Published in final edited form as:

Acad Radiol. 2016 August ; 23(8): 977–986. doi:10.1016/j.acra.2016.02.019.

Association between Breast Parenchymal Complexity and False-positive Recall from Digital Mammography versus Breast Tomosynthesis: Preliminary Investigation in the ACRIN PA 4006 trial

Shonket Ray, PhD¹, Lin Chen, PhD¹, Brad M. Keller, PhD¹, Jinbo Chen, PhD², Emily F. Conant, MD¹, and Despina Kontos, PhD¹

¹Department of Radiology, University of Pennsylvania, 1 Silverstein Bldg., 3400 Spruce St, Philadelphia PA 19104

²Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, 423 Guardian Drive Philadelphia, PA 19104

Abstract

Rationale and Objectives—We investigate associations between measures of mammographic parenchymal complexity and false-positive (FP) recall from screening with digital mammography (DM) versus digital breast tomosynthesis (DBT).

Materials and Methods—We retrospectively analyzed data from 541 women recruited by the ACRIN 4006 trial, designed to evaluate callback and detection rates from screening with DM versus combined DM and DBT. Of these, 68 and 56 were false-positives based on DM alone versus the combined DM/DBT readings, respectively. Mammographic complexity was quantified with computerized texture analysis and percent density (PD). Logistic regression was performed to evaluate associations between extracted features and FP recall, after adjusting for age and number of previous benign biopsies. Odds ratios (OR) and area under the curve (AUC) of the receiver operating characteristic (ROC) were used to assess association strength.

Results—For DM, age, previous benign biopsies and texture features of correlation, inverse difference moment, sum average, and sum variance were deemed as significant predictors ($p < 0.05$) of FP recall, with an AUC=0.77. For DBT, age was the only significant predictor of FP recall with AUC=0.64. Using exploratory ROC thresholds for which no true-positives would be missed, a potential FP reduction of 23.5% and 8.9% was demonstrated, respectively, for DM alone versus DM/DBT.

Corresponding Author: Despina Kontos, PhD, Department of Radiology, University of Pennsylvania, Rm D702 Richards Bldg., 3700 Hamilton Walk, Philadelphia, PA 19104, Phone: 215-746-4064, Fax: 215-573-1811, Despina.Kontos@uphs.upenn.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conclusion—Measures of breast complexity measured on 2D digital mammograms are indicative of the likelihood for FP recall from screening with DM, and could help identify women who could benefit from supplemental screening, including digital breast tomosynthesis.

Keywords

Digital Mammography; Digital Breast Tomosynthesis; Breast Cancer Risk Assessment; Texture Analysis; False-Positive Recall

Introduction

False-positive (FP) recall from breast cancer screening is defined as being called back from screening for further assessment without breast cancer being subsequently detected. Currently, the majority of recalls from screening with digital mammography (DM) are false-positive, contributing to an overall low positive predictive value (PPV) (1) and concerns for over-imaging (1). This issue of false-positive recalls has been cited as a major contributing factor in many recent studies calling into question the true benefits and effectiveness of breast cancer screening programs (2–4). Reducing unnecessary FP events, without compromising the sensitivity of cancer detection, could reduce the overall cost related to any breast cancer screening program (5) due to less diagnostic workup and fewer biopsy procedures, while also reducing unnecessary physical pain, psychological trauma and potentially, scarring from unnecessary biopsy procedures (6). Digital breast tomosynthesis (DBT) is emerging as an alternative modality to complement or even replace DM, which has spurred the investigation of its potential benefits within the screening paradigm, especially in reducing FP recalls (7, 8). In DBT, tomographic breast images are reconstructed from multiple x-ray projections to produce a “quasi-3D” rendition of the breast, allowing superior breast tissue visualization and reducing the effect of tissue superposition inherent to 2D imaging (9–11).

