Texture Features from Mammographic Images and Risk of Breast Cancer

Armando Manduca,[1] Michael J. Carston,[1] John J. Heine,[2] Christopher G. Scott,[1] V. Shane Pankratz,[1] Kathy R. Brandt,[1] Thomas A. Sellers,[2] Celine M. Vachon,[1] and James R. Cerhan[1]

[1]College of Medicine, Mayo Clinic, Rochester, Minnesota and [2]H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida

Abstract

Mammographic percent density (PD) is a strong risk factor for breast cancer, but there has been relatively little systematic evaluation of other features in mammographic images that might additionally predict breast cancer risk. We evaluated the association of a large number of image texture features with risk of breast cancer using a clinic-based case-control study of digitized film mammograms, all with screening mammograms before breast cancer diagnosis. The sample was split into training (123 cases and 258 controls) and validation (123 cases and 264 controls) data sets. Age-adjusted and body mass index (BMI)–adjusted odds ratios (OR) per SD change in the feature, 95% confidence intervals, and the area under the receiver operator characteristic curve (AUC) were obtained using logistic regression. A bootstrap approach was used to identify the strongest features in the training data set, and results for features that validated in the second half of the sample were reported using the full data set. The mean age at mammography was 64.0 ± 10.2 years, and the mean time from mammography to breast cancer was 3.7 ± 1.0 (range, 2.0-5.9 years). PD was associated with breast cancer risk (OR, 1.49; 95% confidence interval, 1.25-1.78). The strongest features that validated from each of several classes (Markovian, run length, Laws, wavelet, and Fourier) showed similar ORs as PD and predicted breast cancer at a similar magnitude (AUC = 0.58-0.60) as PD (AUC = 0.58). All of these features were automatically calculated (unlike PD) and measure texture at a coarse scale. These features were moderately correlated with PD (*r* = 0.39-0.76), and after adjustment for PD, each of the features attenuated only slightly and retained statistical significance. However, simultaneous inclusion of these features in a model with PD did not significantly improve the ability to predict breast cancer. (Cancer Epidemiol Biomarkers Prev 2009;18(3):837–45)

Introduction

Whether defined by the Wolfe parenchymal pattern or measures of percent breast density, the radiographic appearance of the breast is the strongest risk factor for breast cancer after age, atypia on breast biopsy, and inherited mutations in BRCA1 and BRCA2 (1-4). Wolfe (5) defined four parenchymal patterns: N1, characterized by a fatty breast with only very small amounts of dysplasia (areas of increased density) and no visible duct formation; P1, which consisted of the N1 fatty breast with ducts occupying up to 25% of the breast volume; P2, characterized by a more severe linear or nodular ductal pattern, which occupied >25% of the breast volume; and DY, which had no prominent ducts, but significant densities or dysplasia, usually comprising 50% to 75% of the breast volume. Studies have consistently shown that women with the P2 and DY categories have significantly greater risk of breast cancer than women with N1 or P1 categories (6).

Although the Wolfe pattern may be considered as a texture-based system, it is subjective, which limits its usefulness. To more reliably and accurately quantify the mammographic pattern of the breast, mammographic breast densities have been expressed as a quantitative percentage of total breast area or volume, and computerized techniques were developed to quantitate breast percent density (PD; refs. 7, 8). PD has been well characterized (7), is reproducible (6, 7, 9), and consistently shows a strong association with breast cancer risk (6). The Wolfe pattern predicted breast cancer risk as well as PD in one (10) but not a second (11) study.

Although PD is an established risk factor for breast cancer, it is not sufficiently strong enough to use clinically for risk prediction (12, 13). One strategy to improve risk prediction is to incorporate additional aspects of the mammogram image. The observation that the Wolfe pattern does not correlate perfectly with PD suggests that there is additional information contained within the image observable by radiologists that is not being captured by PD alone. Image texture has long been used in image analysis for segmentation and classification, and multiple approaches to estimate and quantify texture exist. First-order features are those that can be calculated from the histogram of the values in the region of interest. These include mean value, variance, skewness, kurtosis, gray level percentile, and entropy (14). Such features, because they are based on the histogram,

Received 7/11/08; revised 12/16/08; accepted 12/29/08; published OnlineFirst 3/3/09.

Grant support: National Cancer Institute grants P01 CA82267 and R01 CA97396 and Department of Defense grant DAMD 17-00-1-0331.

Requests for reprints: James R. Cerhan, Department of Health Sciences Research, Mayo Clinic College of Medicine, 200 1st Street Southwest, Rochester, MN 55905. Phone: 507-284-4156; Fax: 507-266-2478. E-mail: cerhan.james@mayo.edu

doi:10.1158/1055-9965.EPI-08-0631

ignore information about the spatial relationships of the pixels and consider only their intensities. Second-order features do consider spatial relationships. Perhaps the most commonly used are Markovian features, based on the use of cooccurrence matrices, which measure the probability that a pixel of a certain gray level will be positioned at a particular distance and orientation from a pixel of a certain other gray level (15). Other second-order features include run-length measures, which evaluate texture by measuring ''runs'' of consecutive pixels with similar gray levels and calculating statistics based on the number and length of such runs (16). Laws features are based on ''microtexture masks,'' which are specifically designed to detect patterns such as edges or corners at specific orientations (17). Wavelet features measure energy from features at different scales and orientations (18, 19). Finally, Fourier techniques are based on the frequency spectrum of an image or region and can be used to identify preferred periodicities or orientations in a textural pattern or simply to evaluate high-frequency versus low-frequency content (20).

