

# A Probabilistic Model to Support Radiologists' Classification Decisions in Mammography Practice

Jiaming Zeng, Francisco Gimenez, Elizabeth S. Burnside , Daniel L. Rubin\*, and Ross Shachter\*

We developed a probabilistic model to support the classification decisions made by radiologists in mammography practice. Using the feature observations and Breast Imaging Reporting and Data System (BI-RADS) classifications from radiologists examining diagnostic and screening mammograms, we modeled their decisions 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 2 separate institutions with a 1.1% prevalence of malignancy, we trained a probabilistic Bayesian network (BN) to estimate the malignancy probabilities of lesions. For each radiologist, we learned an observed probabilistic threshold within the model. We compared the sensitivity and specificity of each radiologist against the BN model using either their observed threshold or the standard 2% threshold recommended by BI-RADS. We found significant variability among the radiologists' observed thresholds. By applying the observed thresholds, the BN model showed 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 was 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

classification decision, decision support, mammography, observed threshold

Date received: September 20, 2018; accepted: January 18, 2019

The American Cancer Society recommends annual screening mammography for women older than 45 y to detect breast cancer early, when it is most treatable.<sup>1–3</sup> However, the U.S. Preventive Services Task Force recommends less aggressive screening based on literature that asserts that the harms of early and frequent screening outweigh the benefits.<sup>4,5</sup> 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.

Stanford University School of Engineering, Stanford, CA, USA (JZ, RS), Stanford University School of Medicine (Department of Biomedical Data Science, Radiology, and Medicine), CA, USA (FG, DLR) and University of Wisconsin Madison School of Medicine and Public Health, Madison, WI, USA (ESB). The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The authors received no financial support for the research, authorship, and/or publication of this article. \*These authors contributed equally to this work.

## Corresponding Author

Jiaming Zeng, Department of Management Science and Engineering, Stanford University School of Engineering, Huang Engineering Center, 475 Via Ortega, 338J, Stanford, CA 94305-4027, USA. (jjiaming@stanford.edu)

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.<sup>6-9</sup>

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 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.<sup>10-17</sup> While many have been shown to improve the performance of radiologists as well as to reduce their intrareader 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.<sup>18,19</sup> 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

### *Data Set*

For our study, we used a total of 112,433 mammography reports, with 1214 malignant cases, a 1.1% prevalence, collected from 13 radiologists across 2 teaching hospitals, 8 radiologists at Institution I and 5 radiologists at Institution II. The reports included 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 at Institution II from April 5, 1999, to February 9, 2004. We obtained Institutional Review Board approval for this research.

We defined a mammography case as the patient's risk factors, the radiologist's observations, and the pathological ground truth. Each case included 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 were split into 6 levels: 1, 2, or 3 indicated a negative assessment (no immediate follow-up), while 0, 4, or 5 indicated a positive assessment (follow-up imaging or biopsy recommended).<sup>18,19</sup> 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 1 y 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 recognized each case as either false positive, false negative, true positive, or true negative. A summary of the data set is shown in Table 1.

### *Probabilistic Model*

Building on the mammography Bayesian network (BN) model described by Burnside et al.,<sup>20</sup> we trained a BN to estimate a lesion's probability of malignancy based on the features identified by the radiologists. A BN 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.<sup>21</sup>

**Table 1** Our study includes 112,433 diagnostic and screening mammography cases with separate analysis for each of the 13 radiologists at 2 institutions

Radiologist ID	Institution	No. of Cases	Malignant	Benign	False Negative	False Positive	Prevalence (%)
1	I	1759	35	1724	4	225	2.0
2	I	8185	123	8062	10	988	1.5
3	I	11,415	151	11,264	25	1160	1.3
4	I	6144	32	6112	5	686	0.5
5	I	4126	69	4057	14	475	1.7
6	I	15,908	199	15,709	27	1369	1.3
7	I	8217	110	8107	10	992	1.3
8	I	3736	74	3662	10	629	2.0
9	II	23,497	174	23,323	32	2678	0.7
10	II	16,604	148	16,456	19	1840	0.9
11	II	3077	26	3051	3	313	0.8
12	II	3634	30	3604	4	420	0.8
13	II	6131	43	6088	13	701	0.7
Total	—	112,433	1214	111,219	176	12,476	1.1

Using the data set described, we learned both the structure and the parameters of the BN. The structure was learned using Tree-Augmented Nave Bayes.<sup>22,23</sup> 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.<sup>24</sup>

We estimated the posterior probability of malignancy for each case by using an iterative 10-fold model. The data were stratified into 10 folds by the 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 the next section). The operating point was 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.

