Editorial Manager(tm) for International Congress Series
Manuscript Draft

Manuscript Number: 1268276LE205R1

Title: Improving a Bayesian Network's Ability to Predict the Probability of Malignancy of Microcalcifications on Mammography

Article Type: Full Length Article (FLA)

Section/Category:

Keywords: Expert system; Mammography; Bayesian network; Calibration

Corresponding Author: Dr. Elizabeth Suzanne Burnside University of Wisconsin Medical School

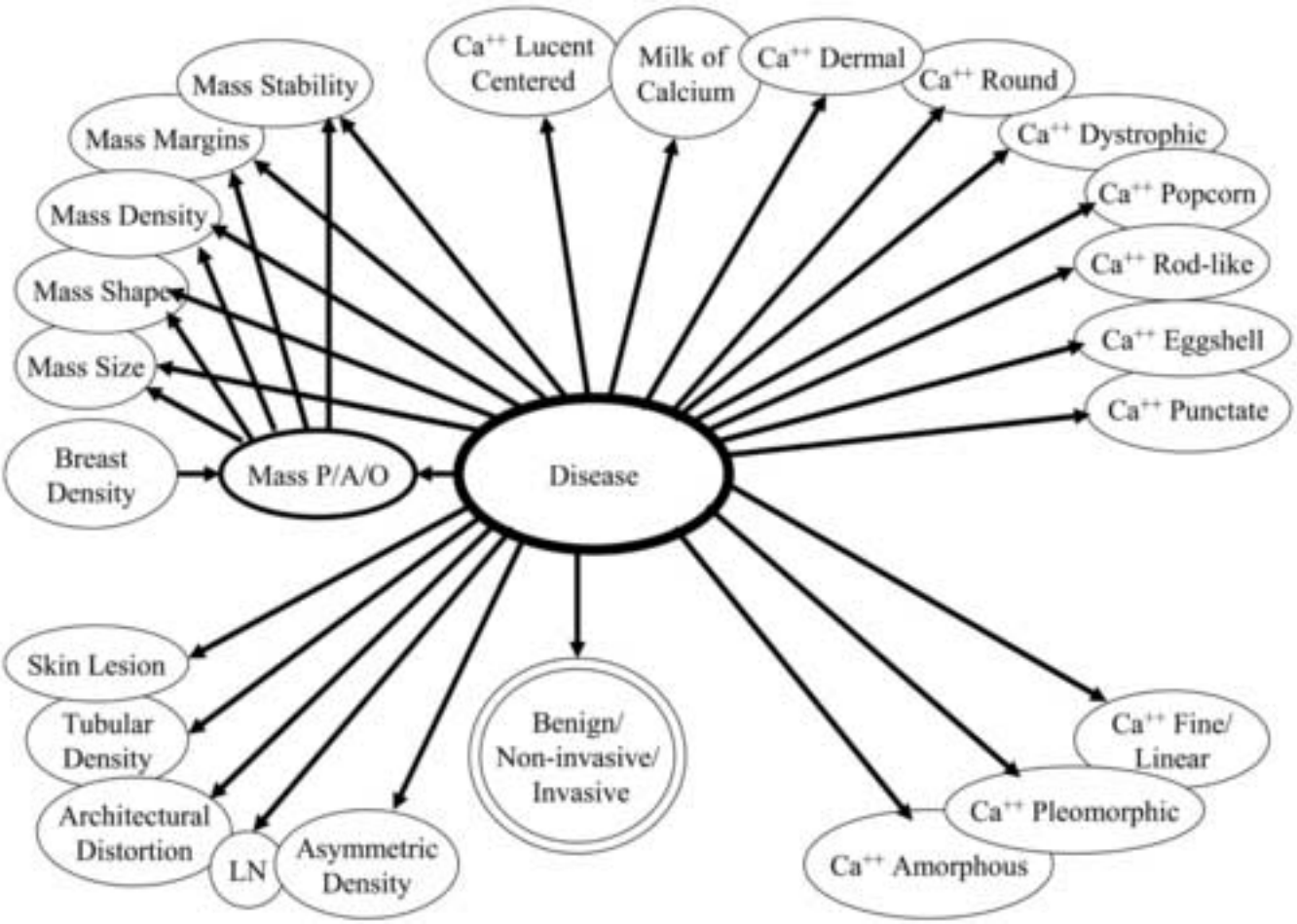
First Author: Elizabeth S. Burnside, MD, MPH, MS

