Breast Imaging Reporting and Data System (BI-RADS) breast composition descriptors: Automated measurement development for full field digital mammography

E. E. Fowler, T. A. Sellers, and B. Lu

Department of Cancer Epidemiology, Division of Population Sciences, H. Lee Moffitt Cancer Center, Tampa, Florida 33612

J. J. Heine^{a)}

Department of Cancer Imaging and Metabolism, H. Lee Moffitt Cancer Center, Tampa, Florida 33612

(Received 12 July 2013; revised 17 September 2013; accepted for publication 20 September 2013; published 21 October 2013)

Purpose: The Breast Imaging Reporting and Data System (BI-RADS) breast composition descriptors are used for standardized mammographic reporting and are assessed visually. This reporting is clinically relevant because breast composition can impact mammographic sensitivity and is a breast cancer risk factor. New techniques are presented and evaluated for generating automated BI-RADS breast composition descriptors using both raw and calibrated full field digital mammography (FFDM) image data.

Methods: A matched case-control dataset with FFDM images was used to develop three automated measures for the BI-RADS breast composition descriptors. Histograms of each calibrated mammogram in the percent glandular (pg) representation were processed to create the new BR_{pg} measure. Two previously validated measures of breast density derived from calibrated and raw mammograms were converted to the new BR_{vc} and BR_{vr} measures, respectively. These three measures were compared with the radiologist-reported BI-RADS compositions assessments from the patient records. The authors used two optimization strategies with differential evolution to create these measures: method-1 used breast cancer status; and method-2 matched the reported BI-RADS descriptors. Weighted kappa (κ) analysis was used to assess the agreement between the new measures and the reported measures. Each measure's association with breast cancer was evaluated with odds ratios (ORs) adjusted for body mass index, breast area, and menopausal status. ORs were estimated as per unit increase with 95% confidence intervals.

Results: The three BI-RADS measures generated by method-1 had κ between 0.25–0.34. These measures were significantly associated with breast cancer status in the adjusted models: (a) OR = 1.87 (1.34, 2.59) for BR_{pg} ; (b) OR = 1.93 (1.36, 2.74) for BR_{vc} ; and (c) OR = 1.37 (1.05, 1.80) for BR_{vr} . The measures generated by method-2 had κ between 0.42–0.45. Two of these measures were significantly associated with breast cancer status in the adjusted models: (a) OR = 1.95 (1.24, 3.09) for BR_{pg} ; (b) OR = 1.42 (0.87, 2.32) for BR_{vc} ; and (c) OR = 2.13 (1.22, 3.72) for BR_{vr} . The radiologist-reported measures from the patient records showed a similar association, OR = 1.49 (0.99, 2.24), although only borderline statistically significant.

Conclusions: A general framework was developed and validated for converting calibrated mammograms and continuous measures of breast density to fully automated approximations for the BI-RADS breast composition descriptors. The techniques are general and suitable for a broad range of clinical and research applications. © 2013 American Association of Physicists in Medicine. [<http://dx.doi.org/10.1118/1.4824319>]

Key words: mammography, breast density, BI-RADS, calibration, differential evolution optimization

1. INTRODUCTION

Mammographic density is an important breast cancer risk factor.¹ Due to measurement difficulties and the lack of measurement automation, breast density is mainly used in research studies of breast cancer etiology, and not for breast cancer risk applications in the clinical environment.² The Breast Imaging Reporting and Data System (BI-RADS) lexicon,³ developed to standardize clinical reporting in mammography, includes a breast composition classification system. This system is a four category ordinal scale comprised of

both a qualitative description and an estimate of the percentage of fibroglandular (glandular) tissue content of the breast defined as follows:³ “1. the breast is almost entirely fatty (<25% glandular); 2. there are scattered fibroglandular densities (approximately 25%–50% glandular); 3. the breast tissue is heterogeneously dense, which could obscure detection of small masses (approximately 51%–75% glandular); and 4. the breast is extremely dense. This may lower the sensitivity of mammography (>75% glandular).” These ratings are reported by the attending radiologist by visual assessment. As indicated by the upper two categories, as breast density

increases the sensitivity of mammography may decrease. In epidemiologic research, these descriptors have been used as a discrete measure of breast density for breast cancer risk assessments.¹ Although extremely useful, evidence shows visual assessments of breast density are inaccurate and reader dependent.⁴ Additionally, manual methods are often not cost-effective when applied en masse.

Various calibrated (i.e., standardized) measures of breast density are under investigation,^{5–15} which may provide several benefits. Such quantitative measures may facilitate the use of breast density in the clinical setting because they are automated, which can support large-scale epidemiologic studies more efficiently, and can be used to develop a measure of the BI-RADS breast composition descriptors.^{16,17} The need for an automated measure for use in the clinical setting is also supported by legislative mandates. Many states have either recently enacted or are considering legislation to inform women at screening if they have dense breasts because of the associated breast cancer risk and potential negative impact on mammographic sensitivity,¹⁸ although the type of breast density measurement has not been specified.

The calibration techniques used in this report were developed previously^{19–21} for FFDM. Briefly, the calibration adjusts for the x-ray acquisition technique differences across mammograms resulting in standardized images. The approach produced validated measures of breast density,^{12–14} one of which captures variation in the mammogram referred to as V_c (i.e., variation measured from calibrated mammograms). We also showed that the variation in raw mammograms (i.e., V_r) was a valid breast density measurement using images from FFDM (Ref. 14) and digitized film.²² In the current report, we present new techniques for converting both calibrated mammograms and continuous measures of breast density variation, individually, into a four-state ordinal measure as automated approximations for the BI-RADS breast composition descriptors.