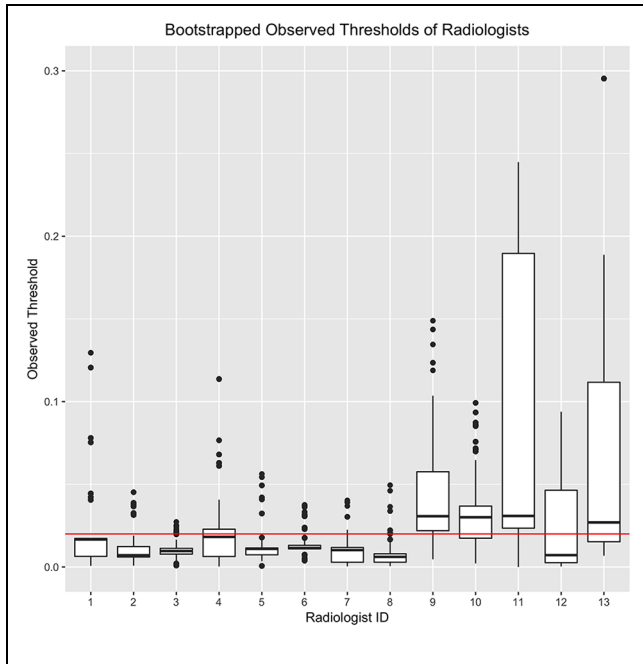
To match a radiologist's sensitivity, we calculated the probability of malignancy for each of the radiologist's

cases and found the largest probabilistic threshold with at least as much sensitivity as the radiologist. For most of our radiologists, this threshold led to an increase in specificity. For each radiologist, we repeated the process 5001 times over bootstrapped samples of the radiologist's cases. The median threshold from these samples was 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 1 more or 1 less false negative. Across all radiologists, there was a net increase of 1 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 was fairly high variability in our sampled thresholds across radiologists. This can be interpreted as interreader variability and lack of consistency of practice. Moreover, most of the radiologists deviated 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 implemented 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 compared the performance of these 3 classifications via sensitivity and specificity. We used McNemar's test of proportions to evaluate the statistical significance between each radiologist and each of the BN thresholds



**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% Breast Imaging Reporting and Data System–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.

for all of the radiologist’s cases. Significance was determined at a 95% confidence level ( $P < 0.05$ ). In addition, we used the two one-sided test to establish the noninferiority margin for sensitivity for each pair of methods. That margin was calculated to be 1.96 times the standard deviation of interreader sensitivity, corresponding to the 95% confidence interval across all radiologists. The confidence interval for each pair of proportions was calculated using the Agresti and Min method.<sup>25</sup>

All statistics were estimated by R 3.1.0.<sup>26</sup> Binary proportions testing was done using the DTComPair package version 1.03.<sup>27</sup> Bootstrapping was performed by the boot package version 1.3-11.<sup>28</sup>

## Results

We compared and analyzed the performance of the 3 methods for classifying mammograms as positive or negative: 1) each radiologist’s BI-RADS assessment, 2) classification using the radiologist’s observed threshold on

**Table 2** Observed thresholds for each radiologist learned from bootstrapped samples using the probabilistic Bayesian network model

