Published in final edited form as:

Med Decis Making. 2019 April ; 39(3): 208–216. doi:10.1177/0272989X19832914.

A probabilistic model to support radiologists' classification decisions in mammography practice

Jiaming Zeng, M.S.*,

Department of Management Science and Engineering, Huang Engineering Center, 338J, 475 Via Ortega, Stanford, CA 94305-4121

Francisco Gimenez, Ph.D.*,

Department of Biomedical Data Science, Radiology, and Medicine (Biomedical Informatics), Medical School Office Building (MSOB), Room X-335, 1265 Welch Road, Stanford CA 94305-5479

Elizabeth S. Burnside, M.D, M.P.H, M.S.,

University of Wisconsin School of Medicine and Public Health, 600 Highland Ave., E3/311 CSC, Madison, WI 53792-3252

Daniel L. Rubin, M.D, M.S*, and

Department of Biomedical Data Science, Radiology, and Medicine (Biomedical Informatics), Medical School Office Building (MSOB), Room X-335, 1265 Welch Road, Stanford CA 94305-5479

Ross Shachter, Ph.D.*

Department of Management Science and Engineering, Huang Engineering Center, Room 337, 475 Via Ortega, Stanford, CA 94305-4121

Abstract

We develop a probabilistic model to support the classification decisions made by radiologists in mammography practice. Using the feature observations and BI-RADS classifications from radiologists examining diagnostic and screening mammograms, we model their decisions in order to understand their judgments. Our model could help improve the decisions made by radiologists using their own feature observations and classifications while maintaining their observed sensitivities. Based on 112,433 mammographic cases from 36,111 patients and 13 radiologists at two separate institutions with a 1.1% prevalence of malignancy, we train a probabilistic Bayesian network (BN) to estimate malignancy probabilities of lesions. For each radiologist, we learn an observed probabilistic threshold within the model. We compare the sensitivity and specificity of each radiologist against the BN model using either their observed threshold or the standard 2% BI-RADS recommended threshold.

*corresponding author jiaming@stanford.edu.

*Authors contributed equally to the work.

The work was done at Stanford University, Department of Management Science and Engineering and Department of Biomedical Data Science, Radiology, and Medicine, and University of Wisconsin School of Medicine and Public Health, Department of Radiology.

We find significant variability among the radiologists' observed thresholds. By applying the observed thresholds, the BN model shows a 0.01% (1 case) increase in false negatives and a 28.9% (3612 cases) reduction in false positives. When using the standard 2% BI-RADS recommended threshold, there is a 26.7% (47 cases) increase in false negatives and a 47.3% (5911 cases) reduction in false positives. Our results show that we can significantly reduce screening mammography false positives with a minimal increase in false negatives.

We find that learning radiologists' observed thresholds provides valuable information regarding the conservativeness of clinical practice and allows us to quantify the variability in sensitivity across and within institutions. Our model could provide support to radiologists to improve their performance and consistency within mammography practice.

Keywords

mammography; classification decision; observed threshold; decision support

Introduction

The American Cancer Society recommends annual screening mammography for women over 45 to detect breast cancer early when it is most treatable [1, 2, 3]. However, the USPSTF recommends less aggressive screening based on literature that asserts the harms of early and frequent screening outweigh the benefits [4, 5]. While a reduction in screening is one possible solution to addressing the issue of false positive detections, it risks missing cancer at an early stage. An alternative solution is to help radiologists reduce their false positive interpretations of mammography.

Like all screening tests, mammography balances sensitivity against specificity, or equivalently, balances false negative against false positive findings. These tradeoffs are determined by a radiologist's personal subjective threshold. A conservative radiologist might practice at a higher sensitivity and corresponding lower specificity, decreasing false negative findings while increasing false positive findings. Such subjectivity results in variability in mammography practice, which is a well-known and unsolved challenge [6, 7, 8, 9].

Computer-aided diagnosis (CADx) systems could potentially diminish subjectivity in the interpretation of mammography using quantitative methods and an objective *operating point*, a particular value of sensitivity and corresponding specificity on the receiver operating characteristic (ROC) curve of the system. Moreover, by adjusting that operating point, it is possible to modify performance in CADx systems, a task that is much more challenging in unassisted human readers. A variety of CADx systems have been developed for mammography [10, 11, 12, 13, 14, 15, 16, 17]. While many have been shown to improve the performance of radiologists as well as to reduce their intra-reader variability in controlled settings, they show less improvement in real world settings. CADx systems seeking to reduce false negatives will trade sensitivity for specificity and subsequently increase false positives, in a manner similar to radiologists. The difference between such systems and radiologists is that the operating point in CADx systems can be explicitly set to maximize the system's performance. In probabilistic CADx systems, a probabilistic threshold solely

determines the operating point. This threshold can be interpreted as the minimum probability of cancer that a lesion must exhibit before it is deemed a positive finding (i.e. recalled).