2. METHODS

2.A. Design overview

An overview of the study design is provided to facilitate the description of the methods before relaying the specific details because the work incorporates multiple endpoints, evolutionary optimization strategies, and new techniques for creating ordinal measures, in addition to calibration. The image data and participant information were obtained from an established case-control study of breast cancer risk.^{12–14,23} For comparison purposes, we used the radiologist-reported BI-RADS breast composition descriptors abstracted from the patient records as *known quantities*. We used two optimization strategies to develop automated approximations for the BI-RADS descriptors: (i) optimization method-1 used the case-control status of the patients as the endpoint without considering the radiologist-reported assessments from the patient records. In this first scenario, the cancer/no-cancer status (case versus control) was used to generate a new ordinal measure that provided maximum case-control separation. This

was achieved in combination with logistic regression (LR) modeling by estimating the parameters of the LR model and determining the ordinal variable within a continual looping operation; and (ii) optimization method-2 attempted to reproduce the *known quantities* when generating a new measure. In this second scenario, the error between the reported assessments and the predicted categories was used as the endpoint and minimized. The two optimization strategies are similar in implementation but vary in the endpoint fitness function. We used differential evolution (DE) optimization²⁴ for both strategies. We considered these two optimization methods to potentially broaden the applicability of our methods, essentially forming two sets of three approximations (i.e., six new measures) for the BI-RADS composition descriptors serving different purposes.

We present findings from calibrated mammograms, as well as from raw images. When calibrating a given mammogram, each pixel is mapped into the normalized percent glandular (pg) representation, accounting for acquisition technique differences across images.^{20,21} The pg theoretical dynamic range is 0–100 measured in pg quantities (unitless). The pg format is a normalized effective x-ray attenuation coefficient representation, making pixel quantities comparable across images. The first new measure was derived from the pg (pixel) representation by integrating the histogram for each image giving an approximation for the respective cumulative distribution for each image in the dataset. In this capacity, each normalized histogram represents the probability distribution function for the respective image. Our optimization procedure determines critical/cut-points (we explain in detail below) using the cumulative distribution for each image; this four-state variable is referred to as BR_{pg} . (BR is short for BI-RADS). We also developed BI-RADS composition descriptor approximations using the V_c and V_r measures by determining cut-points from their respective sample population distributions. We refer to these new measures as BR_{Vc} and BR_{Vr} , respectively.

2.B. Patient population and data

The study population, inclusion criteria, and data collection methods for the original study were described in detail previously^{12–14,23} and are summarized here. This was a matched case-control study developed to quantify risks associated with breast density measures. Cases are first-time unilateral breast cancer patients attending the breast clinics at the Moffitt Cancer Center. The mammograms of the noncancerous breast of the cases define the study image laterality. Controls were individually matched (1:1) to cases on age, hormone replacement therapy usage/duration, screening history, and breast laterality. All mammograms were acquired with the same General Electric Senographe 2000D FFDM unit used for screening in this Center. For a given patient, the image dataset includes the standard four-view screening mammograms. We used the raw images for all analyses in this current report. This system has a 19.2 cm × 23 cm detector field of view, produces images with 100 μ m spatial resolution, and the raw images have 14 bit dynamic range per pixel. Our analysis was restricted to cranial-caudal (CC) views and to

participants that had the BI-RADS breast descriptor assessments available in their medical records, totaling 163 case-control pairs (163 dataset). We refer to this collection of radiologist-reported BI-RADS breast composition descriptors below as the case-report assessments (or case-reports). Before analyzing or calibrating the mammograms, the breast image area is eroded by 25% in a radial direction. The reduced breast region approximates the image area related to the portion of the breast that was in contact with the compression paddle during the image acquisition as outlined previously.^{13,20} This is an approximation to restrict the analysis to the portion of the image area corresponding to where the breast was *uniformly* compressed to reduce unwanted variation. The study was approved by the University of South Florida Institution Review Board.

2.C. Descriptors derived from calibrated pixel distribution: BR_{pg}

The BR_{pg} measure uses the histogram for each calibrated image (i.e., for each patient). We let an arbitrary pg pixel value = x and let the normalized histogram = p (probability distribution) for an arbitrary image. Although we have discrete pixel quantities, we define the cumulative distribution using a continuous approximation for ease of notation and methodology development giving

$$P(x) = \int_{x_{\min}}^x p(x)dx. \quad (1)$$

When evaluating $x = z$, $P(z)$ is the probability of $x \leq z$. We generate $P(x)$ for each image labeled specifically as $P_i(x)$, where the subscript, i , is the patient index. When using either optimization strategy, we determine four unknown parameters: x_c is the critical pg pixel reference value and three function (cut-points) values of $P(x)$ given by s , q , and r such that $s < r < q$, which are used to make comparisons with $P(x_c)$. For an arbitrary observation, the four-state ordinal breast density measure, BR_{pg} , is determined by these four conditions paralleling the BI-RADS classification:

- (1) $P_i(x_c) \geq q$, the sample is in group 1
- (2) $r \leq P_i(x_c) < q$, the sample is in group 2
- (3) $s \leq P_i(x_c) < r$, the sample is in group 3
- (4) $P_i(x_c) < s$, the sample is in group 4

2.D. Descriptors derived from the breast density variation measures: BR_{vc} and BR_{vr}

The process for creating the BR_{vc} and BR_{vr} measures starts with two existing variation breast density measures, calculated from calibrated and raw images referred to as V_c and V_r , respectively. Either measure is calculated as the standard deviation of the pixel values within the eroded breast region. The optimization procedure (both strategies) determines three unknown parameters such that $a < b < c$. These are cut-points from the respective V (i.e., using V generically) measurement

population distribution. The solution to this problem gives the BI-RADS descriptor approximations, BR_{vc} or BR_{vr} , dependent upon the process input. The four-state ordinal variable for the i th observation is determined from these four conditions, also paralleling the BI-RADS classification:

- (1) $V_i \leq a$, the sample belongs to group 1
- (2) $a < V_i \leq b$, the sample belongs to group 2
- (3) $b < V_i \leq c$ the sample belongs to group 3
- (4) $V_i > c$, the sample belongs to group 4