Radiologist ID	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

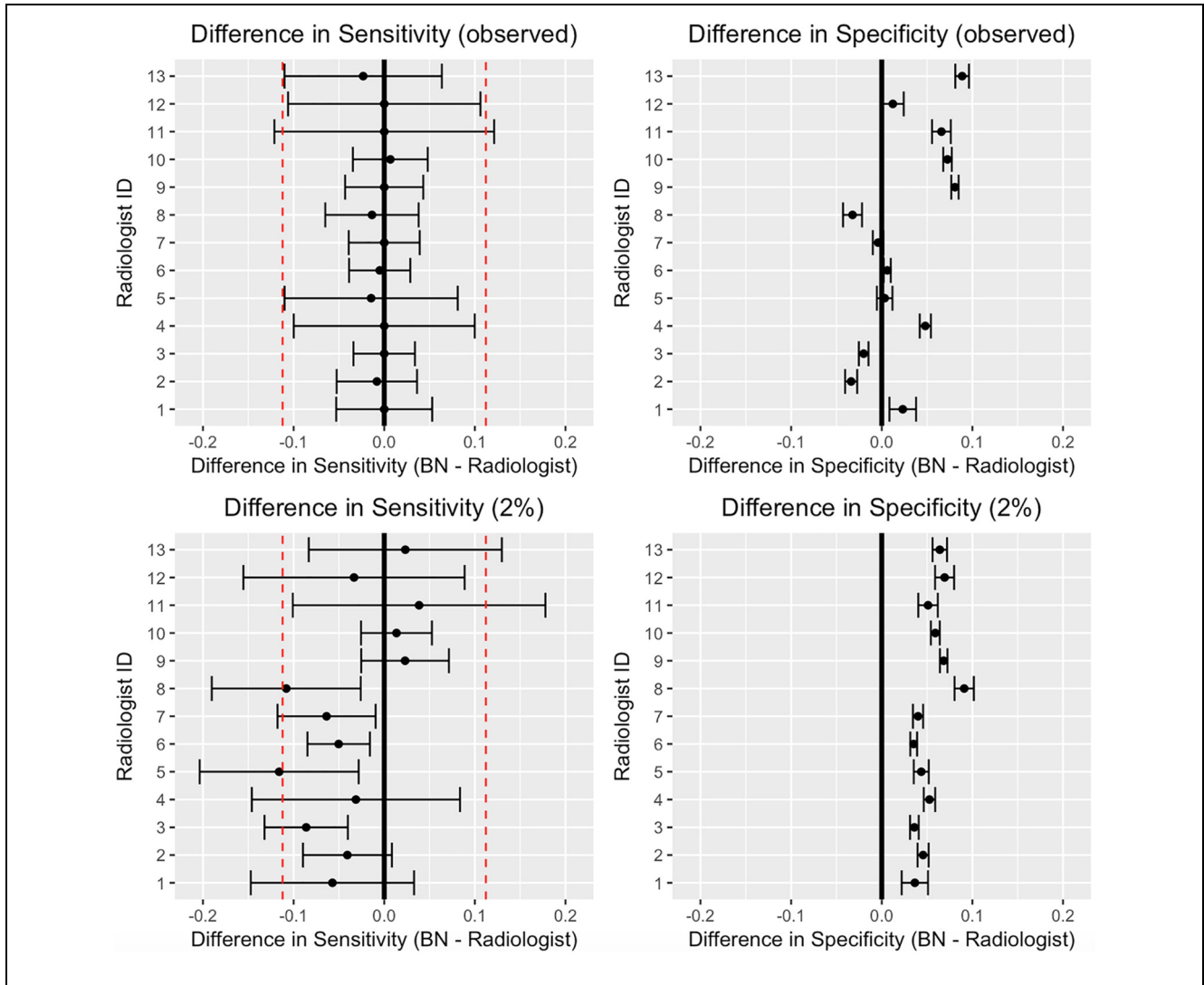
the probabilities estimated by our BN model, and 3) classification using the 2% BI-RADS threshold on those probabilities.

## Comparison of Sensitivity and Specificity

We compared 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 were considered noninferior to each other with respect to sensitivity if the 95% confidence interval of the difference in sensitivity, indicated by the whisker lines, was 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:** The sensitivity performance between radiologists and the BN model with observed thresholds was within the equivalence bounds and noninferior for all but 1 radiologist. The difference in specificity was statistically significant for 9 radiologists, with 6 increasing and 3 decreasing under the BN model. The remaining 4 radiologists showed no significant difference in specificity.
- **2% BI-RADS–recommended threshold:** The sensitivity performance between radiologists and the model’s 2% BI-RADS threshold was within the equivalence bounds for 4 radiologists, and the BN model was noninferior for 6 of the 13 radiologists.



**Figure 2** Comparison of sensitivity and specificity between each radiologist and the Bayesian network (BN) model using either the radiologist’s observed thresholds or the 2% Breast Imaging Reporting and Data System–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.