Although several studies have assessed some of these features, we undertook a more systematic and comprehensive assessment of texture features using a large, community-based screening population with the aim of finding features that may capture an element of risk in addition to breast density. We implemented automated routines to calculate a large number of textural features on mammograms from each class described above and trained these on 123 cases and 258 controls (''training'' data set). We next examined the strongest features using an additional 123 cases and 264 controls (''validation'' data set), and for features that validated, we fit final models in the combined data set. All mammograms were prediagnostic screening mammograms and were available on average 3.7 years before breast cancer diagnosis in the cases.

Materials and Methods

Study Population and Data Collection. The study population has been described previously (9). Briefly, subjects were selected from the Mayo Clinic mammography screening practice in Rochester, Minnesota. Patients who did not provide research authorization for medical record studies (3.6%) were not eligible. Cases (*n* = 373) were women aged 50 y and older who were first diagnosed with breast cancer between 1997 and 2001. Each woman had at least two prior screening mammograms done 2 y before diagnosis and lived within a 120-mile radius of the clinic. Women who had bilateral mastectomies or breast implants before diagnosis were excluded. Controls were also selected from the screening practice, among women with no prior history of breast cancer. Two controls were individually matched to each case on age (within 5 y), final screening exam date (within 4 mo), menopausal status at final exam date, time between initial and final mammogram (within 8 mo), number of prior screening mammograms (within one mammogram), and residence (county).

Risk factor data, including weight, height, and use of hormone replacement therapy, were obtained from a clinical database or abstracted from the medical record for the dates closest to a particular mammogram. Height and weight were used to construct body mass index (BMI) in kilograms per meters squared. All mammograms during the preceding 10-y period were obtained.

For these analyses, we retrieved the earliest mammogram available on all cases and controls and digitized them at high resolution. Based on the need to digitize at high resolution, mammograms were available on 246 of the cases and 522 of the controls. The analysis was based on film mammograms. All mammograms were digitized on a Lumiscan 85 scanner with 12-bit grayscale depth. The pixel size was approximately 0.100 × 0.100 mm^2. Most of the images were 18 × 24 cm, yielding images of approximately 1,800 × 2,400 pixels, whereas ~15% were 24 × 30 cm, yielding images of approximately 2,400 × 3,000 pixels. All four views (left and right mediolateral oblique view and left and right craniocaudal view) were digitized.

Estimation of PD. Percent breast density (dense area divided by total area × 100) and absolute dense area (cm^2) were estimated for each view using a computer-assisted thresholding program (Cumulus) developed specifically to quantify breast density (7). Briefly, two thresholds are set by a single trained technician: one separates the breast from the background and the second separates dense from nondense tissue. We consistently showed high intrareader reliability ($r > 0.90$) for our technician while reading over 500 duplicate images from this study across varying time frames (9).

Batch files comprised both cases and controls with randomly assigned views and sides within a woman to maximize precision of PD estimates (21). A 5% repeat set of images was included within each batch file for assessment of reliability.

Texture Measurements. In imaging terms, texture can be described as the spatial arrangement and variation of intensities (gray values) within an image. We assessed the textural properties of the breast tissue using five broad families of features: Markovian cooccurrence matrices, run-length analysis, Laws features, wavelet decomposition, and Fourier analysis. The skewness and kurtosis of the image histograms were also computed. The breast was isolated from the image by using the thresholds and delineations that were defined in the estimation of PD, and texture computations were limited strictly to the breast tissue.

The breast is compressed before image acquisition, which creates two different regions within the breast. The first region consists of the central area of the breast where the thickness is nearly uniform and is referred to as the constant thickness region (CTR). The other consists of tissue near the edge of the breast where the thickness gradually tapers due to the breast geometry. The CTR was approximated by eliminating a margin 160 pixels wide from the perimeter of the breast region. Markovian, run-length, Laws, and wavelet features were calculated for this approximate CTR. A rectangular region within the CTR was also located with a fully automated method developed previously (22). This method finds the largest rectangular box that can be inscribed within the breast region (CTR-box), and this was used to calculate the Fourier texture features, which operate only on rectangular regions. The distribution of total breast area did not vary significantly between cases (138.8 ± 54.4 cm^2) and controls (139.5 ± 52.4 cm^2; $P = 0.68$), and this was true for the CTR and the CTR-box.

Markovian Features. Markovian texture measurements quantify image texture by analyzing how frequently pixels of given intensities appear at specific distances from pixels of other intensities. Computation of Markovian features begins by first constructing a cooccurrence matrix. The *a,b* entry of the cooccurrence matrix specifies the probability that pixels of intensity *a* and *b* appear in the image separated by a distance *d* in either the x or y direction. The distance *d* is specified before construction of the matrix based on the spatial scale to be explored. To reduce the effects of noise variations in image intensity, the image may be decimated in gray scale by a specified factor *g*. The decimation is done by dividing the intensity of each pixel by *g* and retaining the integer portion of the result. Assuming the original image has *N* gray levels, the resulting image will have *N/g* gray levels. The gray level reduction is fixed before construction of the matrix based on the expected strength of the texture signal and the degree of noise suppression desired.

Markovian texture features are then computed based on measurements derived from the cooccurrence matrix. Pressman (23) described 21 texture features that could be computed from cooccurrence matrices. All 21 features were used in this study and are referred to as Markov 1-21. Seven additional related features proposed in the literature by various authors or developed here were also calculated and are denoted as Markov 22-28 (24, 25). Two additional features derived from cooccurrence matrices were described in a study of mammographic density patterns in women at high risk of developing breast cancer (26). These two features, termed coarseness and contrast in the latter study, are referred to in this article as Markov 29 and Markov 30, respectively (note that contrast is also the common term for Markov 2, but it is calculated with a different formula).

