27 A BAYESIAN NETWORK TO ASSIST MAMMOGRAPHY INTERPRETATION

Daniel L. Rubin[1], Elizabeth S. Burnside[2] and Ross Shachter[3]

[1] Stanford Medical Informatics
Stanford University
Stanford, CA 94305

[2] Department of Radiology
University of Wisconsin
Madison, WI 53792

[3] Department of Management Science and Engineering
Stanford University
Stanford, CA 94305

SUMMARY

Mammography is a vital screening test for breast cancer because early diagnosis is the most effective means of decreasing the death rate from this disease. However, interpreting the mammographic images and rendering the correct diagnosis is challenging. The diagnostic accuracy of mammography varies with the expertise of the radiologist interpreting the images, resulting in significant variability in screening performance. Radiologists interpreting mammograms must manage uncertainties arising from a multitude of findings. We believe that much of the variability in mammography diagnostic performance arises from heuristic errors that radiologists make in managing these uncertainties. We developed a Bayesian network that models the probabilistic relationships between breast diseases, mammographic findings and patient risk factors. We have performed some preliminary evaluations in test cases from a mammography atlas and in a prospective series of patients who had biopsy confirmation of the diagnosis. The model appears useful for clarifying the decision about whether to biopsy abnormalities seen on mammography, and also can help the radiologist correlate histopathologic findings with the mammographic abnormalities observed. Our preliminary experience suggests that this model may help reduce variability and improve overall interpretive performance in mammography.

KEY WORDS

Mammography, Diagnosis, Breast cancer, Bayesian networks

27.1 INTRODUCTION

Breast cancer is the most frequently diagnosed malignancy among American women. It is the second leading cause of cancer death (after lung cancer) among women of all ages and the leading cause of cancer death among women aged 40 to 59 years [1]. Mammography has been shown to be effective in detecting breast cancer before it becomes clinically evident [2]; consequently, routine screening with mammography is now generally accepted as a valuable tool for decreasing mortality from breast cancer.

The benefits of screening mammography are limited by the quality of the image acquired and by the accuracy of image interpretation. Image acquisition quality depends on the quality and operation of the imaging equipment, while interpretation depends on the training and expertise of a human reader (the radiologist). In recent years, standards relating to the imaging equipment such as the Mammography Quality Standards Act (MQSA) [3] have improved the quality of mammogram images at many facilities [4]. However, overall accuracy of mammographic interpretation, in terms of sensitivity and specificity, is a problem because of variability in the training and experience of radiologists interpreting the images [5]. Several studies have reported substantial mammogram interpretation inconsistencies among different radiologists [6-8], which would lead to different followup testing and treatment decisions.

False-negative and false-positive interpretations have been called “risks” of screening mammography and have been cited as arguments against routine screening of various populations of women [9-11]. False negative interpretations are risks because patients having cancer are not detected (reduced efficacy of screening), thus delaying cancer treatment and leading to higher morbidity and mortality. On the other hand, false positive interpretations are risks because patients without cancer undergo unnecessary biopsy (causing anxiety and increased medical costs). Variability in interpretive accuracy among radiologists lowers the average positive predictive value for mammography, which makes it a less effective tool for the early detection of breast cancer. Therefore, strategies to reduce variability in mammographic interpretation are essential to improve patient care.

Some of the variability in interpretative accuracy among radiologists is likely related to training and experience. Some radiologists have subspeciality training in mammography and read these studies exclusively. These individuals are generally considered “experts” in the field. On the other hand, community radiologists read the majority of mammograms in the

context of diverse general practice. Community radiologists have higher biopsy rates and thus lower positive predictive value of malignant disease [5, 12].

One approach to reduce the variation in interpretations among radiologists is to standardize the vocabulary used in mammography reports. The American College of Radiology (ACR) developed BI-RADS, a lexicon of mammogram findings (or "features") and the distinctions that describe them. [13]. The developers of BI-RADS tried to identify those features of mammograms that are most useful for discriminating diseases of the breast. To accomplish this, they performed statistical analysis of the terms ("descriptors") used to describe imaging findings to determine which descriptors best discriminate between a benign or malignant diagnosis [14].

While BI-RADS is an important step in reducing the variation in mammography reporting, it does not solve the problem of how radiologists relate a set of findings they observe on the mammogram to a diagnosis. Specifically, how does the radiologist determine the probability of malignancy given a set of observed findings so as to choose followup tests and treatments? The quality of this determination is the essential difference between an expert and a non-expert, and likely accounts for much of the interpretive variation among radiologists.

Because the BI-RADS findings observed on mammography were selected to discriminate between benign and malignant diseases, they contain precisely the information we want to obtain for diagnosis. Our hypothesis is that we can build a probabilistic model relating diseases to the BI-RADS descriptor findings seen on mammography, and that, given the BI-RADS findings, this model can be used to compute posterior probabilities for the possible breast diseases. Such probabilities can guide the radiologist's decision making.

Our goal has been to build a model that represents these probabilistic relationships among BI-RADS findings and includes other pertinent information (patient risk factors) to standardize how combinations of BI-RADS findings are interpreted. Such a model could also be used to determine the likelihood of breast diseases and to evaluate the agreement (concordance) between biopsy results and the mammographic findings. Our hope is to bring clarity to decision making and reduce suboptimal variability in patient management based on a normative approach.

27.2 METHODOLOGY

27.2.1 Building a model for mammography diagnosis

To represent the probabilistic relationships among findings and diseases, we built a Bayesian belief network. Bayesian networks are graphical probabilistic models of the conditional dependencies among variables of interest [15]. In our application, we are interested in breast diseases that are diagnosed on mammography, the radiological findings that are observed on mammography (in terms of BI-RADS descriptors), and patient risk factors (age, history of hormone treatment, and history of prior breast cancer).
Because a mammogram may contain more than one abnormality ("lesion"), we built a lesion-centric model; if a patient has more than one lesion, the model can be applied to each lesion independently. For the time being, however, we will assume that each patient has at most one lesion.

Diseases From a review of the literature and with the assistance of an expert in mammography, we identified 23 diseases of the breast. These diseases, in addition to a "normal" diagnosis and two combined diagnoses, were selected as the distinctions for a DISEASE node (having 26 states) in the model (Table 27.1).

