Journal of Digital Imaging

VOL 11, NO 3 AUGUST 1998

A Breast Density Index for Digital Mammograms Based on Radiologists' Ranking

John M. Boone, Karen K. Lindfors, Carol S. Beatty, and J. Anthony Seibert

The purpose of this study was to develop and evaluate a computerized method of calculating a breast density index (*BDI*) from digitized mammograms that was designed specifically to model radiologists' perception of breast density. A set of 153 pairs of digitized mammograms (cranio-caudal, CC, and mediolateral oblique, MLO, views) were acquired and preprocessed to reduce detector biases. The sets of mammograms were ordered on an ordinal scale (a scale based only on relative rank-ordering) by two radiologists, and a cardinal (an absolute numerical score) BDI value was calculated from the ordinal ranks. The images were also assigned cardinal BDI values by the radiologists in a subsequent session. Six mathematical features (including fractal dimension and others) were calculated from the digital mammograms, and were used in conjunction with single value decomposition and multiple linear regression to calculate a computerized BDI. The linear correlation coefficient between different ordinal ranking sessions were as follows: intraradiologist intraprojection (CC/CC): $r = 0.978$; intraradiologist interprojection (CC/MLO): $r = 0.960$; and interradiologist intraprojection (CC/CC): $r = 0.968$. A separate breast density index was derived from three separate ordinal rankings by one radiologist (two with CC views, one with the MLO view). The computer derived *BDI* had a correlation coefficient (r) of 0.907 with the radiologists' ordinal *BDI*. A comparison between radiologists using a cardinal scoring system (which is closest to how radiologists actually evaluate breast density) showed $r = 0.914$. A breast density index calculated by a computer but modeled after radiologist perception of breast density may be valuable in objectively measuring breast density. Such a metric may prove valuable in numerous areas, including breast cancer risk assessment and in evaluating screening techniques specifically designed to improve imaging of the dense breast.

KEY WORDS: breast cancer, mammography, breast density, digital mammography, computer aided diagnosis

WOMEN WITH DENSE BREASTS appear to have a four to six fold increase in breast cancer risk,[1-4] yet imaging the dense breast continues to be problematic. Cancers are detected at later stages in dense breasts and radiologists recognize that their diagnostic accuracy is lower in such women. Consequently, efforts to improve the detectability of breast cancer in the dense breast have received increased attention. Refinements in mammography and new techniques including digital mammography,[5] high definition and Doppler ultrasound,[6,7] magnetic resonance imaging,[8-10] positron emission tomography (PET),[11,12] and single photon emission computed tomography imaging (SPECT)[13,14] are all under development. Many of these techniques are aimed at overcoming the limitations of conventional mammography in the radiographically dense breast, yet there is no truly quantitative method for grading breast density. Such a metric would have many uses, including the assessment of the impact of these new modalities on the detection of early cancer in the dense breast.

Wolfe was the first to describe a discrete classification scheme with four classes of mammographic density patterns.[1,2,15-18] More recently, the Breast Imaging Reporting and Data System (BIRADS) was introduced by the American College of Radiology. It also makes use of four classifications of breast density. Although these classifications are helpful for communication of diagnostic sensitivity, they are both subjective and crude. The study presented here was designed to evaluate sets of

From the Department of Radiology, University of California, Davis, UC Davis Medical Center, Sacramento, CA.

Research funded in part by the Breast Cancer Research Program for the United States Army (Grant DAMDI 17-94-J-4424) and the California Breast Cancer Research Program (Grant 1RB-0192).

Address reprint requests to John M. Boone, PhD, Department of Radiology, UC Davis Medical Center, FOLB II E, 2421 45th St, Sacramento, CA 95817.

0897-1889/98/1103-0001$8.00/0

computer-calculated features which could be used to quantify breast density on a continuous scale from digital (or digitized) mammograms. In addition, the breast density index (*BDI*) developed was specifically modeled to adhere to a radiologist's perception of breast density.

The number of useful breast density categories that one can assign a mammogram to is an important consideration; with too many categories, assignment can become less reproducible and arbitrary, while with too few categories, useful density strata would go unappreciated and benefits of breast density classification would be under-realized. Therefore, we have analyzed the many classifications performed in this study in a manner that may shed light on what a reasonable number of categories might be for breast density categorization.

METHODS AND MATERIALS

Case Selection and Film Digitization

A series of normal left mammograms (Cranio-caudal view [CC] and mediolateral oblique [MLO]) of 160 different patients was selected from the breast imaging service at our institution. For each set of films, the patient's name, date of birth, and the examination date was recorded. The patient population at our medical center is representative of the broad ethnic distribution typical of large urban centers in California. Cases were selected serially, and no selection criteria was used to limit incorporation into the study.

The film images were digitized using a Lumisys 200 laser film digitizer (Lumisys, Sunnyvale, CA). The pixel size was 50 μm × 50 μm, and the gray scale was digitized to 12 bits. The large (40 Mbyte) files were cropped using software written for this purpose, eliminating some of the area beyond the silhouette of the breast, and the cropped images were stored at original resolution on a series of optical disks. For realistic manipulation, display, and computation, the images were reduced in size by pixel averaging to 500 μm × 500 μm pixels. At this spatial resolution (for the CC view, the down-sampled images averaged 195.3 ± 4.2 pixels wide and 394.0 ± 4.2 pixels tall), a good overall view of the breast architecture could be appreciated.

Radiologist Ranking Scheme

The radiologist's determination of breast density was used as the gold standard in this study. To rank-order the mammograms in this study, all images needed to be visualized simultaneously by the radiologist. The CC and MLO image sets were therefore replicated in miniature using the following procedure.

The relationship between the digitized gray scale value and the film optical density (OD) was measured by digitizing a sheet of film which contained steps of known optical densities, and the average gray scale value in each region was quantified. The OD as a function of gray scale was fit to a straight line ($r > 0.9999$). The relationship between gray scale value and optical density was also measured for a laser imager, and this relationship was characterized using commercially available software (TableCurve 2D; Jandel Scientific, San Rafael, CA). From these data, a transformation curve was calculated which allowed the digitized mammograms to be printed onto laser film at their original optical densities. Using this method, small replicas of the original mammograms were printed which had the "identical" densities as the original analog film mammograms. Each mammogram replica was approximately 3.5 cm × 8.0 cm, but varied slightly with breast size from image to image. Of the 160 original pairs of mammograms, there were technical difficulties with seven, including digitizer errors (corrupted data files), and lost or duplicate miniature films. Consequently, 153 pairs of mammograms were used in the subsequent analyses. Using a 4-over-1 lightbox placed flat on a countertop (area of view box was 142 cm wide by 43 cm tall), all 153 miniature mammograms could be placed in order with simultaneous visualization of all images for comparison purposes. Whereas the effect of using miniature mammograms was not explicitly evaluated, it is anticipated that this had little or no effect on the results because the present task involved the assessment of breast density only, and not diagnosis.

Both radiologists involved in this study are experienced in the interpretation of mammograms. They were instructed to place in order, from most dense to least dense, the 153 images in each set. The rank ordering process required approximately 2-3 hours for each session. Radiologist 1 (RAD_1) rank-ordered the CC set twice (referred to as RAD_1 CC_1 and CC_2), in sessions that were performed more than 4 months apart. The MLO set was rank ordered by RAD_1 once. To evaluate inter-observer variability, a second radiologist (RAD_2) rank-ordered the CC image set (RAD_2 CC_1) as well.

Rank ordering a series of mammograms with the entire image set in full view of the radiologist is a conceptually different task than viewing an individual mammogram and assigning a density value. To measure the difference between these two distinct tasks, both radiologists assigned a "freehand" breast density index to each image, which was viewed alone and months apart from any other ranking session. This assignment used the scale where 100 corresponded to a very dense breast and 0 was a totally fatty replaced breast. The freehand assignment of breast density will be referred to as the CC_3 ordering session for each radiologist (Rad_1 and Rad_2).

The Breast Density Index (BDI)

To generate a quantitative scale of breast density, a breast density index (BDI) was computed from the radiologist's rank ordering of the images. The BDI was designed to range from 0 to 100 on a continuous scale, where 100 corresponds to an extremely dense breast, and 0 coincides with an extremely non-dense (fatty replaced) breast. The BDI was calculated for each of the ordinal rankings described above (RAD_1 CC_1, RAD_1 CC_2, RAD_1 MLO_1, or RAD_2 CC_1). In order to do this, the ordinal ranking scale was used to produce the cardinal BDI scale. There is justification for going from ordinal to cardinal scales when the number of cases is large.[19] To do this, the maximum rank score (S_{max}), corresponding to the least dense mammogram and the minimum rank score (S_{min}) corresponding to the most dense mammogram were computed from the rank ordering data, and then the BDI_j for image j which received a

rank score of S_j was calculated using the equation:

$$BDI_j = 100 \times \left[1 - \frac{S_j - S_{min}}{S_{max} - S_{min}}\right] \quad \text{[Equation 1]}$$

A consensus score from three separate ordinal ranking sessions from a single radiologist (RAD_1) was used for the "gold standard" BDI (referred to as the standard BDI, s-BDI). The three ranks assigned by radiologist 1 during the CC_1, CC_2 and MLO_1 ordering sessions were summed for each image, and the *s-BDI* was calculated using Equation 1.