Order of Authors: Elizabeth S. Burnside, MD, MPH, MS; Daniel L Rubin, MD, MS; Ross D Shachter, PhD

Abstract: Mammography is the best test we have for the early detection of breast cancer but it is not perfect largely because performance is attenuated by significant variability of practice. We set out to develop a probabilistic expert system that would uniformly improve performance of all radiologists to the level of expert knowledge. This expert system has been found to effectively discriminate between benign and malignant conditions based on individual patient risk factors and mammographic findings. In this experiment, we test whether the expert system can generate well-calibrated probability estimates of malignancy based on mammographic findings for use in decision-making.

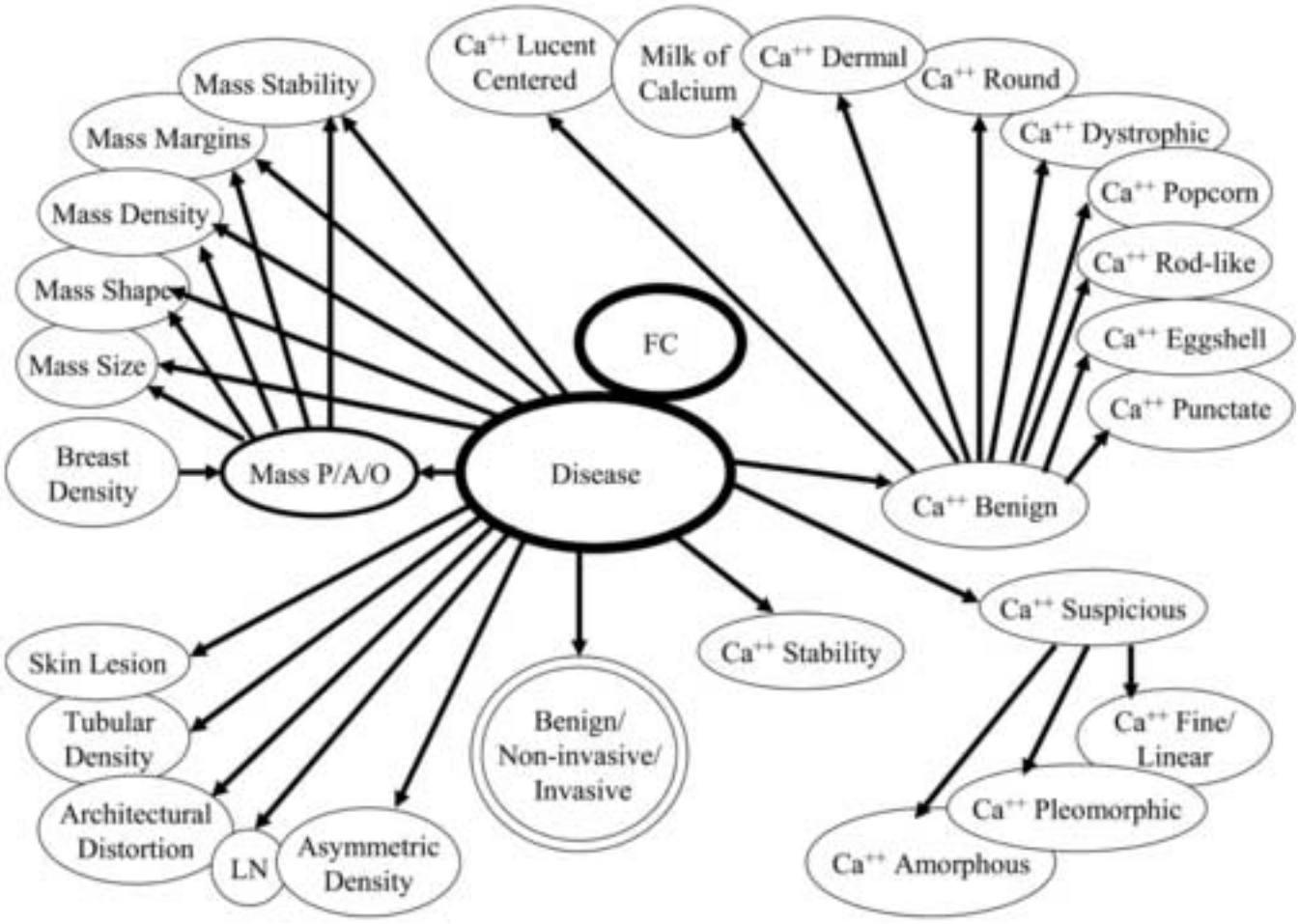
Figure

Click here to download high resolution image



