J Digit Imaging
DOI 10.1007/s10278-017-0022-2

Understanding Clinical Mammographic Breast Density Assessment: a Deep Learning Perspective

Aly A. Mohamed[1] · **Yahong Luo**[2] · **Hong Peng**[1,3] · **Rachel C. Jankowitz**[4,5] · **Shandong Wu**[1,6,7]

Abstract Mammographic breast density has been established as an independent risk marker for developing breast cancer. Breast density assessment is a routine clinical need in breast cancer screening and current standard is using the Breast Imaging and Reporting Data System (BI-RADS) criteria including four qualitative categories (i.e., fatty, scattered density, heterogeneously dense, or extremely dense). In each mammogram examination, a breast is typically imaged with two different views, i.e., the mediolateral oblique (MLO) view and cranial caudal (CC) view. The BI-RADS-based breast density assessment is a qualitative process made by visual observation of both the MLO and CC views by radiologists, where there is a notable inter- and intra-reader variability. In order to maintain consistency and accuracy in BI-RADS-based breast density assessment, gaining understanding on radiologists' reading behaviors will be educational. In this study, we proposed to leverage the newly emerged deep learning approach to investigate how the MLO and CC view images of a mammogram examination may have been clinically used by radiologists in coming up with a BI-RADS density category. We implemented a convolutional neural network (CNN)-based deep learning model, aimed at distinguishing the breast density categories using a large (15,415 images) set of real-world clinical mammogram images. Our results showed that the classification of density categories (in terms of area under the receiver operating characteristic curve) using MLO view images is significantly higher than that using the CC view. This indicates that most likely it is the MLO view that the radiologists have predominately used to determine the breast density BI-RADS categories. Our study holds a potential to further interpret radiologists' reading characteristics, enhance personalized clinical training to radiologists, and ultimately reduce reader variations in breast density assessment.

Keywords Breast density · Deep learning · Digital mammography · Reading behavior · Radiology · Breast cancer

✉ Shandong Wu
wus3@upmc.edu

[1] Department of Radiology, University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA 15260, USA

[2] Department of Radiology, Liaoning Cancer Hospital & Institute, 44 Xiaoheyan Rd, Dadong District, Shenyang City, Liaoning Province 110042, China

[3] Department of Radiology, Chinese PLA General Hospital, 28 Fuxing Rd, Haidian District, Beijing 100853, China

[4] Magee-Womens Hospital of University of Pittsburgh Medical Center, 300 Halket St, Pittsburgh, PA 15213, USA

[5] Department of Medicine, University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA 15260, USA

[6] Department of Biomedical Informatics, University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA 15260, USA

[7] Department of Bioengineering, University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA 15260, USA

Introduction

In mammographic breast cancer screening, breast density is a routine clinical measure visually assessed by breast imaging radiologists. Breast density measures the relative amount of the dense (i.e., fibroglandular) tissue depicted on digital mammogram images. In current clinical workflow, breast density is mainly evaluated in terms of the Breast Imaging and Reporting Data System (BI-RADS) breast density criteria [1], including four qualitative categories, i.e., fatty, scattered density, heterogeneously dense, or extremely dense. There are

Published online: 20 September 2017

computational approaches to determine a quantitative breast density assessment [2–4]. Commercially available software such as Quantra [3] and Volpara [4] can compute a volume-based breast density but these methods function only on the raw ("FOR PROCESSING") digital mammogram images, which are not routinely stored in most medical centers. Current clinical standard of breast density assessment is predominately the BI-RADS-based qualitative categories.

The importance of breast density assessment is mainly due to the fact that breast density has long been shown to be a risk biomarker of developing breast cancer. Comparing women with "extremely dense" breasts to women with "fatty" breasts, their breast cancer risk is about four to six times higher [5]. When women's breast density falls into the category of either "heterogeneously dense" or "extremely dense," they are considered to have "dense breasts"—indicating a higher risk of breast cancer than "non-dense" breasts (i.e., "fatty" or "scattered density")—and that may trigger additional actions such as supplementary screening. In the clinical setting, it is apparently easy to distinguish the "fatty" breasts from the "extremely dense" breasts. However, it is highly confusing and difficult for radiologists to visually and consistently distinguish between the "scattered density" and "heterogeneously dense" categories [6]. There is a notable inter- and intra-reader variability in the visual breast density assessment made by radiologists [5, 6]. It poses a clinical need for radiologists to reliably assign a precise and consistent BI-RADS breast density category.

