A Probabilistic Expert System That Provides Automated Mammographic–Histologic Correlation: Initial Experience

Elizabeth S. Burnside1,2
Daniel L. Rubin3
Ross D. Shachter4
Rita E. Sohlich5
Edward A. Sickles1

OBJECTIVE. We sought to determine whether a probabilistic expert system can provide accurate automated imaging–histologic correlations to aid radiologists in assessing the concordance of mammographic findings with the results of imaging-guided breast biopsies.

MATERIALS AND METHODS. We created a Bayesian network in which Breast Imaging Reporting and Data System (BI-RADS) descriptors are used to convey the level of suspicion of mammographic abnormalities. Our system is a computer model that links BI-RADS descriptors with diseases of the breast using probabilities derived from the literature. Mammographic findings are used to update pretest probabilities (prevalence of disease) into posttest probabilities applying Bayes’ theorem. We evaluated the histologic results of 92 consecutive imaging-guided breast biopsies for concordance with the mammographic findings during radiology–pathology review sessions. First, radiologists with no knowledge of the biopsy results chose BI-RADS descriptors for the mammographic findings. After the histologic diagnosis was revealed, the radiologists assessed concordance between the pathologic results and the mammographic findings. We then input the information gathered from these sessions into the Bayesian network to produce an automated mammographic–histologic correlation.

RESULTS. We had a sampling error rate of 1.1% (1/92 biopsies). Our expert system was able to integrate pathologic diagnoses and mammographic findings to obtain probabilities of sampling error, thereby enabling us to identify the incorrect pathologic diagnosis with 100% sensitivity while maintaining a specificity of 91%.

CONCLUSION. Our probabilistic expert system has the potential to help radiologists in identifying breast biopsy results that are discordant with mammographic findings and discovering cases in which biopsy sampling errors may have occurred.

Correlating histologic results for imaging-guided breast biopsy specimens with mammographic findings is critically important [1–3] in identifying sampling errors and avoiding delays in diagnosing breast cancers. Large-core 14-gauge needle biopsy has been shown to have a rate of discordance of 3.3–6.2%, whereas stereotactic 11-gauge vacuum-assisted biopsy is associated with an 0.8–1.7% rate of discordance [4–6]. Seventy percent of breast cancers missed at biopsy can be recognized promptly through careful correlation of the gross pathologic and histologic data with the imaging findings [6–8].

Experts advocate that the final pathologic results for samples from every percutaneous breast biopsy be directly correlated with breast images obtained before, during, and after the biopsy. In selected cases, close interaction between the radiologist and the pathologist is necessary to correlate the mammographic and histologic characteristics of an abnormality [1, 3]. Although potentially difficult and labor-intensive, accurate imaging–histologic correlation has been shown to help physicians to avoid missing breast cancers. In hopes of improving the accuracy of this important task, we have developed a probabilistic expert system that can provide accurate, automated imaging–histologic correlation to aid radiologists in assessing the concordance of mammographic findings with histologic results from imaging-guided breast biopsies.

We created a Bayesian network that is based on the ability of Breast Imaging Reporting and Data System (BI-RADS) descriptors [9] generated by a radiologist to convey...
the level of suspicion of mammographic abnormalities. Our Bayesian network is a computer model that links BI-RADS descriptors with diseases of the breast applying probabilities derived from the literature. Using the mammographic findings, the system applies Bayes’ theorem to update a patient’s baseline disease risk into posttest probabilities of specific diseases of the breast as well as an overall probability of malignancy. In a previous study, we showed that our Bayesian network succeeded in accurately estimating the probability that imaging findings represent various pathophysiologic conditions [10]. We hypothesize that our system is robust and flexible enough to accurately assess the likelihood of imaging–histologic concordance on the basis of these same core probabilities.