In order for us to define mutually exclusive disease distinctions, we assume that a given lesion on the mammogram represents a single specific disease process. Because of the pathophysiology of breast cancer, it is possible to see two diseases simultaneously within a single malignant lesion. Simultaneous appearance of two diseases occurs when atypical cells transform into malignant cells. For example, "ductal carcinoma *in situ*" (DCIS) contains non-invasive neoplastic cells that may undergo transformation into "ductal carcinoma, not otherwise specified" (DCNOS). Thus, some breast lesions may contain both diseases if only some of the cells have transformed. For this reason, the DISEASE node includes two combined diagnoses, "LC+LCIS" and "DCNOS+DCIS" (Table 27.1).

In order for all 26 states in DISEASE to be collectively exhaustive, we assume that we have modeled all possible diseases (or disease combinations) that may be diagnosed on mammography, including a benign state of no disease (Normal).

Findings and Patient Risk Factors We compiled a list of findings (abnormalities) observed on mammography from the BI-RADS descriptors [13]. BI-RADS consists of 43 descriptors, some of which are organized in a hierarchical taxonomy (Figure 27.1). The hierarchical structure of the descriptors helps the user navigate and select descriptors and their modifiers

Table 27.1 Diagnoses seen on mammography that are incorporated in the DISEASE node in the Bayesian network. LC+LCIS and DCNOS+DCIS each represent combinations of two single diseases.

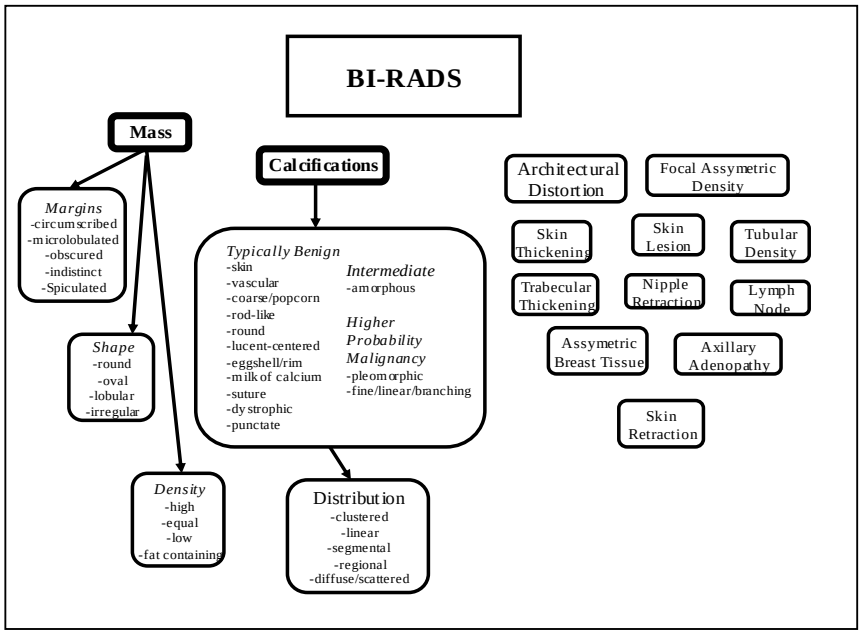
Malignant	**Benign**
Invasive Ductal Carcinoma (DCNOS)	Lobular Carcinoma in situ (LCIS)
Ductal Carcinoma in situ (DCIS)	Cyst
Lobular Carcinoma (LC)	Fibroadenoma
LC+LCIS**	Fibrocystic Change
DCNOS+DCIS**	Hamartoma
Tubular Carcinoma	Focal Fibrosis
Papillary Carcinoma	Fat Necrosis
Medullary Carcinoma	Secretory Disease
Colloid Carcinoma	Post-operative change
Phylloides Tumor	Skin Lesion
Metastasis	Lymph node
	Papilloma
	Radial Scar*
	Atypical Ductal Hyperplasia (ADH)*
	Normal

* Radial Scar and atypical ductal hyperplasia are considered ”high-risk” because they can be associated with in situ or invasive breast cancer. Though controversy surrounds both of these diagnosis, they are currently considered benign.
** These diagnoses represent two individual diagnoses present simultaneously during a process of transformation.

efficiently. For example, once a mass is identified, the user can describe the margins, shape, and density. The mass shape can be characterized by “detailed descriptors” such as round, oval, lobular, or irregular (Figure 27.1). Other examples of detailed descriptors are the modifiers of mass density: high, equal, low, and fat-containing (Figure 27.1).

We incorporated 38 of the BI-RADS descriptors into the model. We excluded five descriptors (skin thickening, trabecular thickening, nipple retraction, skin retraction, and asymmetric breast tissue) because they are rare, late, or non-contributory findings on screening mammography, and because they would have increased the complexity of the model without significantly improving its diagnostic effectiveness. Each descriptor we selected became a node in our Bayesian network. If that descriptor had detailed descriptors, they became distinctions for that node; otherwise the states for the node were “present” and “not-present.” For example, the DENSITY node has states called “high,” “equal,” and “low”; the ROUND CALCIFICATION node has states “clustered,” “linear,” “segmental,”

Figure 27.1 The BI-RADS terminology

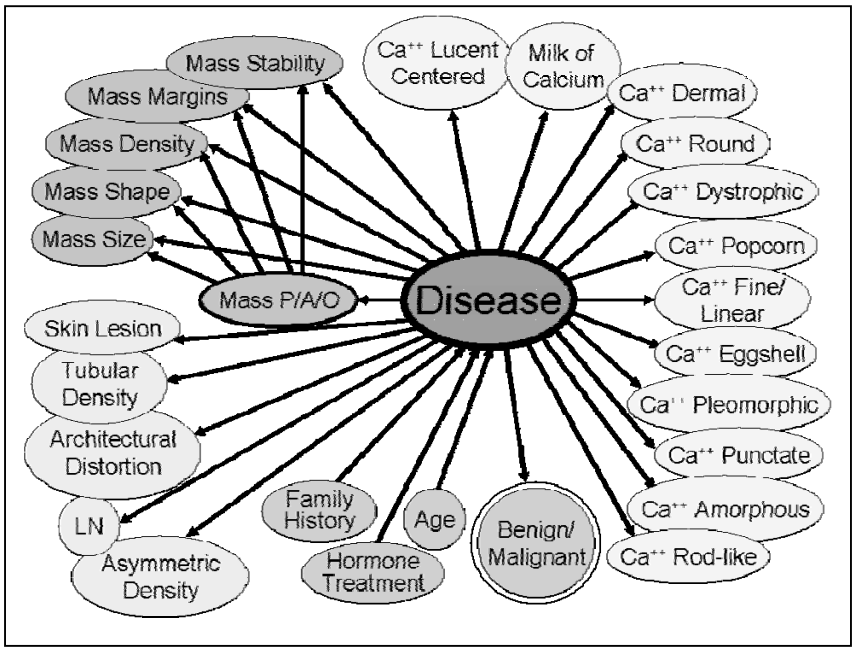


"regional," and "diffuse/scattered"; the ARCHITECTUAL DISTORTION node has states "present" and "not present" (Figure 27.1).