In screening digital mammography, each breast is typically imaged with two different views, i.e., the mediolateral oblique (MLO) view and cranial caudal (CC) view (Fig. 1). The MLO view is taken from the center of the chest outward, while the CC view is taken from above the breast. While CC view depicts the entire breast, MLO view reflects more of the breast in the upper-outer quadrant, giving the best view of the lateral side of the breast, which statistically is the most common place for pathological changes. While both the MLO and CC views were read by radiologists, only one BI-RADS density category is assigned. It has been roughly suspected that the difference of breast density assessment is small between the CC and MLO views. It however remains unclear how radiologists actually rely on or "subconsciously" use the two views in assigning a BI-RADS breast density category from the two views, particularly between the two difficult-to-distinguish categories, i.e., "scattered density" vs "heterogeneously dense."

In this study, we aim to understand radiologists' mammographic reading behaviors or characteristics. More specifically, we propose to leverage a newly emerged machine learning technique, i.e., deep learning, to investigate the underlying mechanism of how radiologists use the two mammogram views (MLO and CC) in giving rise to a BI-RADS breast

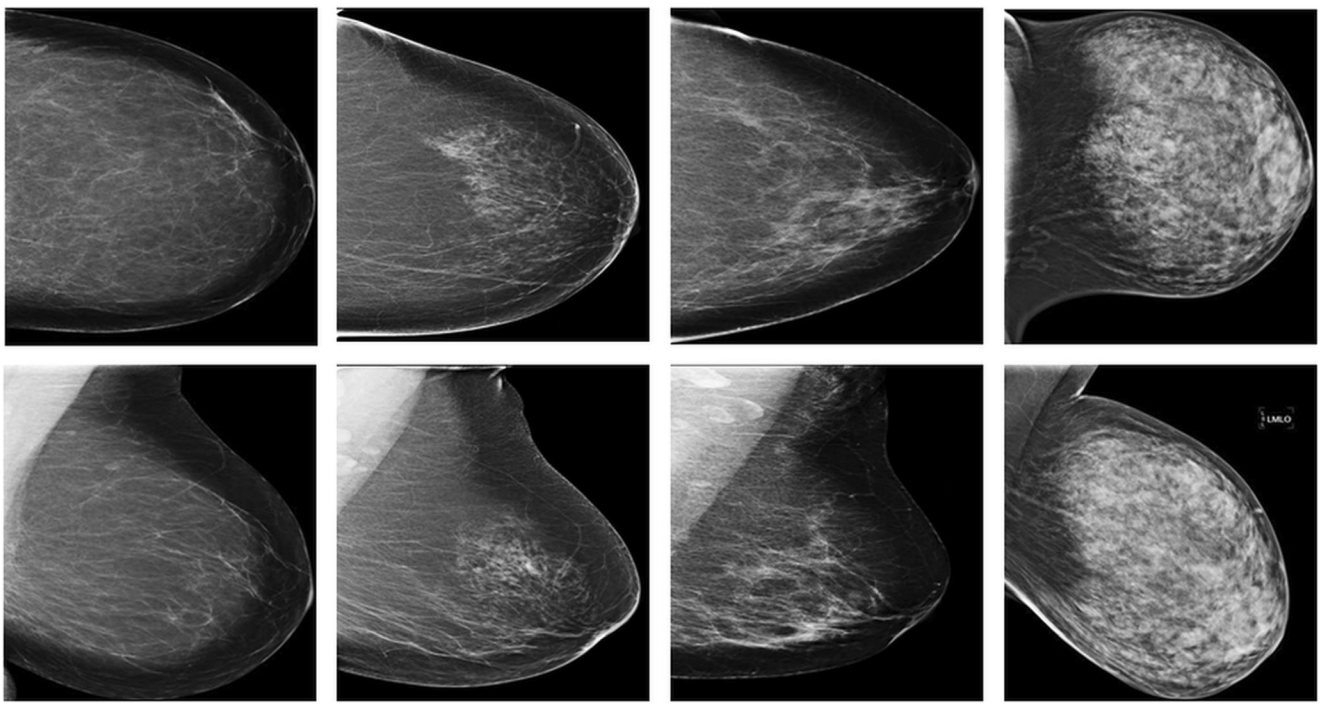

Fig. 1 Examples of the mediolateral oblique (MLO) view (row 1) and cranial caudal (CC) view (row 2) digital mammogram images, illustrating the four qualitative Breast Imaging and Reporting Data System (BI-RADS) breast density categories, i.e., fatty (column 1), scattered density (column 2), heterogeneously dense (column 3), and extremely dense (column 4)

density category. Using clinical visual observation to determine a BI-RADS breast density category is somewhat a "subconscious" and qualitative decision-making process. By employing deep learning-based classification, this process can be modeled "backward" to discover which view may play a predominant or equivalent role in assigning the BI-RADS breast density categories. This is an important clinical question to answer because by gaining knowledge on radiologists' reading patterns, it would help come up potential solutions to enhance their reading consistency and reduce reader variations, and thus, generate a more reliable breast density assessment. However, understanding such reading behaviors is not straightforward by traditional approaches, because the visual and qualitative decision-making process is hard to directly model or interpret. Hence, we propose to employ a novel deep learning-based approach to study this question.