The assignment of BDI values in session CC_3 did not employ a rank ordering (ordinal scale) of images, but rather was a direct assignment of (cardinal) BDI values by the radiologists. Therefore, the BDI values from session CC_3 (Rad_1 and Rad_2) did not make use of the ordinal to cardinal conversion shown in Equation 1.

Image Preprocessing

H and D Curve Correction. While film mammograms were used in this study, the technique described is intended to be applicable for the more general class of digital mammography images. Digital mammography systems for full field imaging may be commercially available in the next few years, and these systems will in general exhibit a linear response to the x-rays incident upon them (the characteristic curve will be a straight line). In order to make the technique described here applicable to linear images, the non-linear influence of the film was removed using the following pre-processing steps.

The characteristic curve of the screen-film system (Dupont Microvision, Wilmington, DE) was measured over approximately 20 steps by varying the x-ray exposure; the film was processed normally, and the optical density of each step was measured using a calibrated densitometer (TBX-U; Tobias Associates, Ivyland, PA). The exposure to the screen-film cassette (in milliroentgen, 1 mR = 2.58×10^{-7} C/kg) as a function of optical density (in OD units) was computer-fit using commercial software (TableCurve 2D; Jandel Scientific, San Rafael, CA) to an eighth-order polynomial ($r > 0.9999$). The gray scale value-to-exposure transform was combined with the linear relationship between OD and digital number (described previously) to create a function which converted the gray scale values of the digitized images (the raw digital numbers from the film digitizer) to the corresponding x-ray exposure (in mR) to the detector. This transform, shown in Fig 1, was applied to each pixel, effectively reversing the nonlinearity caused by the "H and D" curve of the film.

Image Log-Normalization. The next step in the image pre-processing was performed with the intent to make the digital images more dependent upon the physical characteristics of the breast, while reducing the dependency on absolute exposure levels. In the background regions of the image, outside the breast anatomy (where no breast was in the x-ray beam), the exposure theoretically corresponds to the unattenuated x-ray beam intensity, I_o. Under the breast silhouette, the exposure striking the detector is equal to $I_{\mu x} = I_o e^{-\mu x}$, where μx corresponds to the attenuation properties of the voxel of breast tissue corresponding to each pixel: μ is the linear attenuation coefficient of the tissue in the voxel, and x is the thickness of the voxel. In the compressed breast, the voxel thickness is quite uniform towards the center of the breast and the variability in μx is therefore strongly influenced by μ, which is desirable.

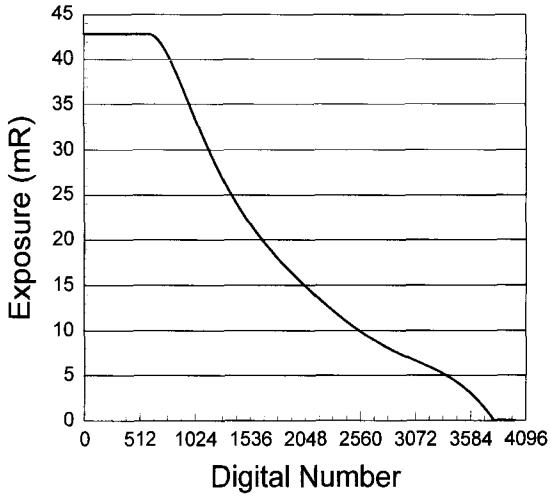

Fig 1. The shape of the functional curve that was used to transform the raw digital number (from digitizing the film) to exposure to the screen-film detector. This curve was generated from the H and D curve of the film, and the (linear) OD-to-digital number response of the laser digitizer used. Using this transform, digital images were corrected for the non-linearities of the screen film detector system.

All of the mammograms had regions outside the breast silhouette which received unattenuated x-ray exposure, I_o, since of course the compressed breasts were approximately semi-circular and the images were rectangular. Each digital image was displayed on an imaging workstation and a small rectangular region outside the breast silhouette in a background area was hand positioned using mouse/cursor software written for this purpose. In this background region of interest, the mean gray scale value was calculated and then transformed using the gray scale-to-exposure curve (shown in Fig 1) to estimate I_o for that image. The value of $I_{\mu x}$ was then calculated for each pixel in the image using the gray scale-to-exposure transform, and the attenuation factor, $\mu x = LN(I_o/I_o e^{-\mu x})$ was calculated, multiplied by 1000 for scaling, and the resulting value was stored as an integer for each pixel. To further clarify the fact that these images were pre-processed images, and are not simply digitized mammograms, the digital images processed as described above will be referred to as "μx-mammograms."

It is acknowledged that the log-normalization of the image is only an estimate of μx, since beam hardening owing to the polyenergetic x-ray spectrum and spatial non-uniformities due to x-ray scatter and other factors were not accounted for. However, this procedure was performed to reduce the dependency of the analyses on absolute exposure and to reduce the dependency of the results on the non-linear response of film. Breast density is intrinsically related to μx; it is therefore only logical to computer-process the images using the available information such that they reflect this quantity to the extent possible.

Image Cropping. When radiologists look at a mammogram, they ignore the background (the region beyond the border of the

breast anatomy) on which the image is projected. This is not automatic, however, for computer analysis and specific efforts have to be taken to focus the computer algorithms on only the breast parenchyma. In order to do this, a threshold of $\mu x = 500$ was set, and based on visual feedback from the images this value was able to segment the breast parenchyma (where in general $\mu x > 500$) from the background periphery (where in general $\mu x < 500$). In some cases, simple thresholding was not sufficient to segment breast from non-breast areas, and therefore each image was inspected and individual image cropping was performed as needed. The predominant structures that required hand cropping were the lead markers ("LCC" and "LMLO"). Cropping was also used to eliminate regions where skin folds resulted in obviously artifactual high attenuation. A final reason for individually editing the images was that, in some images the laser digitizer presented some overshoot near the leading edge of the film, and these areas were cropped out of the μx-mammograms as well.

Regions outside of the breast parenchyma that were segmented out by thresholding and cropping were set to a uniform pixel value of 0 (zero). Since all areas of the image containing breast parenchyma had gray scale values greater than 500, this difference allowed the application of algorithms to only regions of the image containing breast parenchyma.

Computer Algorithms

When a radiologist looks at a mammogram, the human visual (eye-brain) complex applies an incredible array of subjective computations on the image, which can result in the ranking of breast density. For the computer to do this, specific *features* have to be mathematically quantified from each μx-mammogram. A feature is really anything that can be calculated from the images, and there are infinite possibilities of features. Examples of simple features can be the mean gray scale value on the μx-mammograms, or the standard deviation in gray scale values from those images. Much more complicated features can be calculated as well. For each feature, a single numerical value is calculated for each image, using an algorithm specific for that feature applied to each image. Finding the features which most closely correlate with breast density as determined by the radiologist was a principal focus of this study.

In this study, about two hundred features were evaluated for their ability to predict the radiologist's ranking of the images in terms of breast density. During the feature development phase, one third of the data base (51 images) was used for evaluation of features. As the features which were most effective became identified, the full data base was then used for analysis. No single feature that was evaluated was found to correlate with the BDI-standard with a linear correlation coefficient (r) of better than 0.78, calculated over the 153 images in the data base. Therefore, multiple features were combined to improve the computerized determined BDI (referred to as *c-BDI*) fit to the *s-BDI*. This approach required both the delineation of features which performed well, and the identification of the most effective combination of features.

While the calculation of each feature needs to be described mathematically, the details of these calculations may be of interest to only a subset of readers. Therefore, the mathematical description of how each of the six features was calculated is included in the Appendix. A list of the six image features that were ultimately used in the *c-BDI* is given in Table 1. A list of some of the candidate features that were studied but ultimately not incorporated into the Breast Density Index model is given in Table 2.

Table 1. A Brief Description of the Parameters Used to Calculate the Computer-Derived BDI (c-BDI). A Full Discussion of How These Parameters Were Calculated From the μx-Mammograms Is Given in the Appendix

Parameter Number	Parameter Abbreviation	Description	Linear Correlation Coefficient (r)*
1	FD_Th_75	Fractal Dimension with threshold = 75%	−0.7457
2	FD_Th_85	Fractal Dimension with threshold = 75%	−0.6640
3	FD_Sigma	Fractal Dimension of Standard Deviation	−0.0083
4	CD_Yint	Y intercept of Continuous Dimension	0.5715
5	CD_Slope	Slope of Continuous Dimension	0.7327
6	HZ_Proj	Standard Deviation in Horizontal Projections	0.5022

*The linear correlation coefficient was calculated for 153 images, comparing the *s-BDI* with only this parameter value.