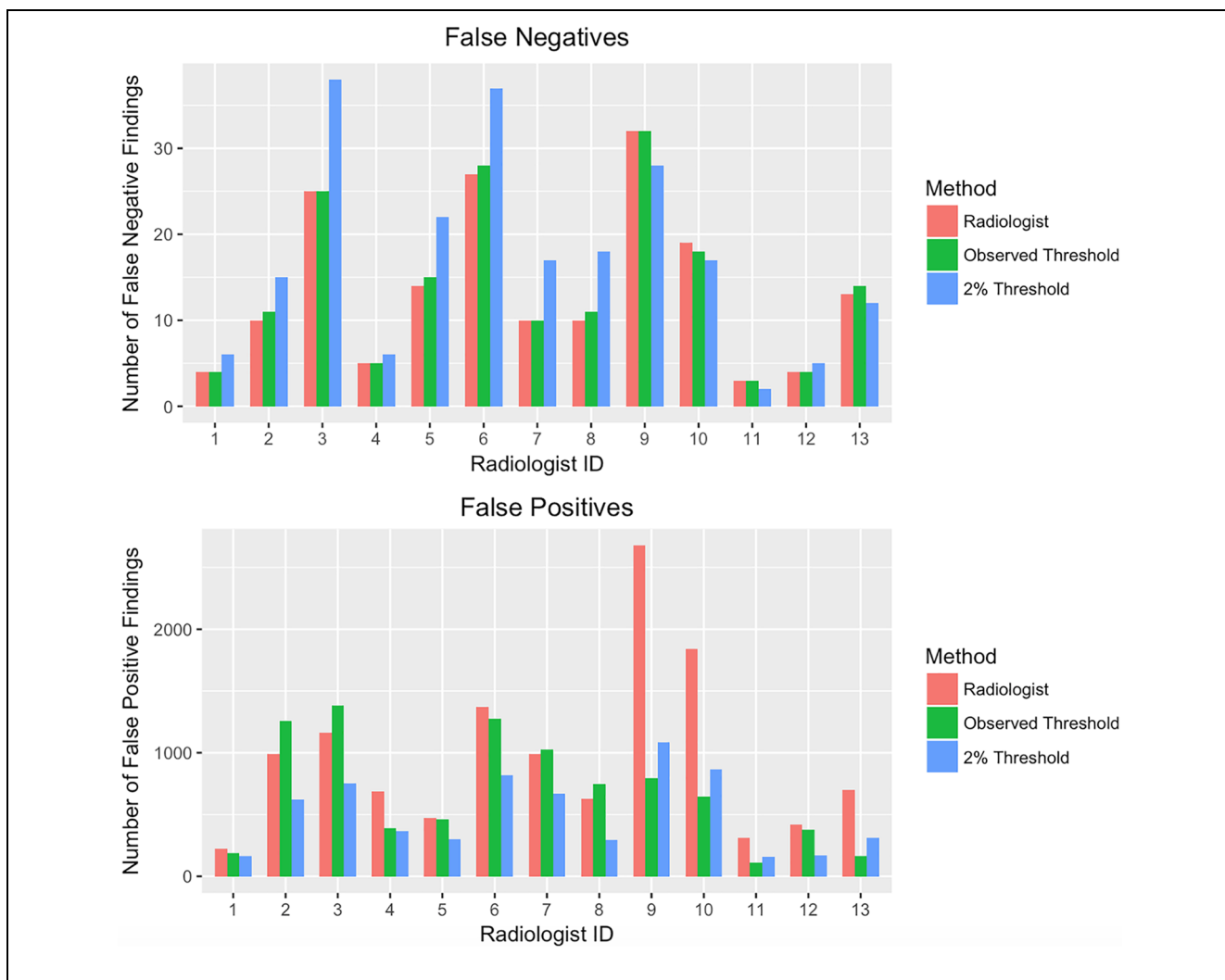
With respect to specificity, the 2% threshold BN model showed statistically significant improvement for all radiologists.

*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 were designed to match each radiologist’s sensitivity, so the false-negative counts for each radiologist were within one of the BN models, as seen in Figure 3. For 9 of the 13 radiologists, there was a decrease in false-positive counts for each, with no net increase in false-negative counts.



**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 Bayesian network model using either the observed threshold or the Breast Imaging Reporting and Data System 2% threshold for each radiologist’s cases.

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

Overall, with the observed thresholds, the BN model showed a net increase of 1 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 showed 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 did 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 demonstrated 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 BN model, we characterized 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 enforced the standard 2% BI-RADS–recommended threshold in our BN model, we saw 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 1 of 7 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,<sup>20,29</sup> and this is one possible reason that numerous studies have found significant variability in the effectiveness of mammography.<sup>7,30,31</sup> 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 was 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 of 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 in which we know the pathological ground truth.

Furthermore, the BI-RADS descriptors for mammography, although 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.<sup>32</sup> 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 radiologists reduce false positives. This indicates that radiologists are incorporating additional information not documented in the radiology reports (e.g., salient nonimage descriptors) when determining their final assessment.<sup>33</sup> In summary, we show that a collaborative decision support system for mammographic classification has the 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 the 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.

## Acknowledgments

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.