While most radiologists strive for a fixed operating point, the holistic and qualitative nature of mammography interpretation makes it difficult to quantify their probabilistic thresholds. The Breast Imaging Reporting and Data System (BI-RADS) designates a probability of malignancy greater than 2% to be a positive result [18, 19]. Unfortunately, there has been no way to measure what threshold radiologists are actually using and thereby understand how conservatively they are practicing.

We propose a methodology to measure a radiologist's effective probabilistic threshold for declaring a positive finding. Furthermore, we show that it is possible to help radiologists increase their specificity without decreasing sensitivity. This strategy allows for a reduction in false positive findings with a minimal increase in false negative findings and could improve the effectiveness of screening mammography.

Methods

Dataset

For our study, we used a total of 112,433 mammography reports, with 1,214 malignant cases, a 1.1% prevalence, collected from thirteen radiologists across two teaching hospitals, eight radiologists at Institution I and five radiologists at Institution II. The reports include both diagnostic and screening mammograms. We included prospectively interpreted consecutive screening and diagnostic mammograms as recorded in our structured reporting software (PenRad Technologies, Buffalo, MN) at Institution I from October 3, 2005 to July 30, 2010 and Institution II from April 5, 1999 to February 9, 2004. We had Institutional Review Board approval for this research.

We define a mammography case as the patient's risk factors, the radiologist's observations, and the pathological ground truth. Each case includes features such as patient demographic risk factors (e.g. age, personal history), BI-RADS observations (e.g. mass size, mass stability), and the radiologist's BI-RADS assessment category. The BI-RADS assessment categories are split into 6 levels - 1 or 2 or 3 indicates a negative assessment (no immediate follow-up), while 0 or 4 or 5 indicates a positive assessment (recommends follow-up imaging or biopsy) [19, 18]. Using this, we treated each radiologist's classification decision on malignancy as a binary outcome of positive or negative.

Pathological ground truth of malignancy was determined through biopsy results or at least one year of clinical follow up. For patients who were not biopsied or did not develop cancer within a year of the mammogram, pathological ground truth was determined by matching them to state cancer registries.

By comparing the radiologist's BI-RADS based decisions to the pathological ground truth, we can recognize each case as either false positive (FP), false negative (FN), true positive (TP), or true negative (TN). A summary of the dataset is shown in Table 1.

Probabilistic Model

Building upon the mammography Bayesian network (BN) model described in Burnside et al. [20], we trained a BN to estimate a lesion's probability of malignancy based on the features identified by the radiologists. A Bayesian network models the joint distribution of many random variables to efficiently update the probability of a malignant case given a radiologist's observation of BI-RADS features, BI-RADS assessment category, and patient risk factors [21]. Using the dataset described, we learned both the structure and the parameters of the BN. The structure was learned using Tree-Augmented Nave Bayes (TAN) [22, 23]. We estimated the conditional probability table parameters using gradient descent. Both the BN structure and parameters were estimated by an iterative 10-fold cross-validated model within the training data. All model learning and classification was done in Norsys Netica 5.14 [24].

We estimated the posterior probability of malignancy for each case by using an iterative 10-fold model. The data was stratified into 10 folds by radiologist and number of malignant cases. For each fold, we used the other 9 folds as the training set to build a probabilistic model for diagnosis, and the fold itself as the test set to measure the probability that a finding was malignant. We then estimated each radiologist's operating point using these posterior probabilities of malignancy (see next section). The operating point is characterized by a probability in the BN model, which we call the *observed threshold*.

Observed Threshold Selection

Finding a radiologist's operating point is challenging because radiologists have different beliefs about the probability of malignancy. Instead, we learn a radiologist's observed threshold by setting the BN model to be as specific as possible given that it is at least as sensitive as the radiologist. In other words, the observed threshold minimizes the false positive rate while matching the radiologist's false negative rate in bootstrapped samples.

To match a radiologist's sensitivity, we calculate the probability of malignancy for each of the radiologist's cases and find the largest probabilistic threshold with at least as much sensitivity as the radiologist.