2.E. Optimization procedures

The DE optimization²⁴ was used to determine the parameter vectors defined in Secs. 2.C and 2.D. Our DE methods were described in detail previously.²⁵ For reference, we used the abbreviations for the DE parameters provided by its founders:²⁴ the vector field population is $NP = 40$ (rule of thumb population size) random vectors; the crossover $CR = 0.1$; and the evolutionary amplification factor $F = 0.5$. The maximum number of generations was fixed giving $G = 1000$. In brief, DE incrementally finds the parameters by either maximizing or minimizing the fitness function (whichever is applicable) by repeated processing of 100 image case-control dataset (100 pair subset described below) with NP parameter-vector competitions at each generation to determine the vectors that populate the next generation, where the process then repeats. This process was initialized with 40 (i.e., NP) random vectors [uniformly distributed random variables with values defined in this range (0, 1)] for a given BI-RADS approximation (i.e., p_v as well as β where applicable). The number of generations could terminate earlier than $G = 1000$ when a preset convergence condition was met. For optimization method-1, the process was terminated early if this condition $|Az_{\max} - Az_{\min}| \leq 0.01$ was reached within a given generation. For optimization method-2, the process was terminated early if this condition $|\Delta_{\max} - \Delta_{\min}| \leq 0.001$ was reached within a given generation.

To introduce variation and mitigate overtraining effects, we developed each new measure by repeatedly choosing random samples of 100 case-control pairs (bootstrap subdatasets) from the 163 dataset with replacement as the inputs to the optimization process. Cases were selected randomly (i.e., non-cancerous breast) and an arbitrary breast side of their matched control was selected randomly. The final comparisons and analyses were based on the 163 dataset discussed in Sec. 2.F.

The BR_{pg} formation process is used as an example to illustrate the implementation of the optimization procedures and strategies. For optimization method-1, we apply DE for two purposes within a sequence. First, DE determines the parameter vector $p_v = [x_c, r, q, s]$, creating the new ordinal breast density measure. This measure is then passed to conditional LR modeling. In the second application, DE then determines the LR coefficient vector $\beta = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4]$, where β_0 is the offset that factors out of the analysis in the matched case-control application, β_1 is the ordinal breast density

measure coefficient, β_2 is BMI coefficient, β_3 is the breast area coefficient, and β_4 is the binary menopausal status coefficient. These two applications of DE are linked in a continual looping mechanism (described below). For the conditional LR modeling we coded the appropriate maximum likelihood estimators²⁶ and validated the application. The area under the receiver operating characteristic curve (Az), estimated from the LR model output, was used as the optimization's fitness function to drive the $\mathbf{p}_v$ determination process (i.e., the Az is passed back to the first DE application). That is, the sequence is driven by maximizing Az.

For optimization method-2, the problem is set up similarly (i.e., the four-state variable conditions cited above are the same) with a modified fitness function without the simultaneous LR modeling. We let the predicted, or estimated, BI-RADS description for the i th patient = BR_{pgi} (i.e., using $\mathbf{p}_v$ components from the optimization procedure) and let the corresponding case-report assessment for the i th patient = BR_i . The fitness function for optimization method-2 is defined as

$$\Delta = \sum_{i=1}^{2n} |BR_i - BR_{pgi}|, \quad (2)$$

where $n = 100$. The reason for using a 100 case-control sample data subset is to prevent overfitting, as discussed above. In this situation, the optimization procedure is driven by minimizing Δ to match the case-report assessments. In contrast with optimization method-1, the resulting ordinal measure is evaluated with LR after the optimization process is terminated. The corresponding optimization procedures for BR_{vc} and BR_{vr} are analogous to those used for BR_{pg} with the proper substitutions: we let $\mathbf{p}_v = [a, b, c]$ and substitute BR_{vci} or BR_{vri} in Eq. (2) for BR_{pgi} as appropriate using the same indexing and optimization strategies.

2.F. Statistical analyses

Conditional logistic regression was used to evaluate a given measure's association with breast cancer. In the final analysis we used the noncancerous breast side for the 163 cases and the matched side of the controls. Each measurement was treated as a four-state ordinal variable. The breast density measurement odds ratios (ORs) are estimated per category increase in the ordinal scale in unadjusted models and models with simultaneous adjustments for body mass index (BMI) measured in kg/m^2 , breast area (BA) measured in cm^2 , and menopausal status (MS). ORs are presented with 95% confidence intervals (CIs). The area under the receiver operating characteristic curve (Az) was used to evaluate a given model's ability to separate cases from controls (i.e., predictive capability).

We compared the distribution for each of the new measures with the case-report assessments using a joint frequency analysis. To summarize the agreement (similarity/dissimilarity) and make comparisons, we used the weighted kappa statistic (κ), presented with 95% CIs, due to ordinal nature of the measures. The value of κ varies over this range $[-1, 1]$. The upper κ bound indicates perfect agreement between two dis-

tributions, zero indicates the distributions are disjoint, and the lower κ bound indicates perfect negative agreement. Because the assessments from the patient reports were made by multiple radiologists, the κ agreement is best interpreted as comparing a given new measure with an *averaged* or composite effect.

In the final analyses, logistic regression modeling and κ analyses were performed with SAS 9.3 (SAS Institute Inc., Cary, NC).

3. RESULTS

3.A. Approximations for the BI-RADS composition descriptors

For optimization method-1, we found $x_c = 23.0$ (pg units) and $[q, r, s] \approx [0.987, 0.700, 0.228]$ for the BR_{pg} development. For illustration purposes, we provide an explicit example to describe the BR_{pg} process because it is a new and more involved methodology. Figure 1 (top) shows clinical display images (surrogates for the raw images for viewing purposes only) from four representative patient examples corresponding to the four BI-RADS categories (i.e., 1 through 4 from left to right). The bottom row shows the respective calibrated images after the erosion process, and Fig. 2 shows the corresponding distributions [i.e., $p_i(x)$]. Figure 3 shows the corresponding cumulative distributions [i.e., $P_i(x)$] determined by integrating the probability distributions shown in Fig. 2 with

FIG. 1. Image examples: The top row shows four clinical display mammograms. We refer to these as examples 1–4. We use these as surrogates for the raw images for viewing purposes only because they are more easily displayed than the raw images. From left to right, the case-report BI-RADS categories are 1, 2, 3, and 4. The bottom row shows the corresponding images in the calibrated percent glandular format with 25% erosion.

