)
Linear	19	10 (53)	1 (5)
Total	111	24 (21.6)	6 (5.4)

Note.—Number in parentheses is the percentage.

* High-risk lesions included four cases of atypical ductal hyperplasia and two cases of lobular carcinoma in situ.

[†] Including round, punctate, dystrophic, and rodlike microcalcifications.

Table 3

Rates of Malignant and High-Risk Lesions according to Distribution Descriptors

Distribution	No. of Lesions	No. of Malignant Lesions	No. of High-Risk Lesions*
Scattered	1	0 (0)	0 (0)
Regional	4	0 (0)	0 (0)
Clustered	86	13 (15)	5 (6)
Segmental	8	3 (38)	1 (12)
Linear-ductal	12	8 (67)	0 (0)
Total	111	24 (21.6)	6 (5.4)

Note.—Number in parentheses is the percentage.

* High-risk lesions included four cases of atypical ductal hyperplasia and two cases of lobular carcinoma in situ.

Discussion

In our pilot study, we demonstrated that a Bayesian network can (a) predict the likelihood that microcalcifications at mammography are malignant, (b) perform as well as a subspecialty-trained mammographer in estimating the probability of malignancy, and (c) improve the PPV of the decision to perform biopsy. It is interesting that when the probabilities of the radiologist and the Bayesian network were averaged, the A_z value improved significantly compared with that for the radiologist alone. Published data support the contention that collaborative detection and decision making improves performance. The literature is replete with articles that demonstrate that double reading improves the sensitivity of screening mammography (20). Use of a computer-assisted detection program as a second reader has been shown to improve sensitivity in the screening set-

ting (21,22). Several systems have also been used to help the radiologist improve his or her decision to sample breast imaging findings for biopsy (23–25).

Our results indicate that the average probability assessment, as a surrogate for the consensus opinion, can better estimate risk than either the radiologist or the Bayesian network alone. The fact that the ROC curve of the average probabilities completely dominated (ie, was to the left of) the curve for the radiologist alone suggests that combined probabilities improve both sensitivity and specificity at all threshold levels.

Both the radiologist and the computer model were able to select patients who had a low likelihood of malignancy and in whom biopsy may not have been warranted. The baseline PPV of 21.6% in our practice is below the recommended guidelines of 25%–40% (1). With the aid of the Bayesian network, at both 2% and 10% thresholds the PPV

was elevated to be within recommended guidelines without missing a breast cancer. Larger studies would be needed to establish that such an increase in PPV could be maintained without missing any cancers. However, given our present results, our Bayesian network has the potential to be used as a decision-support tool that could help underperforming practitioners improve the PPV of biopsy recommendations.

The fact that decision-support tools

that use the BI-RADS lexicon may help improve practice is not surprising. The BI-RADS lexicon was created and has been revised to capture predictive mammographic descriptors in a standardized manner (6,18). For example, in our study, typically benign morphologic descriptors were associated with no malignancies, while the highly suspicious morphologic features of pleomorphic and linear calcifications were associated with a 23% and 53% likelihood of malignancy, respectively. Taking into account population risk differences between studies, our results correspond to those of other studies in which these descriptors have been evaluated (26, 27). Despite the fact that moderate interobserver variability has been documented in the application of BI-RADS descriptors, training in the lexicon has “resulted in improved agreement with the consensus of experienced breast imagers for feature analysis” (28). It is logical that BI-RADS, in concert with computer-aided diagnosis, can be effectively applied to improve performance.

In our pilot study, the radiologist involved in building the Bayesian network (E.S.B.) also provided the BI-RADS features and probabilities for the cases. This evaluation technique has been used before in a study of the early development of a Bayesian network for lymph node disease, called Pathfinder, in which the Bayesian network’s perfor-

mance was compared with that of the same subspecialty-trained pathologist who built it (29). Similar to the performance of the Bayesian network in our study, Pathfinder performed at least as well as the pathologist in 53 selected cases of lymph node disease. A subsequent study revealed that Pathfinder performed significantly better than did pathologists who had less experience in the diagnosis of lymph node disease and were not involved in the creation of the system (30). We plan further evaluation of our Bayesian network by using radiologists who were not involved in the construction of the Bayesian network and may see similar performance improvements.

