

Using a Bayesian Network to Predict the Probability and Type of Breast Cancer Represented by Microcalcifications on Mammography

Elizabeth S. Burnside^a, Daniel L. Rubin^b, Ross D. Shachter^c

^aDepartment of Radiology, University of Wisconsin Medical School, USA

^bStanford Medical Informatics, Stanford University, CA, USA

^cManagement Science and Engineering, Stanford University, CA, USA

Abstract

Since the widespread adoption of mammographic screening in the 1980's there has been a significant increase in the detection and biopsy of both benign and malignant microcalcifications. Though current practice standards recommend that the positive predictive value (PPV) of breast biopsy should be in the range of 25-40%, there exists significant variability in practice. Microcalcifications, if malignant, can represent either a non-invasive or an invasive form of breast cancer. The distinction is critical because distinct surgical therapies are indicated. Unfortunately, this information is not always available at the time of surgery due to limited sampling at image-guided biopsy. For these reasons we conducted an experiment to determine whether a previously created Bayesian network for mammography could predict the significance of microcalcifications. In this experiment we aim to test whether the system is able to perform two related tasks in this domain: 1) to predict the likelihood that microcalcifications are malignant and 2) to predict the likelihood that a malignancy is invasive to help guide the choice of appropriate surgical therapy.

Keywords: Expert system, Bayesian analysis, mammography

Introduction

Early diagnosis of breast cancer through screening mammography is the most effective means of decreasing the death rate from this disease. [1, 2] The widespread adoption of mammography screening in the 1980's introduced the diagnosis and management of clinically occult abnormalities that signified cancer that had never been dealt with before. A large proportion of these abnormalities were microcalcifications. Malignant microcalcifications on mammography most commonly represent ductal carcinoma *in-situ* (DCIS). Prior to the adoption of mammography, DCIS was a rare diagnosis. In the late 1990's DCIS accounted for approximately 18% of breast cancer diagnosis.[3] In fact, in 1993, the total number of DCIS cases in the US was 200% higher than expected based

on trends established in the previous decade; the majority of these cases attributable to mammographic screening.[4] DCIS is a non-invasive malignant condition with a very favorable prognosis. Involvement of the axillary lymph nodes is rare. Surgical therapy consists of lumpectomy without axillary node sampling. Unfortunately, malignant microcalcifications may also indicate the presence (or future potential development) of invasive malignancy. Invasive breast cancer has an increased risk of axillary node metastasis and depending on the size and grade of the malignancy, axillary node sampling is usually necessary. Microcalcifications can also represent several benign conditions including fibrocystic changes, a fibroadenoma and fat necrosis.

It has been surmised and confirmed in the literature that the mammographic appearance as described by the radiologist can predict the histology of breast cancer.[5] Unfortunately, there is significant variability in this predictive ability; subspecialist, fellowship-trained mammographers perform superiorly. We have demonstrated that our probabilistic expert system, a Bayesian network (BN), can predict the most likely diagnoses and therefore the likelihood of malignancy based on demographic factors and mammography findings as well as expert mammographers.[6] Our system uses predictive imaging features to determine the likely underlying breast disease by using the standardized lexicon established in breast imaging, the Breast Imaging Reporting and Data System (BI-RADS), which defines mammogram feature distinctions and the terminology used to describe them.[7] BI-RADS arose in part from a study of the common terms used to describe mammography abnormalities. The descriptors most highly associated with a benign or malignant diagnosis were considered the most predictive.[8] Subsequently, these terms were incorporated in the BI-RADS lexicon.

For these reasons, we believe that our expert system will be able to predict the likelihood of benign and malignant disease underlying microcalcifications on mammography. This is a more challenging task than our first experiment in which we tested the BN on cases in a teaching atlas.[6] The perform-

ance for the teaching cases was equal to that of an expert mammographer as described in the literature but the cases were not representative of true clinical practice. In this experiment, we chose to test our system on a more challenging dataset: a consecutive series of patients selected to undergo biopsy for microcalcifications. This retrospective review of clinical cases tested the hypothesis that our system would be able to accurately predict the likelihood that microcalcifications are malignant and assess whether the microcalcification represent in situ or invasive breast cancer to aid in preoperative planning.