Multiple Linear Regression Technique

A multiple linear regression algorithm using single value decomposition was developed using commercially available subroutines.[20] Using the multiple linear regression technique, given 6 features ($F_1(j), F_2(j), \ldots F_6(j)$) that were calculated for

Table 2. A Brief Description With the Linear Correlation Coefficients of Some of the More Obvious Features That Were Studied Over the Course of This Investigation

Parameter Number	Parameter Description	Linear Correlation Coefficient (r)*
1	Area of the breast	−0.4003
2	Standard Deviation	0.4586
3	Mean GS value	0.0772
4	Median GS value	−0.0045
5	Mean-Median GS value	0.3514
6	Gray Scale value at 95% of Maximum GS value	0.3909
7	Width of image (nipple to chest wall)	−0.2986
8	Height of image (top to bottom)	−0.0951
9	Fractal dimension calculated on image histogram	0.0204
10	Fractal dimension on moments calculated on histogram	0.0546
11	Coefficient of Variation (standard deviation/mean)	−0.2310
12	Total image counts	−0.0230

*The linear correlation coefficients calculated here were calculated on ⅓ of the image set.

Table 3. The Values of the Coefficients Used to Calculate the Breast Density Index ($BDI = A_0 + A_1P_1 + A_2P_2 + A_3P_3 + A_4P_4 + A_5P_5 + A_6P_6$)

Co-efficient	Parameter (P_n) Associated with	Mean[a]	Standard Deviation[b]	Representative Values[c]
A0	(constant)	245.57845	2.7743060	246.07840
A1	FD_Th_75	−28.78644	0.9250735	−28.95025
A2	FD_Th_85	−23.30803	0.8287140	−23.26487
A3	FD_Sigma	−86.79803	2.0627900	−86.80530
A4	CD_Yint	54.64385	0.7178585	54.55355
A5	CD_Slope	406.09454	4.6207106	406.39166
A6	HZ_Proj	1.52770	0.0267123	1.52822

Notes: (a) Mean of 153 jackknifed trials with 152 in each trial; (b) Standard Deviation (1 s) of 153 jackknifed trials with 152 in each trial; and (c) Specific values for the coefficients for the last of 153 jackknifed trials.

each image j, the BDI for that image was calculated using Equation 2:

$$BDI(j) = A_0 + A_1 \cdot F_1(j) + A_2 \cdot F_2(j) + A_3 \cdot F_3(j) + A_4 \cdot F_4(j) + A_5 \cdot F_5(j) + A_6 \cdot F_6(j) \quad \text{[Equation 2]}$$

For a given image j, the six features $F_1(j) - F_6(j)$ were calculated from the image using specific algorithms described in the Appendix. The values for the 7 coefficients in Equation 2, A_0, $A_1, \ldots A_6$ were solved for using the single value decomposition (SVD) multiple linear regression technique. The data needed to solve Equation 2 for the coefficients (A_N) are the 6 feature values calculated from a set of μx-mammograms *and* the corresponding *s-BDI* values [BDI(j)] for the same set of mammograms (Table 3).

There were a total of 153 cases acquired for this study, so $N_{cases} = 153$. The data set used to solve for the 7 coefficients may include as few as 7 cases (this is a constraint of the single value decomposition technique) or may include up to all the 153 cases that were compiled. However, in order to independently demonstrate the feasibility of this method, the available cases need to be divided up into a *training set* and a *testing set.* A training set is a set of a number of cases (N_{train}) that are used to solve for (hence "train") the coefficients ($A_0, A_1, \ldots A_6$) using SVD multiple linear regression. The testing set makes use of a number of cases (N_{test}) that were *not a part* of the training set. The testing set, also called the *validation set,*[21] is used to evaluate the performance of the technique independently of the cases used to find the coefficients.

There are a huge number of permutations in which 153 different cases can be distributed between the two sets, but the validity and applicability of the results are dependent on some of the finer points of the methodology. A typical approach might be to take half of the cases and assign them to the training set, and take the other half for the testing set, however there is no assurance that this is the most efficient split of the data. There are many conflicting views on the "correct" way to allocate the data between training and testing.[21-26] We have therefore attempted to be very thorough in addressing this important issue. There are two general approaches to allocating the available data to the training and testing sets, a *straightforward split approach* and a *jackknife approach.* Both techniques were used, and will be described separately below.

The Straightforward Split Approach. This approach simply splits the available cases (N_{cases}) between the training set and the testing set, such that $N_{train} + N_{test} = N_{cases} = 153$. Let us also stipulate that we keep at least 5% of the cases in either set, meaning that N_{test} or N_{train} cannot be less than 7 cases ($\sim 0.05 \times 153$). There are 140 possible choices for selecting N_{train} and N_{test}. Specifically, these choices are: ($N_{train} = 7$, $N_{test} = 146$), ($N_{train} = 8$, $N_{test} = 145$), ($N_{train} = 9$, $N_{test} = 144$), ... ($N_{train} = 146$, $N_{test} = 7$). However, there are an enormous number of possible distributions of the 153 cases amongst the training and testing sets, for each (N_{train}, N_{test}) point. In this study, all 140 possible choices for N_{train} and N_{test} were examined 1000 times *each,* where a different random distribution of cases between training and testing sets was used. A random number generator[27] was used to randomize the ordering of the cases, and then the N_{train} cases were assigned to the training set and were used to calculate the multiple linear regression coefficients, A_0-A_6. The remaining N_{test} cases were then used to compare the performance of the *c-BDI* approach with the *s-BDI.* The performance metric used in this study was the linear correlation coefficient, *r*. The different case mixes at each (N_{train}, N_{test}) point were used to quantify the mean and standard deviation in the linear correlation coefficients (*r*) at these points.

The Jackknife Approach. The jackknife approach[23,28,29] to separating N_{cases} into training and testing sets is designed to maximize the number of cases in the training set, but to still get N_{cases} independent cases for testing. With this approach, of the available N_{cases}, the first one was placed in the testing set ($N_{test} = 1$), and the remaining ($N_{cases} - 1$) cases were placed into the training set. The SVD multiple linear regression technique was used to solve for the coefficients using the ($N_{cases} - 1$) cases in the training set, and the *c-BDI* of the single case in the testing set was calculated using Equation 2 and stored. This procedure was executed again, except that the second case was placed in the testing set, and all remaining cases were used for training as before. Performance of case 2 was then calculated as above and stored. This process was repeated until each case had its turn sitting out of the training set, and being used in the testing set. The linear correlation coefficient r was then calculated on all N_{cases} independent test cases that were stored using this jackknife procedure. The average linear regression coefficient from all N_{cases} training sessions was also computed.

To evaluate the performance of the jackknife approach as a function of the number of cases used, the jackknife method was run using a number of cases ($N_{jackknife}$) ranging between 20 to 153. For each value of $N_{jackknife}$, 1000 different random samples of cases taken from the 153 total cases were made and evaluated. As $N_{jackknife}$ approached the total number of cases available (153), however, the amount of diversity in terms of case mix was reduced such that when $N_{jackknife} = 153$, all 1000 random realizations were identical and there was no diversity between the 1000 random samples ($\sigma = 0$).

Shrinkage. For both the straightforward split and the jackknife approaches discussed above, the linear correlation coefficient (*r*) was calculated for both the training set (r_{train}) and the testing set (r_{test}). In general, the value of r_{train} was higher than r_{test} because the SVD multiple linear regression procedure is designed to essentially maximize r_{train}. The test set represents new cases, not used in training, which are necessary to

independently verify the performance of the overall technique. As a result of this, r_{test} will usually fall short of r_{train}, since the coefficients were not specifically optimized for that (testing) data set. *Shrinkage* is a general term[21,22,30,31] that refers to the lower performance of the testing set, relative to the performance of the training set. If shrinkage is very small or zero, this implies that the technique was robust and the coefficients that were derived from the training set also worked well with the testing set. This further implies that the overall approach being studied generalizes well to an independent population of cases. To specifically quantify shrinkage for this study, where the linear correlation coefficient was used as the metric of performance, an equation was needed in which shrinkage is zero when $r_{test} = r_{train}$, and increases as the ratio (r_{test}/r_{train}) decreases. The equation which meets these criteria is given below:

$$\text{shrinkage}(\%) = 100 \times \left(1 - \frac{r_{test}}{r_{train}}\right) \quad \text{[Equation 3]}$$

The calculation of shrinkage was used in this study to indicate the degree to which the overall technique is able to generalize to an independent population of cases.

Other Issues

The computer used in this study was a Pentium class PC equipped with image display (DOME Imaging Systems, Waltham, MA; and an NEC 6FG Monitor), and removable WORM drives for data storage. All code was written by the authors, except for the single value decomposition and multiple regression algorithms which were commercially available as source code and were ported to our compiler. All programs were written using the C language, and a 32 bit C compiler (Intel C Code Builder, no longer available commercially). Over 400 computer programs were written specifically for this study, including programs for displaying, cropping and analyzing the images, and others for calculating, analyzing, and graphing breast features, and so on. The SVD/multiple linear regression software developed by the authors was verified for accuracy against other commercially available software capable of this analysis (SigmaStat 1.0; Jandel Scientific, San Rafael, CA). The SVD/multiple linear regression subroutines were executed well over a million times in this study, and therefore it was not feasible to utilize the commercial software directly because each run would have required user interaction. Statistical analysis was also performed using SigmaStat 1.0.