FIG. 2. Calibrated histogram examples: This shows histograms from the four calibrated mammogram examples shown in Fig. 1: (1) example 1 with a solid line; (2) example 2 with short dashes; (3) example 3 with dashes and dots; and (4) example 4 with long dashes. The x-axis represents calibrated pixel values (x = percent glandular quantities) and the y axis is the relative normalized frequency. These normalized histograms approximate the probability distributions for each image.

x_c denoted. The BR_{pg} measure classified these images in the same categories as the case-report assessments. The key to this measure's operation is noting where $P_i(x_c = 23)$ is situated with respect to $[q, r, s]$ for a given patient. Example 4 patient, with $P_i(x)$ denoted with the long-dashes in Fig. 3, has $P_i(x_c) \approx 0.0$ indicating that 100% of the pixels have values greater than x_c and the image was placed in group 4. In contrast, example 1 patient, with $P_i(x)$ denoted by a solid line in Fig. 3, has $P_i(x_c) \approx 0.99$ indicating that 99% of the pixels have values less than or equal to $x_c = 23$, and the image was placed in group 1. Also noted in Fig. 2, $x < 0$ in $p(x)$ should not exist

FIG. 3. BR_{pg} measure examples from optimization method-1. The x-axis represents calibrated pixel values (x = percent glandular quantities). This shows the cumulative distributions determined from the histograms shown in Fig. 2 for the four patient examples. The BR_{pg} processing with optimization method-1 categorized these examples as follows using $x_c = 23$ (vertical dashed line) and $[q, r, s] \approx [0.987, 0.700, 0.228]$: (1) example 1 was placed in category 1 denoted with a solid line; (2) example 2 was placed in category 2 denoted with short dashes; (3) example 3 was placed in category 3 denoted with dashes and dots; and (4) example 4 was placed in category 4 denoted with long dashes.

FIG. 4. The V_r population distribution for the entire case-control dataset and BR_{vr} measurement cutoffs derived with optimization method-1: The vertical dashed lines from left to right show the cut-point parameter values for the BR_{vr} measure with $[a, b, c] \approx [71.9, 151.1, 207.5]$.

in theory. In practice, the presence of negative x values may be due to both a mismatch between the x-ray attenuation of the adipose calibration phantom material and that of adipose breast tissue and possibly inaccurate compressed breast thickness estimations as discussed previously.^{13,20} For BR_{vc} , $[a, b, c] \approx [4.8, 8.5, 14.5]$, and for BR_{vr} , $[a, b, c] \approx [71.9, 151.1, 207.5]$. The BR_{vc} measure placed these examples in the 1, 2, 3, and 3 categories, respectively, whereas the BR_{vr} measure placed these examples in the 2, 2, 3, and 2 categories. Figure 4 shows the population distribution for V_r and the $[a, b, c]$ quantities marked with vertical dashes. The method for converting V_c to the ordinal variable is analogous to that of converting V_r and is, therefore, not shown (no examples provided).

For estimates derived with optimization method-2, the same interpretation used for method-1 applies. For BR_{pg} , we found $x_c = 19.0$ (pg units) and $[q, r, s] \approx [0.99, 0.98, 0.03]$, and the examples were placed in the 1, 3, 3, and 4 categories. For BR_{vc} , $[a, b, c] \approx [2.3, 5.7, 16.5]$, and the examples were placed in 2, 3, 3, and 3 categories. For BR_{vr} , $[a, b, c] \approx [32.0, 97.3, 326.1]$, and the examples were placed in the 2, 3, 3, and 3 categories (same as BR_{vc}). The similarity between the variation measures is expected because of their correlation.¹⁴ The differences between the measures are further assessed by the magnitude of their respective association with breast cancer.

Table I provides the associations with breast cancer for the case-report assessments (top) and for each new measure derived with optimization method-1 (left-side) and method-2 (right-side). In the adjusted models, the findings from BR_{pg} ($OR = 1.87$; $Az = 0.648$) and BR_{vc} ($OR = 1.93$; $Az = 0.663$), derived from calibrated mammograms using method-1, provided significant OR associations and greater Az in comparison with the case-report findings ($OR = 1.49$; $Az = 0.632$), where the OR was not significant. The BR_{vr} findings ($OR = 1.37$; $Az = 0.639$) were similar in scale to that of the case-report measures, although the OR for BR_{vr} was significant in the adjusted model. For optimization method-2,

TABLE I. The associations of BI-RADS composition descriptors with breast cancer and their case-control predictive capability. Findings are presented for both optimization methods. The case-report findings are duplicated on the top row for easy reference and comparison.

Optimization method-1			Optimization method-2		
BI-RADS variable	OR (95% CI)	^a Adjusted OR (95% CI)	BI-RADS variable	OR (95% CI)	^a Adjusted OR (95% CI)
Case-report	1.21 (0.85, 1.72)	1.49 (0.99, 2.24)	Case-report	1.21 (0.85, 1.72)	1.49 (0.99, 2.24)
Az	0.519	0.632	Az	0.519	0.632
BR _{pg}	1.27 (0.99, 1.61)	1.87 (1.34, 2.59)	BR _{pg}	1.26 (0.87, 1.82)	1.95 (1.24, 3.09)
Az	0.557	0.648	Az	0.527	0.634
BR _{vc}	1.35 (1.03, 1.76)	1.93 (1.36, 2.74)	BR _{vc}	1.10 (0.72, 1.70)	1.42 (0.87, 2.32)
Az	0.559	0.663	Az	0.510	0.626
BR _{vr}	1.19 (0.94, 1.50)	1.37 (1.05, 1.80)	BR _{vr}	1.50 (0.93, 2.42)	2.13 (1.22, 3.72)
Az	0.542	0.639	Az	0.538	0.639

Note: OR: Odds Ratio (for one category increase in ordinal scale). CI: Confidence Intervals. Az: Area under the receiver operating characteristic curve.