**ORCID iD**

Elizabeth S. Burnside  <https://orcid.org/0000-0002-6600-435X>

**Supplementary Material**

Supplementary material for this article is available on the *Medical Decision Making* Web site at <http://journals.sagepub.com/home/mdm>.

**References**

- Nyström L, Andersson I, Bjurstam N, et al. Long-term effects of mammography screening: updated overview of the swedish randomised trials. *Lancet*. 2002;359(9310):909–19.
- 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):1599–614.
- 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 Cancer J Clin*. 1997;47(3):134–49.
- Siu AL. Screening for breast cancer: US Preventive Services Task Force recommendation statement. *Ann Intern Med*. 2016;164(4):279–96.
- Kalager M, Adami HO, Bretthauer M, et al. Overdiagnosis of invasive breast cancer due to mammography screening: results from the Norwegian screening program. *Ann Intern Med*. 2012;156(7):491.
- 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):641–51.
- Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by us radiologists: findings from a national sample. *Arch Intern Med*. 1996;156(2):209–13.
- Elmore JG, Wells CK, Lee CH, et al. Variability in radiologists' interpretations of mammograms. *N Engl J Med*. 1994;331(22):1493–9.
- Taplin S, Abraham L, Barlow WE, et al. Mammography facility characteristics associated with interpretive accuracy of screening mammography. *J Natl Cancer Inst*. 2008;100(12):876–87.
- 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):1223–38.
- 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):787–94.
- Eadie LH, Taylor P, Gibson AP. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. *Eur. J Radiol*. 2012;81(1):e70–6.
- Fujita H, Uchiyama Y, Nakagawa T, et al. Computer-aided diagnosis: the emerging of three CAD systems induced by Japanese health care needs. *Comput Methods Programs Biomed*. 2008;92(3):238–48.
- Oliver A, Freixenet J, Marti J, et al. A review of automatic mass detection and segmentation in mammographic images. *Med Image Anal*. 2010;14(2):87–110.
- Bright TJ, Wong A, Dhurjati R, et al. Effect of clinical decision-support systems: a systematic review. *Ann Intern Med*. 2012;157(1):29–43.
- Burnside ES. Bayesian networks: computer-assisted diagnosis support in radiology 1. *Acad Radiol*. 2005;12(4):422–30.
- Burnside E, Rubin D, Shachter R. A Bayesian network for mammography. *Proc AMIA Symp*. 2000:106–10.
- Baker JA, Kornguth PJ, Floyd C Jr. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. *AJR Am J Roentgenol*. 1996;166(4):773–8.
- Liberman L, Menell JH. Breast Imaging Reporting and Data System (BI-RADS). *Radiol Clin*. 2002;40(3):409–30.
- 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):663–72.
- Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Burlington, MA: Morgan Kaufmann; 2014.
- Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. *Machine Learning*. 1997;29(2):131–63.
- Friedman N, Goldszmidt M. Building classifiers using Bayesian networks. In: *Proceedings of the National Conference on Artificial Intelligence*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence; 1996. p. 1277–84.
- Norsys Software Corp. Netica-J [manual]. Version 4.18 and higher. Vancouver, Canada: Norsys Software Corp; 2012.
- Agresti A, Min Y. Simple improved confidence intervals for comparing matched proportions. *Stat Med*. 2005;24(5):729–40.
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2014. Available from: <https://www.R-project.org/>
- Stock C, Hielscher T. Dcompair: comparison of binary diagnostic tests in a paired study design. R package version 1.0.3. 2014. Available from: <http://CRAN.R-project.org/package=DTCOMPAIR>
- Canty A, Ripley B. boot: Bootstrap r (s-plus) functions. R package version 1.3-11. 2014. Available from: <https://cran.r-project.org/package=boot>



29. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. In: Wendt D, Vlek CA, eds. *Utility, Probability, and Human Decision Making*. New York: Springer; 1975. p. 141–62.
30. Elmore JG, Nakano CY, Koepsell TD, et al. International variation in screening mammography interpretations in community-based programs. *J Natl Cancer Inst.* 2003; 95(18):1384–1393.
31. Cox B. Variation in the effectiveness of breast screening by year of followup. *JNCI Monogr.* 1997;1997(22):69–72.
32. Burnside ES, Sickles EA, Bassett LW, et al. The ACR BI-RADS® experience: learning from history. *J Am Coll Radiol.* 2009;6(12):851–60.
33. Sachchamarga W. *Using Conflicting Information and Holistic Judgments in Diagnostic Computer Models* [PhD thesis]. Stanford, CA: Stanford University; 2012.