In an ideal personalized paradigm, a prior test could be used to identify women at high-risk for such FP events, allowing for better implementations of individualized breast cancer screening protocols where supplemental screening tests could be offered to women for whom conventional mammography may have low specificity. Presently, mammographic breast density is a known risk factor for FP events (12). Breast density, which is commonly used to characterize the complexity of the breast tissue, is also known to confound the sensitivity of screening mammography (13). However, the commonly used mammographic density measures (i.e., percent density and/or the ACR BI-RADS categories) are global imaging descriptors that may not adequately capture the spatially localized conspicuities and complexity of the breast parenchymal tissue, which may be driving FP recalls (14). We hypothesize that mammographic features of parenchymal texture, computed in a spatially adaptive manner which characterizes local regions of confounding tissue texture, can provide complementary information to the more global density measures in describing the complexity of the breast tissue and therefore augment the assessment of the likelihood for a FP screening event. Previous literature has suggested potential relationships between parenchymal texture features and breast cancer risk (15–17). Here we use a locally-adaptive texture analysis algorithm to quantify mammographic parenchymal complexity and

investigate associations between such texture features and FP recall from breast cancer screening. We also examine if these features have a stronger association with FP recall from DM versus DBT screening, which would further suggest that DBT could be used as a replacement or supplement to DM for women with more texturally complex breast tissue.

Materials and Methods

Imaging

Anonymized DM images from the American College of Radiology Imaging Network (ACRIN) 4006 trial (18), were retrospectively analyzed. This trial was designed as a reader study to compare screening call-back rates from standard DM imaging versus a trial-specific protocol of low-dose DM acquisition combined with DBT, in an enriched cohort of women. Participants included two groups of patients, consisting of 503 asymptomatic women scheduled to undergo routine screening, who consented to be imaged both with conventional DM and the trial-specific DM and DBT combined protocol (for brevity, called DM/DBT) and 50 women called back from previous routine DM screening (i.e., the enriched recalled population), who also consented to receive additional imaging with the combined DM/DBT protocol as part of the trial. Using an on-site clinical reading protocol, radiologists were randomly assigned to studies and were blinded to complementary imaging. All DM and DBT readers were accredited under the Mammography Quality Standards ACT (MQSA) and had clinical DBT experience. All trial participants signed informed consent. Demographic data including age and number of previous benign biopsies were also collected. A total of 549 studies were available with six cases being confirmed as malignant cases or true-positive (TP) recalls. Of the remaining 543 negative screening studies, 68 and 57 were FP recalls based on recall decision for DM alone versus the combined DM/DBT protocol, respectively. For the purposes of our study, all available bilateral MLO view DM images acquired with the standard clinical dose were analyzed (Selenia Dimensions, Hologic Inc.). Preliminary evaluation with a subset of the trial-specific low-dose DM images has been previously reported (19). For our analysis, images were available in vendor post-processed ("For Presentation") Digital Imaging and Communications in Medicine (DICOM) format at 65–70 μ m resolution and 12-bit gray-level depth.

Mammographic Density Estimation

Mammographic percent density was estimated using a publically-available automated software tool, the Laboratory for Individualized Breast Radiodensity Assessment (LIBRA) (20, 21). This software algorithm has been previously validated against the current standard semiautomated Cumulus method (22), showing good agreement for both raw and vendor postprocessed digital mammograms (23). The algorithm initially segments the breast region and pectoral muscle, followed by an adaptive histogram-based determination of fuzzy c-means (FCM) clustering. This clustering step is based on the image gray-level intensities within the breast region for parenchymal tissue segmentation, and results into sub-clusters of tissue with similar densities, which are merged based on intensity and structural features to identify the entire region of dense tissue within the breast, and its relative ratio to the entire breast area (Figure 1). The bilateral estimates were averaged for each woman, to create a per-study estimate.

Parenchymal Texture Analysis

Thirteen parenchymal complexity features, specifically Haralick texture features (24), were estimated per study using a spatially locally adaptive computerized algorithm (19). The algorithm consists of two main steps: 1) adaptive selection of localized regions of interest (ROIs) and 2) subsequent texture feature extraction. For the first step, before ROI selection, the morphological operation of erosion (25) using a disk-shaped structure with radius set to 0.56mm is applied to the segmented dense tissue to remove noise-like small segmented pseudo-dense regions. Subsequently, the morphological flood-filling operation (26) is applied to fill small holes and join narrow breaks in addition to smoothing the segmented dense tissue (Figure 2). Once this step is established, 8-pixel connected adjacency (27) is used to automatically group the segmented dense regions by finding the set of all pixels connected to any given foreground pixel. Based on these connected components, a sequence of path is computed to identify the location of different regions. The ROIs are constructed around each labeled region as a bounding box with a maximum limit of 20 ROIs set per image view to cap the total number of ROIs for computational efficiency and to prevent image over-segmentation. To remove erroneous ROIs (typically containing very little parenchymal tissue and/or located on the skin/nipple/pectoral regions), three primary criteria are further applied (Figure 3). First, the overlapping ratio between each ROI size and the size of the corresponding foreground pixels is computed as:

$$r_i = \frac{|R_i^{\text{foreground}}|}{|R_i|} \quad (\text{Eq. 1})$$

where R_i is the i -th ROI rectangular box, $R_i^{\text{foreground}}$ is the region corresponding to the foreground pixels of R_i , and $|\cdot|$ returns the region (or area) of a given input argument. If r_i is less than a prefixed threshold (set to 0.15 based on empirical evidence (19)) for a specific ROI, then the ROI is marked as erroneous, and not further used in localized texture analysis. The second criteria used is estimating the area of each ROI and computing its square root (assuming the window is square). If this length measurement is outside of a specified mm value range (i.e., here set to 1.12 to 17.92 mm based on prior studies (28–30) to prevent the selection of unrealistically small ROIs or overly large ROI essentially comprising the whole image), the selected ROI is discarded. The final criteria applied calculate a ratio defined as the percent of the ROI within its corresponding processed breast mask region. If this ratio is under 50%, the ROI is dismissed because it likely covers too much air, skin or pectoral regions.

On a per-study basis, thirteen Haralick texture features (see Appendix A for related mathematical notations) were extracted based on Gray Level Co-occurrence Matrix (GLCM) computations and obtained by averaging the features computed from all ROIs from each MLO view. These features are based on second-order statistics that consider the spatial relationships of pixel-wise intensity values. Each GLCM measures the probability that a pixel of certain gray-level intensity appears at a specific distance and orientation from a pixel of another given intensity. Most of the texture feature calculations are weighted averages of normalized GLCM cell contents, designed to quantitatively “describe” the

appearance of certain patterns within a given image. For example, energy is a measure of intensity homogeneity of an image, correlation measures the linear dependency of grey levels between adjacent pixels and entropy measures the gray-level distribution randomness. These features have been frequently used to characterize breast parenchymal patterns in association to breast cancer risk (31–33).

Statistical Analysis

A total of sixteen variables were evaluated in this study: two representing the established FP risk factors of age and number of previous benign biopsies, breast percent density (PD) and the thirteen extracted texture features. Age and number of prior benign biopsies were considered as the “baseline” feature set. Breast PD was then added to construct a second augmented predictor set, and the final set contained the combination of all available variables, including the texture features. Prior to combining them, all texture features were categorized into quartiles with variance inflation factors (VIF) applied to remove potential multicollinearity (34). Backwards feature selection was then performed to further reduce the number of texture features from thirteen to five with VIF values below the rule-of-thumb threshold of 10 (i.e., higher VIF values indicate higher colinearity (34)). Stepwise logistic regression using bidirectional elimination was then implemented to finalize feature selection (35). Odds ratios (OR) were used to measure association strength between predictors and FP recall, and area under the curve (AUC) of the receiver operating characteristic (ROC) was used to measure predictiveness.

Finally exploratory analysis was performed to examine the feasibility for a threshold of a potential prior test for which a reduction in FP recall would not compromise cancer detection. For this, we used the model-generated likelihood for FP recall for the six biopsy-confirmed cancer cases (i.e., the true positives in the same screening trial) in conjunction with the overall predicted probability distributions for the entire cohort, to determine a probability threshold for the DM versus the DM/DBT stepwise regression model, for which a potential decision on FP determination could be made, without falsely dismissing any true-positive cancer cases. DeLong’s test (36) was applied to compare the ROC AUCs, where study significance was set to $p=0.05$. All related statistical analysis was performed with the R software package (v.3.1.0) (37).

This study was compliant with the Health Insurance Portability and Accountability Act and received institutional review board (IRB) approval with a waiver of consent.