Materials and Methods

The Expert System

To build the mammography Bayesian network, we first identified 25 diseases of the breast from the literature, 11 of which are malignant and 14 of which are benign. The complex pathophysiology of benign and malignant breast disease presented certain challenges in developing our model. For example, two of the benign diseases have a clinically significant association with or tendency to progress to malignancy: atypical ductal hyperplasia and radial scar [11, 12]. Standard management for either of these diagnoses has historically been surgical excision [12–18]. Excision continues to be used in cases of atypical ductal hyperplasia, but recent clinical evidence has shown that some radial scars can be safely diagnosed with percutaneous biopsy [19]. Because both of these diagnoses can be associated with malignancy, we consider atypical ductal hyperplasia and radial scar to be high-risk conditions.

A challenging aspect of building a Bayesian network for breast cancer is the pathophysiology of breast malignancy because there is a transition from in situ to invasive disease. To simplify the model, we assumed that there was a single uncertain variable, “disease,” that can take on one value—either “normal” or a value corresponding exactly to one of the 25 diseases. We assumed that it is impossible for two distinct breast diseases to arise together, although in situ and invasive breast cancers are commonly encountered simultaneously. In the network, the coexistence of two different diseases would violate the assumption that the disease variable has mutually exclusive and collectively exhaustive states [20]. Our solution was to include a mixed diagnosis for ductal and lobular neoplasms to fulfill the requirement of mutual exclusivity (Table 1). We did not include other simultaneous lesions such as atypical ductal hyperplasia occurring with ductal carcinoma in situ (DCIS) or DCIS occurring with lobular carcinoma in situ. In addition, we assumed that benign and malignant diagnoses such as fibrocystic change and DCIS did not coexist in a single area. We believe this assumption did not significantly affect the performance of our model, but we plan future study in this area. Although not all possible diseases of the breast are included in our system, those excluded are extremely rare and were not seen in our series.

The BI-RADS lexicon is the foundation for the observable features in our Bayesian network [9]. BI-RADS consists of 43 descriptors organized into a hierarchical lexicon (Fig. 1). For this initial experiment, we excluded five descriptors to simplify our model: skin thickening, trabecular thickening, nipple retraction, skin retraction, and asymmetric breast tissue. These findings are either late signs of breast cancer or benign features rarely observed in the population of patients undergoing percutaneous biopsy. None of the features excluded from our model was observed by the radiologists in our series.

In a Bayesian network, uncertain variables that affect the probability of disease are represented as “nodes” (Fig. 2), which are data structures that store probabilities and can be understood by both humans and computers. In our system, the disease node represents the diagnoses of the breast enumerated in Table 1. This node stores the prior probabilities of disease (the prevalence of each breast disease) determined by specific risk factors such as the patient’s age, family history, and use of hormone replacement therapy. Each of the remaining nodes in the network represents possible findings on a mammogram and contains a conditional probability table that relates the findings to the variables that affect them. The conditional probability table has a row for each possible combination of parent values. For example, the disease node is a parent of “calcifications (Ca++)” fine or linear node (and, of course, the Ca++ fine or linear node is a child of the disease node).

The conditional probability table for Ca++ fine or linear node contains a row for each disease of the breast. The structure of the model is composed of directional arcs (Fig. 2) that encode the conditional dependence and independence relationships among the variables. The absence of an arc represents conditional independence. Each arc implies an influence (or in some cases, a causal link) between the nodes joined by that arc.

To construct our computer model and perform inference, we used the GeNe modeling environment developed by the Decision Systems Laboratory of the University of Pittsburgh [21]. The model is structured on the assumption that all the BI-RADS descriptors except breast density are children of the disease node. We modeled the calcification descriptors as conditionally independent manifestations of disease. The distribution descriptors of each type of calcification are the mutually exclusive states of the corresponding calcification nodes. If calcifications are described, distribution choices are available (e.g., fine or linear, clustered, segmental, regional, diffuse, or scattered) in the conditional probability table of each calcification node. The structure of our model reflects the hierarchical structure of the BI-RADS lexicon (Fig. 1): if a mass is the finding of interest, the underlying mass descriptors are available to further modify the finding of “mass.” The descriptors themselves (e.g., in the case of the shape of the mass, the descriptors are “round,” “ovoid,” “lobular,” or “irregular”) are stored in the conditional probability table with an associated probability distribution. We model special cases and associated findings as conditionally independent expressions of disease. We use the breast density descriptors in the BI-RADS lexicon to influence the visualization of masses (i.e., the breast density node influences whether a mass is present, absent, or obscured). The node with a double border is a deterministic node, meaning it assigns the respective disease types into three categories: benign, high-risk, and malignant. The value of this node, which is the probability of malignancy, can provide decision support to guide case management.