^aThe final models are simultaneously adjusted for BMI, breast area, and menopausal status.

the BR_{pg} (OR = 1.95; Az = 0.634) and BR_{vr} (OR = 2.13; Az = 0.639) measures provided significant ORs, whereas the BR_{vc} (OR = 1.42; Az = 0.626) association was not significant.

In summary, for optimization method-1, BR_{vc} was most predictive of breast cancer status (across all measures) and the BR_{pg} and BR_{vc} (calibrated measures) provided stronger OR associations than given by the BR_{vr} (raw image) measure. For optimization method-2, both BR_{pg} and BR_{vr} measures provided comparable predictive capability and stronger OR associations than BR_{vc} and the case-report measures. The new measures all compared well against the predictive capability and OR associations provided by the case-report measures.

3.B. Agreement with the case-report assessments

The comparisons and concordance of the new measures derived from optimization method-1 with the case-report assessments are provided in Table II. All three measures are related to the case-report assessments with $\kappa = 0.25$ for BR_{pg}, $\kappa = 0.34$ for BR_{vc}, and $\kappa = 0.27$ for BR_{vr}. The BR_{vc} measure provided the closest agreement with the case-report classification. There are relatively few observations in the first category ($n = 6$) and many in the third category ($n = 182$) according to the case-report classification. In contrast, the new measures spread the patients across the categories more generally due to the optimization fitness function. The corresponding comparisons and concordance for the new measures derived from optimization method-2 are shown in Table III. All new measures provided similar agreement with the case-report assessments with $\kappa = 0.42$ for BR_{pg}, $\kappa = 0.45$ for BR_{vc}, and $\kappa = 0.42$ for BR_{vr}. When considering the diagonal elements of Table III (method-2), the agreement between BR_{pg}, BR_{vc}, and BR_{vr} with the reported assessments was 63%, 68%, and 68%, respectively. In contrast with optimization method-1, the measures derived from optimization method-2 tended to cluster the patients into the second and third categories attempting to match the case-report classifications.

4. DISCUSSION

Two types of automated ordinal breast density measures were generated for approximating the radiologist-reported

TABLE II. The joint frequency distribution of the four-state ordinal variables derived from optimization method-1 and the case-report assessments. The weighted κ statistic is provided below each respective comparison with 95% confidence intervals. For reference and comparison purposes, the percentages of observations in each of the BI-RADS categories (category 1 thorough category 4) for each of the measures are provided: (a) case-report BI-RADS gave 1.8%, 30.4%, 55.8%, and 11.9%; (b) BR_{pg} gave 31.9%, 32.5%, 21.5%, and 14.1%; (c) BR_{vc} gave 16.3%, 35.6%, 37.7%, and 10.4%; and (d) BR_{vr} gave 15.9%, 42.6%, 25.1%, and 16.2%.

Case-report BI-RADS					
BR _{pg}	1	2	3	4	<i>n</i>
1	5	60	38	1	104
2	1	27	73	5	106
3	0	9	50	11	70
4	0	3	21	22	46
<i>n</i>	6	99	182	39	326
$\kappa: 0.25 (0.19, 0.31)$					
Case-report BI-RADS					
BR _{vc}	1	2	3	4	<i>n</i>
1	6	38	9	0	53
2	0	42	70	4	116
3	0	18	86	19	123
4	0	1	17	16	34
<i>n</i>	6	99	182	39	326
$\kappa: 0.34 (0.27, 0.41)$					
Case-report BI-RADS					
BR _{vr}	1	2	3	4	<i>n</i>
1	5	35	11	1	52
2	1	47	78	13	139
3	0	12	63	7	82
4	0	5	30	18	53
<i>n</i>	6	99	182	39	326
$\kappa: 0.27 (0.21, 0.34)$					

TABLE III. The joint frequency distribution of the four-state ordinal variables derived from optimization method-2 and the case-report assessments. The weighted κ statistic is provided below each respective comparison with 95% confidence intervals. For reference and comparison purposes, the percentages of observations in each of the BI-RADS categories (category 1 thorough category 4) for each of the measures are provided: (a) Case-report BI-RADS gave 1.8%, 30.4%, 55.8%, and 11.9%; (b) BR_{pg} gave 4.3%, 18.7%, 69.4%, and 8.6%; (c) BR_{vc} gave <1%, 26.1%, 69.0%, and 4.6%; and (d) BR_{vr} gave 0%, 28.5%, 69.9%, and 1.5%.

Case-report BI-RADS					
BR_{pg}	1	2	3	4	n
1	2	9	3	0	14
2	3	38	20	0	61
3	1	51	150	21	223
4	0	1	9	18	28
n	6	99	182	39	326
κ : 0.42 (0.34, 0.50)					
Case-report BI-RADS					
BR_{vc}	1	2	3	4	n
1	1	0	0	0	1
2	5	56	23	1	85
3	0	42	155	28	225
4	0	1	4	10	15
n	6	99	182	39	326
κ : 0.45 (0.37, 0.54)					
Case-report BI-RADS					
BR_{vr}	1	2	3	4	n
1	0	0	0	0	0
2	6	61	24	2	93
3	0	37	157	34	228
4	0	1	1	3	5
n	6	99	182	39	326
κ : 0.42 (0.34, 0.50)					