Materials and Methods

The Model

Some of the details of the construction of our BN have been reported previously, but are repeated here in part for the convenience of the reader.[6] We subsequently refined our system by modifying our probability assessments. From the literature, we identified 26 diseases of the breast (Table 1) that represent the most likely diagnoses to be made on mammography. Twelve of these diseases are malignant and fourteen are benign.

Table 1 – Diagnoses in the Bayes net

Malignant	Benign
Ductal carcinoma (DC) ^c	Cyst
Ductal carcinoma <i>in situ</i> (DCIS) ^b	Fibroadenoma
DC/DCIS ^{a, c}	Papilloma
Lobular carcinoma (LC) ^c	Fibrocystic change
LC/LCIS ^{a, c}	Hamartoma
Tubular carcinoma ^c	Lymph node
Papillary carcinoma ^c	Focal fibrosis
Medullary carcinoma ^c	Fat necrosis
Colloid carcinoma ^c	Secretory disease
Phyllodes tumor ^c	Post-operative change
Metastasis	Skin lesion
	Radial scar
	Atypical ductal hyperplasia
	Lobular Carcinoma <i>in situ</i> (LCIS)

^a Signify two individual diagnoses present simultaneously.

^b Represents in situ disease. ^c Represents invasive disease.

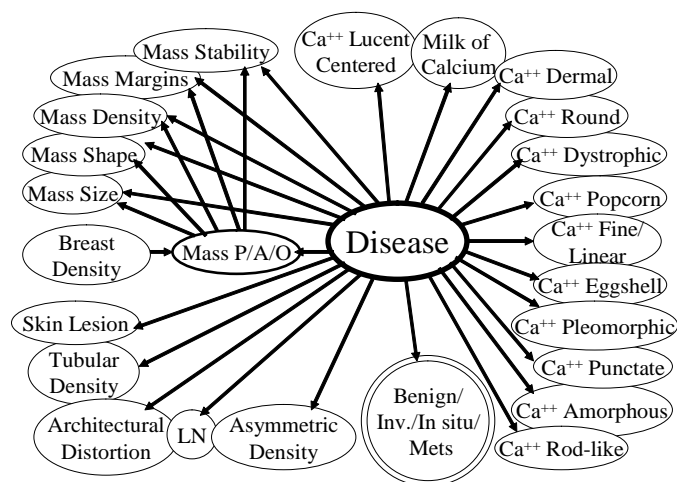
We assume that there is a single uncertain variable, “Disease,” which can take on one value corresponding to exactly one of the 26 diseases or “Normal.” We assume that it is impossible that two unrelated breast diseases occur concomitantly, but *in situ* and invasive breast cancers exist on a spectrum and are commonly present simultaneously. For example, the most common breast malignancy, ductal carcinoma (DC), is gener-

ally thought to develop from ductal carcinoma *in situ* (DCIS). Though the rate of transformation is not well known, the causal relationship between these entities is accepted. We therefore represent these two diseases in our model as three mutually exclusive states in the disease node: DCIS, DC, and DC/DCIS. The third state represents a case in which DC and DCIS are both present in the lesion seen on mammography. Similarly, lobular carcinoma and its noninvasive counterpart lobular carcinoma *in situ* (LCIS) exhibit the same pathophysiology.

The standardized lexicon for breast imaging, BI-RADS, consists of descriptors organized in a hierarchy. These terms describe the density of the breast tissue, and all possible findings on mammography. The most common findings on mammograms are microcalcifications and masses. In our experiment, the characterization of microcalcifications is of interest. When microcalcifications are identified, the radiologist must describe the morphology of the microcalcifications as well as their distribution in the breast.

To construct our belief net and perform inference we used the GeNIe modeling environment developed by the Decision Systems Laboratory of the University of Pittsburgh (<http://www.sis.pitt.edu/~dsl>). We began construction assuming that all of the BI-RADS descriptors except breast density would be children of the disease node. (Figure 1) We modeled the calcification descriptors as conditionally independent manifestations of disease. The distribution, or spatial orientation, descriptors of each type of calcifications are the mutually exclusive states of the corresponding calcification nodes when appropriate. The deterministic (double bordered) node in the belief network has four states, “Benign,” “Non-invasive,” “Invasive,” and “Mets.” The probabilities of non-invasive (analogous to in situ disease), invasive disease, and metastases comprise the total probability of malignancy.