The Markovian features were computed for gray-scale reduction values (*g*) of 1, 2, 4, 8, 16, 32, and 64 and for distance values (*d*) of 1, 2, 4, 8, and 16. These reduction values were selected to cover a wide range of spatial scales and feature strengths relative to noise. Thus, a total of 7*5*30 = 1,050 Markovian texture features were generated for each image.

Run-Length Features. As the name implies, run-length texture features examine runs of similar gray values in an image. Runs may be labeled according to their length, gray value, and direction (either horizontal or vertical). Long runs of the same gray value correspond to coarser textures, whereas shorter runs correspond to finer textures.

Run information from an image is recorded in the run distribution matrix. The *a,b* entry of the matrix specifies the number of runs of length *b* with gray value *a* that occur in the image. Similar to the Markovian texture features, the number of gray levels may be decimated to reduce noise variations in image intensity. Both the gray level decimation and run direction are fixed before construction of the matrix.

Texture content is quantified by computing eight features derived from the run-length distribution matrix. The first five features, described by Galloway (16), examine the length and uniformity of the runs and are invariant with respect to intensity. These features are referred to as RL1 through RL5. Two additional features were computed that weight the texture content according to pixel intensity and are referred to as RL6 and RL7 (27). A final feature, referred to as RL8, simply counts the number of runs within the breast region.

Run-length textures were computed for gray level reduction values of 1, 2, 4, 8, 16, 32, and 64 to explore a range of feature strengths relative to the noise level. Runs were tallied and analyzed separately for the horizontal and vertical directions. Thus, a total of 7*2*8 = 112 run-length features were generated for each image.

Laws Features. Laws features (17) are constructed from a set of five one-dimensional filters, each designed to respond to a different type of structure in the image. These one-dimensional filters are denoted E5 (edges), S5 (spots), R5 (ripples), W5 (waves), and L5 (low pass, or average gray value). By applying a one-dimensional filter in the horizontal direction followed by a (possibly different) one-dimensional filter in the vertical direction, results from 25 different two-dimensional filters can be calculated. These filters were applied to the image at full resolution and to the image downsized by factors of 2, 4, 8, and 16 using cubic spline interpolation. The downsized images allow the Laws filters to detect increasingly broader and coarser textures. The second and fourth moments within the CTR for each of the 25 filters were then calculated. Thus, a total of 25*5*2 = 250 Laws features were calculated for each image.

Wavelet Features. The discrete wavelet transform iteratively decomposes an image into four components based on frequency content and orientation (18, 19). Each iteration splits the image both horizontally and vertically into low-frequency (low pass) and high-frequency (high pass) components. Thus, four components are generated: a high-pass/high-pass component consisting of mostly diagonal structure, a high-pass/low-pass component consisting mostly of vertical structures, a low-pass/high-pass component consisting mostly of horizontal structure, and a low-pass/low-pass component that represents a blurred version of the original image. Subsequent iterations then repeat the decomposition on the low-pass/low-pass component from the previous iteration. These subsequent iterations highlight broader diagonal, vertical, and horizontal textures. The frequency content of a given iteration, termed an octave, is one half that of the previous iteration.

Five iterations of the symlet 12 wavelet transform (18) were done, yielding 15 diagonal, horizontal, and vertical components across five octaves. Each component was then isolated in the image by setting the other components to zero and inverting the transform, resulting in a filtered version of the original. The second and fourth moments of the intensities within the CTR region were then computed. Thus, a total of 5*2*3 = 30 wavelet features were generated for each image.

Fourier Features. The two-dimensional Fourier transform decomposes an image into the sum of sinusoidal waves of increasing frequency. A Fourier transform was done on the CTR-box region for each image. The transform was then broken up into 40 equally spaced annular regions (discarding the central circle) and also a

Table 1. Characteristics of the study population by case and control status

Characteristic	Cases (n = 246)		Controls (n = 522)	
	n	Mean (SD) or percent distribution	n	Mean (SD) or percent distribution
Age at mammogram (y)	246	63.9 (10.1)	522	64.1 (10.2)
Interval between mammogram and breast cancer diagnosis (cases) or exam date (controls), y	246	3.8 (1.0)	522	3.7 (1.0)
PD	245	28.9 (13.1)	520	25.1 (13.9)
BMI (kg/m^2)	235	27.9 (5.0)	510	27.6 (5.3)
Menopausal status				
Premenopausal	21	9%	40	8%
Postmenopausal	220	91%	472	92%

''corner'' region consisting of the region outside of the largest inscribed circle. The energy (second moment) within each region was calculated, resulting in 41 Fourier features for each image. The innermost annular region was labeled region 1 and corresponds to the lowest frequency content of the image. Higher numbered regions represent annuli closer to the outside of the transform and correspond to higher frequency content in the image. Large energy values for lower-numbered regions indicate the presence of broader, coarser textures, whereas large energy values for higher-numbered regions correspond to finer textures. These quantities were calculated both as raw numbers and as relative numbers normalized to the total energy in all the annuli adding to unity (22).

Frequency content was also summarized by estimating the power law spectrum of the Fourier transform as previously suggested (28). The power law spectrum relates the power (energy, or second moment) of the Fourier transform to frequency by finding a coefficient β such that Power $\propto 1/(\text{Radial Freq})^{\beta}$. The coefficient β was estimated using ordinary least-squares regression. Comparisons of previous work in mammographic spectral analysis show that the power law model holds for mammograms acquired from various sensors and different data representations (22).

Histogram Measurements. In addition to the textural features specified above, the sample skewness and kurtosis of the image histograms were also computed (20). The computations were limited to the CTR.