BI-RADS descriptors. The first type was derived from the cumulative distribution of each calibrated image without requiring an established, or pre-existing, breast density measurement. The cumulative distribution approach applies to calibrated data only and illustrates a benefit from establishing a calibration system; initially, we attempted to use a similar formulism with histograms from the raw images but the approach produced little and was not pursued further (data not shown). The second type was derived from continuous measures of breast density, with or without calibration, as demonstrated with V_r and V_c . We used DE optimization by considering two dissimilar endpoints (i.e., different fitness functions) for developing both measurement types. Optimization method-1 used case-control status as the endpoint and does not require *a priori* BI-RADS assessments, making it a desirable approach in situations where either more or less categories are suitable, or where matching previously determined descriptors is not appropriate. We note, increasing categories in the observer based BI-RADS assessments correspond with increased breast cancer risk, although with some degree of ambiguity.^{1,27,28} Optimization method-1 creates a new measure that shares characteristics of the BI-RADS composition's

percentage of density component with increased risk with increasing breast density category, as defined by the logistic regression modeling interpretation. Although optimization method-2 required existing BI-RADS assessments for endpoint matching purposes for this specific application, the approach could be modified to incorporate some other established endpoint as well. As demonstrated, the measures generated by optimization method-2 provided closer agreement with the case-report assessments, whereas the measures generated from optimization method-1 produced stronger predictive capability. When considering the performance metrics (i.e., κ , ORs, and Az), we conclude the new measures are at least equivalent with the case-report assessments. The strength of our methodology is that no assumptions are required within its framework other than the four-state variable imposition, which can also be modified easily. Moreover, the fitness function and endpoints can be modified easily as demonstrated.

There are several characteristics of our work worth noting. Because the patients were selected over a five-year time span (2007–2011), the case-report assessments were provided by multiple radiologists that may have various levels of experience, which could add interoperator variability into the development and impact the derivation of the new measures. Moreover, the clinical display images were used by the radiologists for the clinical assessments, whereas our methods operated on either the calibrated or raw images. The differences between these image formats could contribute additional variability in the comparisons. However, limited by our sample size, we were unable to evaluate the impact of reader variability on the new measurement developments. Related work²⁹ in FFDM shows that the percentages for the BI-RADS descriptors are 9.5%, 45.6%, 35.3%, and 9.6% for the first through fourth categories, respectively, as estimated from a relatively large study population. We note the percentages from the case-report assessments for our dataset (i.e., 1.8%, 30.4%, 55.8%, and 11.9%) differ considerably in the first and third categories in comparison with this larger study. In comparison, the percentages were 16.3%, 35.6%, 37.7%, and 10.4% for the 1–4 categories based on BR_{vc} estimated with optimization method-1 in our study, demonstrating the closest agreement with this related report.²⁹ The current work was restricted to the CC views to eliminate interference from the chest wall and pectoral muscle. Algorithm modifications are required to include either the mediolateral (ML) or mediolateral oblique (MLO) views in future studies. We used randomness in the training to mitigate overfitting. Because only a few parameters were estimated in the measure generation, overfitting is likely not a concern in our study. However, validation with independent datasets is required.

Calibration is a more recent approach for estimating breast density with relatively few published studies evaluating its merits. Potential benefits derived from calibration in our study are exemplified through (a) the demonstrated agreement between the new BR_{pg} and BR_{vc} measures and the conventional descriptors in the case-reports, (b) improved efficiency from automation, and (c) the consistent and significant associations of the new measures with breast cancer. Previous

calibration studies have shown mixed results in developing measurements that produce breast cancer associations stronger than provided by the operator-assisted percentage of breast density (PD) measure,^{9,10,12–15} which is often used as a standard. The close similarities between our approach and other calibrated breast density techniques were discussed in detail in our previous work.²⁰ Our continuous calibrated measures presented previously produced breast cancer associations at least equivalent with those provided by PD (Refs. 12–14) and are generally similar to those presented in this report for the new measures, although different analysis methods were used previously, precluding one-to-one comparisons.

Creating discrete measures from continuous variables may be one approach to develop an automated measure for clinical purposes. The new techniques and the ordinal measures presented in this report are parallel developments with our previous calibration work and may be useful for addressing the current mandatory breast density reporting requirements. For example, this clinical application may require some type of more easily interpretable binary yes–no measure to inform patients whether they have dense breasts¹⁸ because additional screening may be appropriate³⁰ or whether they have a high risk for breast cancer in contrast with providing a specific probability figure of merit derived from a continuous measure. A related calibration study converted a continuous breast density measurement to the BI-RADS composition descriptors and then to a binary (high vs low) density scale with visual assessments to determine the breast density cut-point,¹⁶ although breast cancer status was not considered in the analysis. In contrast, our methodology does not require operator intervention to determine cut-points or other parameter estimates. More general comparisons between our methodology and those used in this related study¹⁶ cannot be made because both the measurement formalism and experimental procedures are different. Other researchers¹⁷ used a calibrated breast density measurement to develop BI-RADS composition descriptors with cut-points derived from the continuous measurement distribution, which is similar to our approach for the BR_{vc} measure (method-2). Our three measurements from method-2 provided similar agreement with the case-report assessments as those reported in the related calibration study.¹⁷ Although our calibration methods have similarities with this related work, there are differences. Our calibrated breast density measurement (i.e., variation metric) and the endpoint for method-1 are different. We note, the BR_{pg} was derived from the integrated distributions for each patient without the requirement of a pre-existing breast density measurement. This is a new approach in comparison with discrete measure derivations that require a pre-existing breast density measure as the starting point. In contrast with the related calibrated research, the BR_{vr} measure was derived without calibration.

Although the efficacy of calibration is still under investigation, the commonly used standard of comparison may not be the best or only metric. In general, after establishing a calibration system, breast density measurements can be acquired automatically, whereas the percentage of breast density measure

is most often obtained with operator-assistance. Establishing a calibration framework may be the cost (i.e., the required effort) for breast density measurement automation. We posit that with the added benefit of automation, calibrated measurements may not have to produce stronger associations with breast cancer than the operator-assisted approach but provide near equivalency.