The fact that the Bayesian network method generates an actual probability estimate is a great advantage because the output can be intuitively applied to decision making. Obtaining a probability for malignancy gives the radiologist, patient, and referring physician the opportunity to engage in shared decision making. There is increasing interest in shared decision making in the radiology community in relation to screening tests (31,32). Each patient has a unique risk tolerance and comorbidities to weigh in the decision as to whether to perform breast biopsy. A patient can understand a probability and be better informed about her own individual risk. The availability of a posttest probability of malignancy would allow decisions to be based on personal preference in the context of discussions with radiologists, referring physicians, and others. Whether this type of collaboration can be used in the future to help decision making will depend on success in prospective trials.

There were several limitations to our study. First and foremost, our study was a retrospective analysis, with the participation of only one radiologist; this may limit the generalizability of our results to their prospective clinical application over a diverse group of radiologists. Our study was planned as a pilot study to validate our model with one individual and eliminate other confounders such as interobserver variability. We predict that there will be differences in performance between observ-

Figure 3

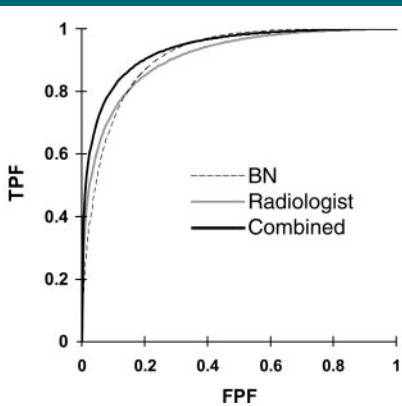


Figure 3: Graph shows ROC curves constructed from the probabilities of the radiologist alone and the Bayesian network (BN) alone and from the average of the probabilities for the radiologist and the Bayesian network. FPF = false-positive fraction (1 – specificity), TPF = true-positive fraction (sensitivity).

Table 4

PPV at Distinct Probability Threshold Levels

Parameter	Baseline*	2% Threshold		10% Threshold	
		BN	Radiologist	BN	Radiologist
Biopsy [†]	111	90	111	77	80
No biopsy [‡]	0	21	0	34	31
Malignant cases	24	24	24	24	24
Benign cases	87	66	87	53	56
PPV (%)	21.6	26.7	21.6	31.2	30.0
95% Confidence interval of PPV (%)	14.4, 30.4	17.8, 37.0	14.4, 30.4	21.1, 42.7	20.2, 41.2
P value [§]006	.99	<.001	<.002

Note.—BN = Bayesian network.

* Total number of patients and recommended biopsies in our population.

[†] Number of biopsies performed if no patients below the assigned probability threshold underwent biopsy.

[‡] Number of biopsies avoided if no patients below the assigned probability threshold underwent biopsy.

[§] See Appendix for explanation of calculated P values; PPV comparisons were all made to the baseline performance.

ers with and without the model when we test it over a broader range of radiologists. Second, our study population represents a group of patients selected because they had suspicious findings. Therefore, we have demonstrated that our computer model performs well only for the subset of patients with microcalcifications recommended for breast biopsy.

Third, in the ROC analysis, there were three pairwise comparisons. We did not correct for these multiple tests when assessing statistical significance (which may increase the likelihood of type I error) because the point of this purely investigational study was to demonstrate that our method shows promise. Further investigation is needed to obtain more definitive results. Finally, in a reader study, it is possible that cases may not represent actual clinical practice. We attempted to minimize the possibility of such a bias by including consecutive biopsy cases in our series, which to some degree ensured that the cases analyzed were no more or less difficult than would be expected in any population of calcification cases referred for biopsy. Our case characteristics and assessed descriptors can shed light on whether our case mix was appropriate.

It is interesting that our patient population consisted of women with denser-than-average breasts as assessed by our reader. Although this might reflect uneven application of the density descriptors, it may also indicate that our patient population had a high proportion of mammographically dense tissue, possibly making our cases more difficult than would be expected and thereby attenuating our performance results. On the other hand, the fact that the PPV of our practice is comparable with other practices in a large survey of mammography practices in the United States (2) supports the contention that our case mix was similar to that of other practices. Although we believe that there is minimal risk of systematic bias in our study, documentation of truly generalizable results requires prospective testing in a larger patient population with multiple readers at multiple institutions.

However, we are encouraged by our promising preliminary results.