Statistical Analyses. The study population was randomly divided into training (123 cases and 258 controls) and validation sets (123 cases and 264 controls) while retaining the matched nature of the data. Although controls had been matched to the cases, unconditional logistic regression was used to test for associations between the mammographic features and breast cancer case-control status. This approach provided a more natural structure under which to do our entire suite of analyses, including the calculation of concordance statistics and repeated bootstrap sampling and analysis. Breaking the matches, and doing analysis using unconditional logistic regression, provided results that were comparable with what would have been observed in the matched analyses because the distributions of the matching variables remained consistent between cases and controls even after the explicit matching was dissolved. We verified the comparability of results from conditional and unconditional logistic regression approaches on selected analyses.

A series of logistic regression–based analyses were carried out to identify features that were associated with breast cancer case-control status. Initial analyses were done within each class of features (i.e., Markovian, run length, Laws, wavelets, and Fourier), and later analyses were done across these classes. Due to the strong correlation among some features within each class, a bootstrap-based variable selection procedure was used to identify the subset of features most strongly associated with breast cancer. Five hundred bootstrap samples of the same size as the original collection of cases and controls in the training set were selected with replacement from the training set. For each bootstrap sample, a stepwise logistic model selection procedure was carried out. This identified the subset of features most strongly associated with breast cancer in the bootstrap sample. Features in this model were recorded and the process

Table 2. Summary of the results for the strongest validated features

	Training data set		Validation data set	Combined data set			
				Unadjusted		Adjusted for age and BMI	
Feature	OR*	AUC	AUC	OR* (95% CI)	AUC	OR* (95% CI)	AUC
PD	1.35	0.587	0.581	1.32 (1.14-1.54)	0.584	1.49 (1.25-1.78)	0.604
Laws (E5L5 M2 scale 16)	1.47	0.597	0.573	1.32 (1.14-1.54)	0.584	1.44 (1.22-1.71)	0.610
Markovian (#23 S8 level 1)	1.51	0.608	0.578	1.36 (1.16-1.58)	0.594	1.36 (1.17-1.60)	0.611
Run length (X direction #3 level 64)[†]	1.45	0.599	0.583	1.37 (1.16-1.63)	0.590	1.44 (1.20-1.73)	0.609
Wavelet (O5 D M2)	1.52	0.623	0.554	1.34 (1.15-1.56)	0.588	1.37 (1.17-1.61)	0.608
Fourier (normal 2)	1.57	0.627	0.568	1.41 (1.21-1.66)	0.598	1.50 (1.27-1.77)	0.622
Power law[†]	1.35	0.589	0.579	1.35 (1.16-1.58)	0.584	1.49 (1.25-1.78)	0.613

*ORs, per SD change.

[†] OR originally <1.0 (inverted for consistency and ease of comparison).

Table 3. Selected descriptive characteristics for the strongest validated features

Variable	Mean ± SD	Min	Max	Correlation with age	Correlation with BMI
PD	26.3 ± 13.8	0	74.7	−0.23	−0.38
Laws (E5L5 M2 scale 16)	594.3 ± 468.6	40.3	5045.1	−0.31	−0.18
Markovian (#23 S8 level 1)	2766.9 ± 862.8	0	8798.1	−0.09	−0.02
Run length (X direction #3 level 64)	0.07 ± 0.03	0	0.31	0.23	0.14
Wavelet (O5 D M2)	246.3 ± 100.8	41.5	979.9	−0.15	−0.06
Fourier (normal 2)	0.09 ± 0.02	0.03	0.17	−0.23	−0.19
Power law	2.55 ± 0.37	2.00	3.40	−0.30	−0.29

was repeated for each bootstrap sample. The number of times each feature appeared in the final stepwise-selected logistic regression model was tabulated.

The features that were most frequently retained in the models specific to each bootstrap sample were considered to be candidate variables for the specific class of image features being considered. These features were then entered into a multiple logistic regression model, which was subsequently simplified using a backward elimination procedure: the least significant features were removed, one at a time, until all variables in the final model were significantly ($P < 0.05$) associated with breast cancer case-control status. The features identified through these model selection procedures were then assessed in the validation data set. This was done by applying the regression coefficient, estimated from the training data set, from the identified feature to the measurements of that feature from the images in the validation data set. The resulting score was assessed for significance in a simple logistic regression model, and the validation concordance statistic was computed.

For the features that validated within each class, models were then refit on the entire data set. These models were used to obtain estimates of odds ratios (OR), confidence intervals (CI), and concordance statistics. Because age, BMI, and PD are strongly associated with risk of breast cancer, ORs and c-statistics were also estimated after adjusting for these factors. To measure the predictive precision of each feature, the concordance or c-statistic, also known as the area under the receiver operator characteristic (ROC) curve (AUC), for a logistic regression model was calculated. This statistic measures how often the model correctly identifies the case in a random case-control pair as having a higher risk. Values of this statistic range from 0.0 (perfect incorrect prediction) to 0.5 (chance) to 1.0 (perfect prediction).

In an attempt to combine all features across the different classes into one model, a final bootstrap procedure similar to the one used within each class of features described above was done where the final validated features from each class were considered to be the only candidate features. As in the individual classes of features, the top selected features were removed by a backward elimination procedure until only those with $P < 0.05$ remained. Two-sided P values were obtained for all tests. Statistical Analysis System software (SAS) was used for all analysis.

Results

Table 1 presents descriptive information on the 246 cases and 522 controls included in this report. The time interval between the baseline mammogram and the date of cancer diagnosis in cases or date of last mammogram in controls was 3.7 ± 1.0 years on average (range, 2.0-5.9) and was >3 years for over 73% of the participants. The matching algorithm was quite effective, as evidenced by the similarity of cases and controls with regard to the design variables. The data set was randomly split into training (n = 381, 123 cases and 258 controls) and validation (n = 387, 123 cases and 264 controls) data sets. There were no differences between the two data sets with respect to age or BMI at time of mammography, PD, or time from mammography to diagnosis (cases) or selection (controls; data not shown).