Conventional machine learning is based on a strong feature engineering, i.e., using hand-crafted descriptors and prior expert knowledge of the data to build strong features. This process is time-consuming and hard for many scenarios; in particular, for the studied question of this work, it is difficult to directly summarize prior knowledge from radiologists' qualitative reading practice. On the other hand, deep learning can extract features directly and automatically from original data. Based on a large training dataset, deep learning has shown promising performance in many recent artificial intelligence applications. In biomedical imaging analyses, deep learning demonstrated impressive capabilities in thoraco-abdominal lymph node detection and interstitial lung disease classification [7], chest pathology identification [8, 9], real-time 2D/3D image registration [10], mammographic lesion detection and diagnosis [11], image segmentation [12], etc.

In this paper, we applied a deep learning architecture, convolutional neural network (CNN), to build a two-class breast density classifier aiming at distinguishing the breast density categories, using separately the MLO view and CC view of the same patient cohort. The goal is to evaluate the CNN model's classification accuracy with respect to the use of the MLO and CC views of a large mammogram imaging dataset, such that gain insights into the potential role of the MLO and CC view images in BI-RADS-based clinical breast density assessment.

Materials and Methods

Study Cohort and Imaging Data

This study received institutional review board (IRB) approval and was compliant to the Health Insurance Portability and Accountability Act (HIPAA). Informed consent from patients was waived. From a retrospectively identified cohort who underwent mammographic breast cancer screening at our institution, we identified a cohort of 963 women who underwent standard digital mammography screening from 2005 to 2016 with a total of 15,415 negative or breast cancer-free digital mammogram images. In average, there are 4 (range 1–7) mammogram examinations per patient. Out of the 15,415 images, there are 1135 "fatty" images, 6600 "scattered density" images, 6600 "heterogeneously dense" image, and 1080 "extremely dense" images, respectively. Each examination includes the MLO and CC views of the left and right breasts (i.e., 4 images). For each examination, the BI-RADS-based breast density categories assessed in standard clinical procedures by radiologists specialized in breast imaging were retrieved from mammogram reports. In this study, our analyses focused on the processed (i.e., "FOR PRESENTATION") mammogram images because the raw (i.e., "FOR PROCESSING") images were not routinely stored in our clinical setting.

CNN-Based Breast Density Classifier

We built two deep learning-based classification models each with two output classes. In the first model, the two classes correspond to the two most confusing BI-RADS categories of "scattered density" and "heterogeneously dense," respectively. In the second model, the two classes correspond to the "non-dense breasts" (combination of the fatty and scattered density images) and "dense breasts" (combination of the heterogeneously dense and extremely dense images), respectively. Note that here the "dense" vs "non-dense" classification represents specific clinical demand because the "dense breasts" are more concerned by both clinicians and patients in terms of elevated breast cancer risk and possible supplementary screening.