Figure 1 – Bayesian network structure



Note: P/A/O = present, absent, or obscured; Ca⁺⁺ = calcifications; Inv. = invasive breast cancer, Mets = metastasis

We made probability assessments from the medical literature and expert opinion. We obtained pretest probabilities, the age specific and risk factor specific distribution of diseases from census data and large randomized trials. We derived many of the joint probabilities from studies of the radiologic/pathologic correlation of individual breast diseases.

Study Design

Our study included 44 consecutive image-guided biopsies performed for microcalcifications detected and deemed suspicious by radiologists. The patient population consisted of women between the ages of 26 and 71 (mean=53.9; SD=10.1). Patients undergoing biopsy procedures between November 2001 and March 2002 were analyzed. 11-gauge stereotactic biopsies and needle localizations done for diagnosis were included in this project. Patients with a known cancer diagnosis undergoing therapeutic needle localization were excluded. Other exclusion criteria included: 1) the patient's films not available for review, 2) calcifications not identified in the histologic specimen, and 3) mammographic follow-up of at least 12 months not available. These criteria ensured that accurate and complete evaluation of the abnormality of interest occurred and the chance of sampling error of the abnormality and possible progression were minimized

Cases included in the study were reviewed in a blinded manner by a fellowship-trained mammographer. The radiologist used a Web-based interface to input mammography findings and her estimate of the likelihood of malignancy into the BN. The structured entry system mandates the use of BI-RADS descriptors. Given mammography findings, our system provides post-test probabilities formulated as a differential diagnosis. For the purposes of this experiment, the system also provides the probabilities associated with the mutually exclusive possibilities of benign changes, invasive malignancy, *in situ* disease, or metastases.

Study Endpoints

Surgical pathology at the time of the patient's ultimate surgical intervention is the gold standard in this study. We considered ultimate surgical intervention to be either lumpectomy with established negative margins or mastectomy. The reason that we considered definitive surgical therapy the gold standard was to avoid the possibility of sampling error at percutaneous biopsy. Using this gold standard, we evaluated the ability of the Bayes net to predict the outcomes of interest: the probability of malignancy of these microcalcifications as well as the likelihood of invasive disease for surgical planning.

For evaluation of the performance of both the radiologist and the expert system in the task of predicting the probability of malignancy, we created receiver operating characteristic (ROC) curves. The areas under each ROC curve (AUC) were also calculated and compared.[9] The AUC can be used to measure the performance of a diagnostic tool in discriminating

between patients with breast cancer from those without it for all possible cutoff values.

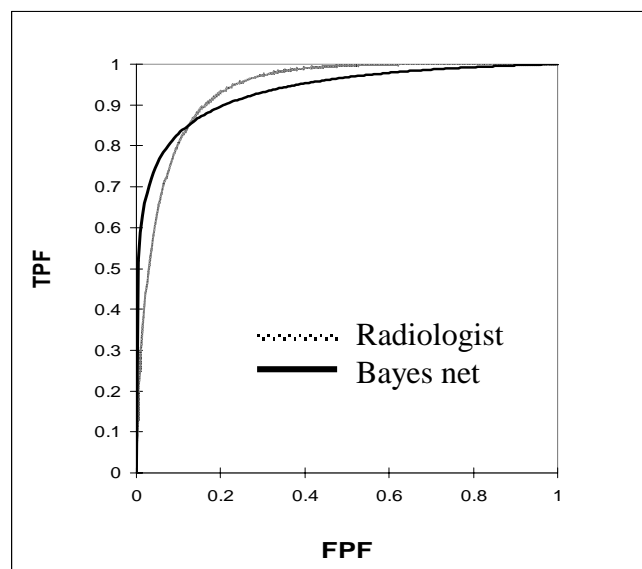
We also created a calibration curve for the radiologist and the Bayes net. This type of graphical representation has been proposed to measure the calibration or reliability of a system in demonstrating the relationship between observed and predicted outcome events. While a calibration curve does not provide a quantitative measure of reliability of probability predictions, it gives a graphical representation to capture the intuitive meaning of calibration of a given system.[10]

For evaluation of the performance of the expert system in distinguishing between invasive breast cancer and *in situ* disease we created a second ROC curve and calculated the AUC.