Table 2 shows the main results for PD and the strongest texture features that validated within each class (details of these features are given in the Appendix), and Table 3 shows descriptive characteristics and correlations with age and BMI for each feature. In the combined data set, the unadjusted ORs were in the range of 1.3 to 1.4 per SD change, which is consistent with previous studies. Adjustment for age and BMI either did not change or actually strengthened the ORs and AUCs. AUCs for the age- and BMI-adjusted models of PD and the individual texture features were similar and in the range of 0.60 to 0.62. There were many other features within each feature class that validated and had AUC statistics nearly as large as those listed in Table 2. These features were highly correlated with the top validated feature in each class and were therefore not selected.

Table 4. Spearman correlation coefficients for strongest validated features and PD

	Markovian	Run length	Wavelet	Fourier	Power law	PD
Laws (E5L5 M2 scale 16)	0.61	−0.84	0.54	0.74	0.76	0.59
Markovian (#23 S8 level 1)		−0.50	0.91	0.68	0.48	0.39
Run length (X direction #3 level 64)			−0.39	−0.74	−0.78	−0.59
Wavelet (O5 D M2)				0.65	0.46	0.43
Fourier (normal 2)					0.77	0.64
Power law						0.76

NOTE: *, all $P < 0.0001$.

Table 5. OR and 95% CI for breast cancer case-control status from models that include age, BMI, a selected feature, and PD

Feature	Feature OR (95% CI)	PD OR (95% CI)	AUC
Laws (E5L5 M2 Scale 16)	1.27 (1.06-1.54)	1.36 (1.12-1.65)	0.623
Markovian (#23 Sep8 level 1)	1.26 (1.07-1.47)	1.40 (1.17-1.68)	0.633
Run length (X direction RL 3 level 64)*	1.26 (1.03-1.54)	1.35 (1.11-1.64)	0.618
Wavelet (O5 D M2)	1.24 (1.05-1.46)	1.38 (1.15-1.66)	0.626
Fourier (normal 2)	1.31 (1.08-1.60)	1.28 (1.04-1.58)	0.627
Power law*	1.27 (1.00-1.61)	1.28 (1.01-1.62)	0.619

NOTE: ORs, per SD change.
*OR originally <1.0 (inverted for consistency and ease of comparison).

As shown in Table 4, the validated texture features were correlated with PD (Spearman correlation coefficients, 0.39-0.76) and with each other (0.46-0.91). When each validated texture feature was individually included in a model with PD, all ORs remained statistically significant, whereas the ORs for PD attenuated from 6% to 14% and the ORs for texture features attenuated from 11% to 18% (Table 5). In these models, the AUC increased only slightly (to 0.62 to 0.63) over models with just the texture feature or PD alone. Finally, in an attempt to combine across the feature families, the bootstrap procedure was repeated using all validated features. The results showed that the Fourier (normalized, region 2) and wavelet (order 5 diagonal, 2nd moment) features were the two most frequently selected. When these two features were fit together in the same model, only the Fourier feature remained statistically significant (OR, 1.31; 95% CI, 1.07-1.60).

The top validated features were all at or near the lowest spatial resolution (coarsest scale) used in these calculations, and a notable general trend for all features of a given class was that the differences between cases and controls (larger intensity variations in the cases) progressively declined toward finer scales. For example, the single best Laws feature measured differences in vertical edge strength at the lowest resolution considered; in this case, the differences were measured at a width of ~40 pixels. As shown in Fig. 1, for all Laws features, the average AUC progressively decreased from 0.60 to 0.50 with progressively finer scales. Notice also that, as mentioned above, many Laws features have AUC results similar to the single best Laws feature. Similar behavior is seen for the wavelet and Fourier features. The wavelet feature showed differences in diagonally oriented tissue structures at the coarsest scale considered, corresponding to 32 to 64 pixels. The Fourier feature is measuring energy at a very low spatial frequency annulus, which varies with the size of the CTR-box but corresponds roughly to 27 to 40 pixels, and the Fourier power law exponent shows that there is relatively more energy at lower frequencies in the cases compared with controls. Figures 2 and 3(for the wavelet and Fourier features, respectively) show that as one moves to lower wavelet orders or higher frequency bins (i.e., toward finer scales in both cases), the ability to discriminate between cases and controls degrades gradually. The Markovian feature, which measures a variant of image contrast, examined differences 8 pixels apart, which was the coarsest scale considered for this family. The run-length features were calculated only at one scale, so no general statement can be made about scale, but the relationship found indicates a similar trend: the cases have a less uniform image, with shorter runs of adjacent pixels with similar gray levels.

Discussion

Following a comprehensive evaluation of >1,000 individual measurements from five different classes of texture features, we identified texture features derived from mammographic images that predict breast cancer risk at the same magnitude as PD. When individual features were included in the same model as PD, both the feature and PD remained significant predictors, although the strength of the association (OR) weakened slightly for both variables, likely due to their moderate correlation. Including variables from across feature classes did not improve prediction ability. More importantly, none of these features added significantly to risk prediction beyond PD alone.