The BI-RADS breast composition descriptors defined in the lexicon include percentage categories as well as a qualitative description related to overall image-texture. There is arbitrariness to this definition because one type of measurement may not capture both components simultaneously in all situations. Additionally, the density percentage component of the BI-RADS descriptors may be an artificial construct. As indicated by related calibration research, the average breast is comprised of approximately 19% fibroglandular tissue,³¹ which is remarkably similar to the critical values determined in this report for the BR_{pg} measure (i.e., $x_c = 23$ and $x_c = 19$). Despite a different formalism used in this related work,³¹ we have shown previously²⁰ that their measure is similar to our average percent glandular measure. We evaluated two types of measurements that capture breast density or texture content, which are different image characteristics. The BR_{pg} measure is related to breast density content but does not capture texture or variation. In contrast, the variation measures (BR_{vc} and BR_{vr}) capture a broad range of texture information as defined by their descriptions from Fourier analysis, but not the degree, or amount, of dense breast tissue explicitly. At present, it is not clear whether a breast density measurement designed to indicate that a lesion could have been missed is the same as a measure designed to optimize its association with breast cancer or vice versa. Future work includes exploring techniques for combining these approximation measures to capture the degree of breast density and the texture components simultaneously. The most appropriate method will require further investigation.

5. CONCLUSION

A general methodology was presented for converting calibrated mammograms and continuous measurements of breast density into approximations for the standard BI-RADS breast composition descriptors. Both breast density measures and the development of automated BI-RADS composition descriptors are active fields of research. Our work establishes a general platform for developing ordinal measures that can be adapted easily for both research and clinical applications using calibrated or noncalibrated mammograms. The work demonstrates the benefits of applying a calibration methodology, while also showing calibration is not essential to the BI-RADS approximation derivation. Our work also relied heavily on DE optimization. DE was employed in multiple capacities with different endpoints indicating the applicability of our methods to other imaging platforms and endpoints not considered in this report as well, without undue modification. The work presented in this report may be useful for addressing the legislative breast density mandates as well. However, further analyses are required to make our measures

suitable for the clinical setting, such as evaluating the impact of inter-reader variability on our measure developments. Methods to include ML or MLO views in our analyses will require further algorithm developments. Because our methods and findings pertain to a specific indirect x-ray conversion FFDM technology³² and patient dataset, evaluation of our calibration approach and measurement developments across imaging platforms and other populations is required.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health Grant No. R01 CA114491. The Moffitt Cancer Center has a pending patent application for this work. The principals on this application are J.J.H., T.A.S., and E.E.F.

^{a)} Author to whom correspondence should be addressed. Electronic mail: john.heine@moffitt.org

¹ V. A. McCormack and I. dos Santos Silva, "Breast density and parenchymal patterns as markers of breast cancer risk: A meta-analysis," *Cancer Epidemiol. Biomarkers Prev.* **15**, 1159–1169 (2006).

² V. Brower, "Breast density gains acceptance as breast cancer risk factor," *J. Natl. Cancer Inst.* **102**, 374–375 (2010).

³ C. J. D'Orsi, L. W. Bassett, W. A. Berg, *et al.* BI-RADS: Mammography, 4th edition in: C. J. D'Orsi, E. B. Mendelson, D. M. Ikeda *et al.*: Breast Imaging Reporting and Data System: ACR BI-RADS – Breast Imaging Atlas (American College of Radiology, Reston, VA, 2003).

⁴ M. B. Lobbes, J. P. Cleutjens, V. Lima Passos, C. Frotscher, M. J. Lahaye, K. B. Keymeulen, R. G. Beets-Tan, J. Wildberger, and C. Boetes, "Density is in the eye of the beholder: Visual versus semi-automated assessment of breast density on standard mammograms," *Insights Imaging* **3**, 91–99 (2012).

⁵ R. Highnam and M. Brady, *Mammographic Image Analysis* (Kluwer Academic Publishers, Boston, MA, 1999).

⁶ R. Highnam, X. Pan, R. Warren, M. Jeffreys, G. Davey Smith, and M. Brady, "Breast composition measurements using retrospective standard mammogram form (SMF)," *Phys. Med. Biol.* **51**, 2695–2713 (2006).

⁷ M. Jeffreys, R. Warren, R. Highnam, and G. D. Smith, "Initial experiences of using an automated volumetric measure of breast density: The standard mammogram format," *Br. J. Radiol.* **79**, 378–382 (2006).

⁸ M. Jeffreys, R. Warren, R. Highnam, and G. Davey Smith, "Breast cancer risk factors and a novel measure of volumetric breast density: Cross-sectional study," *Br. J. Cancer* **98**, 210–216 (2008).

⁹ N. Boyd, L. Martin, A. Gunasekara, O. Melnichouk, G. Maudsley, C. Peressotti, M. Yaffe, and S. Minkin, "Mammographic density and breast cancer risk: Evaluation of a novel method of measuring breast tissue volumes," *Cancer Epidemiol. Biomarkers Prev.* **18**, 1754–1762 (2009).

¹⁰ J. Ding, R. Warren, I. Warsi, N. Day, D. Thompson, M. Brady, C. Tromans, R. Highnam, and D. Easton, "Evaluating the effectiveness of using standard mammogram form to predict breast cancer risk: Case-control study," *Cancer Epidemiol. Biomarkers Prev.* **17**, 1074–1081 (2008).

¹¹ M. Jeffreys, J. Harvey, and R. Highnam, "Comparing a New Volumetric Breast Density Method (Volpara™) to Cumulus," in *Digital Mammography: 10th International Workshop on Digital Mammography (IWDM2010)*, edited by J. Martí, A. Oliver, J. Freixenet, and R. Martí (Springer, Girona, Spain, 2010), pp. 408–413.