We incorporated the following patient risk factors into our model: age (40-44, 45-50, 51-54, 55-60, 61-64, 65-70), history of breast cancer (strong, minor, or no history), and history of hormone treatment (less than 5 years, more than 5 years, or no history).

Bayesian Network To implement the model and perform inference given particular observations, we used the GeNIe modeling environment developed by the Decision Systems Laboratory of the University of Pittsburgh (http://www2.sis.pitt.edu/~genie/). In defining the structure of the model, we consulted with two experts in mammography to define the conditional dependencies among findings and diseases (Figure 27.2). The prior probability of disease is dependent on the patient risk factors, and these factors were believed to be conditionally independent of disease; thus disease has a separate parent for each risk factor. All BI-RADS descriptors except for those relating to a mass were believed to be conditionally independent manifestations of disease, so each descriptor has disease as a parent (Figure 27.2). The descriptors relating to a mass all depend on the

Figure 27.2 Bayesian network model of mammography diagnosis



parent (Figure 27.2). The descriptors relating to a mass all depend on the presence of a mass, and we assume they become conditionally independent given the disease and whether the mass is present.

Normal structures of the breast can obscure masses on a mammographic image. This obscuration is more common in younger women and women on estrogen replacement therapy because they tend to have relatively dense breast tissue. While obscuration of the finding does not change the probability of disease given findings about the mass, it does decrease the probability that the mass and its features will be recognized. To model obscuration, we added an "obscured" state to joint distribution of the "mass" descriptor. Thus, a mass on the mammogram may be present, absent, or obscured, represented by the node "MASS P/A/O" which depends on the disease (Figure 27.2). (In later versions of the model, MASS P/A/O also depends on the patient's age and hormone treatment.)

The Bayesian network includes a deterministic node labeled "Benign/Malignant" which categorizes the diseases into these two distinct categories (Figure 27.2 and Table 27.1). This is useful because knowing the type of disease is important in determining correct management. For

example, standard practice requires that all malignant diseases receive definitive therapy including excision. Some diseases such as radial scar and papilloma are controversial. These benign disease entities are sometimes associated with malignancy and therefore management depends on many factors. We have classified these diseases in the "benign" category pending further data (Table 27.1).

The initial values for the joint probability distributions were defined by consulting the experts and by reviewing the literature. We obtained the prior probabilities, age-specific and risk factor-specific distributions of diseases from census data and large randomized trials [16-18]. We derived many of the joint probabilities from studies of radiological/pathological correlation of individual breast diseases [19, 20]. Some of these initial probabilities evolved over time as we evaluated the model with test cases in which the correct diagnosis was known.

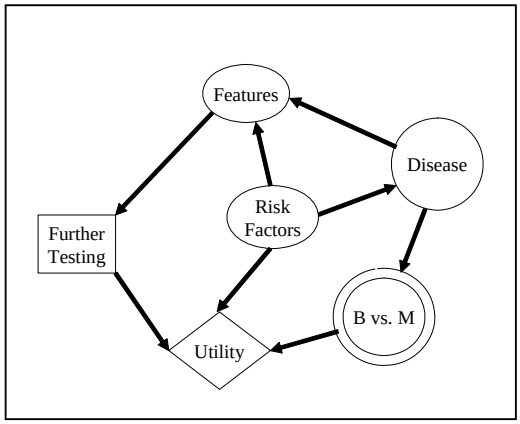
The Bayesian network we built calculates the probability of disease given these findings. The calculation depends on the pre-test probability of disease *d* (prevalence), given patient risk factors, and the joint probability distributions associated with the patient risk factors, disease, and the BI-RADS descriptor nodes in the Bayesian network (Figure 27.2). For a particular patient, values for risk factors and mammography findings are entered as observed evidence in the Bayesian network. If a particular finding or patient risk factor is not reported, then the corresponding node is unobserved. The joint probability distributions can then be updated using Bayes rule, giving a posterior probability distribution over the diseases, $P\{d|f\}$. This posterior probability can be used in several ways. Below we describe two ways we have used this information thus far: (1) to make a diagnosis on mammography, and (2) in evaluating concordance of mammography findings with biopsy results.

27.2.2 Using the model to make a diagnosis on mammography

Mammography is a screening test used to recognize breast cancer as early as possible when therapies are most effective and least debilitating. In the screening setting, observations are made to differentiate "benign" or "malignant" disease, and this diagnosis affects patient management decision making (Figure 27.3). If the screening mammogram has features suspicious for malignant disease, then the patient is called back for a more detailed diagnostic mammogram. If the probability of malignancy given all of the information available is high enough, then the radiologist will perform a biopsy for histological diagnosis. The influence diagram [21, 22] shown in Figure 27.3 represents these decisions. The critical distinction, represented

by the deterministic node, B vs. M, is whether the disease is benign or malignant.

Figure 27.3 Influence diagram showing the radiologist's management decision whether to perform further tests which may lead to treatments



At the time of the "Further Testing" choice (whether to recall the patient for a diagnostic x-ray or to biopsy) the radiologist has observed features visible on the available mammogram(s). We assume that the patient's utility only depends on the malignancy of the underlying disease and whether the patient receives sufficient testing to confirm the diagnosis and initiate subsequent treatment. It is therefore critical that the set of findings observed on mammography be correctly translated into a probability of malignancy so that the correct decision about patient management can be made. Integrating the findings into a "benign vs. malignant" diagnosis is likely responsible for much of the variation among radiologist practice effectiveness in mammography.

Our Bayesian network model can be used to formulate a differential diagnosis (a ranked list of diagnoses in decreasing order of $P\{d|f\}$) by entering the patient risk factors and findings seen on mammography and calculating the posterior probability distribution over diseases. If the model puts a very high probability on a particular disease, this indicates that the model believes this is the mammographic diagnosis. If more than one disease shares similar probability mass, then the model suggests more than one diagnosis should be considered in the mammographic diagnosis. The model can also give the probability of benign or malignant disease from the probability distribution in the BENIGN/MALIGNANT node (Figure 27.2)

which corresponds to the "B vs. M" node in the influence diagram (Figure 27.3).