In conclusion, we have demonstrated that use of the BI-RADS lexicon, when coupled with our Bayesian network model, can produce an accurate probability estimate of the risk of malignancy and match the performance of a subspecialty-trained breast imager. Furthermore, this system may have the potential to improve the PPV of biopsy and help practices to perform within the national guidelines for biopsy recommendation. Finally, by generating an accurate estimate of the posttest probability of malignancy, a Bayesian network for mammography may provide the opportunity for intuitive and collaborative decision making between patients and physicians in the future. Ultimately, we hope that, with further testing and use, probabilistic models will aid decision making in mammography practice.

Appendix

This appendix provides the theoretical justification for the test of the equivalence of two PPVs derived from related data in two stages. Specifically, if one is interested in comparing the results from the first-stage test with those from the second-stage test, where the positive tests from the second stage are refined by using a second test, we demonstrate a method and rationale for applying the Fisher exact test. In our study, stage 1 refers to the original biopsy recommendation and stage 2 refers to the judgments of the radiologist and the Bayesian network.

Let PPV1 be the PPV for the first-stage test only, and let PPV2 be the PPV of the combined two-stage procedure. Let TP1 and FP1 be the probability of a true-positive and the probability of a false-positive result at stage 1, respectively. Let TP2 be the probability of a true-positive test at both stages 1 and 2, and let FP2 be the probability of a false-positive test at both stages 1 and 2. Let TN2 be the probability of a false-positive test at stage 1 and a true-negative test at stage 2, and let FN2 be the probability of a true-positive test at stage 1 and a false-negative test at stage 2. By

definition, $PPV1 = TP1/(TP1 + FP1)$, and $PPV2 = TP2/(TP2 + FP2) = (TP1 - FN2)/(TP1 + FP1 - FN2 - TN2)$.

If one estimates PPV1 and PPV2 by plugging in the sample proportions for TP1, FP1, TP2, FP2, TN2, and FN2, one cannot naively use the Fisher test to evaluate the equivalence of PPV1 and PPV2 because of complex correlations in PPV1 and PPV2. One can show algebraically that $PPV1 = PPV2$ implies that $FN2/TP1 = TN2/FP1$. That is, the false-negative rate from stage 2 based on true-positives from stage 1 equals the true-negative rate from stage 2 based on false-positives from stage 1. This result permits use of a standard Fisher test for the equivalence of PPV1 and PPV2, because the number of individuals with false-negative tests at stage 2 and true-positive tests at stage 1 and the number of individuals with true-negative tests at stage 2 and false-positive tests at stage 1 are independent binomial random variables conditioned on the outcomes of stage 1 testing.

The mathematic argument is as follows:

$$\begin{aligned} &\rightarrow PPV 1 = PPV 2 \\ \rightarrow TP1/(TP1 + FP1) &= (TP1 - FN2)/ \\ &(TP1 + FP1 - FN2 - TN2) \\ \rightarrow TP1(TP1 + FP1 - FN2 - TN2) & \\ &= (TP1 + FP1)(TP1 - FN2) \\ \rightarrow TP1^2 + TP1(FP1 - TP1(FN2 & \\ &+ TN2)) = TP1^2 - TP1(FN2) \\ &+ FP1(TP1) - FP1(FN2) \\ \rightarrow TP1(TN2) &= FP1(FN2) \\ \rightarrow TN2/FP1 &= FN2/TP1. \end{aligned}$$

References

1. Bassett LW, Hendrick RE, Bassford TL, Butler PF, Carter D, DeBor M. Quality determinants of mammography. Clinical Practice Guideline No. 13. AHCPR publication no. 95-0632. Rockville, Md: Agency for Health Care Policy and Research, Public Health Service, U.S. Department of Health and Human Services, October 1994.