RESULTS

Radiologist Intraobserver Variability

The intraobserver variability for determining breast density is shown in Fig 2A. The breast density index for the second ranking of the CC images is plotted as a function of the *BDI* calculated from the first ranking, where both rankings were performed by a single radiologist (RAD_1). An excellent fit is illustrated ($r = 0.978$), demonstrating very reproducible performance. A histogram showing the deviation from the linear regression (best fit) line is inset in Fig 2A. In the histogram, the "*BDI* Residual" is the difference between a plotted data point and the best fit line. Breast density is an attribute that is related to the breast, and should therefore be relatively independent of the x-ray projection through the breast. The *BDI* determined from the MLO projection images is plotted as a function of the *BDI* for the first CC ranking in Fig 2B. RAD_1 performed both rankings. The correlation coefficient calculated between x-ray projections ($r = 0.960$) was only slightly less than that calculated using repeated rankings of the same projection. The very obvious correlations ($P < 0.001$) with both the CC_1/CC_2 and the MLO_1/CC_1 comparisons lends support to the notion that the BDI is relatively projection-independent. However, there is a statistically significant difference in the *precision* (reproducibility) obtained from the intraprojection (CC_1/CC_2) comparisons and the interprojection (MLO_1/CC_1) comparison ($P < .01$, F test on ratio of variances[32]).

Radiologist Interobserver Variability

The *BDIs* resulting from the rank-ordering performed by two different radiologists on the same data set (CC_1) are compared in Fig 2C. The interobserver variability is quite low, as demonstrated by a very high correlation coefficient of $r = 0.968$. This degree of correlation is only slightly lower than the $r = 0.978$ value found for intraobserver variability, suggesting that these two radiologists apparently make use of very much the same criterion in their ranking of breast density. Despite the excellent match between radiologists seen in Fig 2C, there was a statistically significant difference in precision between interradiologist classification performance and intraradiologist performance ($P \approx .01$, F test on the ratio of variances).

Figure 2C shows the comparison between two radiologists *ordinal* ranking of the images ($r = 0.968$), whereas Fig 2D shows the comparison between the two radiologist's *cardinal* scoring of the breast density of the same image set ($r = 0.913$). In the ordinal ranking the radiologists ranked all the mammograms together, while in the cardinal scoring the radiologist simply assigned a density value while looking at only one image at a time. The cardinal scoring is more akin to how mammographers currently assess breast density. There is a significant difference in the precision between ordinal ranking and cardinal scoring ($P < .01$, F test on the ratio of variances) of the breast density.

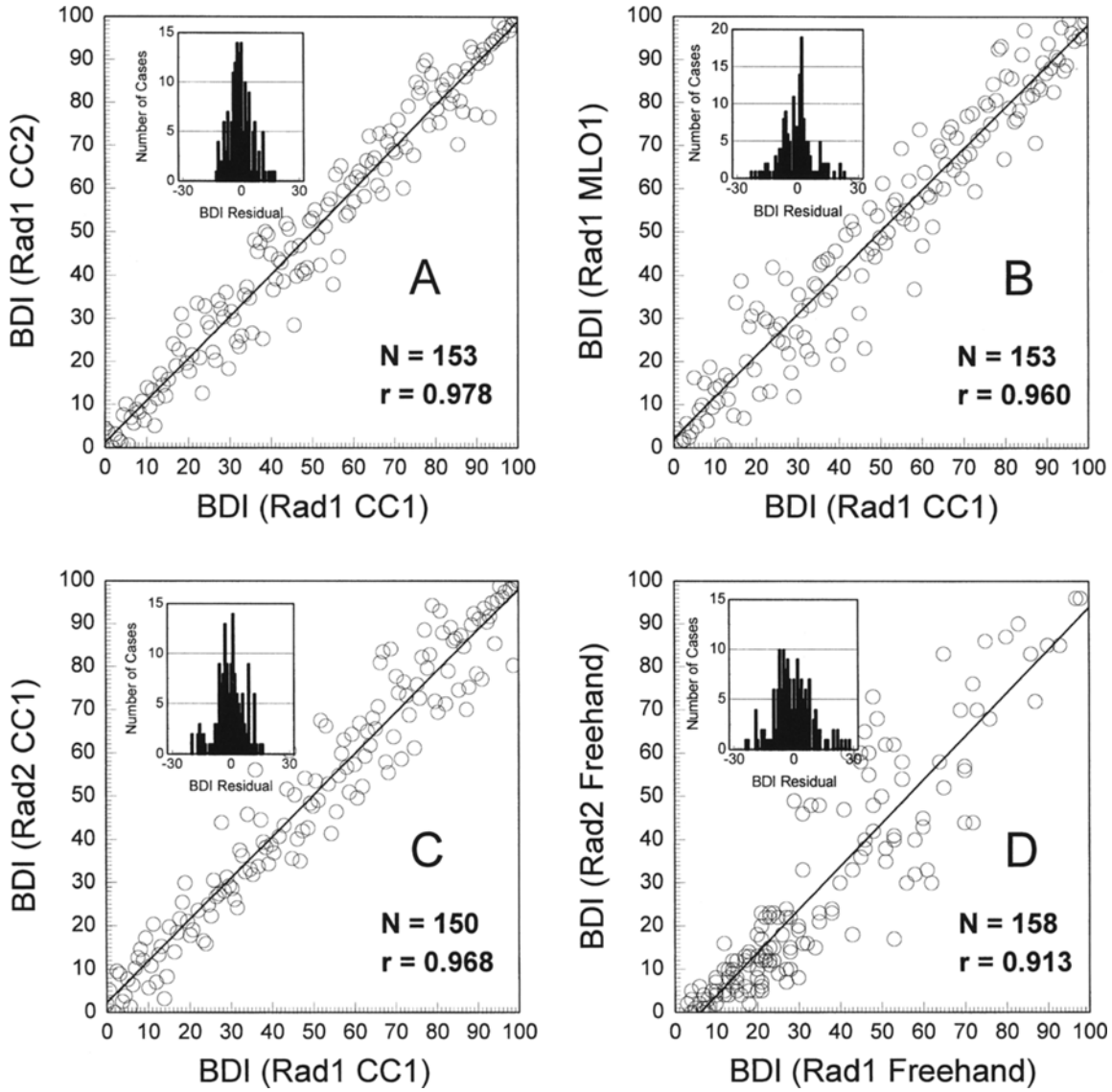

Fig 2. (A) The performance of a single radiologist, on the same set of CC mammograms, ordered in two different sessions (CC1 and CC2) four months apart. (B) The performance of a single radiologist, on CC and MLO mammograms from the same patients. The numbering scheme between the two sets of miniature mammograms was randomized to reduce bias, and the two orderings were performed 3 months apart. (C) A comparison of the *BDI* assigned by rank ordering, between two different radiologists on the same set of miniature CC mammograms is shown. The rank data were converted to a *BDI* scale using Equation 1. (D) Two different radiologists operating on the same CC data set, viewed each image independently (alone) and assigned a (cardinal) density value ranging between 0 and 100. This is similar to how radiologists assign breast density currently using the 4 classification scheme of the BIRADS, except that the scale was expanded.

Computer Determined BDI Performance

Each of the six features used in the computer-determined *BDI* (*c-BDI*) is plotted as a function of the radiologist determined gold standard (*s-BDI*) in Fig 3. It can be seen from the figure that some of the features demonstrated good correlation with the BDI standard, others showed only poor correlation. It is noted that by combining six features in a multiple regression fit, one is actually striving for some of the features to correlate with the *residuals* between the other features and the *BDI*.

The training and testing correlation values for the straightforward split analysis paradigm are shown in Fig 4A, as a function of the percentage of the 153 case data set that was used in the training set. Towards the left hand side of the plot, for