### TABLE I Diseases of the Breast Considered in the Bayesian Network

<table>
<thead>
<tr>
<th>Malignant (n = 11)</th>
<th>Benign (n = 14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Invasive ductal carcinoma not otherwise specified</td>
<td>Cyst</td>
</tr>
<tr>
<td>Ductal carcinoma in situ</td>
<td>Fibroadenoma</td>
</tr>
<tr>
<td>Ductal carcinoma in situ and ductal carcinoma not otherwise specifieda</td>
<td>Papilloma</td>
</tr>
<tr>
<td>Lobular carcinoma</td>
<td>Fibrocystic change</td>
</tr>
<tr>
<td>Lobular carcinoma and lobular carcinoma in situa</td>
<td>Hamartoma</td>
</tr>
<tr>
<td>Tubular carcinoma</td>
<td>Lymph node</td>
</tr>
<tr>
<td>Papillary carcinoma</td>
<td>Focal fibrosis</td>
</tr>
<tr>
<td>Medullary carcinoma</td>
<td>Fat necrosis</td>
</tr>
<tr>
<td>Colloid carcinoma</td>
<td>Secretory disease</td>
</tr>
<tr>
<td>Phyllodes tumor</td>
<td>Postoperative change</td>
</tr>
<tr>
<td>Metastasis</td>
<td>Skin lesion</td>
</tr>
<tr>
<td></td>
<td>Radial scarb</td>
</tr>
<tr>
<td></td>
<td>Atypical ductal hyperplasiab</td>
</tr>
<tr>
<td></td>
<td>Lobular carcinoma in situ</td>
</tr>
</tbody>
</table>

aTwo individual diseases presenting simultaneously.

bConsidered high-risk findings.
Automated Mammographic–Histologic Correlation

Originally, this system was designed to use mammographic findings and demographic factors to predict the likelihood of malignancy [10], but one of the advantages of a Bayesian network is that it can be used for other reasoning tasks in the same domain. In this experiment, we used this model for a different purpose: imaging–histologic correlation. Thus, some of the features important for mammographic interpretation (such as the probability that calcifications may be dermal) become unimportant in the context of imaging–histologic correlation (because it is difficult to mistakenly biopsy skin calcifications). The importance of each node in the model is simply a consequence of the different circumstances under which we use the expert system. Performance is not affected because the fundamental probabilities and relationships are generalizable if the outcome of interest is appropriately calculated. The probability of sampling error, which is the output of interest in our experiment, is based on a probability calculation that we now describe.

The system was initially designed to apply Bayes’ theorem to the pretest probability, \( P(d) \), of disease, \( d \)—in other words, the prevalence of disease, \( P(d) \)—to compute the probability of disease associated with given imaging findings, \( f \). This probability is represented by \( P(d | f) \). Our system functions in this same manner, assessing the concordance of imaging findings with histologic results. This concordance is directly related to the probability of sampling error, \( P(\text{miss}) \). The value of this parameter can be based on the data reported in the literature or an audit of an individual practice. We chose to use 3.3% as the probability for two reasons. First, percutaneous biopsies in our practice involve a mix of sonographically guided large-core 14-gauge needle biopsies and stereotactic 11-gauge vacuum-assisted biopsies; according to reports in the literature, 3.3% is an average rate of discordance for this patient mix. In addition, a retrospective review of the data in our series reveals a rate of discordance of 3.3% [4–8]. We are interested in minimizing the probability of a radiologist mistakenly accepting an erroneous biopsy result. Given a specific histologic diagnosis \( d \) and a constellation of findings \( f \), \( P(\text{miss} | d, f) \) represents the case-specific chance that a sampling error has occurred. If we assume that when sampling errors occur, the resultant histologic diagnosis represents the pretest prevalence of breast disease \( P(d) \), we can compute this probability using data available from our Bayesian network as follows:

\[
P(\text{miss} | d, f) = \frac{1}{1 + \frac{P(d | f) \times P(\text{miss})}{P(d)}}
\]

Therefore, our expert system can generate a number that indicates the likelihood of a sampling error for a histologic diagnosis and mammographic findings.

Study Design

Sonographically guided 14-gauge biopsies, stereotactic 11-gauge biopsies, and needle localizations performed for diagnosis were included in this project. Although there is little literature investigating...
the rate of sampling error in excisional biopsies, we included patients who underwent this procedure because sampling error is possible. Patients with a known cancer diagnosis who were undergoing therapeutic needle localization were excluded. At our institution, the radiologists review results of each imaging-guided breast biopsy procedure in conjunction with the imaging results. We evaluated the pathologic results for 92 consecutive imaging-guided breast biopsies for our experiment, and the radiology or pathology review sessions were conducted to reach consensus on both the BI-RADS descriptors and imaging–histologic concordance. Participants included one attending radiologist, two breast imaging fellows, and one or two radiology residents. In these sessions, the participants had access to available demographic information, such as the patient’s age, family history, hormone replacement status, prior surgeries, and personal history of breast cancer and selected the appropriate BI-RADS descriptors without knowledge of the histologic diagnosis. After the histologic diagnosis was revealed, the radiologists assessed the concordance between the histologic results and the mammographic findings. Consensus regarding BI-RADS descriptors and concordance was reached through discussion. In most cases, unanimous consensus was reached. In the few cases in which there was disagreement, the attending radiologist made the final determination. We input the information gathered in these sessions and the demographic risk factors into the Bayesian network to evaluate the automated imaging–histologic concordance in these cases predicted by the model.

Study End Points

Determining sampling error at imaging-guided biopsy was the end point of our study. We defined sampling errors as those cases resulting in a false-negative diagnosis: patients with benign biopsy results who were later found to have breast cancer at or adjacent to the biopsy site. We considered cases in which a high-risk lesion discovered at percutaneous biopsy was upgraded at excision to be underestimations rather than sampling errors [8]. Therefore, the gold standard for detecting sampling error in our study was clinical follow-up 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 negative biopsy result.

Using this gold standard, we compared the ability of our expert system as compared with the radiologists to predict sampling error. Parameters used by the radiologists to establish concordance included histologic documentation of microcalcifications when the mammographic abnormality contained microcalcifications, histologic explanation for the imaging pattern (e.g., explanation for a mass such as fibroadenoma or focal fibrosis as opposed to benign breast tissue), and histologic explanation for abnormalities with a high pretest probability of cancer (either a diagnosis of cancer or specific histologic explanation of the suspicious mammographic findings) [3]. In addition, radiologists sometimes code a finding as a BI-RADS
category 5 abnormality (highly suggestive of malignancy) for which only a diagnosis of cancer can be accepted as concordant. Therefore, the BI-RADS category can contribute to the assessment of imaging-histologic concordance. Our expert system does not currently include BI-RADS categories. It does, however, calculate a high posttest probability of malignancy for these highly suspicious findings that parallels the BI-RADS category, making a benign abnormality in these situations discordant even without considering the BI-RADS category per se. The probabilistic analysis of our system for our experiment was based on an analysis of findings rather than the radiologists’ interpretation as represented by the BI-RADS category. In our series, all cases were coded prospectively as BI-RADS category 4. Therefore, inclusion of the BI-RADS category of the findings would not have affected the assessments of concordance for the radiologists or the expert system in this series. In general, the radiologists’ assessment of concordance was based on a combination of the guidelines, the literature, and experience.