To simplify the process of entering test cases into the network, we created a web-based data entry form (Figure 27.4). The web form corresponds to all observations that might be made for a particular lesion in a patient. The user enters the observations that apply to a patient and then submits the web form. We make a distinction between a feature that is observed to be present, a feature that is observed to be absent, and the lack of an observation about the feature. The evidence is submitted to our model and the posterior probability distribution is reported back to the user as a ranked differential diagnosis list, with the most probable diagnosis at the top of the list (Figure 27.5).

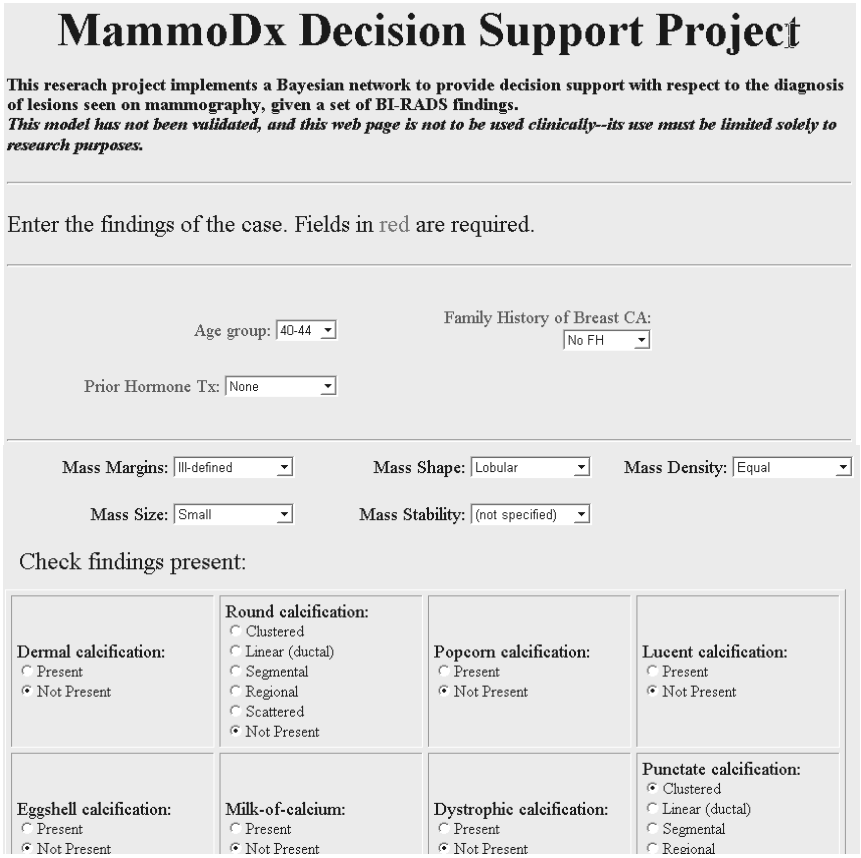
We evaluated the quality of mammographic diagnoses made by the model by entering several mammography cases in which the actual diagnosis was known (established by biopsy). In addition, we entered 105 cases from a teaching atlas of mammography [23] that contains sufficient clinical information and mammographic descriptors to enter these cases into the Bayesian network. To summarize the varying sensitivity and specificity at different probability thresholds, we built a receiver operating characteristic (ROC) curve using the ROCKIT 0.9B software (http://www-radiology.uchicago.edu/krl/toppage11.htm).

27.2.3 Using the model to evaluate concordance of mammography with biopsy results

Once a patient has a mammogram and the findings have been recognized by the radiologist, we can compute a post-test probability of malignancy. If that probability is high enough, then the radiologist will perform a biopsy so that a pathologist can make a more definitive, histological diagnosis.

Unfortunately, the biopsy test is imperfect and sampling error might occur. If the biopsy does not contain a sample of the lesion or the pathologist fails to observe the lesion cells, then the pathologist might fail to recognize malignant disease. Therefore, it is very important to correlate the histologic results from breast biopsy with the mammography findings [24-26]. The error rate can be as high as 3.3-6.2% in 14-gauge large-core needle biopsy, and 70% of these errors can be recognized immediately through careful correlation of the gross and/or histologic data with the mammography imaging findings [27-29]. Another tissue sampling technique, 11-gauge stereotactic vacuum-assisted biopsy, is associated with a lower but still significant sampling error rate, 0.8-1.7%. These sampling errors are also immediately detectable with careful imaging-histologic correlation [29-31].

Figure 27.4 A web form used to submit data on a patient to the Bayesian network model of mammography diagnosis (only a portion of the form is shown). Any finding that is not completed is treated as unobserved evidence.



Thus, breast imaging experts recommend that mammography images should be correlated with the pathology results [24-26]. This can be onerous in high volume settings, so an automated method to correlate biopsy results with mammography findings would be highly desirable.

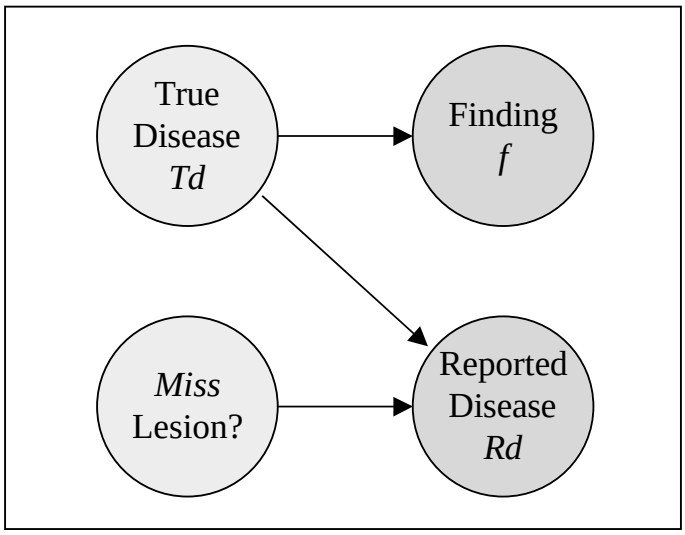
Our model can assess concordance of mammographic image findings with histologic results from biopsy by recognizing that there can be a biopsy sampling error, which we denote by *miss*. Our model then uses the radiological findings, *f*, and the pathologist's reported disease, *Rd*, as evidence, and computes a posterior probability, $P\{miss|Rd,f\}$. This prob-

Figure 27.5 A list of diseases and the posterior probabilities. The diseases are ranked with the most likely disease first (an incomplete list is shown). These results are generated from the observations entered into the form in Figure 27.4.