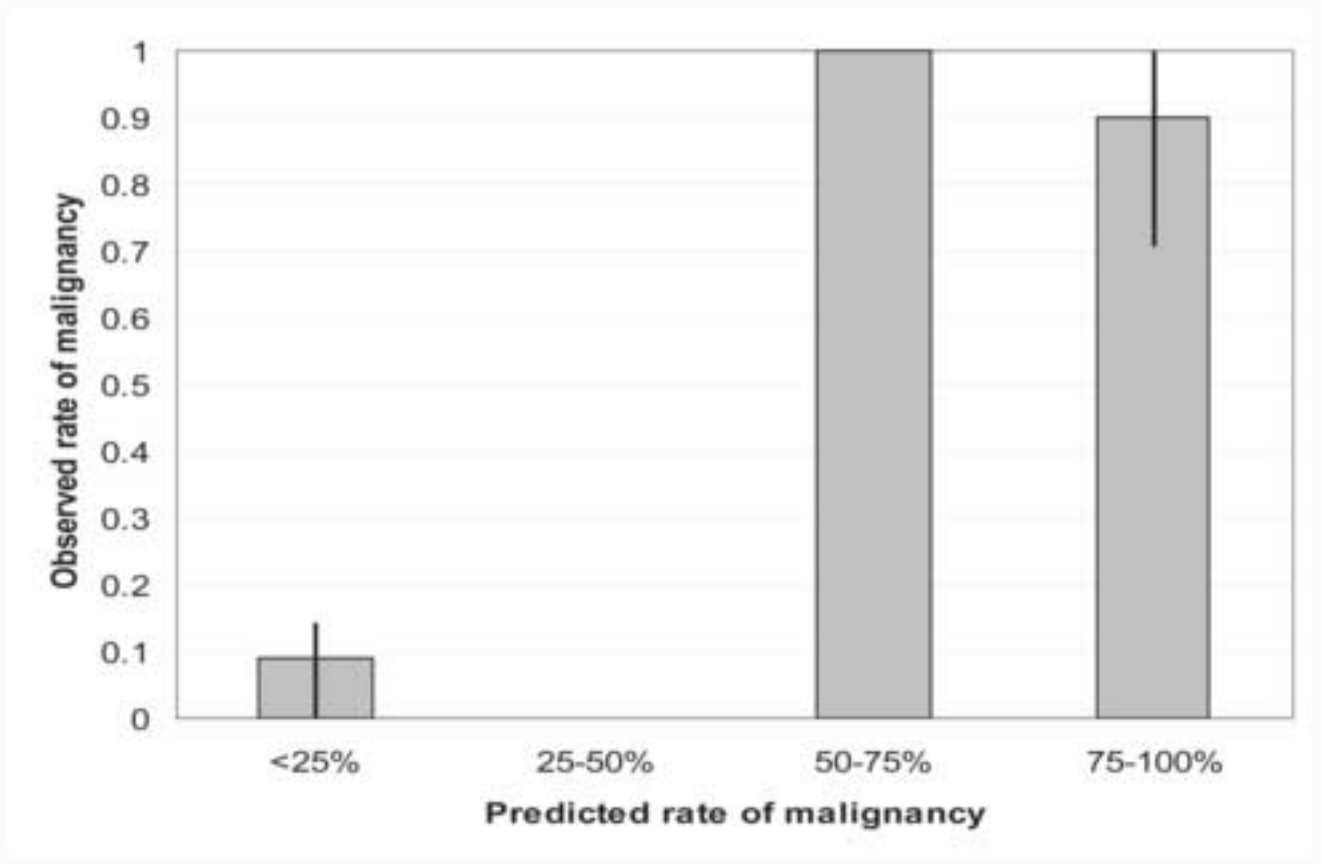
Figure

Click here to download high resolution image



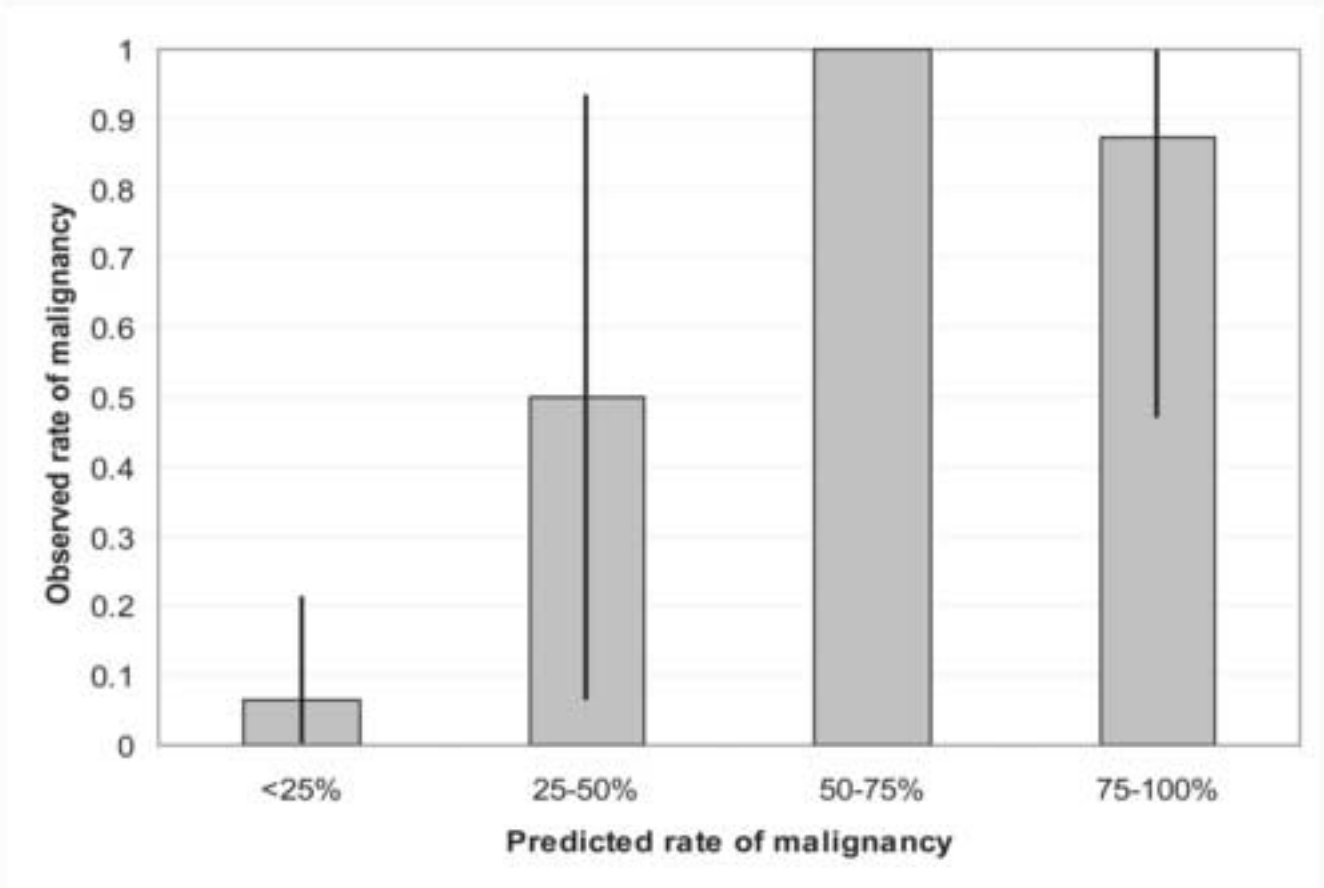
Figure

Click here to download high resolution image



Figure

Click here to download high resolution image



Improving a Bayesian Network's Ability to Predict the Probability of Malignancy of Microcalcifications on Mammography

Elizabeth S. Burnside[a,*], Daniel L. Rubin[b], Ross D. Shachter[c]