Results

For DM interpretation, both texture features and established risk factors had a significant association with FP recall (Figure 4). For the baseline model, age (OR = 0.96, 95% CI: 0.93–0.98, $p<0.05$) and previous benign biopsies (OR = 0.46, 95% CI: 0.23–0.93, $p<0.05$) were both deemed as significant predictors of FP recall, yielding an AUC of 0.67 (95% CI: 0.60–0.73). Adding breast PD increased the AUC to 0.69 (95% CI: 0.63–0.75), with age (OR = 0.95, 95% CI: 0.92–0.97, $p<0.05$), previous benign biopsies (OR = 0.47, 95% CI: 0.23–0.95, $p<0.05$) and PD (OR = 0.97, 95% CI: 0.95–1.00, $p<0.05$) being significant predictors. When combining all variables, including the texture descriptors through stepwise

selection, the features of correlation (OR = 4.26, 95% CI: 1.55–11.72, $p < 0.05$), inverse difference moment (OR = 2.78, 95% CI: 1.06–7.33, $p < 0.05$), sum average (OR = 2.80, 95% CI: 1.20–6.55, $p < 0.05$ for 4th versus 1st quartile as a reference), sum variance (OR = 0.31, 95% CI: 0.10–0.92, $p < 0.05$ for 3rd versus 1st quartile as a reference) were also deemed significant predictors in addition to age (OR = 0.94, 95% CI: 0.92–0.97, $p < 0.05$) and number of prior benign biopsies (OR = 0.46, 95% CI: 0.22–0.94, $p < 0.05$). This combined feature set had the highest AUC of 0.77 (95% CI: 0.71–0.83) (Table 1), which was statistically significantly different ($p < 0.05$) from the AUC values obtained from the baseline model and the model augmented only with breast PD. All ROC curves had AUC p -values < 0.001 , therefore statistically significant at the study's $\alpha = 0.05$ level.

For the models based on the combined DM/DBT readings (Figure 5), age and previous benign biopsies had an AUC of 0.64 (95% CI: 0.56–0.71). Adding breast PD to this baseline model did not further increase the AUC. When combining all variables, including the texture features via stepwise selection, only age (OR = 0.96, 95% CI: 0.93–0.99, $p < 0.05$) was deemed as a significant predictor (Table 2). The AUC of this ROC curve was not statistically significantly different ($p > 0.05$) from the AUC values obtained from the baseline model and the model augmented only with breast PD. Compared to the corresponding combined model based on the DM-alone readings, this AUC was statistically significant different ($p < 0.05$).

As an exploratory analysis, we compared the DM FP recall probability histograms for both FP recalls and true negatives (TN) populations with the discrete probabilities of the six individual cancer cases (Figure 6.a). The six cancer cases ranged in predicted probability from 0.005 to 0.369, indicating a low likelihood of being false-positives. Using the maximum of these probabilities as an exploratory threshold for determining FP recalls where ideally no cancers would be missed, a total of 16 out of the 68 false-positive screens, representing 23.53% of the total FP recalls, would have been dismissed as true negatives. We then compared the FP recall probability histograms for both FP and TN recalls based on the combined DM/DBT protocol readings, with the discrete probabilities of the six individual cancer cases (Figure 6.b). The six cancer cases ranged respectively in predicted probability from 0.080 to 0.176, similarly suggesting low likelihood of representing false-positive recalls. Again, using the maximum of these two probabilities as a potential exploratory threshold value for which a decision on FP recall could be made and ideally no cancers would be missed, a total of 5 out of the 56 false-positives would have been dismissed, representing 8.93% of the total FP recalls.

Discussion

Our preliminary analysis suggests that measures of breast complexity are indicative of the likelihood for false-positive recall from screening with digital mammography, and therefore could help identify women for whom mammographic specificity is low who may benefit from supplemental screening. Established risk factors, including breast density, age and number of previous benign biopsies are also shown to have significant associations with FP recall from screening. The addition of mammographic texture features was shown to strengthen the observed associations. Specifically, parenchymal texture features of correlation, a measure of linear dependency of local tissue structure patterns (38), was

associated with higher odds of FP recall. Likewise, inverse difference moment, a measure of local tissue homogeneity (39), was also shown to be associated with higher odds of FP recall when comparing women with the lowest texture homogeneity, and therefore the lowest tissue uniformity, to those with the highest texture homogeneity and therefore with the most uniformity in their parenchymal tissue structure (i.e., comparing the fourth to first quartile). The other two highlighted texture features of sum average and sum variance, reflect the mean and variance, respectively, calculated from local image intensities (i.e., contrast patterns) (24). When these texture features were collectively considered as FP recall risk factors, breast PD was not selected by stepwise logistic regression and ceased to retain statistical significance as a FP risk factor, possibly as a result of the texture features already capturing breast parenchymal details inherent to breast density. This supports our hypothesis that refined measures of local mammographic texture may be better markers than global density descriptors in capturing intrinsic parenchymal complexity, and therefore important indicators of the confounding effect that this may have on mammographic specificity (Figure 7).