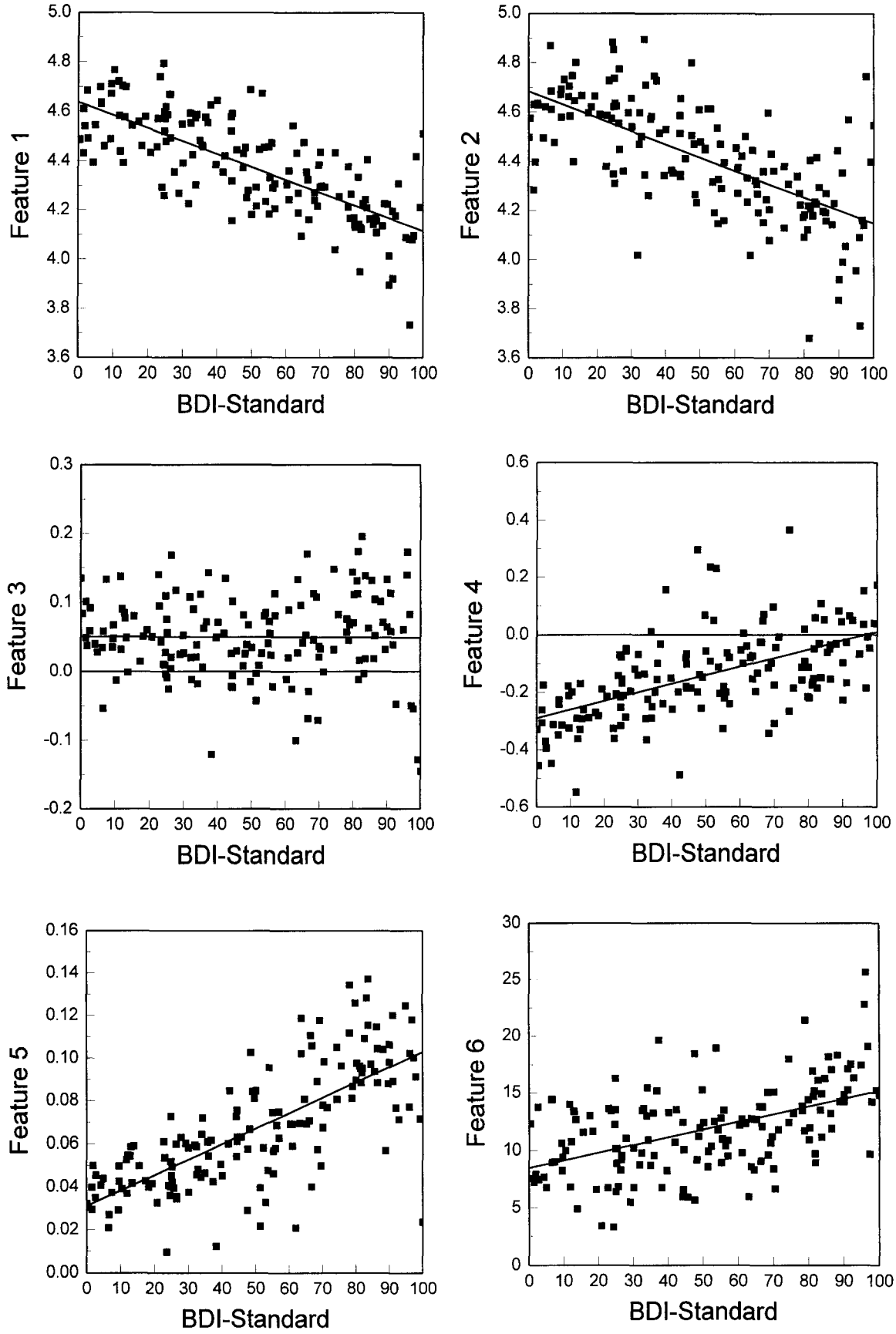

Fig 3. The six parameters are plotted as a function of the *BDI* standard. None of the individual correlations shown exceed $r = 0.74$. The parameters and their respective correlation coefficients are: Parameter 1: FD_TH_75 ($r = -0.756$), Parameter 2: FD_TH_85 ($r = -0.664$), Parameter 3: FD_Sigma ($r = -0.00827$), Parameter 4: CD_Yint ($r = 0.572$), Parameter 5: CD_Slope ($r = 0.733$), Parameter 6: HZ_Proj ($r = 0.502$).

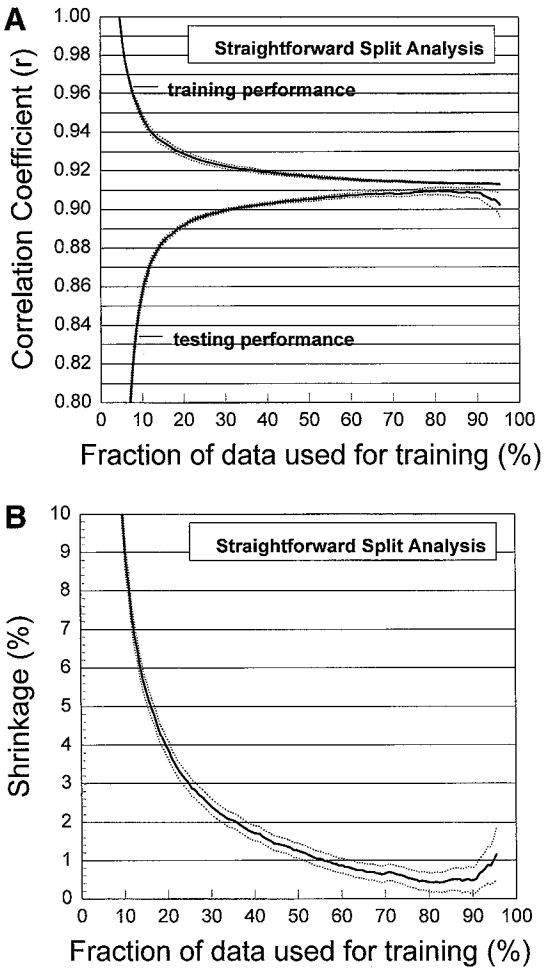

Fig 4. (A) This graph demonstrates both the training and the testing correlation coefficient (*r*) for the straightforward split paradigm, as a function of the fraction of the 153 cases used in training set. At 10%, N_{train} = 15 and N_{test} = 146, at 50%, N_{train} = 76 and N_{test} = 77, and at 90% N_{train} = 146 and N_{test} = 15. As the number of cases in the training set increases, the multiple regression fit needs to accomodate a more diverse group and r_{train} drops with increasing N_{train}. The performance of the fit parameters on the testing group increases as N_{train} increases (and N_{test} decreases), since using more cases in training usually increases the generalizability of the coefficients. The error bars shown are ± 2 σ. At each point along the abscissa, 1000 different random drawings from the 153 cases were used to assign cases to the training and testing sets at each pair of (N_{train}, N_{test}). The mean r of these 1000 samples (solid lines) with the ±2σ error bars (dotted lines) are shown for each point. (B) The shrinkage is shown for the straightforward split analysis. For these data, shrinkage is seen to be near a minimum at and abscissa value of 85%, where N_{train} = 130 and N_{test} = 23. The mean shrinkage at this distribution between the training and testing sets was 0.52%, indicating that the multiple linear regression equation was capable of robust generalization.

example at the 10% value on the abscissa, 15 images were used for the training and the remaining 138 images were used for testing. Because only a small number of cases were used in training, the multiple linear regression algorithm was able to fit the data points quite well ($r = 0.949 \pm 0.0020$). However, because the relatively few cases used in training were not representative of the wide array of variations in the testing data set, the *c-BDI* was not able to generalize well, and the testing performance at this point (abscissa = 10%) was relatively low ($r = 0.855 \pm 0.0035$). The error bars shown in this and all related figures show the 95% confidence limits (±2 σ), based on 1000 different case distributions. Looking towards the right of Fig 4A, for example where the abscissa value is 80% (N_{train} = 122, N_{test} = 31), the training correlation coefficient is lower ($r_{train} = 0.914 \pm 0.0004$) compared with r_{train} = 0.949 at the 10% point on the abscissa, because there were more points in the training set and a wider case variation was seen. However, with this relatively large number of points used in training, the *c-BDI* embodied a wider variation in data, and its ability to generalize was better as demonstrated by a higher correlation coefficient for testing ($r = 0.910 \pm 0.0019$). Towards the right of the 80% point, the number of test cases becomes too few and the occasional bad fit in the testing set is not counterbalanced by the mostly good fits, and so the testing correlation value suffers.

Shrinkage, defined previously, is a measure of how well the *c-BDI* may be expected to generalize. As the testing performance approaches the training performance, the shrinkage is reduced and the applicability of the technique to the "general" case improves. The shrinkage for the straightforward split paradigm is shown in Fig 4B, along with the ± 2 σ error bars. At the 80% point on the abscissa, shrinkage is near a minimum at 0.43%. This indicates that the 80% training-20% testing case mix may be near optimal for this experiment, and that may be an interesting methodological observation to some. More importantly, the low 0.4% shrinkage indicates that the results demonstrated for the *c-BDI* technique may be representative of a broader patient population.

A second paradigm for distributing training and testing cases is the jackknife method. The training and testing performance using the jackknife approach is illustrated in Fig 5A. In this graph, the abscissa represents the number of cases used in total for the entire training and testing procedure. At an abscissa value of 20, this means that only 20 cases were used for both training and testing. The