[a]*Department of Radiology, University of Wisconsin Medical School, USA*
[b]*Stanford Medical Informatics, Stanford University, CA, USA*
[c]*Management Science and Engineering, Stanford University, CA, USA*

Abstract. Mammography is the best test we have for the early detection of breast cancer but it is not perfect largely because performance is attenuated by significant variability of practice. We set out to develop a probabilistic expert system that would uniformly improve performance of all radiologists to the level of expert knowledge. This expert system has been found to effectively discriminate between benign and malignant conditions based on individual patient risk factors and mammographic findings. In this experiment, we test whether the expert system can generate well-calibrated probability estimates of malignancy based on mammographic findings for use in decision-making.

Keywords: Expert system; Mammography; Bayesian network; Calibration

---

1. Purpose

The interpretation of a mammogram and decisions based on it involve reasoning and management under uncertainty. The wide variation of training and practice among radiologists results in significant variability in screening performance with attendant cost and efficacy consequences. We built a probabilistic expert system in order to improve the performance of radiologists in this domain and attempt to bring mammographic decision-making to the level of experts for all practitioners. We have created a Bayesian network to integrate patient risk factors and the findings from a mammogram. We have tested the Bayesian network's ability to predict the likelihood that microcalcifications detected on mammography are malignant. We found the network to have excellent discriminatory capabilities but found that it was not well calibrated, i.e., the probabilities are not accurately scaled. In this project we refine our model in order to improve its calibration.

---

* Corresponding author. *Email address:* bburnside@mail.radiology.wisc.edu

2. Methods

To build our model, we created a Bayesian network establishing the probabilistic relationships between diseases of the breast and findings on mammography. Our initial model makes strong assumptions of conditional independence. We first identified 26 diseases of the breast from the literature that represent the most likely diagnoses to be identified on mammography. Twelve of these diseases are malignant and fourteen are benign. We assume that there is a single uncertain variable, "disease," which is either exactly one of the 26 diseases or "normal." To represent the findings from a mammogram in the Bayesian network, we used the standardized lexicon for breast imaging, the Breast Imaging Reporting and Data System (BI-RADS) [1]. This lexicon consists of descriptors organized in a hierarchy. These terms describe the density of the breast tissue, and all possible findings from mammography using unique terms. We made our probability assessments from the medical literature and expert opinion. We obtained our pretest probabilities from census data and from results of large randomized trials. We derived many of our conditional probabilities from studies of the radiologic/pathologic correlation of individual breast diseases [2].

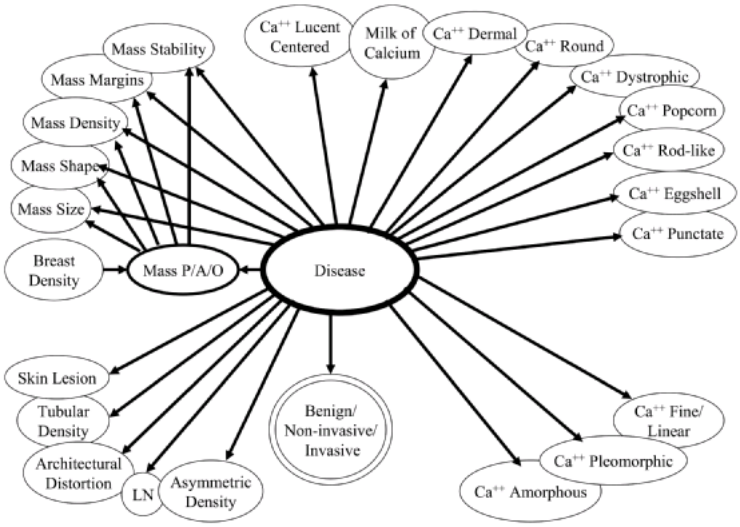


Fig. 1. Original Bayesian network structure

The most common findings on mammograms are microcalcifications and masses. In our experiment, the characterization of microcalcifications is of particular interest. When microcalcifications are identified, the radiologist must describe their morphology as well as their distribution in the breast. 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 there was accurate and complete evaluation of the abnormality of interest and that the chance of sampling error of the abnormality and possible progression were minimized.