In our study, the radiologists assessed concordance as a binary variable; the histologic diagnosis was either concordant with the imaging findings, or it was not. In contrast, the output from our expert system was a continuous variable, a probability between 0 and 100 that conveyed the likelihood of sampling error.

Results

The study group consisted of 92 consecutive patients (mean age ± SD, 58.2 ± 10.5 years; range, 37–84 years) who underwent mammographically or sonographically guided breast biopsy between October 2001 and March 2002. Using P(miss / d, f), we generated a receiver operator characteristic (ROC) curve to quantify the performance of the Bayesian network in terms of the prediction of discordance (Fig. 3). This ROC curve enabled us to identify the optimal threshold for accepting a given histologic diagnosis. In breast imaging, the primary focus is to avoid missing a breast cancer. Therefore, our system is most acceptable when it detects all discordant results as defined by the radiologists (100% sensitivity). The ROC curve shows 100% sensitivity at a specificity of 93%. Using this threshold (which corresponds to P(miss / d, f) = 0.4), we found nine pathologic results that were below the level sufficient to be deemed discordant (Table 2).

The biopsy types and resultant pathologic categories included in our study are summarized in Table 3. Eight patients were lost to follow-up after 6–9 months. The remainder of patients had follow-up data available for 12 months or longer. A total of 63 patients had biopsy results that were benign without high-risk histologies. These patients were considered to be candidates for sampling error. The median follow-up period for all patients with benign biopsies was 24 months (range, 6–36 months). Of the 63 patients, 47 had mammographic follow-up (median follow-up period, 23 months; range, 6–33 months). The remaining 16 patients had clinical follow-up only (median, 25 months; range, 6–36 months). Of the patients judged by the radiologists to have discordant results, only one—a 61-year-old woman—had an excisional biopsy, which revealed benign breast tissue. The cases of two other patients—a 75-year-old woman and a 59-year-old woman—were reviewed with the pathologists who, in retrospect, recognized fibrosis that could account for a mass seen on mammography. Each of these three patients was healthy and had normal findings on follow-up mammograms.

Only one patient, a 49-year-old woman, was found to have sampling error in our series, which translates into a sampling error of 1.1% (1/92 cases) or a rate of missed cancers of 4.3% (1/23 cancers). This patient had a history of DCIS in the right breast. The mammogram of the right breast obtained 6 months after diagnosis and treatment (6-month follow-up is standard in our practice for patients who have undergone lumpectomy) revealed residual microcalcifications adjacent to the lumpectomy site. The patient underwent needle localization and excisional biopsy that showed microcalcifications in benign ducts associated with fibrocystic changes. When she returned in 11 months (5 months late for the subsequent 6-month follow-up examination), she was found to have developed more diffuse microcalcifications. A repeated biopsy revealed comedo and solid high-grade DCIS. At the subsequently performed mastectomy, seven separate foci of DCIS in multiple quadrants were found. The Bayesian network succeeded in identifying the patient as having a high likelihood of sampling error, but the panel of radiologists did not.

Accurately assessing sampling error was our outcome of interest. Cases evaluated as having discordant findings that at follow-up were found to be sampling errors were considered true-positive (radiologists, no cases; expert system, one case). Cases with findings deemed concordant that at follow-up showed no subsequent breast cancer were considered true-negative (radiologists, 88 cases; expert system, 83 cases). Cases evaluated as having discordant findings that within 12 months had a breast cancer detected were considered false-negative (radiologists, one case; expert system, no cases). Finally, cases evaluated as having discordant findings that had no subsequent breast cancer were considered false-positive (radiologists, three cases; expert system, eight cases). Our expert system succeeded in identifying the solitary case of sampling error with 100% sensitivity (95% confidence interval [CI] of 5–100%) while maintaining a specificity of 91% (95% CI, 88–94%).