Our results consistently show that texture features at low spatial frequencies (i.e., coarser mammographic textures) provided the strongest predictors of future breast cancer risk. The general trend is that the cases seem to have stronger intensity variations (i.e., more energy) across coarse scales than the controls. Many features from different classes have similar ability to discriminate case and control status, and this discrimination degrades as features are calculated at finer scales. Indeed, the results suggest that a wide variety of methods of measuring intensity differences over scales of several mm or so may yield predictive power comparable with PD. We hypothesize that this may

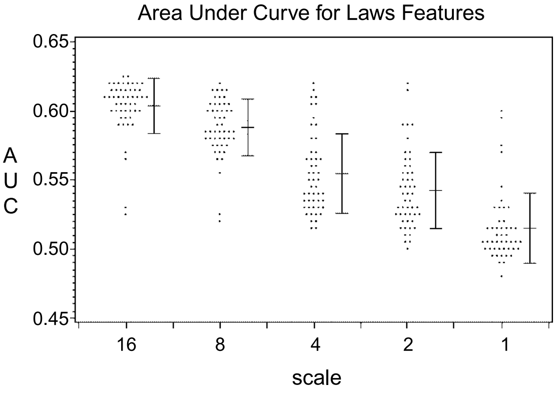

Figure 1. AUC by decreasing scale for Laws features.

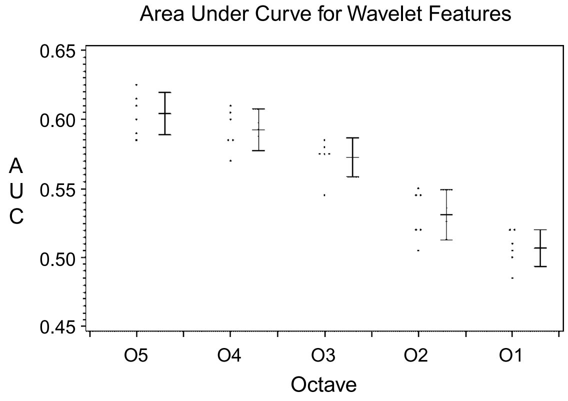

Figure 2. AUC by decreasing octave level for wavelet features.

primarily be a reflection of the simple presence of breast density itself. Breast density is a measure of what percentage of the breast has intensity above some threshold that is meant to separate dense fibroglandular tissue from darker fatty tissue. Because the breast density for most women is well under 50%, an increase in density usually means that there are more bright areas in an image that is still primarily dark. In this scenario, texture features would show increased energy at low frequencies as density increases because pixel differences would increase at broad spatial scales. This would explain why so many features have predictive power roughly equivalent to PD, but the predictive power does not significantly increase when they are combined with PD. As noted by Pepe et al. (29), for the AUC to be increased substantially by the addition of new risk factors, the magnitude of association for at least one of the added variables needs to be quite large. Unfortunately, the ORs corresponding to the features identified here were not of a magnitude that could lead to a marked improvement in the AUC.

It is also possible that the textural differences between mammogram images from cases and controls may be correlated with the Wolfe classifications. Both the high-risk P2 and DY categories consist of broad, coarse patterns that are likely to exhibit strong energy at lower frequencies. Furthermore, our texture measurements, similar to the Wolfe patterns, offer little discrimination capability above and beyond PD. However, we did not have the Wolfe classification on our study population, and additional study would be required to establish a definitive link between the texture measurements and the Wolfe classifications.

Several previous studies have assessed texture features as predictors of breast cancer risk. Torres-Mejia (10) found a 39% decreased risk of breast cancer per 1 SD increase in lacunarity after adjusting for SD and other confounders. A decrease in lacunarity at a given density corresponds to a more regular or homogeneous arrangement of objects versus gaps in the image (i.e., a more homogeneous mammographic parenchymal pattern) and is associated with a higher risk of breast cancer. Because this analysis was done in conjunction with a fractal dimension analysis and calculated a single lacunarity coefficient across all scales, it is difficult to compare this result to ours in terms of strength of features at different scales.

Byng et al. (30) and Yaffe et al. (31) reported a negative correlation between skewness and cancer risk, but not as strong as PD itself, and probably due to the high correlation between skewness and PD. Nagao et al. (32) studied an automated measure based on the location of the peak of the histogram and also saw a result similar to but not as strong as PD.

Other studies have also used mammographic texture analyses to identify differences among women at different levels of risk for developing breast cancer. Notably, Huo et al. (26) used a linear combination of four mammographic texture and histogram features to distinguish between mammograms of a small sample of BRCA1/BRCA2 mutation carriers (n = 30), of whom the majority (n = 17) also had breast cancer, and age-matched noncarriers (n = 60). The latter study achieved a very high AUC (0.91) for distinguishing mutation carriers from low-risk noncarriers. Both their Markovian features (coarseness and contrast) and their histogram features (skewness and kurtosis) were analyzed in our data set but did not show a significant ability to predict breast cancer risk in our average-risk population. Huo et al. (26) concluded, however, that BRCA1/BRCA2 carriers tend to have mammograms that are low in contrast and are coarse in texture, which is consistent with our findings for discriminating case and control status.