The deep learning-based classifier was implemented using the CNN structure [13] and an improved AlexNet model [14], which is not trained with the relighting data-augmentation. The CNN structure consists of five convolutional layers, three max-pooling layers, and three fully connected layers with a final two-way softmax function. The deep learning platform we used is Caffe running on a graphics processing unit (GPU)-accelerated desktop computer (Intel® Core™ i7-4790 CPU@3.60 GHZ with 8 GB RAM and a Titan X Pascal GPU). We also used rectified linear units (ReLU) in place of the traditional tangent function and the sigmoid function as the activation function [13] to speed up training.

In each image, the whole-breast region was first segmented automatically [15] from non-breast regions and used as input data of the CNN model. In order to calibrate the intensity contrast of the images, we applied histogram equalization to all images as a preprocessing step. We also generated the mean image of training data and subtracted the mean image from each input image to ensure that every feature pixel has a zero mean. CNN training was based on six-fold cross-

validation, with 70% of the entire images of each BI-RADS category for training, 15% for validation, and 15% for testing. In each time, the training and validation samples were randomly selected from the entire images and the remaining samples accordingly for testing. Receiver operating characteristic (ROC) analysis was used with computing the area under the ROC curve (AUC) as the performance metric of the breast density classifier [16]. We repeated the training and testing processes for ten times and the averaged classification AUCs were reported. Delong test [17] was used to assess the statistical significance for the differences of AUCs.

Results

As shown in Fig. 2, for the CNN model distinguishing the two categories of "scattered density" vs "heterogeneously dense," the AUC of the classification model is 0.95 when using only the MLO view images. In comparison, the AUC is 0.88 when using only the CC view images. It is observed that the AUC difference is significant ($p \leq 0.001$) when comparing the MLO vs. CC view. When both the MLO and CC view images were combined as a single dataset and the whole training and testing experiments were repeated on this combined dataset, the AUC is 0.92.

Similarly, the CNN model's accuracy of classifying the two classes of "dense breasts" vs "non-dense breasts" was shown in Fig. 3. As can be seen, the AUC is 0.97 or 0.92 when using only the MLO or CC view images (the AUC difference is significant with $p \leq 0.001$). AUC is 0.95 on the combined dataset of the CC and MLO view images.

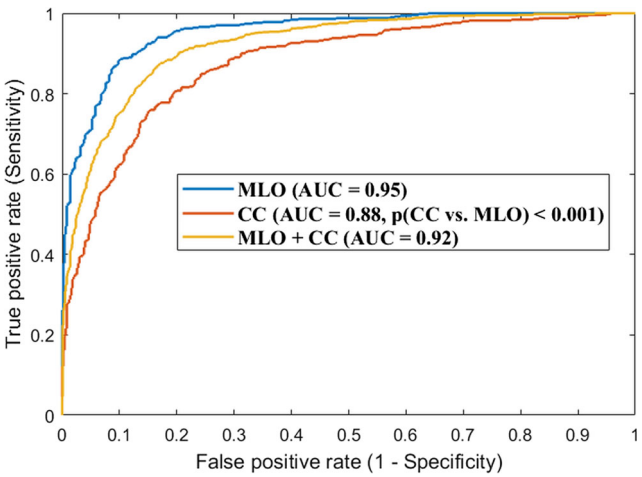

Fig. 2 Convolutional neural network (CNN)-based breast density classification accuracy between "scattered density" and "heterogeneously dense," when using the MLO view images alone, the CC view images alone, and the combination of the two view images. This result showed that the classification using MLO view is more predominate in comparison to that using the CC view