Discussion

The correlation of imaging findings and histologic data involves reasoning and management of uncertainty. A well-constructed Bayesian network integrates probabilities to logically and reproducibly manage this un-
The purpose of this expert system is not to replace the radiologist in determining imaging–histologic concordance but to assign a probability of sampling error on the basis of the imaging and histologic findings in the context of the patient’s inherent risk factors.

Among the cases reviewed, our system identified biopsy results that were likely to be concordant. For example, a 39-year-old woman with no demographic risk factors underwent biopsy of a lobulated, partially obscured solid mass. The histologic examination revealed a fibroadenoma. Our Bayesian network assessed the chance of sampling error to be less than 0.2%. More than half of the patients in our study had a probability of sampling error of less than 1%.

On the other hand, our system also reliably predicted which results were likely to be discordant and therefore likely to require careful review by the radiologists. All the cases deemed to have discordant results by the radiologists were included among the cases that according to our system had a significant probability of sampling error. Because our Bayesian network generates a probability of discordance, a threshold can be defined above which a case warrants special attention. In our study, the performance of our system was optimal using a sampling error estimate of 40% as a threshold. At this threshold, our Bayesian network identified only nine cases as discordant. This group included the three cases of discordant diagnoses assigned by the radiologists. More important, it identified the one case of sampling error.

---

**Note.**—Stereotactic = stereotactic 11-gauge vacuum-assisted biopsy, excisional = surgical excision guided by needle localization.

- aConsidered positive if the patient had one or more first-degree relatives (mother, sister, or daughter) who had breast cancer. Otherwise, family history was considered negative.
- bPatient’s use of exogenous hormonal replacement therapy (HRT) at time of mammography is indicated as positive. No use of HRT is indicated as negative.
- cBI-RADS = Breast Imaging Reporting and Data System [9].
- dDerived from formula $P(\text{miss} | d, f)$, where $P(\text{miss}) = \text{probability of sampling error}$, $d = \text{disease}$, and $f = \text{imaging findings}$.
- eMulticentric ductal carcinoma in situ was diagnosed 11 months after the benign biopsy findings described.

---

### TABLE 2

<table>
<thead>
<tr>
<th>Patient Age (yr)</th>
<th>Family History of Breast Cancer a</th>
<th>HRT Status b</th>
<th>Finding on Mammography</th>
<th>BI-RADS c Descriptors of Lesion</th>
<th>Histologic Diagnosis</th>
<th>Findings as Assessed by Radiologists</th>
<th>Probability of Sampling Error d</th>
<th>Biopsy Type</th>
<th>Follow-Up (mo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>Negative</td>
<td>Negative</td>
<td>Mass</td>
<td>III-defined, irregular, high-density</td>
<td>Lymph node</td>
<td>Concordant</td>
<td>99.9</td>
<td>Stereotactic</td>
<td>22</td>
</tr>
<tr>
<td>75</td>
<td>Positive</td>
<td>Negative</td>
<td>Mass</td>
<td>III-defined, irregular, low-density</td>
<td>Fibrocystic change</td>
<td>Discordant</td>
<td>99.8</td>
<td>Stereotactic</td>
<td>23</td>
</tr>
<tr>
<td>59</td>
<td>Negative</td>
<td>Negative</td>
<td>Mass</td>
<td>III-defined, irregular, isodensity</td>
<td>Fibrocystic change</td>
<td>Discordant</td>
<td>96.8</td>
<td>Stereotactic</td>
<td>24</td>
</tr>
<tr>
<td>69</td>
<td>Negative</td>
<td>Negative</td>
<td>Mass</td>
<td>III-defined, irregular, isodensity</td>
<td>Fibrocystic change</td>
<td>Concordant</td>
<td>82.0</td>
<td>Stereotactic</td>
<td>24</td>
</tr>
<tr>
<td>48</td>
<td>Negative</td>
<td>Negative</td>
<td>Calcifications</td>
<td>Clustered, amorphous, pleomorphic, dystrophic</td>
<td>Fibrocystic change</td>
<td>Concordant</td>
<td>71.6</td>
<td>Stereotactic</td>
<td>22</td>
</tr>
<tr>
<td>49</td>
<td>Positive</td>
<td>Negative</td>
<td>Calcifications</td>
<td>Clustered, pleomorphic</td>
<td>Fat necrosis</td>
<td>Concordant</td>
<td>69.6</td>
<td>Excisional</td>
<td>11 b</td>
</tr>
<tr>
<td>79</td>
<td>Negative</td>
<td>Negative</td>
<td>Calcifications</td>
<td>Clustered, amorphous, pleomorphic</td>
<td>Fibroadenoma</td>
<td>Concordant</td>
<td>60.5</td>
<td>Stereotactic</td>
<td>24</td>
</tr>
<tr>
<td>56</td>
<td>Negative</td>
<td>Positive</td>
<td>Other</td>
<td>Focal asymmetric density</td>
<td>Postoperative changes</td>
<td>Concordant</td>
<td>49.9</td>
<td>Stereotactic</td>
<td>26</td>
</tr>
<tr>
<td>61</td>
<td>Positive</td>
<td>Negative</td>
<td>Other</td>
<td>Architectural distortion</td>
<td>Lobular carcinoma in situ</td>
<td>Discordant</td>
<td>41.1</td>
<td>Excisional</td>
<td>22</td>
</tr>
</tbody>
</table>