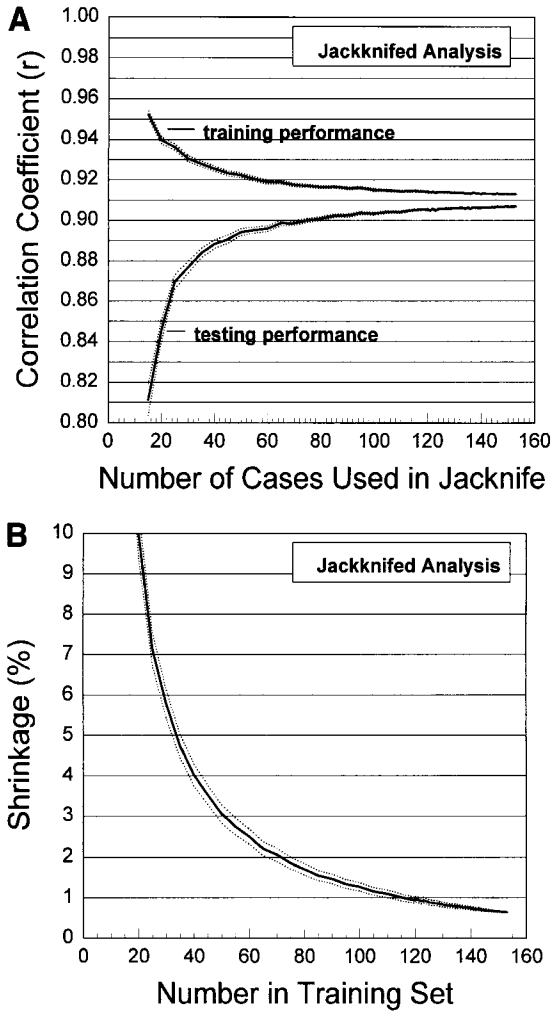

Fig 5. (A) This figure is analogous to the data shown in Fig 6A, except here the results for the jackknifed analysis paradigm are shown. Here, the number on the abscissa is equal to the total number of cases used at each point, $N_{jackknife} = N_{train} + N_{test}$. One case is set aside, and the multiple linear regression technique is run $N_{jackknife}$ times. The mean r from the $N_{jackknife}$ training sessions, and the testing performance is the correlation coefficient calculated from fitting the $N_{jackknife}$ test cases results. At each point along the abscissa corresponding to a specific value of $N_{jackknife}$, 1000 different random samplings from the 153 total cases available were used. The mean (solid line) and 95% confidence limits ($\pm 2\sigma$) were calculated from the 1000 sessions run at each point along the abscissa. As $N_{jackknife}$ approaches N_{cases} towards the right side of this graph, the actual diversity achieved in the different random samplings decreases, to the point where when $N_{jackknife} = N_{cases}$ (the right-most data point), the exact same set of 153 cases was used 1000 times. This is why the error bars approach zero towards the right of the graph. (B) The shrinkage is shown as a function of $N_{jackknife}$ in this figure, demonstrating that for the jackknifed analysis, the shrinkage is at a minimum when $N_{jackknife} = N_{cases}$. It is seen in this figure that the minimum shrinkage value is 0.6%, and that the curve appears to be approaching zero assymtoptically.

error bars were calculated by randomly varying the case mix (from the pool of 153 cases) in these 20 cases, 1000 times. The point of this analysis is to demonstrate the convergence between the training and testing performance as the number of jackknife cases increases.

Figure 5B illustrates the shrinkage for the jackknife analysis. As the number of jackknifed cases increases, the number of cases used in training also increases and the shrinkage is seen to decrease. For the case where all 153 available cases were used in jackknifed approach, the measured shrinkage was 0.65%.

Figure 6 demonstrates the c-BDI as a function of the *s-BDI* for the jackknife approach. The Pearson correlation coefficient is $r = 0.9069$, meaning that 82.2% ($100\% \times r^2$) of the variance of the *c-BDI* seen in Fig 6 is attributable to its relationship with s-BDI. This implies that 17.7% of the fluctuation seen in the figure is unaccounted for. The excellent reproducibility (ie, precision) and correlation of results between radiologists as seen in Fig 2A-2C gives substantial credibility to the consistency and the quality of their ranking. How much of the 17.7% fluctuation in Fig 6 can be attributed to radiologist imprecision? Recalling the radiologist results shown in Figs 2A-2C, the percent of variance attributable to radiologist imprecision was 4.35% ($100 \times [1 - r^2]$) in the best case (the CC_1 versus CC_2 for RAD_1), 7.8% in the worst case

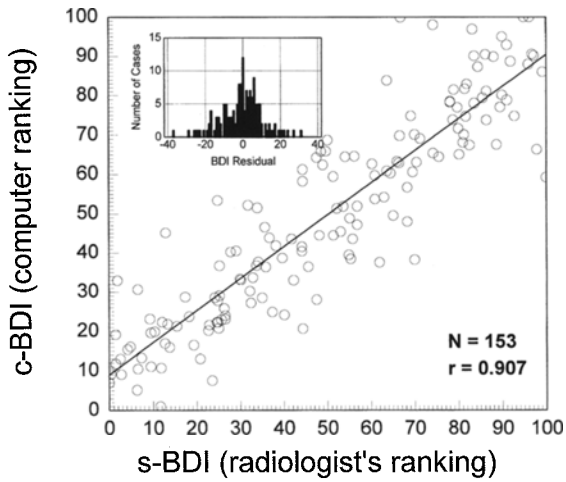

Fig 6. The *c-BDI* results are shown plotted as a function of the *s-BDI*. While the computer was not able to achieve the same level of agreement (based on the correlation coefficient) with the radiologists as the radiologists achieved, good performance is nevertheless shown. It is noted that the computer calculated *BDI* has a reproducability with $r = 1.000$.

(RAD_1, CC_1 versus MLO_1), and 6.3% in the intermediate case (RAD_1 CC_1 versus RAD_2 CC_1), for these ordinal rankings. Averaging these values, we approximate that about 6% to 7% of the fluctuations seen in Fig 6 are attributable to variation in the *s-BDI* (radiologist scoring). Therefore, the remaining 10% to 11% of the variance must lay with the *inaccuracy* of the computer to replicate the decisions made by the radiologists in determining breast density (this is not due to computer imprecision, since the *c-BDI* is utterly reproducible).

The breast density index presented in this study was chosen over a continuous scale ranging from 0 to 100, however this does not imply that it is practical or desirable to define 100 different categories for breast density. For comparison, the Wolfe grade methodology and the current American College of Radiology recommendations make use of 4 different categories of breast density. The number of categories that breast density can be meaningfully assigned into is related to the precision that one can actually determine breast density. To evaluate this, an analysis was performed whereby two classification schemes were compared to see what fraction of the cases would be assigned to the same classification category. The number of categories was varied between 2 and 50 in the analysis. Figure 7A shows the percentage of cases that were assigned to the exact same category, plotted as a function of the number of categories used for assignment. For example, for 2 categories the assignment is either "dense" or "not dense," and the percentage of cases that were correctly tabulated into these categories is relatively high. As the number of categories increases, the number of divisions between categories increases, and a smaller percentage of cases end up being classified into the same category. The open circles compare intra-radiologist (comparing the CC_1 and CC_2 sessions of Rad_1) ordinal classification performance. The crosses show inter-radiologist ordinal performance (Rad_2 CC_1 versus *s-BDI*, which was an average of Rad_1 CC_1, CC_2 and MLO_1 sessions). The filled squares show the comparison between the cardinal scores, the computer *c-BDI* and the radiologist's *s-BDI*. The open diamonds show the comparison between the freehand BDI assignment of Rad_1 (CC_3) and the freehand assignment of Rad_2 (CC_3). In Figure 7A, only cases that were assigned into exactly the same category were tallied as "correct," whereas in Fig 7B cases that were assigned into the exact same or the next adjacent categories (on either side) were counted as "correct." This relaxation of the definition of "correct" improves performance as is apparent in Fig 7B. The definition of "correct" is relaxed to include the surrounding 2 categories in Fig 7C, and categorization performance improves even further.

DISCUSSION

There have been previous efforts to develop numerical estimates of breast density reported in the literature, most notably by Wolfe.[1] The BI-RADs classification schema has been adopted by the American College of Radiology as a standard for breast density characterization. Other investigators have reported using computerized techniques employing planimetry[33,34] and computer-derived image features.[28,35]

In one reported study,[33] radiologists ranked mammograms into 6 discrete categories depending on their estimate of the "proportion of breast volume occupied by the radiological signs of 'ductal prominence' or "mammographic dysplasia."[33] A planimeter was used which required human input (about one minute per film), and essentially used the computer to calculate the fractional area of the breast which was dense, based on hand traced areas of the dense breast regions and the total breast area. The focus of that study was primarily to evaluate the reproducibility of human estimates versus human-planimeter estimates of breast density. As such, the computer was not used to identify mammographic features per se, but only to integrate the radiologist-traced areas. For the 6 category scale used in the study, the investigators found 52.4% exact agreement between radiologist and planimeter estimates of densities. In the study of Saftlas et al,[34] planimetry was used essentially as above and showed 77% agreement based on a 5 category scale of density.

Caldwell et al[28] pioneered the use of the fractal dimension as a feature which correlates well to breast density, as defined by the 4-category Wolfe grade classification scheme (N1, P1, P2, and DY). In terms of categorization reliability, inter-radiologist agreement (3 radiologists compared) in the 4 category scale ranged between 66% and 74% for exact agreement. The computerized assignments to the density categories agreed exactly with the radiologists between 57% and 67%. The computer assignments demonstrated had minor disagreement

(plus exact agreement) in 88% of the 70 cases studied.