Ranked probabilities of diseases:

n	Disease	p
1	FA	0.97423
2	DCNOS	0.00738
3	FC	0.00722
4	DCISDCNOS	0.00453
5	FF	0.00434
6	DCIS	0.00123
7	Cy	0.00071
8	Pap	0.00015
9	RS	0.00007
10	PapCA	0.00004
11	ADH	0.00002

Figure 27.6 Belief network showing a constant chance of biopsy sampling error, “Miss Lesion?”, and the relationship between the true disease and the observed radiological findings and the pathologist’s disease report. The radiologic finding and the disease report are observed, while the other nodes are not observed.



ability is based on the same relationships between diseases and findings that we discussed earlier. The radiologist should consider performing another biopsy when *Rd* for a benign disease has a high enough $P\{miss|Rd,f\}$, but if that probability is low enough or *Rd* is malignant then the radiologist can be confident that malignant disease has not been overlooked due a sampling error.

To obtain the formula for $P\{miss|Rd,f\}$, we need to make several assumptions, some of which are shown in the belief network (Figure 27.6): (1) the finding, *f*, and disease report, *Rd*, are observed while the true disease, *Td*, and whether there was a sampling error, *miss*, are not; (2) the finding is independent of the disease report and error if we knew the true disease; the sampling error is equally likely to happen with any patient and any findings, denoted $P\{miss\}$; and (4) if there is no sampling error, then *Rd*=*Td* and otherwise the disease will be observed at the prevalence rate, $P\{Rd\}=P\{Td\}$. In that case, given any finding, *f*, there are two possibilities: a sampling error which produces report *Rd* with probability $P\{miss\}P\{d\}$ or a concordant diagnosis which yields report *Rd*=*Td* with probability $(1-P\{miss\})P\{d|f\}$.

$$P\{miss \mid Rd, f\} = \frac{P\{miss\}P\{d\}}{P\{miss\}P\{d\} + (1 - P\{miss\})P\{d \mid f\}} = \frac{1}{1 + \dfrac{(1 - P\{miss\})P\{d \mid f\}}{P\{miss\}P\{d\}}}$$

Thus, our model can produce a probability indicating how likely a biopsy samples a lesion seen on mammography (i.e., whether the biopsy is concordant with mammography findings). This can be very useful to the radiologist to help identify those cases that are likely not to be concordant, and thus require further evaluation.

We evaluated the ability of our model to assess the concordance of breast biopsy results with mammography by entering cases into our model that had breast biopsy. We included 92 consecutive cases having 14-gauge and 11-gauge biopsies. A panel of expert radiologists reviewed each case and determined the concordance between the pathology and the mammography findings as concordant ("C") or non-concordant ("N"). The experts used the following guidelines that are generally used in assessing concordance: (1) histologic documentation of microcalcifications when the mammographic abnormality contained microcalcifications; (2) histologic explanation for the imaging pattern (e.g., histologic explanation for a mass such as

fibroadenoma or focal fibrosis in contrast to benign breast tissue); and (3) histologic explanation for abnormalities with a high pre-test probability of cancer (either a diagnosis of cancer or specific histology explaining the suspicious mammography findings) [26]. In our series of 92 cases, condition (1) was satisfied in all cases; thus, concordance hinged on agreement between the pathology report and the mammographic findings.

We used the concordance determination from this expert panel as our gold standard for testing our model, and compared these assessments with the probability *P{miss|d,f}* produced by the Bayesian network. Since different values of *P{miss|d,f}* can be used as a threshold for categorizing a case as "C" or "N," we constructed an ROC curve to quantify the performance of the Bayesian network across different thresholds in the concordance assessment task.

27.3 RESULTS

27.3.1 Using the model to make a diagnosis on mammography

We tested several cases (in which the diagnosis was known) to evaluate the behavior of the model. Table 27.2 shows the probability distribution for the following cases as well as the probability for the categorized diagnosis of "benign" and "malignant" disease. No probability is truly zero but many are rounded to zero when we only display four decimal places.

Case 1 A 40 year old female with no family history or hormone use has a mammogram which demonstrates a spiculated mass with associated linear and branching calcifications. According to literature and expert opinion, a spiculated mass is typical for ductal carcinoma. The branching calcifications suggest an intraductal component. In this case our model generates the following probabilities: DCNOS+DCIS diagnosis is most likely with a 95% post-test probability. DCNOS alone has a post-test probability of 4.5%, and DCIS alone is unlikely. Variations of this scenario illustrate how the probabilities change as features differ.

Case 2 A patient with similar demographic characteristics has a spiculated mass without calcifications detected on her mammogram. This finding elicits an increased post-test probability of DCNOS to 88%. DCNOS+DCIS decreases to 2.9%, and again DCIS is unlikely.

Table 27.2 Differential diagnosis as well as summation into management categories with associated post-test probabilities for example cases. Boldface type indicates the most likely diagnoses.

Disease	Pre-test	Case 1	Case 2	Case 3	Case 4
DCNOS	0.0090	**0.0451**	**0.8851**	0.0011	0.0047
DCIS	0.0019	0.0003	0.0000	**0.7053**	0.0000
DCNOS+ DCIS	0.0012	**0.9502**	**0.0285**	**0.0892**	0.0001
LC	0.0009	0.0000	0.0000	0.0007	0.0000
LCIS	0.0008	0.0000	0.0000	0.0007	0.0000
LC/ LCIS	0.0001	0.0000	0.0000	0.0001	0.0000
TubCA	0.0001	0.0005	0.0093	0.0000	0.0000
PapCA	0.0002	0.0000	0.0009	0.0014	0.0004
MedCA	0.0001	0.0000	0.0002	0.0000	0.0040
CollCA	0.0001	0.0000	0.0002	0.0000	0.0040
Phy	0.0010	0.0000	0.0000	0.0000	0.0005
Mets	0.0010	0.0024	0.0474	0.0001	0.0001
RS	0.0010	0.0000	0.0000	0.0008	0.0000
Cy	0.0700	0.0000	0.0002	0.0122	**0.7749**
FA	**0.1200**	0.0014	0.0274	0.0218	**0.1815**
FC	**0.1300**	0.0000	0.0003	0.1098	0.0000
Ham	0.0001	0.0000	0.0000	0.0000	0.0000
FF	0.0050	0.0000	0.0000	0.0014	0.0137
FN	0.0050	0.0000	0.0000	0.0042	0.0000
SecDis	0.0010	0.0000	0.0000	0.0008	0.0000
POC	0.0010	0.0000	0.0000	0.0008	0.0000
SL	0.0230	0.0000	0.0000	0.0181	0.0130
LN	0.0300	0.0000	0.0000	0.0253	0.0026
Pap	0.0020	0.0000	0.0002	0.0000	0.0004
Normal	**0.5945**	0.0000	0.0002	0.0051	0.0000
Benign	**0.9738**	**0.0014**	**0.0284**	**0.2007**	**0.9867**
Malignant	**0.0262**	**0.9986**	**0.9716**	**0.7973**	**0.0133**