---

**Note.**—Stereotactic = stereotactic 11-gauge vacuum-assisted biopsy, excisional = surgical excision guided by needle localization.

- aConsidered positive if the patient had one or more first-degree relatives (mother, sister, or daughter) who had breast cancer. Otherwise, family history was considered negative.
- bPatient’s use of exogenous hormonal replacement therapy (HRT) at time of mammography is indicated as positive. No use of HRT is indicated as negative.
- cBI-RADS = Breast Imaging Reporting and Data System [9].
- dDerived from formula $P(\text{miss} | d, f)$, where $P(\text{miss}) = \text{probability of sampling error}$, $d = \text{disease}$, and $f = \text{imaging findings}$.
- eMulticentric ductal carcinoma in situ was diagnosed 11 months after the benign biopsy findings described.

---

### TABLE 3

<table>
<thead>
<tr>
<th>Biopsy Type</th>
<th>Pathologic Results</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Malignant</td>
<td>Benign</td>
</tr>
<tr>
<td></td>
<td>No.</td>
<td>%</td>
</tr>
<tr>
<td>Stereotactic 11-gauge vacuum-assisted core</td>
<td>19</td>
<td>28</td>
</tr>
<tr>
<td>Sonographically guided 14-gauge needle core</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>Excisional</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>25</td>
</tr>
</tbody>
</table>

- aHigh-risk findings were atypical ductal hyperplasia or radial scar.
Examining those cases deemed to have discordant findings by the expert system gives us insight into the system itself. The Bayesian network assigned a high likelihood of sampling error to particular imaging–histologic combinations. The cases that the Bayesian network identified as having discordant findings revealed unexpected results, even if the findings were ultimately deemed by the radiologists to be concordant. In general, the expert system calculated a high probability of sampling error when a histologic diagnosis was associated with an uncommon imaging presentation (e.g., a case in which a lymph node in a 48-year-old woman presented as an ill-defined mass); when a suspicious mammographic finding was found to be benign (e.g., a case in which architectural distortion in a 61-year-old woman yielded the histologic diagnosis of lobular carcinoma in situ or a case in which a mass with amorphous pleomorphic calcifications in a 69-year-old woman was found to be fibrocystic change); or when a patient with a history of prior breast surgery, a parameter not recorded in the Bayesian network, showed postoperative changes (a 49-year-old woman) or fat necrosis (a 56-year-old woman) at biopsy.