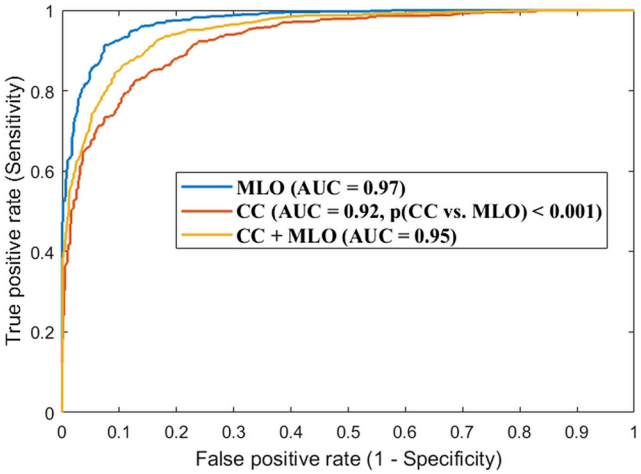

Fig. 3 Convolutional neural network (CNN)-based breast density classification accuracy between "dense breasts" and "non-dense breasts," when using the MLO view images alone, the CC view images alone, and the combination of the two view images. This result showed that the classification using MLO view is more predominate in comparison to that using the CC view

Discussion and Conclusions

In this study, we investigated a deep learning-based clinical application to understand how radiologists may use MLO and CC view mammogram images in assessing a patient's breast density by the BI-RADS-based density categories. In breast cancer screening, a large amount of digital mammogram examinations are acquired annually and it is a routine clinical need to assess breast density for every examination. However, a radiologist may not be able to reproduce his/her assessment, and amongst different radiologists, there are substantial discrepancies on assessing a breast either as "scattered density" or "heterogeneously dense." Reducing the reading variations on breast density represents a real clinical demand. In order to achieve so, it is critical to first get an understanding on some of the aspects of how radiologist actually "read" the images, like how the different views of a mammogram are used. We realize that there is essentially no "ground truth" for the BI-RADS-based breast density assessment. However, breast density assessment represents a significant clinical question and routine requirement in mammographic reading. While there is no satisfactory truth, we chose to study real clinical data and aimed to improve current existing clinical workflow.

Currently, there are more than 30 states in the USA that have enacted lawful breast density notification [18], requiring delivering some level of information of breast density assessment to patients after a screening mammogram. This notification may help patients understand the implications of breast density in perceiving their breast cancer risk and developing potential supplementary screening strategies. Therefore, reducing reader variations and improving consistency of breast density assessment hold important necessity and value in

better informing patients. One way to address this is to better understand the actual clinical mammogram reading behaviors of breast imaging radiologists. By studying classification accuracy of MLO and CC view images individually, and in combination, with respect to the actual clinical outcome, i.e., the assigned BI-RADS density categories, it will help discover radiologists' reading patterns to potentially help calibrate the reading to be more consistent.

In this study, we used a large real-world digital mammogram imaging dataset and standard clinical BI-RADS assessment as an effective end point. In our results, it is shown that the AUC of using the MLO view images is significantly higher than that using the CC view images. This indicates that most likely it is the MLO view that the radiologists have predominately used to determine the breast density BI-RADS categories. This is a clinically valuable finding because this knowledge can be educational in calibrating/training radiologists' reading towards generating a more consistent BI-RADS density assignment. In this regard, we believe further exploration is needed in future work. It should be noted that this is a single-center retrospective study, and therefore, the generalizability of our findings warrants further evaluation by a potential multi-center dataset.

In this work, the studied images were read by many radiologists during a long time period of the past several years. This is advantageous because the results are less likely to be driven by a certain radiologist's reading pattern. On the other hand, it is beyond the efforts that we could afford to track for all images that which images were read by which specific radiologists. If we were to have such information, it would enable us to examine the reading behaviors of a mixture of individual readers, and as a result, the insights we gained from that would bring us significant values in providing personalized training on improving image reading for clinical breast density assessment. This process has a great potential to be integrated in clinical quality control and enhance education to the radiologists.

While we focused on clinically used BI-RADS-based categories for breast density assessment, we acknowledge that using objective and quantitative measures of breast density will enhance the work. Unfortunately, the Volpara and Quantra softwares are not currently available at our institution or laboratory so we were not able to generate quantitative density measures using them. Moreover, the raw mammogram images were not routinely stored in our clinics, which prevented us from retrospectively computing quantitative density measures even if we had these softwares at hand. We expect the preliminary but promising results shown in this study will help us pursue future studies to look into more data analysis using quantitative density measures.