Case 3 The only finding, in a similar patient, is linear calcifications in a clustered distribution. The post-test probability for DCIS increases to 70%. DCNOS+DCIS has a post-test probability of 8.9% and DCNOS alone is .001%. These probabilities are consistent with the pathophysiology of the disease as described above.

Case 4 A 50 year old patient has a mammogram that demonstrates a round, circumscribed mass. This is our example of a "probably benign" finding. Our system reveals that with these findings there is a 1.3% chance of malignancy in this setting. This is consistent with the radiology literature [32].

We also conducted an evaluation of the Bayesian network in a larger number of test cases selected from a teaching atlas (Section 27.3.2). The performance of the model on these test cases is summarized by the ROC curve shown in Figure 27.7. The area under the ROC curve is 0.95, which compares favorably to that of an earlier Bayesian network model, 0.88 [33]. In fact, our system compares favorably with several other computer diagnostic aids developed in the domain of screening mammography in which a similar area under the ROC curve methodology was used to evaluate these systems. Two different studies tested neural network models, reporting area under the ROC curve (Az) values of 0.85 [34] and 0.76 [35]. Finally, a survey of US radiologists evaluating performance used a test set containing 79 cases of which 45 were malignant (56% malignant). The average Az value for these radiologists was .85 [36]. We realize that the Az value can be influenced by the composition of the test set used to generate the ROC curve, but this was the benchmark used for the other procedures and allows a first-order comparison of the different methodologies. We believe that our results compared with the prior studies are encouraging and suggest that our model can assist in making a mammographic diagnosis.

27.3.2 Using the model to evaluate concordance of mammography with biopsy results

Of the 92 total cases evaluated for concordance, 3 were non-concordant. Thus, the non-concordance rate was 3.3%, which is comparable to that reported previously [24-29]. In most of the cases that the expert panel determined to be concordant, the model generated an extremely low $P\{miss|d,f\}$, strongly predicting concordance (Figure 27.8). The model's assessment of concordance between mammography and pathology, $P\{miss|d,f\}$, was extremely high ($P\{miss|d,f\}$ less than 0.02) in 75 of the 92 cases (all concordant cases). $P\{miss|d,f\}$ was 2-7% in 5 cases, and 23-28% in 3 cases; all of these cases were also concordant. In the remaining 9 cases, $P\{miss|d,f\}$ was 41% and greater; 3 of these cases were the ones considered

Figure 27.7 ROC analysis of 105 teaching cases. TPF: true positive fraction; FPF: false positive fraction.

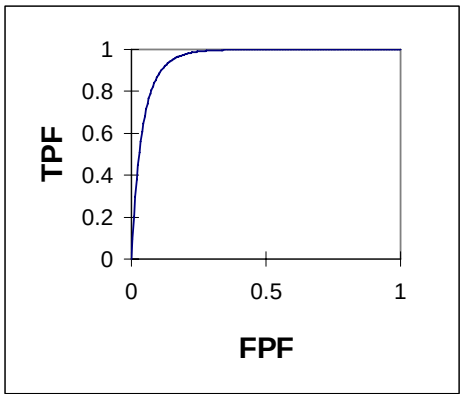
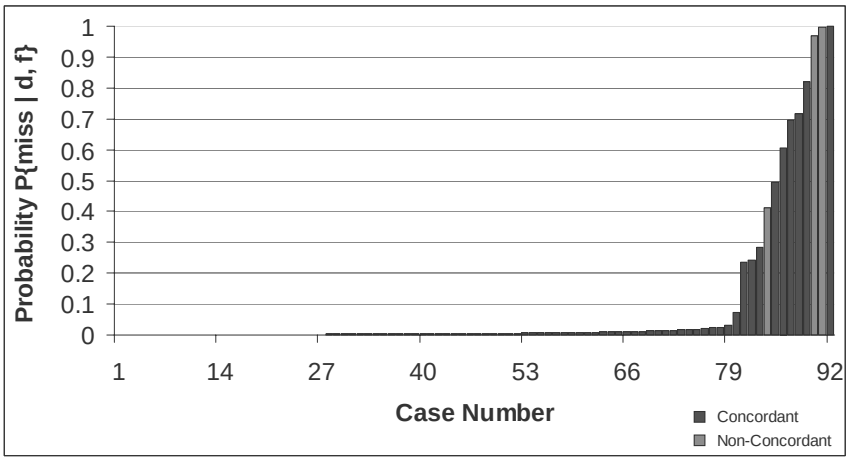
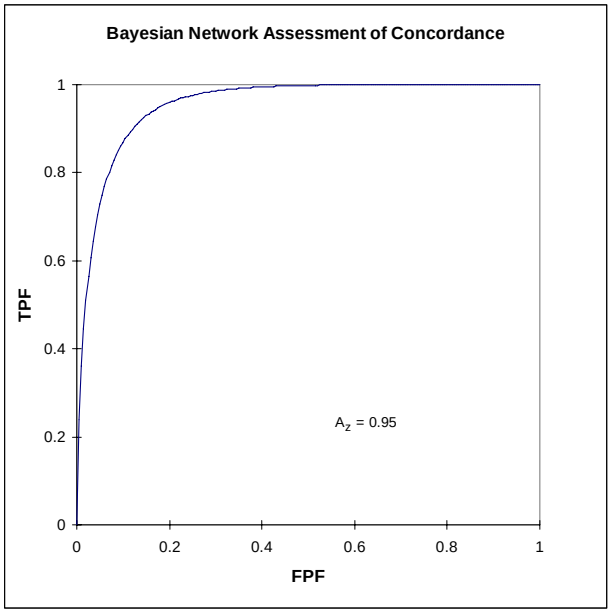


Figure 27.8 Histogram of 92 cases of mammography-biopsy correlation. In most of the cases that are concordant, the model predicts the probability of sampling error (*P{miss|d,f}*) is extremely low.



non-concordant. Using a threshold on *P{miss|d,f}* of 40% and assuming that detecting a non-concordant case is a "positive" case, there were 3 true positives, 83 true negatives, 6 false positives, and 0 false negatives (or 100% sensitivity and 93% specificity). Actually, it is more likely that a radiologist would prefer a lower threshold on *P{miss|d,f}*, such as 7%, for predicting non-concordance, which would lower the specificity to 90% with 100% sensitivity.