For most of our radiologists, this threshold leads to an increase in specificity. For each radiologist, we repeat the process 5001 times over bootstrapped samples of the radiologist's cases. The median threshold from these samples is selected as the observed threshold. While this bootstrapping provides robust estimates of the observed thresholds, it can fail to match the number of false negatives exactly, sometimes yielding one more or one less false negative. Across all radiologists, there is a net increase of one false negative. In Figure 1, the box plot shows the spread of the thresholds for the bootstrapped samples, with outliers shown as black dots. We note that there is fairly high variability in our sampled thresholds across radiologists. This can be interpreted as inter-reader variability and lack of consistency of practice. Moreover, most of the radiologists deviate substantially from the 2% BI-RADS recommended threshold. The learned observed threshold values are also shown in Table 2.

Statistical Analysis

Using the BN model, we also implement the 2% BI-RADS recommended threshold and examine its results. In the Results section, we compare the classification decisions by each radiologist with the BN model using either the radiologist's observed threshold or the 2% BI-RADS recommended threshold.

We compare the performance of these three classifications via sensitivity and specificity. We use McNemar's test of proportions to evaluate the statistical significance between each radiologist and each of the BN thresholds for all of the radiologist's cases. Significance is determined at a 95% confidence level ($p < 0.05$). Additionally, we use the two one-sided test (TOST) to establish the non-inferiority margin for sensitivity for each pair of methods. That margin is calculated to be 1.96 times the standard deviation of inter-reader sensitivity, corresponding to the 95% confidence interval across all radiologists. The confidence interval for each pair of proportions is calculated using the Agresti & Min method [25].

All statistics are estimated by R 3.1.0 [26]. Binary proportions testing is done using the DTComPair package version 1.03 [27]. Bootstrapping is performed by the boot package version 1.3-11 [28].

Results

We compare and analyze the performance of the three methods for classifying mammograms as positive or negative: 1) each radiologist's BI-RADS assessment, 2) classification using the radiologist's observed threshold on the probabilities estimated by our BN model, and 3) classification using the 2% BI-RADS threshold on those probabilities.

Comparison of Sensitivity and Specificity

We compare the difference in sensitivity and specificity between each radiologist and the BN model with either that radiologist's observed threshold or the 2% BI-RADS recommended threshold in Figure 2. The methods are considered non-inferior to each other with respect to sensitivity if the 95% confidence interval of the difference in sensitivity, indicated by the whisker lines, is completely within the "equivalence bounds" across radiologists, indicated by the dashed red lines. Those equivalence bounds correspond to the 95% confidence interval of sensitivity, based on the variability among the radiologists.

In Figure 2, we compare the performance of each radiologist and our methods on the same cases in terms of sensitivity and specificity.

Observed Thresholds: Sensitivity performance between radiologists and the BN model with observed thresholds is within the equivalence bounds and non-inferior for all but one radiologist.

The difference in specificity is statistically significant for 9 radiologists, with 6 increasing and 3 decreasing under the BN model. The remaining 4 radiologists show no significant difference in specificity.

2% BI-RADS Recommended Threshold: Sensitivity performance between radiologists and the model's 2% BI-RADS threshold is within the equivalence bounds for 4 radiologists and the BN model is non-inferior for 6 of the 13 radiologists.

With respect to specificity, the 2% threshold BN model shows statistically significant improvement for all radiologists.

Comparison of False Negative and False Positive Counts

In Figure 3, we compare the false negative and false positive counts for each radiologist with those from the model with either the observed thresholds or the 2% BI-RADS recommended threshold. We then evaluate the impact of these different thresholds.

The observed thresholds are designed to match each radiologist's sensitivity, so the false negative counts for each radiologist are within one of the BN model, as seen in Figure 3. For 9 of the 13 radiologists, there is a decrease in false positive counts for each with no net increase in false negative counts.

Using the 2% BI-RADS recommended threshold, there is a reduction in the false negative counts for 4 radiologists and an increase for the other 9 radiologists. The false positive counts decrease for all radiologists.

Overall, with the observed thresholds, the BN model shows a net increase of one false negative (0.01%) and a net decrease of 3612 false positives (28.9%) relative to the radiologists' assessments. With the 2% BI-RADS recommended threshold, the BN model shows a net increase of 47 false negatives (26.7%) and a net decrease of 5911 false positives (47.3%) relative to the radiologists.