Results

The AUC of our expert system in predicting whether microcalcifications are malignant, .935, is similar to that found for prediction in the teaching atlas.[6] This is comparable to the AUC of .938 achieved by the radiologist. (Figure 2) There was no statistically significant difference between the AUC of the radiologist and the Bayes net. Therefore, the radiologist and the Bayes net demonstrate similar abilities to predict the likelihood of malignancy of microcalcifications.

Figure 2 – ROC curve measuring discrimination of malignant disease



The calibration curves on the other hand show a different level of prediction reliability between the radiologist and the Bayes Net. The calibration curve for this experiment shows the predictions divided into quartiles. The graphs illustrate the relationship between observed and predicted outcome event rates. The error bars represent 95% confidence intervals. Ideal calibration would show that each quartile has an equal predicted and observed probability (x-axis would equal the y-

axis). Figure 3 shows that the radiologist is fairly well calibrated, while the expert system is not. The Bayes net tends to predict extreme probabilities: high probabilities are overestimated and low probabilities are underestimated. Although the small population size included in this study causes us to view this analysis with caution, it gives a preliminary view of the reliability of the predictions of the radiologist and the model.

Figure 3 – Calibration curve for Radiologist

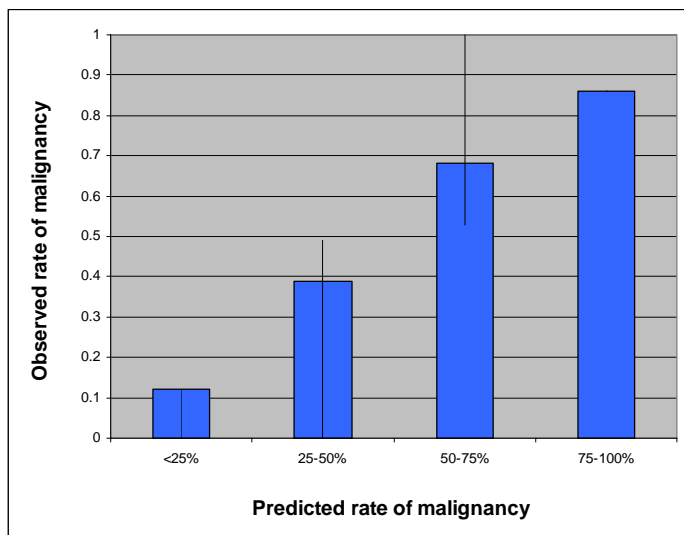
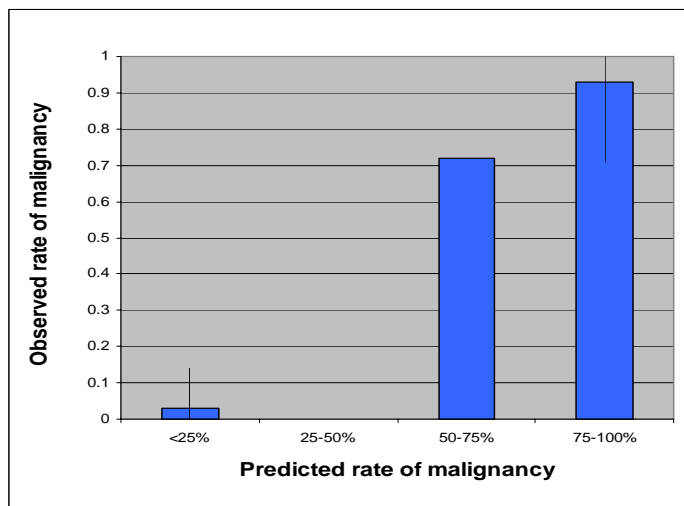