When comparing these associations between screening with DM alone, versus the combined DM/DBT imaging, no significant associations were observed between the texture features and FP recall, suggesting that the confounding effect of the superimposed complex breast tissue in projection mammography can be alleviated when screening is supplemented with DBT. This, in turn, would further suggest that women with inherently complex breast tissue could benefit from screening with three-dimensional imaging, such as with DBT, by sparing them unnecessary recalls that are more likely with DM alone screening, in which superimposed normal breast tissues may falsely appear as suspicious pseudo-lesions. In addition, age for both DM and DBT along with the number of previous benign biopsies for DM, were shown to be associated with a lower odds of FP recall.

We also investigated the feasibility of a hypothetical diagnostic test (40) that could help identify women at possible higher risk for FP recall, based on their parenchymal complexity. As a preliminary evaluation, we selected an exploratory probability cutoff threshold equivalent to the maximum predicted probability of FP recall for the six true-positive cancers, to therefore prevent any misclassifications of cancers in our cohort, which would have represented a decrease of 23.5% in FP recall for DM-alone interpretation. Similarly for the DM/DBT readings, a decrease of 8.9% in FP recall was demonstrated as potentially feasible. Although preliminary and relatively small in magnitude, this suggests a potential for clinically important reduction in FP recalls, based on these parenchymal complexity measures, especially when considering an extrapolation of such a reduction at the millions of women screened in the general population.

The primary limitation of our study is the relatively small number of screening exams available. Considering the limited scope of our analysis, larger studies are needed to validate our findings, while also including additional established clinical risk factors. Therefore, our study primarily serves as a proof-of concept investigation, especially in regards to the evaluation of the locally adaptive parenchymal texture measures. Nevertheless, it is important to note that in our analysis we were able to take into account two critical factors for statistical modeling accuracy and stability: feature overfitting and multicollinearity.

Performing the VIF selection in conjunction with stepwise regression allowed us to identify features that are potentially inherently collinear and could be eliminated using both selection criteria, along with reducing overfitting by decreasing the overall number of included predictors. In addition, the likelihood test demonstrated that all listed models were stable and at least one predictor independently contributed to the modeling outcome.

It is important to recognize, however, that direct clinical interpretation of all these texture features, which are fundamentally mathematical and based on image-processing techniques, may be difficult to perform. Even with certain texture features identified as significant predictors with meaningful OR values, precisely defining and identifying the “complexity” of the observed parenchymal patterns may still be challenging to interpret quantitatively. Future descriptors could be developed by further taking into consideration the clinical perspective that influences recall decision based on screening imaging and examinations. Finally, further validation could lead to the incorporation of such parenchymal complexity descriptors, along with breast density and other established risk factors into a comprehensive quantitative Breast Complexity Index (BCI), for incorporation into precision-medicine clinical decision-aid tools. Ideally, such a BCI would be predictive of higher odds of false positive recalls, as well as the potential for false negative screens, adding supplemental information to global measures of breast density.

Conclusions

In conclusion, the inclusion of breast parenchymal complexity features derived from a locally-adaptive texture analysis algorithm with established clinical risk factors and breast density demonstrated significant associations to FP recall from breast cancer screening. Additionally, these features were shown to have stronger associations with FP recall based on DM interpretation alone, compared to a combined DBT-DM imaging protocol. Future investigations could incorporate such texture features into an imaging index that accurately characterizes breast parenchymal complexity, allowing for better identification of women for whom mammography may have low specificity and a high-risk for false-positive recall, who may potentially benefit from supplemental screening that incorporates tomosynthesis imaging.

Acknowledgments

Sources of Support:

This work was supported in part by the National Institutes of Health (NIH) 5R01CA161749-02 and U54CA163313-01 grants, and a Breast Cancer Alliance Young Investigator Grant. We also thank the American College of Radiology Imaging Network for the access to the 4006 trial data.