¹² J. J. Heine, K. Cao, and D. E. Rollison, "Calibrated measures for breast density estimation," *Acad. Radiol.* **18**, 547–555 (2011).

¹³ J. J. Heine, K. Cao, D. E. Rollison, G. Tiffenberg, and J. A. Thomas, "A quantitative description of the percentage of breast density measurement using full-field digital mammography," *Acad. Radiol.* **18**, 556–564 (2011).

¹⁴ J. J. Heine, E. E. Fowler, and C. I. Flowers, "A comparison of calibrated and non-calibrated breast density measurements with full field digital mammography," *Acad. Radiol.* **18**, 1430–1436 (2011).

¹⁵ J. A. Shepherd, K. Kerlikowske, L. Ma, F. Duewer, B. Fan, J. Wang, S. Malkov, E. Vittinghoff, and S. R. Cummings, "Volume of mammographic density and risk of breast cancer," *Cancer Epidemiol. Biomarkers Prev.* **20**, 1473–1482 (2011).

¹⁶ S. Ciatto, D. Bernardi, M. Calabrese, M. Durando, M. A. Gentilini, G. Mariscotti, F. Monetti, E. Moriconi, B. Pesce, A. Roselli, C. Stevanin, M. Tapparelli, and N. Houssami, "A first evaluation of breast radiological density assessment by QUANTRA software as compared to visual classification," *Breast* **21**, 503–506 (2012).

¹⁷ R. Highnam, N. Sauber, S. Destounis, J. Harvey, and D. McDonald, "Breast Density into Clinical Practice," in *Breast Imaging: 11th International Workshop on Digital Mammography*, edited by A. D. A. Maidment, P. R. Bakic, and S. Gavenonis (Springer, Philadelphia, PA, 2012), pp. 466–473.

¹⁸ J. Pushkin, "Breast Density Inform: 2013 and Beyond," in *Imaging Technology News, Vol. July/August 2013* (Scranton Gillette Communications, Arlington Heights, IL, 2013).

¹⁹ J. J. Heine, K. Cao, and C. Beam, "Cumulative sum quality control for calibrated breast density measurements," *Med. Phys.* **36**, 5380–5390 (2009).

²⁰ J. J. Heine, K. Cao, and J. A. Thomas, "Effective radiation attenuation calibration for breast density: Compression thickness influences and correction," *Biomed. Eng. Online* **9**, 73–98 (2010).

²¹ J. J. Heine and J. A. Thomas, "Effective x-ray attenuation coefficient measurements from two full field digital mammography systems for data calibration applications," *Biomed. Eng. Online* **7**, 13–24 (2008).

²² J. J. Heine, C. G. Scott, T. A. Sellers, K. R. Brandt, D. J. Serie, F. F. Wu, M. J. Morton, B. A. Schueler, F. J. Couch, J. E. Olson, V. S. Pankratz, and C. M. Vachon, "A novel automated mammographic density measure and breast cancer risk," *J. Natl. Cancer Inst.* **104**, 1028–1037 (2012).

²³ C. M. Vachon, E. E. Fowler, G. Tiffenberg, C. G. Scott, V. S. Pankratz, T. A. Sellers, and J. J. Heine, "Comparison of percent density from raw and processed full-field digital mammography data," *Breast Cancer Res.* **15**, R1 (2013).

²⁴ K. V. Price, R. M. Storn, and J. A. Lampinen, *Differential Evolution: A Practical Approach to Global Optimization* (Springer, Berlin, 2005).

²⁵ M. Behera, E. E. Fowler, T. K. Owonikoko, W. H. Land, W. Mayfield, Z. Chen, F. R. Khuri, S. S. Ramalingam, and J. J. Heine, "Statistical learning methods as a preprocessing step for survival analysis: Evaluation of concept using lung cancer data," *Biomed. Eng. Online* **10**, 97–111 (2011).

²⁶ D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed. (John Wiley & Sons Inc., Hoboken, 2000).

²⁷ J. J. Heine, M. J. Carston, C. G. Scott, K. R. Brandt, F. F. Wu, V. S. Pankratz, T. A. Sellers, and C. M. Vachon, "An automated approach for estimation of breast density," *Cancer Epidemiol. Biomarkers Prev.* **17**, 3090–3097 (2008).

²⁸ K. Kerlikowske, A. J. Cook, D. S. Buist, S. R. Cummings, C. Vachon, P. Vacek, and D. L. Miglioretti, "Breast cancer risk by breast density, menopause, and postmenopausal hormone therapy use," *J. Clin. Oncol.* **28**, 3830–3837 (2010).

²⁹ J. M. Lewin, R. E. Hendrick, C. J. D'Orsi, P. K. Isaacs, L. J. Moss, A. Karellas, G. A. Sisney, C. C. Kuni, and G. R. Cutter, "Comparison of full-field digital mammography with screen-film mammography for cancer detection: Results of 4,945 paired examinations," *Radiology* **218**, 873–880 (2001).

³⁰ R. J. Hooley, K. L. Greenberg, R. M. Stackhouse, J. L. Geisel, R. S. Butler, and L. E. Philpotts, "Screening US in patients with mammographically dense breasts: Initial experience with Connecticut Public Act 09-41," *Radiology* **265**, 59–69 (2012).

³¹ M. J. Yaffe, J. M. Boone, N. Packard, O. Alonzo-Proulx, S. Y. Huang, C. L. Peressotti, A. Al-Mayah, and K. Brock, "The myth of the 50-50 breast," *Med. Phys.* **36**, 5437–5443 (2009).

³² U. Bick and F. Diekmann, "Medical radiology diagnostic imaging and radiation oncology," in *Continuation of Handbuch der medizinischen Radiologie Encyclopedia of Medical Radiology*, edited by A. L. Baert, L. W. Brady, H.-P. Heilmann, H. Hricak, M. Knauth, M. Molls, C. Nieder, and M. F. Reiser (Springer-Verlag, Berlin, 2010).