The technique of rank-ordering of mammograms used in this study is fundamentally different than assigning breast density using a small number of categories. The rank order data can be retrospectively divided up into a large number of different categories. Rebinning our results to the 4 breast density classification categories used by Caldwell,[28] the computer and radiologist agreed (exactly) 67% of the time, in excellent agreement with Caldwell's 57% to 67%. The intraradiologist exact agreement for 4 categories observed in this study was calculated (circles on Fig 8A) as 84%, which compares well with Caldwell's interradiologist agreement levels of 66% to 74%. The computer-derived *c-BDI* compared with the gold standard radiologist *s-BDI* scoring agreed in 99% of the cases when minor disagreements are disregarded (ie, when assignment in just-adjacent categories are considered as "agreement"), compared to Caldwell's 88% agreement. This improvement of the *c-BDI* technique over Caldwell's results may be explained by the fact that the *c-BDI* technique presented in this study used 6 features derived from the μx-mammograms, as opposed to 2 features used by Caldwell et al. Furthermore, while Caldwell et al briefly explored the role that the screen-film characteristic curve played on their computer derived classification parameters, their results quoted above did not include corrections for the characteristic curve or exposure levels that our results include. This may be another factor contributing to the slightly better

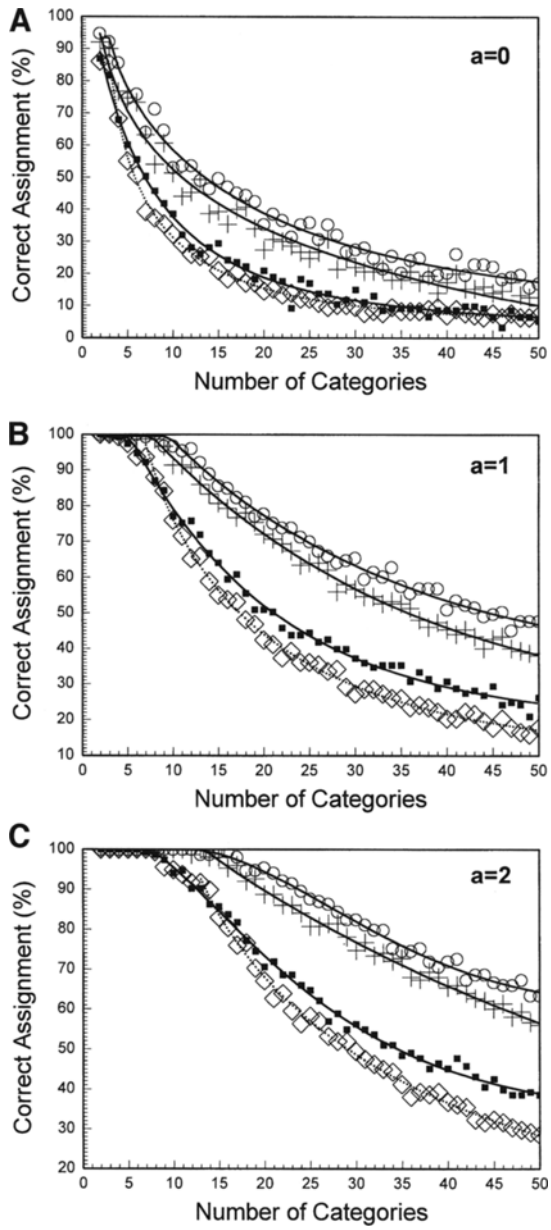

Fig 7. (A) The precision with which breast images can be ranked in order of breast density relates to the number of meaningful categories that breasts can be classified into. In comparing the ranking between two approaches, this graph shows the percentage of assignments into the same categories between the two approaches, as a function of the number of categories. The open circles show the comparison between ordinal ranking sessions RAD_1 CC_1 and RAD_1 CC_2 (intraradiologist). The crosses show comparisons of the ordinal rankings between radiologists (Rad_1's *s-BDI* and Rad_2 CC_1). The solid squares are the (cardinal) *c-BDI* results compared with the cardinalized *s-BDI*, and the open diamonds show the (cardinal) freehand sessions (CC_3) between radiologists (Rad_1 versus Rad_2). (B) This plot is the same as Fig 7A, except that the definition of correct is relaxed to include both the exact and the next adjacent category assignments. This relaxation of the agreement increases the meaningful number of breast density categories that could be used. The open circles show the comparison between ordinal ranking sessions RAD_1 CC_1 and RAD_1 CC_2 (intraradiologist). The crosses show comparisons of the ordinal rankings between radiologists (Rad_1's *s-BDI* and Rad_2 CC_1). The solid squares are the (cardinal) *c-BDI* results compared with the cardinalized *s-BDI*, and the open diamonds show the (cardinal) freehand sessions (CC_3) between radiologists (Rad_1 versus Rad_2). (C) In this plot, mammograms that were reproducably assigned within ± 2 categories were counted as correct. This easing of the constraint further increases the number of meaningful categories that mammograms could be assigned to in terms of breast density. The open circles show the comparison between ordinal ranking sessions RAD_1 CC_1 and RAD_1 CC_2 (intraradiologist). The crosses show comparisons of the ordinal rankings between radiologists (Rad_1's *s-BDI* and Rad_2 CC_1). The solid squares are the (cardinal) *c-BDI* results compared with the cardinalized *s-BDI*, and the open diamonds show the (cardinal) freehand sessions (CC_3) between radiologists (Rad_1 versus Rad_2).

performance of the *c-BDI* in the minor disagreement category.

Figures 7A-7C demonstrate that rank-ordering a large series of mammograms results in greater precision than that achievable by assigning cardinal *BDI* values. For example, focusing on Figure 7B (where minor disagreement by plus or minus one category is considered "agreement") using 90% correct assignment as a threshold requirement, intraradiologist ordinal ranking could achieve 13 meaningful categories, and interradiologist ordinal ranking could achieve 11 meaningful categories of breast density. Cardinal ranking techniques proved less precise. For the computer results (*c-BDI*) compared against the *s-BDI*, it was found that 7 meaningful breast categories could be distinguished. Interradiologist cardinal scoring could also produce about 7 meaningful categories. These results suggest that radiologist organizations considering future modifications to the BIRADs breast density scale may want to consider increasing the number of categories from 4 to 7. It is conceivable that if radiologists were to make use of an atlas showing a series of mammograms covering the full range of breast densities (thus using an ordinal scale), a higher level of precision may be achievable. This technique would be similar to the use of the Greulich and Pyle atlas for skeletal age determination.[36] Alternately, when digital mammography becomes the norm, algorithms such as that reported for the *c-BDI* here could be used to calculate breast density.

CONCLUSION

The link between breast density and the risk of breast cancer that was first made by Wolfe[1,2,16-18,37] has begun to be appreciated by the medical community as a whole.[3,34,38-45] Quantification of a patient's breast density using the *BDI* developed in this study would allow for more precise evaluation of breast cancer risk, which may influence the optimal choice of screening strategy for each patient. For example, patients with moderate breast densities might be evaluated more frequently using mammography, but those with very high *BDIs* might be screened routinely using modalities in addition to mammography, such as ultrasound or MRI.[6,7,46-48] Furthermore, the efficacy of using alternate modalities such as ultrasound or MRI could be studied in terms of the proposed continuous scale for breast density. A more precise metric for quantifying breast density would also allow closer monitoring of changes in breast density due to menopause or hormone replacement therapy.[49-52] In addition, a continuous *BDI* scale may permit better technique optimization for serial mammography screening, even with automatically adjusting mammography systems such as the DMR (General Electric; Milwaukee, WI). A priori knowledge of a precise BDI value would allow an automatic technique system to initiate the technique closer to the optimum level, possibly minimizing exposure total time and reducing motion unsharpness in the mammogram. Finally, the availability of a standardized breast density index such as that proposed here may permit the a priori application of different sets of algorithms for computer aided diagnosis, each set optimized for a specific range of breast density.

ACKNOWLEDGMENT

We would like to thank Ms. Dorene Bishop for her efforts in digitizing the mammograms used in this study.

APPENDIX

In this section, a description is given as to how each of the features listed in Table 1 was calculated. In all cases, the features were only calculated on regions of the μx-mammograms where actual breast parenchyma was imaged; the image background was excluded from the calculation using a "mask." The image served as its own mask, since all background and cropped areas on the images were set to a gray scale value of 0, and all the areas on the image where breast parenchyma was present the gray scale values were > 500.

Feature 1: FD_Th_75

The fractal dimension has been recognized for some time to be a good indicator of breast density.[28] The first step in calculating fractal properties was to produce a series of N_k images with formats decreasing by factors of 2. For example, if the original μx-mammogram was N_x pixels wide by N_y pixels tall, for k = 1 the image is still $N_x \times N_y$, for k = 2 the image is reduced to $N_x/2 \times N_y/2$ in size, for k = 3, the image size is $N_x/4 \times N_y/4$, and for k = 4, the image size is $N_x/8 \times N_y/8$. The images are reduced to smaller formats by averaging gray scale values.

The next step that was applied was to *trinarize* the μx-mammograms. Background pixels in the image were kept zero, pixels that were in the breast but were below a certain gray scale value were set to 1, and pixels in the breast above the threshold value were set to 2. The threshold value was calculated based on a percentile of the range of gray scale in the μx-mammogram. The histogram of the image was calculated, and the gray scale value corresponding to the 75th percentile was chosen as the threshold value for feature 1, FD_Th_75 (for comparison, the median gray scale would correspond to the 50th percentile).

The next step in calculating the fractal features is to calculate the feature of interest, and here the integrated gradient was

calculated using:

$$\text{Gradient}_k = \sum_x \sum_y \sum_{x'=x-1}^{x'=x+1} \sum_{y'=y-1}^{y'=y+1} [IM(x,y) - IM(x',y')] \quad [A\text{-}1]$$

All pixels having gray scale values of zero were excluded from the above summation. This operation was performed on 4 images ($N_k = 4$). For images k = 1, 2, 3 and 4 (which were increasingly smaller), pairs of (x, y) values were calculated as: ($LOG_{10}(½ k)$, $LOG_{10}(\text{gradient}_k)$). This set of 4 (x, y) points was then fit to a straight line using linear regression, and the value of the feature was calculated as: FD_Th_75 = 2 − slope. Feature 1 is referred to as the fractal dimension of the image thresholded at 75%, abbreviated as FD_Th__75.