Figure 27.9 ROC analysis of performance of the model in assessing concordance of mammography findings with pathology. TPF: true positive fraction; FPF: false positive fraction.



An alternative summary of the performance of the model for assessing concordance is ROC analysis (Figure 27.9). The area under the curve on this graph is 0.95.

These results suggest that our model can discern those patients whose biopsy results are concordant with mammography findings. Consequently, many mammography-histologic correlations can be accurately assessed using the model, reserving only those cases where the model is uncertain (e.g., $P\{miss|d,f\} > 7\%$) for review by the radiologist. With the threshold of 7%, 80 of the 92 cases (87%) would not have required the radiologist to manually correlate mammography findings with histopathology. This is an important benefit because these correlations are labor-intensive and consume a considerable amount of radiologist time. In some busy practices, such correlation is not even routinely done. Even if the radiologist were very conservative with the model's predictions and reviewed any case for which the probability of sampling error given by the model is 1% or greater, more than half of the cases (55%) would not require manual review.

The probability value, $P\{miss|d,f\}$, produced by our system can be useful beyond establishing a threshold. First, it may encourage the pathologist to be more specific in benign disease diagnosis, to allow more concordance with the radiological findings. Second, a high probability suggests discordance and possible sampling error, prompting further review by the radiologist. We reviewed the cases that the model predicted to be discordant with high probability, but were called concordant by the expert panel. These cases were particularly complicated clinical scenarios, such as a diagnosis with an uncommon imaging presentation. These cases require expert evaluation, so a discordant assessment by the model is actually desirable behavior. Thus, our model can be useful in categorizing cases where concordance evaluation by a physician is needed.

27.4 AVENUES FOR FUTURE RESEARCH AND CONCLUSION

Our preliminary experience using a Bayesian network to model the uncertainties associated with mammography diagnosis appear promising. The model may provide several benefits. First, the model provides a normative approach to integrating findings observed by the radiologist into a ranked list of diagnoses. Second, the probabilities of disease given observed findings can be integrated with biopsy results to predict which cases are likely to be discordant, assisting patient management. Third, the model makes the mammographic decision making process explicit, providing radiologists of all levels of expertise a basis for communication and practice improvement.

The benefit of a normative approach in mammography diagnosis is greater consistency in mammography interpretation and subsequent improved health outcomes. Theoretically, if two patients have the same risk factors and the same abnormalities on mammography, they should have the same differential diagnosis and subsequent workup. However, previous studies have shown great variation in mammography practice [6-8]. Much of this variation can be attributed to how the patient risk factors and mammography findings are integrated into a differential diagnosis. Our model will produce consistent results with consistent inputs, so we would expect this to reduce some of the variation in mammography practice currently observed. Of course, this assumes that the radiologist is able to consistently detect the pertinent abnormalities on the mammogram in the first place, a fundamental task in radiology. There will still likely remain variation among radiologists with respect to identifying abnormalities on mammography images and assigning BI-RADS descriptors, but at least the variation in decisions based on these findings can be minimized using our Bayesian network.

Beyond assisting mammography diagnosis, the model's predictions and other information such as biopsy results can be integrated to assist with concordance assessment. We have shown that our model can identify those patients whose results are so likely to be concordant that they do not need to be manually reviewed by the radiologist. This would be particularly helpful in practices that currently do not do imaging-histologic correlation due to time constraints.

Our probabilistic approach is designed to support, rather than supplant, physician decision making. The radiologist is the ultimate decision maker regarding imaging-histologic correlation in difficult cases; our model can help identify those cases that are most suspicious and would benefit most from the expertise of the radiologist.

We have shown that our model performs well using the gold standard of a panel of expert radiologists. The gold standard for detecting breast biopsy sampling error is long term imaging and clinical follow-up to ensure that patients do not subsequently develop breast cancer. We plan to conduct studies evaluating our model using this preferred gold standard, which will allow us to ascertain better how well our model performs in assessing mammographic concordance. In the future, with refinement, our system may be able to improve the radiologist's ability to detect sampling error, using the gold standard of long term follow-up.

The radiology community has only incorporated a small portion of the BI-RADS descriptors into the decision-making process in this field. The entire lexicon, with its probabilistic underpinnings, when coupled with our Bayesian model has great potential to communicate quantitative probabilistic information that will aid management decisions. Our model relates benign and malignant breast diseases to BI-RADS descriptors and allows us to integrate radiological observations in a principled fashion. In addition, BI-RADS descriptors are crisply defined, with atlases showing examples of their proper usage, helping to reduce variability.

We are pursuing several other directions. First, we are collecting a larger series of confirmed cases to validate the model. This is important because we are continuing to improve our conditional probability distributions. Second, we are embarking on a large prospective data collection project in which routine mammogram cases will be compiled and their findings recorded to establish more accurately the probabilities of particular findings given disease. This information may be used to update conditional probability distributions in our Bayesian network that had little supporting data when it was originally built. Third, we will be evaluating the value of

information of BI-RADS descriptors in the model which can suggest to the radiologist particular features on the mammogram that should be checked.

Finally, we wish to compare the diagnostic performance of the model directly with experts and non-experts to determine how well it can elevate performance of mammographers. Ultimately, our goal is to refine our system as an aid in normative decision making and education. We hope to demonstrate that the accuracy and quality of medical practice is elevated among practitioners of varying experience using this approach. We believe that with further testing and use our model will help to elevate the standard of all mammography practice and improve the quality of patient care.

References

[1] Greenlee, R.T., M.B. Hill-Harmon, T. Murray, and M. Thun (2001). Cancer statistics, 2001. *Cancer Journal for Clinicians*, 51, 15-36.