Abbreviations

FP	false-positive
DM	digital mammography
PPV	positive predictive value

DBT	digital breast tomosynthesis
ACR BI-RADS	American College of Radiology Breast Imaging-Reporting and Data System
IRB	institutional review board
ACRIN	American College of Radiology Imaging Network
TP	true-positive
MLO	mediolateral oblique
DICOM	Digital Imaging and Communications in Medicine
PD%	percent density
LIBRA	Laboratory for Individualized Breast Radiodensity Assessment
FCM	fuzzy c-means
ROI	region of interest
GLCM	Gray Level Co-occurrence Matrix
VIF	variance inflation factors
OR	odds ratio
AUC	area under the curve
ROC	receiver operating characteristic
CI	confidence interval
TN	true negative
BCI	Breast Complexity Index

References

1. Christiansen CL, Wang F, Barton MB, Kreuter W, Elmore JG, Gelfand AE, Fletcher SW. Predicting the cumulative risk of false-positive mammograms. *Journal of the National Cancer Institute*. 2000; 92(20):1657–66. [PubMed: 11036111]
2. Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. *New England Journal of Medicine*. 2012; 367(21):1998–2005. [PubMed: 23171096]
3. Screening IUPoBC. The benefits and harms of breast cancer screening: an independent review. *The Lancet*. 2012; 380(9855):1778–86.
4. Miller AB, Wall C, Baines CJ, Sun P, To T, Narod SA. Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *Bmj*. 2014;348.
5. Schousboe JT, Kerlikowske K, Loh A, Cummings SR. Personalizing mammography by breast density and other risk factors for breast cancer: analysis of health benefits and cost-effectiveness. *Annals of internal medicine*. 2011; 155(1):10–20. [PubMed: 21727289]

6. Brett J, Austoker J. Women who are recalled for further investigation for breast screening: psychological consequences 3 years after recall and factors affecting re-attendance. *Journal of Public Health*. 2001; 23(4):292–300.
7. Houssami N, Skaane P. Overview of the evidence on digital breast tomosynthesis in breast cancer detection. *The Breast*. 2013; 22(2):101–8. [PubMed: 23422255]
8. Skaane P, Bandos AI, Gullien R, Eben EB, Ekseth U, Haakenaasen U, Izadi M, Jebsen IN, Jahr G, Krager M. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. *Radiology*. 2013; 267(1):47–56. [PubMed: 23297332]
9. Friedewald SM, Rafferty EA, Rose SL, Durand MA, Plecha DM, Greenberg JS, Hayes MK, Copit DS, Carlson KL, Cink TM. Breast cancer screening using tomosynthesis in combination with digital mammography. *Jama*. 2014; 311(24):2499–507. [PubMed: 25058084]
10. McCarthy AM, Kontos D, Synnestvedt M, Tan KS, Heitjan DF, Schnall M, Conant EF. Screening outcomes following implementation of digital breast tomosynthesis in a general-population screening program. *Journal of the National Cancer Institute*. 2014; 106(11):dju316. [PubMed: 25313245]
11. Conant EF. Clinical implementation of digital breast tomosynthesis. *Radiologic clinics of North America*. 2014; 52(3):499–518. [PubMed: 24792652]
12. Lehman C, White E, Peacock S, Drucker M, Urban N. Effect of age and breast density on screening mammograms with false-positive findings. *AJR American journal of roentgenology*. 1999; 173(6):1651–5. [PubMed: 10584815]
13. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, Jong RA, Hislop G, Chiarelli A, Minkin S. Mammographic density and the risk and detection of breast cancer. *New England Journal of Medicine*. 2007; 356(3):227–36. [PubMed: 17229950]
14. Sickles E, D'Orsi C, Bassett L. *Radiology ACo. ACR BI-RADS® Mammography. ACR BI-RADS Atlas, Breast Imaging Reporting and Data System*. 2013:141–67.
15. Li H, Giger ML, Olopade OI, Lan L. Fractal analysis of mammographic parenchymal patterns in breast cancer risk assessment. *Academic radiology*. 2007; 14(5):513–21. [PubMed: 17434064]
16. Li H, Giger ML, Olopade OI, Margolis A, Lan L, Chinander MR. Computerized Texture Analysis of Mammographic Parenchymal Patterns of Digitized Mammograms. *Academic Radiology*. 2005; 12(7):863–73. [PubMed: 16039540]
17. Wolfe JN. Breast patterns as an index of risk for developing breast cancer. *American Journal of Roentgenology*. 1976; 126(6):1130–7. [PubMed: 179369]
18. Conant E, Maidment A, Copit D, Olson C, Heckel M, Gatsonis C. OT2-05-04: ACRIN PA 4006: Comparison of Full-Field Digital Mammography with Digital Breast Tomosynthesis Image Acquisition in Relation to Screening Call-Back Rate. *Cancer Research*. 2011; 71(24 Supplement 3)
19. Ray S, Choi JY, Keller BM, Chen J, Conant EF, Kontos D. Application of computer-extracted breast tissue texture features in predicting false-positive recalls from screening mammography. *SPIE Medical Imaging: International Society for Optics and Photonics*. 2014 p. 90351X-X-7.
20. Keller, BM.; Hsieh, MMK. [Last Accessed, February 8, 2016] LIBRA: Laboratory for Individualized Breast Radiodensity Assessment. Available from: <https://www.cbica.upenn.edu/sbia/software/LIBRA/index.html>
21. Gastounioti, A.; Oustimov, A.; Keller, BM.; Pantalone, L.; Hsieh, M-K.; Conant, EF.; Kontos, D. Associations of dense and fatty breast-tissue heterogeneity with breast cancer risk: Preliminary evaluation using parenchymal texture measurements driven by breast anatomy. *Radiological Society of North America (RSNA) Annual Meeting; Chicago, IL*. 2015.
22. Byng JW, Boyd N, Fishell E, Jong R, Yaffe MJ. The quantitative analysis of mammographic densities. *Physics in medicine and biology*. 1994; 39(10):1629. [PubMed: 15551535]
23. 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. *Medical physics*. 2012; 39(8):4903–17. [PubMed: 22894417]

24. Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*. 1973; 6:610–21.
25. Dougherty, ER.; Lotufo, RA. *Hands-on Morphological Image Processing*. Bellingham, WA: SPIE press; 2003. p. 4-8.
26. Soille, P. *Morphological Image Analysis: Principles and Applications*. Secaucus, NJ: Springer-Verlag New York, Inc; 2013. p. 173-5.
27. Haralick, RM.; Shapiro, LG. *Computer and Robot Vision*. Boston, MA: Addison-Wesley Longman Publishing Co., Inc; 1992. p. 28-48.
28. Li H, Giger ML, Huo Z, Olopade OI, Lan L, Weber BL, Bonta I. Computerized analysis of mammographic parenchymal patterns for assessing breast cancer risk: effect of ROI size and location. *Medical physics*. 2004; 31(3):549–55. [PubMed: 15070253]
29. Zheng Y, Keller BM, Ray S, Wang Y, Conant EF, Gee JC, Kontos D. Parenchymal texture analysis in digital mammography: A fully automated pipeline for breast cancer risk assessment. *Medical physics*. 2015; 42(7):4149–60. [PubMed: 26133615]
30. Keller BM, Oustimov A, Wang Y, Chen J, Acciavatti RJ, Zheng Y, Ray S, Gee JC, Maidment AD, Kontos D. Parenchymal texture analysis in digital mammography: robust texture feature identification and equivalence across devices. *Journal of Medical Imaging*. 2015; 2(2):024501. [PubMed: 26158105]
31. Daye D, Keller B, Conant EF, Chen J, Schnall MD, Maidment AD, Kontos D. Mammographic parenchymal patterns as an imaging marker of endogenous hormonal exposure: a preliminary study in a high-risk population. *Academic radiology*. 2013; 20(5):635–46. [PubMed: 23570938]
32. Kontos D, Bakic PR, Carton A-K, Troxel AB, Conant EF, Maidment AD. Parenchymal texture analysis in digital breast tomosynthesis for breast cancer risk estimation: a preliminary study. *Academic radiology*. 2009; 16(3):283–98. [PubMed: 19201357]
33. Kontos D, Ikejimba LC, Bakic PR, Troxel AB, Conant EF, Maidment AD. Analysis of parenchymal texture with digital breast tomosynthesis: comparison with digital mammography and implications for cancer risk assessment. *Radiology*. 2011; 261(1):80–91. [PubMed: 21771961]
34. Kutner, MH.; Nachtsheim, C.; Neter, J. *Applied Linear Regression Models*. New York, NY: McGraw-Hill/Irwin; 2004. p. 406-8.
35. Menard, S. *Applied Logistic Regression Analysis*. 2. Thousand Oaks, CA: Sage Publications Inc; 2002. p. 63-6.
36. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988:837–45. [PubMed: 3203132]
37. Team RC. *R foundation for Statistical Computing*. 2005. R: A language and environment for statistical computing.
38. Van Der Sanden J, Hoekman D. Review of relationships between grey-tone co-occurrence, semivariance, and autocorrelation based image texture analysis approaches. *Canadian Journal of Remote Sensing*. 2005; 31(3):207–13.
39. Kumar RM, KS. A Survey on Image Feature Descriptors. *International Journal of Computer Science and Information Technologies (IJCSIT)*. 2014; 5(6):7668–73.
40. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ: British Medical Journal*. 2001; 323(7305):157. [PubMed: 11463691]

Figure 1.