One might wonder how the radiologists managed to correctly judge some of these problematic findings to be discordant, whereas the Bayesian network found them discordant. Radiologists have the ability to take into account factors that are not modeled by our Bayesian network. For example, an abnormality found in the axillary tail of a 48-year-old woman may have been seen in retrospect to have a small fatty notch, a characteristic typical of an intramammary lymph node. Although the indistinct margins made the finding suspicious enough to require biopsy, retrospective analysis of the appearance and location of the finding made the histologic results acceptable to the radiologist. In a 79-year-old woman, the findings of clustered amorphous and pleomorphic microcalcifications in a fibroadenoma were unexpected. Fortunately, the postprocedural mammogram revealed that the small cluster of calcifications had been completely removed, making sampling error highly unlikely. In the cases of a 59-year-old woman and a 69-year-old woman, the radiologists’ consultation with a pathologist revealed significant fibrosis in addition to fibrocystic change, thereby accounting for the mass seen on the mammogram. We will use observations from this experiment to improve the Bayesian network by adding important variables such as the location of findings and history of prior breast surgery.

There are limitations to our study. First, we did not distinguish among 14-gauge, 11-gauge, or excisional biopsies in our population because we wanted to determine whether our system could be generalized to a population representative of true clinical practice. The intuitive belief that sampling error is more likely to occur when a smaller needle is used has been proven and reported in the literature [4, 7, 22]. By stratifying the probability of sampling error on the basis of the biopsy type, we may be able to improve performance when we test our system on a larger number of patients. Second, the histologic diagnoses recorded in our system are broad categorizations of the descriptors used in pathologic reports; unfortunately, there is no standardized lexicon for breast pathology. For example, the diagnosis of fibrocystic change usually includes a diverse collection of pathologic conditions, such as apocrine metaplasia and mammary sclerosing adenosis. Although these conditions can present as a mass on mammography, a consultation with the interpreting pathologist is usually required to determine whether this common underlying condition is the explanation for the imaging finding [1]. Because our system does not subclassify fibrocystic change, an imaging finding of a mass with this histologic diagnosis is judged to be discordant, which is appropriately conservative; further evaluation of the case by a radiologist is needed. Finally, cancers missed at biopsy are rare, and our study is small; therefore, we encountered only one case of sampling error. We realize that because there was only a single case of sampling error, the 95% CI for sensitivity is quite wide, and the result should be interpreted cautiously. However, we are encouraged by our promising preliminary results. We hope to test a larger population of patients to further validate our model in breast imaging–histologic correlation.

The results of this experiment indicate that our probabilistic expert system can identify biopsy findings that most likely represent sampling errors. We believe that the automated oversight provided by our system can help radiologists to identify discordant cases, and it may decrease the incidence of sampling errors not detected during review sessions of imaging or histologic results. The cases assigned a high likelihood of sampling error by either the radiologists or the Bayesian network mandate careful review by the radiologists, close imaging follow-up, or resampling. Furthermore, this Bayesian network can easily use data routinely captured in a structured mammographic report to assess concordance automatically, providing routine oversight to the task of imaging–histologic correlation.

We do not intend for our probabilistic approach to replace the radiologist as the ultimate decision-maker in determining imaging–histologic correlation of difficult cases but rather for it to act as a safeguard to ensure adequate recognition of possibly discordant results. Ultimately, it must be the radiologist who recommends surgical excision of atypical ductal hyperplasia and, when appropriate, percutaneous sampling of radial scars [3]. It is also the radiologist who provides guidance in surgical planning after imaging-guided biopsy.

We have found that a Bayesian network designed to predict the probability of malignancy on the basis of demographic risk factors and mammographic findings is robust and flexible enough to predict other important clinical factors such as imaging–histologic concordance. We believe this system has the potential to calculate and express probabilities related to breast imaging in a more systematic and accurate manner than is currently available. It is our hope that a probabilistic expert system such as ours may play a future role in aiding the radiologist in the complex task of imaging–histologic correlation.

References