Discussion

There are three main contributions of our study: (1) training a probabilistic model to predict malignancy given the mammography imaging observations, (2) discovering a radiologist's effective probabilistic threshold, and (3) exploring how we can help radiologists improve their performance. Our paper differs from earlier work in that we do not try to predict malignancy or the BI-RADS assessment category. Instead, we use the BI-RADS descriptors and assessment category to help revise the clinical decision. We believe our technique is novel, clinically important, and fills an important gap in the literature.

We demonstrate that we can use a probabilistic model based on the features identified and classifications assessed by radiologists to estimate the operating point a radiologist effectively uses for mammography classification. Without a probabilistic model, radiologists make holistic qualitative assessments, resulting in variability across practices. With our learned probabilistic Bayesian network (BN) model, we characterize each radiologist by an observed threshold that matches their sensitivity. The observed thresholds allow us to quantify previously unidentified sources of variability in practice and significantly reduce false positives with a single additional false negative overall. When we enforce the standard 2% BI-RADS recommended threshold in our BN model, we see a significant reduction in the number of false positive findings. Our results suggest that a decision support system

using our BN model could help some radiologists more effectively evaluate screening mammograms.

Moreover, our approach enables mammography assessments to be more consistent and could significantly improve diagnostic accuracy. In BI-RADS, radiologists determine the risk of malignancy by reporting one of seven assessment categories. While each category should indicate a quantitative estimate of malignancy, they are often used as a qualitative assessment of the radiologist's suspicion in practice. Even if radiologists tried to adhere to quantitative assessments of malignancy, it is well established that practitioners make errors in estimating probability [29, 20], and this is one possible reason that numerous studies have found significant variability in the effectiveness of mammography [7, 30, 31]. Our results show that using a learned BN model for estimating the probability of malignancy can significantly reduce the rate of false positives by some radiologists, and thus boost their positive predictive value and the quality of practice.

While our results are promising, there are limitations to our work. First, because our study is developed based on a structured reporting system, our BN will not perform as reliably with missing information that may occur in narrative reports that are not guided by a structured reporting system. Many radiology reports do not use all the BI-RADS descriptors. Moreover, radiologists often omit features that would not change their classification decisions. Thus, many reports have missing BI-RADS descriptors, resulting in inaccurate malignancy probability estimates from the BN model. Finally, our method is able to learn a radiologist's observed threshold from their reports on cases where we know pathological ground truth.

Furthermore, the BI-RADS descriptors for mammography, though comprehensive, may not sufficiently describe all relevant cancer lesion characteristics. In addition, our studies are from the era of analog mammography, a technology that has been largely replaced by digital mammography or tomosynthesis. New descriptors are defined over time by the mammography community, and BI-RADS is continually evolving [32]. A decision support system based on our method would need to be updated as the descriptors change. However, the benefits of using a quantitative image-based method to estimate malignancy probability and reduce false positives should persist.

A BN trained with the radiologist's BI-RADS assessment category produces better results than one without it. Even when the assessment is not included, the model can help the radiologists reduce false positives. This indicates that radiologists are incorporating additional information not documented in the radiology reports (e.g., salient non-image descriptors) when determining their final assessment [33]. In summary, we show that a collaborative decision support system for mammographic classification has potential to aid radiologists in refining and adhering to an optimal threshold. We believe such models would be strengthened by augmenting radiologist extracted features with quantitative data from image processing, and other sources.

Conclusion

With further validation, demonstration of generalizability, and refinement of human computer interaction, our system has the potential to provide decision support to improve radiologists' classification decisions. In addition, our model could be used retrospectively to measure compliance with clinical threshold targets and standards. For both potential real-world applications, our results suggest notable reductions in false positives with a minimal increase in false negatives. Although there may be many challenges in implementation, introducing these methods into clinical practice could improve the quality of care in mammography screening while reducing practice variability.

Acknowledgements

We express our thanks to grants from the National Cancer Institute and National Institutes of Health. The following grants were used to fund the research: 1U01CA190214, 1U01CA187947, and K24CA194251. We thank the reviewers for their helpful suggestions.

Financial Grants: 1U01CA190214, 1U01CA187947, K24CA194251, and NRSA F31 1F3CA171789-01A1.