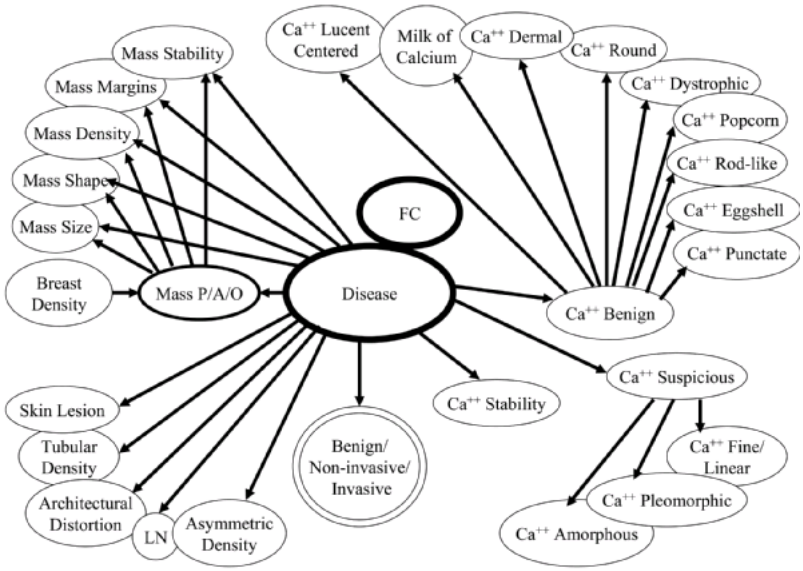


Fig. 2. Bayesian network after changes were made

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 for each case. The structured entry system mandates the use of BI-RADS descriptors. Given patient age, family history, hormone therapy status, and mammography findings, our system provides post-test probabilities for all diseases formulated as a differential diagnosis and as a post-test probability of malignancy. Once we made these analyses, we created a database of these findings and outcomes for evaluation.

We made three changes to the Bayesian network model to improve the accuracy of its post-test probabilities. First, we added variables to represent the presence of either benign or suspicious microcalcifications. The presence of each depends on the disease and conditions the presence of specific types of microcalcifications. For example, suspicious

microcalcifications must be present for there to be either amorphous, pleomorphic, or linear microcalcifications present. Similarly, benign calcifications must be present for there to be other types of microcalcifications such as round, punctate, or popcorn. In this fashion, the presence of two or more suspicious microcalcifications are not overcounted as independent indications of disease.

Second, we added a separate variable for fibrocystic change, so that it can be present by itself or in combination with any of the other diseases [3]. This relaxes the strong assumption in our original model that there is at most one disease. While fibrocystic disease can often present with findings similar to other diseases, it is a diffuse and bilateral process that often coexists with more focal processes such as breast cancer. The third modification was to add a finding quantifying the stability of microcalcifications. This "stability" variable depends on the disease.

To measure the performance of the initial and the modified model in the task of predicting the probability of malignancy, we created receiver operating characteristic (ROC) curves. We calculated the area under each ROC curve (AUC) and compared them. The AUC can be used to measure the performance of a diagnostic tool in discriminating between patients with breast cancer from those without breast cancer for all possible cutoff values.

We also created a calibration curve for both Bayesian networks. 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. The calibration curve gives a graphical representation to capture the intuitive meaning of the calibration of a given system [4, 5].

3. Results

The AUC of our initial expert system in predicting whether microcalcifications are malignant was .921. The calibration of the initial model was found to be poor. There is no parameter to quantify this poor calibration from the curve, but it showed that the initial model severely underestimated low probabilities and overestimated high probabilities. Using the original model, 32 out of 44 cases were estimated to be below 5% or above 95%.

The AUC of the modified model was .909. There was no statistically significant difference between the AUC of initial model and that of the modified model. The modified model showed improved calibration. Using the modified model, 26 out of 44 cases were estimated to be below 5% or above 95%.