The texture measurements presented here may be useful in a multivariate risk prediction model. Although the predictive capabilities of these features are moderate at best, they are similar to the widely used Gail model, including the more recent models that incorporate mammographic density (12, 13). However, density estimation is both subjective and laborious, which may limit its use in clinical practice. The texture measurements presented in this work, though, are both objective and automated, which is an important advantage. However, numerical values of texture features tend to vary with differences in acquisition, so standardizing across institutions and scanners may be difficult. It might be possible to standardize for variables such as compression force, thickness, kVp, and mAs with techniques such as the representation proposed by Highnam and Brady (33). Accounting for these influences may not only improve predictive capability but may also generalize

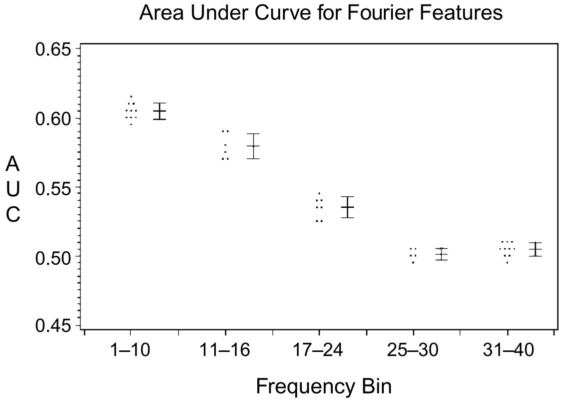

Figure 3. AUC curve by increasing frequency bin for Fourier features.

the measurements to films from other institutions. Unfortunately, acquisition variables were not available for the vast majority of the films in this data set. It will also be important to confirm the significance of these features in the full-field digital mammography environment, where these features could automatically be assessed in the digital data at the time of examination.

Our study design had several key strengths. We used a relatively large case-control study that allowed us to directly assess the association of PD and texture features from screening mammograms with subsequent breast cancer risk. The case-control study included careful control of age, mammogram characteristics (timing and frequency), menopausal status, and residence. We also had data on weight, height, and other breast cancer risk factors at the time of mammogram. We used a quantitative estimate of PD that has consistently shown association with risk (6) and we had excellent intrareader reliability. Our approach to evaluating texture features was comprehensive and systematic, with >1,000 texture features from five different classes analyzed. Our statistical analysis used a robust bootstrapping algorithm that was able to handle the large number of correlated independent variables. Both a split sample validation as well as cross-validation in the training set were used to minimize overly optimistic predictions.

There are some limitations that affect interpretation of our results. Our study sample had limited ethnic diversity, and whether these results will hold in other groups is not known. However, PD seems to be a strong risk factor in several populations, including Asian Americans, African-Americans, and Native Hawaiians (34-36). Although this was not a population-based study, by restricting to the 120-mile radius and requiring serial mammograms on all women, the study population was close to a community-based rather than a referral or high-risk population and the internal validity is high. Finally, all mammograms were acquired from a single institution, and therefore, future studies will need to assess the robustness of these results across different institutions and imaging platforms.

In summary, our results show that breast cancer cases tend to have stronger textural properties at coarse scales in their screening mammograms than controls. The ability to predict individual breast cancer risk from these features was modest, although it is similar to both PD and the widely used Gail model. Further analyses that account for the image acquisition variables may improve predictive ability. The objective and automated nature of these features may make them useful within a multivariate risk prediction model. Finally, this work only shows the relationships with the breast cancer risk but does not attempt to relate the findings to the underlying biology or disease etiology. Such knowledge would be useful in developing better risk prediction models and perhaps identifying novel interventions to reduce risk of breast cancer.

Appendix A

The details of the specific features cited in Tables 2 to 4 are described below. Note, however, that we do not consider these features to be unusually predictive; as discussed above, many other features from each of the classes at coarse scales are almost equally predictive.

The Laws feature (E5L5 M2 scale 16) is the second moment (energy) in an image obtained by decimating the original image by a factor of 16 in each direction and convolving the result with the E5 ''edge'' filter (1 -2 0 2 1) in the x direction and the L5 ''low-pass'' filter (1 4 6 4 1) in the y direction. This can be interpreted as a measure of vertical edge strength at a coarse scale.

The Markovian feature (M23 S8 level 1) is calculated from the original intensities (no gray level decimation). We term the quantity we calculate the ''gray level difference variance,'' and it is simply the variance of the absolute values of the differences between pixels and their neighbors; in this case, neighbors 8 pixels away in x and y. This is related to the classic Markovian quantities M2 and M11 but differs from the ''difference variance'' feature M10 (23). It is also related to the analysis by Chandrasekaran et al. (25): if $p(i)$ is the histogram whose ith component is the probability that two neighboring pixels have an absolute intensity difference of i, then the mean of this histogram is $M22 = \sum_{i=1}^{N} i * p(i)$ (also calculated here), and the corresponding variance is $M23 = \sum_{i=1}^{N} (i - M22)^2 * p(i)$. Despite its simplicity, this feature has not been proposed before to our knowledge.

The run-length feature (X direction #3 level 64) calculates the run distribution matrix $p(i,j)$ for runs of identical pixels in the x direction after the image intensities are divided by 64 and calculates the gray level nonuniformity GLN from this matrix according to Galloway (16):

$$\mathrm{GLN} = \sum_i \left(\sum_j p(i,j) \right)^2 / \sum_i \sum_j p(i,j)$$

The wavelet feature (O5 D M2) applies five iterations of the symlet 12 wavelet transform (18), keeps the values for the diagonal band of this fifth octave, sets all other wavelet coefficients to zero, inverts the wavelet transform, and calculates the 2nd moment of the image intensities within the CTR.

The Fourier feature (norm 2) is the normalized energy in the second annular bin of the Fourier transform, as described in Materials and Methods. The Fourier power law feature is fully described there as well.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank Fang Fang Wu who worked tirelessly to digitize and estimate density on thousands of films for this study and Sondra Buehler for secretarial support.

References

1. Warner E, Lockwood G, Tritchler D, Boyd NF. The risk of breast cancer associated with mammographic parenchymal patterns: a meta-analysis of the published literature to examine the effect of method of classification. Cancer Detect Prev 1992;16:67–72.