Figure 4 – Calibration curve for Bayes net



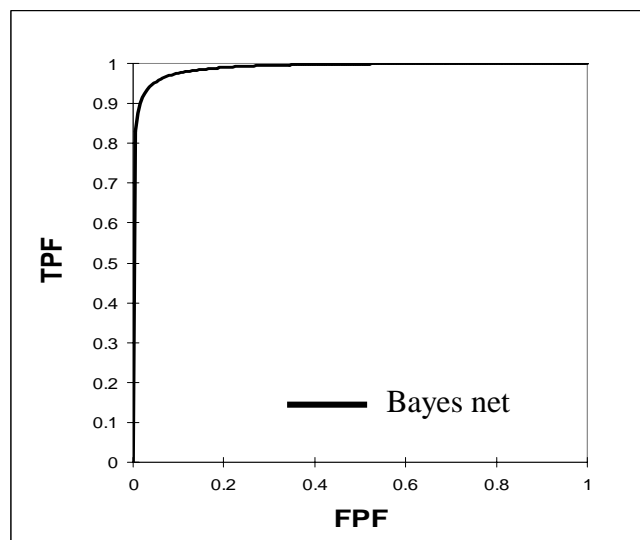
Finally, the ROC curve created to determine the ability of the Bayes net to discriminate between invasive malignancy and in situ disease is shown in Figure 5. The AUC is .990 demonstrating almost perfect discrimination.

Discussion

Our goal in this experiment was to confirm that our BN for mammography 1) is capable of predicting the likelihood of malignancy for microcalcifications on mammography, 2) can

predict the likelihood of invasive disease as opposed to in situ changes in order to help guide appropriate surgical management, and 3) is well-calibrated to the task of predicting malignancy. In addition, through this project, we hoped to identify ways that we could improve our system.

Figure 5 – ROC curve measuring discrimination of invasive disease



Our analysis shows that our expert system has approximately the same ability as a sub-specialist mammographer to discriminate benign and malignant disease in actual patients as opposed to teaching cases as demonstrated previously.[6] In addition, it also performs as well as a full-time, fellowship-trained mammographer in assessing the significance of microcalcifications. The system does this by using the predictive value of BI-RADS descriptors included in the model. We hope to improve performance of the system by incorporating additional descriptors that are likely predictive. For example, our model does not include a descriptor for the temporal evolution of microcalcifications. Often, the radiologist can infer more aggressive disease if microcalcifications are new or increasing.

The system can discriminate which patients are likely to have invasive disease and will likely require removal and analysis of axillary lymph nodes for appropriate staging and prognosis. This represents an area where the model has the potential to contribute to clinical decision-making. Often, percutaneous biopsy of microcalcifications can underestimate the actual disease present (i.e. percutaneous biopsy will indicate in situ disease but excisional biopsy will reveal invasive malignancy).[11] Unfortunately, in these cases, the patient has to be taken back to the operating room for lymph node sampling. It is possible that our expert system may be able to decrease the rate at which patients are forced to return for a second surgery by predicting the necessity of lymph node sampling prospectively. Minimally invasive sampling of lymph nodes using sentinel lymph node biopsy makes this possibility even more realistic.

While the model can ascertain the likelihood that microcalcifications are malignant as well as a fellowship-trained mammographer, the system may not be as well calibrated a mammographer in this domain. Several articles have questioned the ability of a model to simultaneously maximize discrimination and reliability.[10, 12] In the case of a probabilistic expert system, it is important that the predicted probabilities reflect those probabilities ultimately observed if it is to be used in actual patient care. Currently, our system underestimates the likelihood of malignancy when the probability is low and overestimates when the probability is high as reflected by the calibration curves (Figure 4). We believe our model requires additional predictive features and the ability to model coexistent conditions to improve its calibration accuracy.