Financial support for this study was provided entirely by a grant from National Cancer Institute and National Institutes of Health. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

References

- [1]. Nyström L, Andersson I, Bjurstam N et al. Long-term effects of mammography screening: updated overview of the swedish randomised trials. *The Lancet* 2002; 359(9310): 909–919.
- [2]. Oeffinger KC, Fontham ET, Etzioni R et al. Breast cancer screening for women at average risk: 2015 guideline update from the american cancer society. *Jama* 2015; 314(15): 1599–1614. [PubMed: 26501536]
- [3]. Smart CR, Byrne C, Smith RA et al. Twenty-year follow-up of the breast cancers diagnosed during the breast cancer detection demonstration project. *CA: a cancer journal for clinicians* 1997; 47(3): 134–149. [PubMed: 9152171]
- [4]. Siu AL. Screening for breast cancer: Us preventive services task force recommendation statement. *Annals of internal medicine* 2016; 164(4) : 279–296. [PubMed: 26757170]
- [5]. Kalager M, Adami HO, Bretthauer M et al. Overdiagnosis of invasive breast cancer due to mammography screening: results from the norwegian screening program. *Annals of internal medicine* 2012; 156(7): 491. [PubMed: 22473436]
- [6]. Elmore JG, Jackson SL, Abraham L et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology* 2009; 253(3): 641–651. [PubMed: 19864507]
- [7]. Beam CA, Layde PM and Sullivan DC. Variability in the interpretation of screening mammograms by us radiologists: findings from a national sample. *Archives of internal medicine* 1996; 156(2): 209–213. [PubMed: 8546556]
- [8]. Elmore JG, Wells CK, Lee CH et al. Variability in radiologists' interpretations of mammograms. *New England Journal of Medicine* 1994; 331(22): 1493–1499. [PubMed: 7969300]
- [9]. Taplin S, Abraham L, Barlow WE et al. Mammography facility characteristics associated with interpretive accuracy of screening mammography. *Journal of the National Cancer Institute* 2008; 100(12): 876–887. [PubMed: 18544742]
- [10]. Garg AX, Adhikari NK, McDonald H et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *Jama* 2005; 293(10): 1223–1238. [PubMed: 15755945]

- [11]. Jiang Y, Nishikawa RM, Schmidt RA et al. Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications. *Radiology* 2001; 220(3): 787–794. [PubMed: 11526283]
- [12]. Eadie LH, Taylor P and Gibson AP. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. *European journal of radiology* 2012; 81(1): e70–e76. [PubMed: 21345631]
- [13]. Fujita H, Uchiyama Y, Nakagawa T et al. Computer-aided diagnosis: The emerging of three cad systems induced by Japanese health care needs. *Computer methods and programs in biomedicine* 2008; 92(3): 238–248. [PubMed: 18514362]
- [14]. Oliver A, Freixenet J, Marti J et al. A review of automatic mass detection and segmentation in mammographic images. *Medical image analysis* 2010; 14(2): 87–110. [PubMed: 20071209]
- [15]. Bright TJ, Wong A, Dhurjati R et al. Effect of clinical decision-support systems a systematic review. *Annals of internal medicine* 2012; 157(1): 29–43. [PubMed: 22751758]
- [16]. Burnside ES. Bayesian networks: Computer-assisted diagnosis support in radiology1. *Academic radiology* 2005; 12(4): 422–430. [PubMed: 15831415]
- [17]. Burnside E, Rubin D and Shachter R. A bayesian network for mammography In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, p. 106.
- [18]. Baker JA, Kornguth PJ and Floyd C Jr. Breast imaging reporting and data system standardized mammography lexicon: Observer variability in lesion description. *AJR American journal of roentgenology* 1996; 166(4): 773–778. [PubMed: 8610547]
- [19]. Liberman L and Menell JH Breast imaging reporting and data system (birads). *Radiologic Clinics* 2002; 40(3) : 409–430. [PubMed: 12117184]
- [20]. Burnside ES, Davis J, Chhatwal J et al. Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings. *Radiology* 2009; 251(3): 663–672. [PubMed: 19366902]
- [21]. Pearl J Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, 2014.
- [22]. Friedman N, Geiger D and Goldszmidt M. Bayesian network classifiers. *Machine learning* 1997; 29(2): 131–163.
- [23]. Friedman N and Goldszmidt M. Building classifiers using bayesian networks. In *Proceedings of the national conference on artificial intelligence*. pp. 1277–1284.
- [24]. Manual NJ. Version 4.18 and higher, norsys software corp, 2012.
- [25]. Agresti A and Min Y. Simple improved confidence intervals for comparing matched proportions. *Statistics in medicine* 2005; 24(5) : 729–740. [PubMed: 15696504]
- [26]. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2014 URL <https://www.R-project.org/>.
- [27]. Stock C and Hielscher T. Dtcompare: comparison of binary diagnostic tests in a paired study design. r package version 1.0. 3, 2014.
- [28]. Canty A and Ripley B. boot: Bootstrap r (s-plus) functions. R package version 2012; 1(7).
- [29]. Tversky A and Kahneman D. Judgment under uncertainty: Heuristics and biases In *Utility, probability, and human decision making*. Springer, 1975 pp. 141–162.
- [30]. Elmore JG, Nakano CY, Koepsell TD et al. International variation in screening mammography interpretations in community-based programs. *Journal of the National Cancer Institute* 2003; 95(18): 1384–1393. [PubMed: 13130114]
- [31]. Cox B Variation in the effectiveness of breast screening by year of followup. *JNCI Monographs* 1997; 1997(22): 69–72.
- [32]. Burnside ES, Sickles EA, Bassett LW et al. The acr bi-rads® experience: learning from history. *Journal of the American College of Radiology* 2009; 6(12): 851–860. [PubMed: 19945040]
- [33]. Sachchamarga W Using Conflicting Information and Holistic Judgments in Diagnostic Computer Models. PhD Thesis, Stanford University, 2012.