An example of (a) an original digital mammogram, (b) the segmented dense parenchymal tissue region (green contour) using the segmentation algorithm, and (c) the resulting binary mask image with segmented parenchymal tissue.

Figure 2.

An example of (a) a region of interest (ROI) in a digital mammogram, (b) the corresponding binary mask image produced using the segmentation algorithm, and (c) resulting binary mask image after applying erosion and flood-filling to reduce segmentation “noise”.

Figure 3.
Example of a digital mammogram superimposed with all the identified regions of interest (ROIs) in boxes, where red indicates the ROIs included and blue indicates the ROIs excluded from subsequent analysis, after applying the related criteria thresholds.

Figure 4. ROC curves of the three feature sets used for predicting FP recall based on DM readings, showing a statistically significant increase in the AUC when parenchymal texture features are also included as predictors.

Figure 5.

ROC curves of the three feature sets for predicting FP recall based on the combined DM/DBT reading, showing that parenchymal texture does not statistically significantly increase the predictive AUC value.

Figure 6.

FP recall risk probability histograms of FP recalls (light red) and TN non-recalls (light blue) including individual cancer cases with discrete probability values for (a) DM and (b) the combined DM/DBT protocol readings, showing that a portion of the FP distribution could potentially be identified without missing any true-positive (vertical lines) cancer cases.

Figure 7.

Example of left MLO digital mammograms of (a) a 63 year old woman with a false-positive recall from screening (PD = 11.30% and BIRADS density = “heterogeneously dense”) and (b) a 56 year old woman not recalled from screening (PD = 13.39% and BIRADS = “heterogeneously dense”); both have the same BIRADS score and a similar low percent density (PD) but visually different complexity patterns with the FP recall case having higher complexity (texture inverse difference moment = 0.579) than the non-recalled case (texture inverse difference moment = 0.557).

Summary table for logistic regression analysis of FP recall decision from DM interpretation alone, for the stepwise regression feature set. The texture features of correlation (f2), inverse difference moment (f5), sum average (f6), sum variance (f7), difference variance (f10) are coded as categorical with each predictor's 2nd, 3rd or 4th quartile being compared to the reference 1st quartile for its corresponding values.

Table 1

Feature Predictor	Quartile	Coefficient Estimates (B)	p-values	OR	OR 95% CI
Age*	N/A	-0.058	0.000	0.943	0.915 – 0.973
Number of previous benign biopsies*	N/A	-0.784	0.033	0.457	0.222 – 0.940
	2	0.304	0.514	1.355	0.543 – 3.381
	3*	1.448	0.005	4.225	1.545 – 11.717
	4	0.657	0.312	1.930	0.539 – 6.907
f5 (inverse difference moment)	2*	1.024	0.038	2.783	1.057 – 7.332
	3	0.259	0.638	1.296	0.440 – 3.819
	4	0.444	0.477	1.559	0.459 – 5.301
	2	-0.034	0.942	0.967	0.388 – 2.408
f6 (sum average)	3*	1.036	0.011	2.817	1.263 – 6.282
	4*	1.031	0.017	2.804	1.200 – 6.551
	2*	-1.341	0.008	0.261	0.098 – 0.701
	3*	-1.189	0.036	0.305	0.100 – 0.924
f7 (sum variance)	4	-0.657	0.360	0.518	0.127 – 2.114
	2	0.739	0.113	2.093	0.839 – 5.224
	3	-0.110	0.863	0.896	0.259 – 3.096
	4	-0.512	0.498	0.600	0.136 – 2.634

* denotes statistical significance at $\alpha=0.05$ level.

Table 2

Summary table for logistic regression analysis of FP recall decision from combined DM/DBT interpretation for the stepwise selection feature set with statistically significant predictors indicated by asterisks. Out of all predictors evaluated, only age was significant.

Feature Predictor	Coefficient Estimates (B)	p-values	OR	OR 95% CI
Age *	-0.043	0.003	0.958	0.932 – 0.985

* denotes statistical significance at $\alpha=0.05$ level.