Observing the individual cases for which the expert system performed poorly is instructive. Two cases in particular demonstrated that the Bayes net can incorrectly estimate the likelihood of malignancy. The first case, a 53 year old female underwent an 11 gauge stereotactic core biopsy for pleomorphic, linear, dystrophic, clustered microcalcification. Pathology revealed fibrocystic changes including apocrine metaplasia as well as a coexistent papilloma. Malignancy was estimated as 50% and 90% by the radiologist and the Bayes net respectively. The second case involved a 61 year old patient who underwent needle localization and excisional biopsy for clustered pleomorphic microcalcifications accompanied by scattered punctate microcalcifications. The patient was diagnosed with both DCIS and fibrocystic changes. In this case, malignancy was estimated as 20% and .4% by the radiologist and the Bayes net respectively. In both cases, the expert system had more difficulty accurately characterizing the constellation of findings because it does not model the possibility of coexistent diseases. Both of these cases demonstrated that two concomitant pathologic diagnoses can be present in the same area of the breast. More specifically, benign or malignant disorders can co-exist with the very common underlying condition of fibrocystic change and its variants sclerosing adenosis and apocrine metaplasia. This has also recently been reflected in the literature.[13] This violates the mutual exclusivity mandated by the disease node in our model. In future work, we plan to improve the performance of our system by modeling this feature of breast disease.

Conclusion

We believe this small retrospective analysis is encouraging. We have now demonstrated that the BI-RADS lexicon, when coupled with our Bayesian model, has great potential to communicate quantitative probabilistic information beyond teaching cases to actual patients. Our model relates the benign and malignant breast diseases to BI-RADS descriptors and allows us to integrate radiological observations in a principled fashion to discriminate between benign and malignant microcalcifications. We have identified areas to work on in the future, including improving the calibration of the model and incorporating the coexistence of the common underlying condition of

fibrocystic change with concomitant breast disease. Ultimately, we hope that with further testing and use our model will help aid decision-making in mammography.

References

- [1] Tabar L, Yen MF, Vitak B, Chen HH, Smith RA, Duffy SW. Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening. *Lancet* 2003;361(9367):1405-10.
- [2] Andersson I, Janzon L. Reduced breast cancer mortality in women under age 50: updated results from the Malmo Mammographic Screening Program. *J Natl Cancer Inst Monogr* 1997(22):63-7.
- [3] Ernster VL, Ballard-Barbash R, Barlow WE, Zheng Y, Weaver DL, Cutter G, et al. Detection of ductal carcinoma in situ in women undergoing screening mammography. *J Natl Cancer Inst* 2002;94(20):1546-54.
- [4] Ernster VL, Barclay J, Kerlikowske K, Grady D, Henderson C. Incidence of and treatment for ductal carcinoma in situ of the breast. *Jama* 1996;275(12):913-8.
- [5] Thurfjell MG, Lindgren A, Thurfjell E. Nonpalpable breast cancer: mammographic appearance as predictor of histologic type. *Radiology* 2002;222(1):165-70.
- [6] Burnside ES, Rubin DL, Shachter RD. A Bayesian network for screening mammography. *Proc AMIA Symp* 2000:106-10.
- [7] Breast Imaging Reporting and Data System (BI-RADS). 3rd ed. Reston, VA: American College of Radiology; 1998.
- [8] Swets J, Getty D, Pickett R, D'Orsi C, Seltzer S, McNeil B. Enhancing and Evaluating Diagnostic Accuracy. *Medical Decision Making* 1991;11(1):9-16.
- [9] Metz CE, Herman BA, Roe CA. Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. *Med Decis Making* 1998;18(1):110-21.
- [10] Diamond GA. What price perfection? Calibration and discrimination of clinical prediction models. *J Clin Epidemiol* 1992;45(1):85-9.
- [11] Jackman RJ, Nowels KW, Rodriguez-Soto J, Marzoni FA, Jr., Finkelstein SI, Shepard MJ. Stereotactic, automated, large-core needle biopsy of nonpalpable breast lesions: false-negative and histologic underestimation rates after long-term follow-up. *Radiology* 1999;210(3):799-805.
- [12] Habbema DF, Hilden J, Bjerregaard B. The measurement of performance in probabilistic diagnosis. *Meth Inform Med* 1981;20:97-100.
- [13] Gill HK, Ioffe OB, Berg WA. When is a diagnosis of sclerosing adenosis acceptable at core biopsy? *Radiology* 2003;228(1):50-7.

Address for correspondence:

Elizabeth S. Burnside, MD, MPH
G3/101-1840, 600 Highland Ave.
Madison, WI 53792
bburnside@mail.radiology.wisc.edu