The distribution of BI-RADS assessment categories for our cases are shown in Table 3 to demonstrate that each assessment category is represented in our dataset.

Figure 1.

Observed thresholds of radiologists learned from bootstrapped samples. The observed thresholds presented in Table 2 are estimated by the median of the corresponding sampled thresholds. The red line shows the 2% BI-RADS recommended threshold. The radiologists from Institution I practice at a more conservative level than the 2% threshold while most from Institution II have a less conservative practice. The black dots show outliers among the sampled thresholds.

Figure 2.

Comparison of sensitivity and specificity between each radiologist and the BN model using either the radiologist's observed thresholds or the 2% BI-RADS recommended threshold. The observed threshold comparison is shown in the top graphs and the 2% threshold comparison in the bottom graphs, while the difference in sensitivity is shown on the left and the difference in specificity on the right. In each chart, the heavy black vertical line represents no change in performance and positive numbers indicate improved performance using the BN model. In each chart the lines with whiskers show the 95% confidence interval for the comparison for each radiologist. The dashed red lines in the charts on the left indicate the 95% confidence interval for variation in sensitivity across all of the radiologists.

Figure 3. Comparison of the counts of false negatives (top graphs) and false positives (bottom graph) in the classification decisions made by radiologists and by the BN model using either the observed threshold or the BI-RADS 2% threshold for each radiologist's cases.

Table 1.

Our study includes 112,433 diagnostic and screening mammography cases with separate analysis for each of the 13 radiologists at two institutions.

Rad.	Inst.	Cases	Malignant	Benign	FN	FP	Prevalence (%)
1	I	1,759	35	1,724	4	225	2.0
2	I	8,185	123	8,062	10	988	1.5
3	I	11,415	151	11,264	25	1,160	1.3
4	I	6,144	32	6,112	5	686	0.5
5	I	4,126	69	4,057	14	475	1.7
6	I	15,908	199	15,709	27	1,369	1.3
7	I	8,217	110	8,107	10	992	1.3
8	I	3,736	74	3,662	10	629	2.0
9	II	23,497	174	23,323	32	2,678	0.7
10	II	16,604	148	16,456	19	1,840	0.9
11	II	3,077	26	3,051	3	313	0.8
12	II	3,634	30	3,604	4	420	0.8
13	II	6,131	43	6,088	13	701	0.7
Total	-	112,433	1,214	111,219	176	12,476	1.1

Table 2.

Observed thresholds for each radiologist learned from bootstrapped samples using the probabilistic BN model.

Radiologist	Institution	Observed Threshold (%)
1	I	1.7
2	I	0.7
3	I	1.0
4	I	1.8
5	I	1.1
6	I	1.1
7	I	1.0
8	I	0.6
9	II	2.8
10	II	3.0
11	II	3.1
12	II	0.7
13	II	2.8

Table 3.

We show the BI-RADS assessment category distribution of our dataset. Note the dataset contains both screening and diagnostic mammograms.

BI-RADS Category	Count	Percentage
0	11,494	10.2%
1	60,018	53.4%
2	26,245	23.3%
3	12,656	11.3%
4	681	0.6%
4A	567	0.5%
4B	285	0.3%
4C	117	0.1%
5	370	0.3%
Total	112,433	100.0%