In addition, in our dataset, there were multiple examinations per patient. The correlations between these multiple examinations may have to do with the classification performance. The relationship will be worth of an in-depth analysis in our future work, when we have an adequately large number of patients where each patient has only one examination to compare with.

The principle of this study lies in the excellent capability of deep learning and CNN in directly learning relevant traits/features from annotated imaging data [13]. Because the BI-RADS-based density categorization is a qualitative process made by radiologists reflecting their visual observation and subconscious perception, it would be hard to directly interpret the radiologists' reading behaviors by traditional feature engineering of imaging. In this work, we demonstrated how we attempted to gain understanding of mammogram reading by the novel deep learning-based CNN models. This study represents a good example of showing strength of deep learning in identifying image reading patterns from a large data fed to the neural network. Potentially, we anticipate that our study will enhance current clinical assessment of breast density.

Acknowledgements This work was supported by a National Institutes of Health (NIH)/National Cancer Institute (NCI) R01 grant (#1R01CA193603), a Radiological Society of North America (RSNA) Research Scholar Grant (#RSCH1530), a Precision Medicine Pilot Award (#MR2014-77613) from the University of Pittsburgh Cancer Institute-Institute for Precision Medicine, and a Biomedical Modeling Pilot Award from the Clinical and Translational Science Institute of the University of Pittsburgh. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal graphics processing unit (GPU) used for this research. We thank Brenda F. Kurland for the helpful discussion related to this work.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest.

References

1. Obenauer S, Hermann KP, Grabbe E: Applications and literature review of the BI-RADS classification. Eur Radiol 15:1027–1036, 2005
2. Laboratory for Individualized Breast Radiodensity Assessment (LIBRA), available online: https://www.cbica.upenn.edu/sbia/software/LIBRA/index.html. Accessed date: August 1, 2016
3. Quantra, available online: http://www.hologic.com/en/breast-screening/volumetric-assessment/
4. Volpara, available online: http://www.volparadensity.com/
5. Huo CW, Chew GL, Britt KL et al.: Mammographic density—a review on the current understanding of its association with breast cancer. Breast Cancer Res Treat 144:479–502, 2014
6. Berg WA, Campassi C, Langenberg P, Sexton MJ: Breast imaging reporting and data system: inter-and intraobserver variability in feature analysis and final assessment. Am J Roentgenol 174(6):1769–1777, 2000
7. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura DJ, Summers RM: Deep convolutional neural networks for

computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 35(5):1285–1298, 2016

8. Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan. Chest pathology detection using deep learning with non-medical training. In 2015 I.E. 12th International Symposium on Biomedical Imaging (ISBI), pages 294–297. IEEE, 2015.
9. B. van Ginneken, A. A. Setio, C. Jacobs, and F. Ciompi. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In 2015 I.E. 12th International Symposium on Biomedical Imaging (ISBI), pages 286–289. IEEE, 2015.
10. Miao S, Wang ZJ, Liao R: A CNN regression approach for real-time 2D/3D registration. IEEE Trans Med Imag 35(5):1352–1363, May 2016
11. Cheng, Jie-Zhi et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Scientific Reports 6 (2016): 24454. PMC. Web. 25 Apr. 2017.
12. Kallenberg M et al.: Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. IEEE Trans Med Imag 35(5):1322–1331, 2016
13. LeCun Y, Bengio Y, Hinton G: Deep learning. Nature 521:436–444, 2015
14. Caffe, available online: https://github.com/BVLC/caffe/tree/master/models
15. Keller BM, Nathan DL, Wang Y, Zheng Y, Gee JC, Conant EF, Kontos D: Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation. Med Phys 39(8):4903–4917, 2012
16. Metz CE: ROC methodology in radiologic imaging. Invest Radiol. 21(9):720–733, 1986 Sep
17. DeLong ER, DeLong DM, Clarke-Pearson DL: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics:837–845, 1988
18. Breast density notification legislation, available online: http://densebreast-info.org/legislation.aspx