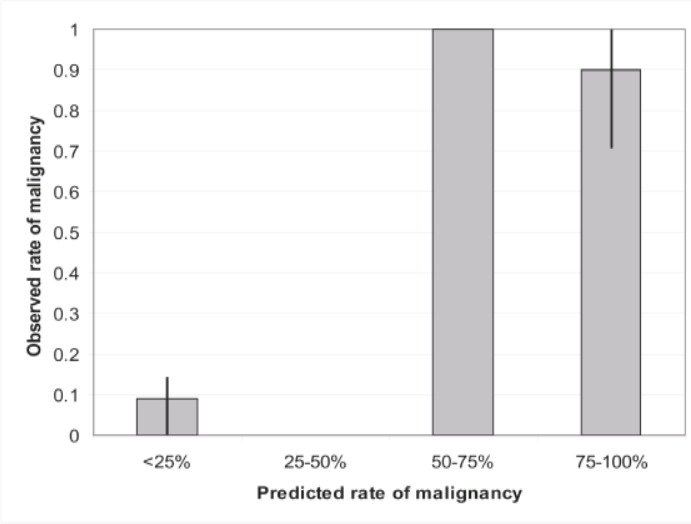


Fig. 3. Calibration curve for the original Bayesian network

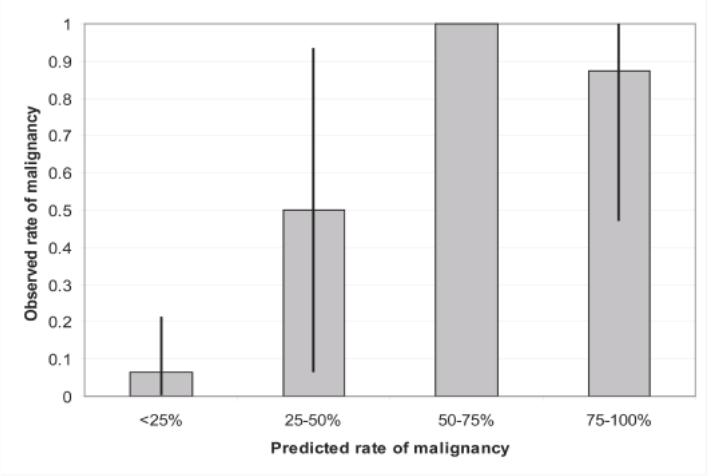


Fig. 4. Calibration curve for the revised Bayesian network

4. Conclusion

If an expert system is to be used to convey the risk of breast cancer to physicians and patients, the probabilities must be accurate. In this experiment, we attempt to improve the accuracy of probability assessments generated from a Bayesian network. The most

commonly used measurement of accuracy for diagnostic tests, an ROC curve, measures the discriminatory abilities of a predictive model but says nothing about the calibration of the model, i.e. the accuracy of the predicted probabilities. A calibration curve attempts to measure the accuracy of probability assessments but is a non-parametric test and, thus, makes comparison between predictive models difficult.

This experiment demonstrates that the changes that we made in the model improved calibration but the amount of improvement is difficult to convey. It has been proposed that any improvement in calibration comes at the cost of discrimination [5]. Though no statistically significant difference was detected between the AUC of the initial and the modified model, the discrimination of the original (less calibrated) model was slightly better than the modified (more calibrated) model. It would be useful to test whether discrimination and calibration are inversely related with a larger sample size.

Improved calibration may be accomplished through additional structural modifications as in this experiment or through improved probabilities within the model. We plan to pursue both of these methods to improve our model's prediction accuracy, especially since we made some simplifying assumptions in estimating the probabilities for the modifications we introduced. Training the network on data has the potential to effectively modify the structure and the probabilities to enable better calibration. To this end, we hope to develop or discover a more robust test for calibration that provides a quantitative evaluation of the accuracy of predicted outcomes. In the future, we hope to create a system that will provide radiologists and patients with accurate estimates of their breast cancer risk based on personal risk factors and imaging findings.

References

[1] Breast Imaging Reporting And Data System (BI-RADS). Reston VA: American College of Radiology; 1998.

[2] Burnside E, Rubin D, Shachter R. A Bayesian network for mammography. Proc AMIA Symp 2000:106-10.

[3] Gill HK, Ioffe OB, Berg WA. When is a diagnosis of sclerosing adenosis acceptable at core biopsy? Radiology 2003;228(1):50-7.

[4] Cook RM. Experts in uncertainty. New York: Oxford University Press; 1991.

[5] Diamond GA. What price perfection? Calibration and discrimination of clinical prediction models. J Clin Epidemiol 1992;45(1):85-9.