2. Saftlas AF, Szklo M. Mammographic parenchymal patterns and breast cancer risk. Epidemiol Rev 1987;9:146 – 74.
3. Oza AM, Boyd NF. Mammographic parenchymal patterns: a marker of breast cancer risk. Epidemiol Rev 1993;15:196 – 208.
4. Boyd NF, Lockwood GA, Byng JW, Tritchler DL, Yaffe MJ. Mammographic densities and breast cancer risk. Cancer Epidemiol Biomarkers Prev 1998;7:1133 – 44.
5. Wolfe JN. Risk for breast cancer development determined by mammographic parenchymal pattern. Cancer 1976;37:2486 – 92.
6. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiol Biomarkers Prev 2006;15:1159 – 69.
7. Boyd NF, Byng JW, Jong RA, et al. Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study. J Natl Cancer Inst 1995; 87:670 – 5.
8. Byng JW, Boyd NF, Little L, et al. Symmetry of projection in the quantitative analysis of mammographic images. Eur J Cancer Prev 1996;5:319 – 27.
9. Vachon CM, Brandt KR, Ghosh K, et al. Mammographic breast density as a general marker of breast cancer risk. Cancer Epidemiol Biomarkers Prev 2007;16:43 – 9.
10. Torres-Mejia G, De Stavola B, Allen DS, et al. Mammographic features and subsequent risk of breast cancer: a comparison of qualitative and quantitative evaluations in the Guernsey prospective studies. Cancer Epidemiol Biomarkers Prev 2005;14: 1052 – 9.
11. Brisson J, Diorio C, Masse B. Wolfe's parenchymal pattern and percentage of the breast with mammographic densities: redundant or complementary classifications? Cancer Epidemiol Biomarkers Prev 2003;12:728 – 32.
12. Barlow WE, White E, Ballard-Barbash R, et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. J Natl Cancer Inst 2006;98:1204 – 14.
13. Chen J, Pee D, Ayyagari R, et al. Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. J Natl Cancer Inst 2006;98:1215 – 26.
14. Garra BS, Krasner BH, Horii SC, Ascher S, Mun SK, Zeman RK. Improving the distinction between benign and malignant breast lesions: the value of sonographic texture analysis. Ultrason Imaging 1993;15:267 – 85.
15. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Syst Man Cybern 1973;3:610 – 21.
16. Galloway MD. Texture classification using gray level run length. Comput Graph Image Processing 1975;4:172 – 9.
17. Laws K. Textured image segmentation [dissertation]. Los Angeles (CA): University of Southern California; 1980.
18. Daubechies I. Ten lectures on wavelets. Philadelphia (PA): Society for Industrial and Applied Mathematics; 1992.
19. Chui CK. An introduction to wavelets. San Diego (CA): Academic Press; 1992.
20. Russ JC. The image processing handbook. Boca Raton (FL): CRC Press; 1992.
21. Stone J, Gunasekara A, Martin LJ, Yaffe M, Minkin S, Boyd NF. The detection of change in mammographic density. Cancer Epidemiol Biomarkers Prev 2003;12:625 – 30.
22. Heine JJ, Velthuizen RP. Spectral analysis of full field digital mammography data. Med Phys 2002;29:647 – 61.
23. Pressman NJ. Markovian analysis of cervical cell images. J Histochem Cytochem 1976;24:138 – 44.
24. Lachmann F, Barillot C. Brain tissue classification for MRI data by means of texture analysis. SPIE Med Imaging VI Image Processing 1992;1652:72 – 83.
25. Chandrasekaran K, Aylward PE, Fleagle SR, et al. Feasibility of identifying amyloid and hypertrophic cardiomyopathy with the use of computerized quantitative texture analysis of clinical echocardiographic data. J Am Coll Cardiol 1989;13:832 – 40.
26. Huo Z, Giger ML, Olopade OI, et al. Computerized analysis of digitized mammograms of BRCA1 and BRCA2 gene mutation carriers. Radiology 2002;225:519 – 26.
27. Chu A, Sehgal CM, Greenleaf JF. Use of gray value distribution of run lengths for texture analysis. Patt Recogn Lett 1990;11:415 – 20.
28. Li H, Giger ML, Olopade O. Power spectral analysis of mammographic parenchymal patterns. Proc SPIE 2006;6144:61445J1 – 4.
29. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 2004; 159:882 – 90.
30. Byng JW, Yaffe MJ, Lockwood GA, Little LE, Tritchler DL, Boyd NF. Automated analysis of mammographic densities and breast carcinoma risk. Cancer 1997;80:66 – 74.
31. Yaffe MJ, Boyd NF, Byng JW, et al. Breast cancer risk and measured mammographic density. Eur J Cancer Prev 1998;7 Suppl 1:S47 – 55.
32. Nagao Y, Kawaguchi Y, Sugiyama Y, Saji S, Kashiki Y. Relationship between mammographic density and the risk of breast cancer in Japanese women: a case-control study. Breast cancer (Tokyo) 2003;10: 228 – 33.
33. Highnam R, Brady M, Shepstone B. A representation for mammographic image processing. Medical Image Analysis 1996;1:1 – 18.
34. Maskarinec G, Meng L. A case-control study of mammographic densities in Hawaii. Breast Cancer Res Treat 2000;63:153 – 61.
35. Ursin G, Ma H, Wu AH, et al. Mammographic density and breast cancer in three ethnic groups. Cancer Epidemiol Biomarkers Prev 2003;12:332 – 8.
36. Maskarinec G, Pagano I, Lurie G, Kolonel LN. A longitudinal investigation of mammographic density: the multiethnic cohort. Cancer Epidemiol Biomarkers Prev 2006;15:732 – 9.

Texture Features from Mammographic Images and Risk of Breast Cancer

Armando Manduca, Michael J. Carston, John J. Heine, et al.

Cancer Epidemiol Biomarkers Prev 2009;18:837-845.