2. Brown ML, Houn F, Sickles EA, Kessler LG. Screening mammography in community practice: positive predictive value of abnormal findings and yield of follow-up diagnostic procedures. *AJR Am J Roentgenol* 1995; 165:1373-1377.
3. Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology* 2002;224:861-869.
4. Eklund GW. Shortage of qualified breast imagers could lead to crisis. *Diagn Imaging (San Franc)* 2000;22:31-33.
5. Gur D, Wallace LP, Klym AH, et al. Trends in recall, biopsy, and positive biopsy rates for screening mammography in an academic practice. *Radiology* 2005;235(2):396-401.
6. American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS). Reston, Va: American College of Radiology, 1998.
7. Swets JA, Getty DJ, Pickett RM, D'Orsi CJ, Seltzer SE, McNeil BJ. Enhancing and evaluating diagnostic accuracy. *Med Decis Making* 1991;11:9-18.
8. Thurffjell MG, Lindgren A, Thurffjell E. Non-palpable breast cancer: mammographic appearance as predictor of histologic type. *Radiology* 2002;222:165-170.
9. Tabar L, Tony Chen HH, Amy Yen MF, et al. Mammographic tumor features can predict long-term outcomes reliably in women with 1-14-mm invasive breast carcinoma. *Cancer* 2004;101:1745-1759.
10. Nakayama R, Uchiyama Y, Watanabe R, Katsuragawa S, Namba K, Doi K. Computer-aided diagnosis scheme for histological classification of clustered microcalcifications on magnification mammograms. *Med Phys* 2004;31:789-799.
11. Burnside E, Rubin D, Shachter R. A Bayesian network for mammography. *Proc AMIA Symp* 2000;106-110.
12. Burnside ES, Rubin DL, Shachter RD, Sohlich RE, Sickles EA. A probabilistic expert system that provides automated mammographic-histologic correlation: initial experience. *AJR Am J Roentgenol* 2004;182: 481-488.
13. Kahn CE Jr, Roberts LM, Wang K, Jenks D, Haddawy P. Preliminary investigation of a Bayesian network for mammographic diagnosis of breast cancer. *Proc Annu Symp Comput Appl Med Care* 1995;208-212.
14. U.S. Census Bureau. Projections of the total resident population by 5-year age groups and sex with special age categories: middle series, 2001 to 2005. Washington, DC: Population Projections Program, Population Division, U.S. Census Bureau, 2000.
15. Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the International Workshop on Screening for Breast Cancer. *J Natl Cancer Inst* 1993;85:1644-1656.
16. Slattery ML, Kerber RA. A comprehensive evaluation of family history and breast cancer risk: the Utah Population Database. *JAMA* 1993;270:1563-1568.
17. Stanford JL, Weiss NS, Voigt LF, Daling JR, Habel LA, Rossing MA. Combined estrogen and progestin hormone replacement therapy in relation to risk of breast cancer in middle-aged women. *JAMA* 1995;274:137-142.
18. American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS). Reston, Va: American College of Radiology, 2003.
19. 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.
20. Kopans DB. Double reading. *Radiol Clin North Am* 2000;38:719-724.
21. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology* 2001;220: 781-786.
22. Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology* 2000;215:554-562. [Published correction appears in *Radiology* 2000;216(1):306.]
23. Hadjiiski L, Chan HP, Sahiner B, et al. Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: an ROC study. *Radiology* 2004; 233:255-265.
24. Huo Z, Giger ML, Vyborny CJ, Metz CE. Breast cancer: effectiveness of computer-aided diagnosis observer study with independent database of mammograms. *Radiology* 2002;224:560-568.
25. Floyd CE Jr, Lo JY, Tourassi GD. Case-based reasoning computer algorithm that uses mammographic findings for breast biopsy decisions. *AJR Am J Roentgenol* 2000;175: 1347-1352.
26. Berg WA, Arnoldus CL, Teferra E, Bhargavan M. Biopsy of amorphous breast calcifications: pathologic outcome and yield at stereotactic biopsy. *Radiology* 2001;221: 495-503.
27. Liberman L, Abramson AF, Squires FB, Glassman JR, Morris EA, Dershaw DD. The Breast Imaging Reporting and Data System: positive predictive value of mammographic features and final assessment categories. *AJR Am J Roentgenol* 1998;171:35-40.
28. Berg WA, D'Orsi CJ, Jackson VP, et al. Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography? *Radiology* 2002;224:871-880.
29. Heckerman DE, Nathwani BN. An evaluation of the diagnostic accuracy of Pathfinder. *Comput Biomed Res* 1992;25:56-74.
30. Nathwani BN, Clarke K, Lincoln T, et al. Evaluation of an expert system on lymph node pathology. *Hum Pathol* 1997;28:1097-1110.
31. Chan EC. Promoting an ethical approach to unproven imaging tests. *J Am Coll Radiol* 2005;2:311-320.
32. Hillman BJ. Informed and shared decision making: an alternative to the debate over unproven screening tests. *J Am Coll Radiol* 2005;2:297-298.