[2] Baker, L.H. (1982). Breast Cancer Detection Demonstration Project: five-year summary report. *Cancer Journal for Clinicians*, 32, 194-225.

[3] Houn, F., M.L. Elliott, and J.L. McCrohan (1995). The Mammography Quality Standards Act of 1992. History and philosophy. *Radiology Clinics of North America*, 33, 1059-1065.

[4] Pisano, E.D., *et al.* (2000). Has the Mammography Quality Standards Act affected the mammography quality in North Carolina? *American Journal of Roentgenology*, 174, 1089-1091.

[5] Sickles, E.A., D.E. Wolverton, and K.E. Dee (2002). Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology*, 224, 861-869.

[6] Ciccone, G., P. Vineis, A. Frigerio, and N. Segnan (1992). Inter-observer and intra-observer variability of mammogram interpretation: a field study. *European Journal of Cancer*, 28A, 1054-1058.

[7] Elmore, J.G., et al. (2002). Screening mammograms by community radiologists: variability in false-positive rates. *Journal of the National Cancer Institute*, 94, 1373-1380.

[8] Elmore, J.G., C.K. Wells, C.H. Lee, D.H. Howard, and A.R. Feinstein (1994). Variability in radiologists' interpretations of mammograms. *New England Journal of Medicine*, 331, 1493-1499.

[9] Sirovich, B.E. and H.C. Sox, Jr. (1999). Breast cancer screening. *Surgery Clinics of North America*, 79, 961-990.

[10] Harris, R. (1997). Variation of benefits and harms of breast cancer screening with age. *Journal of the National Cancer Institute Monographs*, 139-143.

[11] Christiansen, C.L., et al. (2000). Predicting the cumulative risk of false-positive mammograms. *Journal of the National Cancer Institute*, 92, 1657-1666.

[12] Brown, M.L., F. Houn, E.A. Sickles, and L.G. Kessler (1995). Screening mammography in community practice: positive predictive value of abnormal findings and yield of follow-up diagnostic procedures. *American Journal of Roentgenology*, 165, 1373-1377.

[13] American College of Radiology (1998). *Breast Imaging Reporting and Data System (BI-RADS).* American College of Radiology, Reston, VA.

[14] Swets, J.A., et al. (1991). Enhancing and evaluating diagnostic accuracy. *Medical Decision Making*, 11, 9-18.

[15] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, San Mateo, CA.

[16] Colditz, G.A., et al. (1995). The use of estrogens and progestins and the risk of breast cancer in postmenopausal women. *New England Journal of Medicine*, 332, 1589-1593.

[17] Ries, L.A.G. and National Cancer Institute (U.S.). Division of Cancer Prevention and Control. (1997). *SEER Cancer Statistics Review, 1973-1996.* National Cancer Institute, Bethesda, MD.

[18] Slattery, M.L. and R.A. Kerber (1993). A comprehensive evaluation of family history and breast cancer risk. The Utah Population Database. *Journal of the American Medical Association*, 270, 1563-1568.

[19] Monsees, B.S. (1995). Evaluation of breast microcalcifications. *Radiology Clinics of North America*, 33, 1109-1121.

[20] Evans, W.P. (1995). Breast masses. Appropriate evaluation. *Radiology Clinics of North America*, 33, 1085-1108.

[21] Howard, R.A. and J.E. Matheson (1984). Influence diagrams. In *The Principles and Applications of Decision Analysis*, R.A. Howard and J.E. Matheson, Eds., Strategic Decisions Group, Menlo Park, CA.

[22] Shachter, R.D. (1986). Evaluating influence diagrams. *Operations Research*, 34, 871-882.

[23] Tabár, L. and P.B. Dean (1983). *Teaching Atlas of Mammography.* Thieme Medical Publishers, New York.

[24] Berg, W.A., et al. (1996). Lessons from mammographic-histopathologic correlation of large-core needle breast biopsy. *Radiographics*, 16, 1111-1130.

[25] Ioffe, O.B., W.A. Berg, S.G. Silverberg, and D. Kumar (1998). Mammographic-histopathologic correlation of large-core needle biopsies of the breast. *Modern Pathology*, 11, 721-727.

[26] Liberman, L., et al. (2000). Imaging-histologic discordance at percutaneous breast biopsy. *Cancer*, 89, 2538-2546.

[27] Jackman, R.J., et al. (1999). Stereotactic, automated, large-core needle biopsy of nonpalpable breast lesions: false-negative and histologic underestimation rates after long-term follow-up. *Radiology*, 210, 799-805.

[28] Lee, C.H., L.E. Philpotts, L.J. Horvath, and I. Tocino (1999). Follow-up of breast lesions diagnosed as benign with stereotactic core-needle biopsy: frequency of mammographic change and false-negative rate. *Radiology*, 212, 189-194.

[29] Liberman, L. (2000). Centennial dissertation. Percutaneous imaging-guided core breast biopsy: state of the art at the millennium. *American Journal of Roentgenology*, 174, 1191-1199.

[30] Philpotts, L.E., N.A. Shaheen, D. Carter, R.C. Lange, and C.H. Lee (1999). Comparison of rebiopsy rates after stereotactic core needle biopsy of the breast with 11-gauge vacuum suction probe versus 14-gauge needle and automatic gun. *American Journal of Roentgenology*, 172, 683-687.

[31] Burbank, F. (1997). Stereotactic breast biopsy: comparison of 14- and 11-gauge Mammotome probe performance and complication rates. *American Surgeon*, 63, 988-995.

[32] Sickles, E.A. (1995). Management of probably benign breast lesions. *Radiology Clinics of North America*, 33, 1123-1130.

[33] Kahn, C.E., Jr., L.M. Roberts, K. Wang, D. Jenks, and P. Haddawy (1995). Preliminary investigation of a Bayesian network for mammographic diagnosis of breast cancer. *Proceedings of the Annual Symposium on Computing Applied to Medical Care*, 208-212.

[34] Baker, J.A., P.J. Kornguth, J.Y. Lo, M.E. Williford, and C.E. Floyd, Jr. (1995). Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. *Radiology*, 196, 817-822.

[35] Jiang, Y., et al. (1999). Improving breast cancer diagnosis with computer-aided diagnosis. *Academic Radiology*, 6, 22-33.

[36] Beam, C.A., P.M. Layde, and D.C. Sullivan (1996). Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. *Archives of Internal Medicine*, 156, 209-213.