Feature 2: FD_Th_85

Feature 2 was calculated exactly as described above for feature 1, except that the image was thresholded at the 85% level instead of the 75% level.

Feature 3: FD_Sigma

Feature 3 was calculated exactly as described above for feature 1, except that the root mean square (RMS) standard deviation was used as the feature calculated. In this case the image was not trinarized. Linear regression was performed as described above, and the slope of the straight line fit was determined. The feature value was calculated again as F_3 = (2-slope).

Feature 4: CD_Yint

Each image was high-pass filtered with a series of 5 different filters, producing 5 different filtered images. The high-pass filtering was performed using so-called blurred mask subtraction, where a square convolution kernel (all elements of the kernel equal to S^{-2}) of S × S pixels was convolved with the original image, smoothing it. The smoothed image was then subtracted pixel-by-pixel from the original, and an offset of 2000 was added to the image. For the five different images (k = 1, 2, 3, 4, 5), the side length S of the convolution kernel was 5, 9, 13, 17, 21 (ie, S = 4k + 1).

The integrated gradient for each high-pass filtered image was calculated using Equation A-1. Pairs of points (LOG[k], LOG-[gradient_k]) were produced from k = 1 to k = 5 and these 5 pairs of points were submitted to linear regression analysis. The fit was to a straight line, $Y = \alpha + \beta X$. The fit parameter α is the y-intercept of the line, and this continuous dimension y-intercept was used as feature 5, CD_Yint.

Feature 5: CD_Slope

Feature 5 was calculated as described for Feature 4, except the slope (the β in $Y = \alpha + \beta X$) of the linear fit was used instead of the y-intercept. This feature was the continuous dimension slope, CD_Slope.

Feature 6: HZ_Proj

All of the images in the data set were oriented and displayed with the nipple-to-chest wall axis running horizontal, with the nipple on the left. The gray scale values on the μx-mammograms along horizontal lines in the image (or pixel *rows,* running in the x dimension) were summed, producing a profile (or vector) Z(i = 1, N_y) which has as many elements as the image is tall (N_y). Only rows containing more than 10 non-zero pixels (those with breast parenchyma) were summed, all others were set to 0. The Root Mean Square (RMS) standard deviation of all non-zero projection values was calculated, and used for the horizontal projection feature, HZ_Proj.

REFERENCES

1. Wolfe JN: Breast patterns as an index of risk for developing breast cancer. AJR Am J Roentgenol 126:1130-1137, 1976
2. Wolfe JN: Risk for breast cancer development determined by mammographic parenchymal pattern. Cancer 37:2486-2492, 1976
3. 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 87:670-675, 1995
4. Feig SA, Yaffe MJ: Digital mammography, computer-aided diagnosis, and telemammography. Radiol Clin North Am 33:1205-1230, 1995
5. Yaffe MJ: Direct digital mammography using a scanned-slot CCD imaging system. Med Prog Technol 19:13-21, 1993
6. Feig SA: The role of ultrasound in a breast imaging center. Semin Ultrasound CT MR 10:90-105, 1989
7. Guyer PB, Dewbury KC: Ultrasound of the breast in the symptomatic and X-ray dense breast. Clin Radiol 36:69-76, 1985
8. Graham SJ, Bronskill MJ, Byng JW, et al: Quantitative correlation of breast tissue parameters using magnetic resonance and X-ray mammography. Br J Cancer 73:162-168, 1996
9. Dash N, Lupetin AR, Daffner RH, et al: Magnetic resonance imaging in the diagnosis of breast disease. AJR Am J Roentgenol 146:119-125, 1986
10. Turner DA, Alcorn FS, and Adler YT: Nuclear magnetic resonance in the diagnosis of breast cancer. Radiol Clin North Am 26:673-687, 1988
11. Jansson T, Westlin JE, Ahlstrom H, et al: Positron emission tomography studies in patients with locally advanced and/or metastatic breast cancer: A method for early therapy evaluation? J Clin Oncol 13:1470-1477, 1995
12. Wahl RL, Zasadny K, Helvie M, et al: Metabolic monitoring of breast cancer chemohormonotherapy using positron emission tomography: Initial evaluation. J Clin Oncol 11:2101-2111, 1993
13. Rambaldi PF, Mansi L, Procaccini E, et al: Breast cancer detection with Tc-99m tetrofosmin. Clin Nucl Med 20:703-705, 1995
14. Takahashi T, Moriya E, Miyamoto Y, et al: The usefulness of 201T1C1 scintigraphy for the diagnosis of breast tumor. Nippon Igaku Hoshasen Gakkai Zasshi 54:644-649, 1994
15. Wolfe JN: Developments in mammography. Am J Obstet Gynecol 124:312-323, 1976
16. Wolfe JN, Albert S, Belle S, et al: Breast parenchymal patterns: Analysis of 332 incident breast carcinomas. AJR Am J Roentgenol 138:113-118, 1982
17. Wolfe JN, Albert S, Belle S, et al: Breast parenchymal patterns and their relationship to risk for having or developing carcinoma. Radiol Clin North Am 21:127-136, 1983

18. Wolfe JN, Saftlas AF, Salane M: Mammographic parenchymal patterns and quantitative evaluation of mammographic densities: a case-control study. AJR Am J Roentgenol 148:1087-1092, 1987

19. Rockette HE, Gur D, Metz CE: The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques. Invest Radiol 27:169-172, 1992

20. Press WH, Flannery BP, Teukolsky SA, et al: Numerical Recipes in C: The Art of Scientific Computing. Cambridge, Cambridge University Press, 1988

21. Astion ML, Wilding P: The application of backpropagation neural networks to problems in pathology and laboratory medicine. Arch Pathol Lab Med 116:995-1001, 1992

22. Gurney JW: Neural networks at the crossroads: Caution ahead. Radiology 193:27-28, 1994

23. Wu Y, Giger ML, Doi K, et al: Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer. Radiology 187:81-87, 1993

24. Floyd CE, Tourassi GD: An artificial neural network for lesion detection on single-photon emission computed tomographic images. Invest Radiol 27:667-672, 1992

25. Asada N, Doi K, MacMahon H, et al: Potential usefulness of an artificial neural network for differential diagnosis of interstitial lung diseases: Pilot study. Radiology 177:857-860, 1990

26. Tourassi GD, Floyd CE, Sostman HD, et al: Artificial neural network for diagnosis of acute pulmonary embolism: Effect of case and observer selection. Radiology 194:889-893, 1995

27. Morin RL: Monte Carlo Simulation in the Radiological Sciences. Boca Raton, FL, CRC Press, 1988

28. Caldwell CB, Stapleton SJ, Holdsworth DW, et al: Characterisation of mammographic parenchymal pattern by fractal dimension. Phys Med Biol 35:235-247, 1990

29. Dorfman DD, Berbaum KS, Metz CE: Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. Invest Radiol 27:723-731, 1992

30. Boone JM: Neural networks at the crossroads. Radiology 189:357-359, 1993

31. Boone JM: Sidetracked at the crossroads. Radiology 193:28-30, 1994

32. Glantz SA: Primer of Biostatistics (3rd ed). New York, NY, McGraw Hill, 1992

33. Lee-Han H, Cooke G, Boyd NF: Quantitative evaluation of mammographic densities: A comparison of methods of assessment. Eur J Cancer Prev 4:285-292, 1995

34. Saftlas AF, Hoover RN, Brinton LA, et al: Mammographic densities and risk of breast cancer. Cancer 67:2833-2838, 1991

35. Taylor P, Hajnal S, Dilhuydy MH, et al: Measuring image texture to separate "difficult" from "easy" mammograms. Br J Radiol 67:456-463, 1994

36. Greulich WW, Pyle SI: Radiographic atlas of the skeletal development of the hand and wrist (2nd ed). Stanford, CA, Stanford University Press, 1959

37. Wolfe JN: Risk of developing breast cancer determined by mammography. Prog Clin Biol Res 12:223-238, 1977

38. Boyd NF, Jensen HM, Cooke G, et al: Relationship between mammographic and histological risk factors for breast cancer. J Natl Cancer Inst 84:1170-1179, 1992

39. Byrne C, Schairer C, Wolfe J, et al: Mammographic features and breast cancer risk: Effects with time, age, and menopause status. J Natl Cancer Inst 87:1622-1629, 1995

40. Ciatto S, Zappa M: A prospective study of the value of mammographic patterns as indicators of breast cancer risk in a screening experience. Eur J Radiol 17:122-125, 1993

41. Beisang AA, Geise RA, Ersek RA: Radiolucent prosthetic gel. Plast Reconstr Surg 87:885-892, 1991

42. Kato I, Beinart C, Bleich A, et al: A nested case-control study of mammographic patterns, breast volume. Cancer Causes and Control 6:431-438, 1995

43. Brisson J, Morrison AS, Khalid N: Mammographic parenchymal features and breast cancer in the breast cancer detection demonstration project. J Natl Cancer Inst 80:1534-1540, 1988

44. Saftlas AF, Wolfe JN, Hoover RN, et al: Mammographic parenchymal patterns as indicators of breast cancer risk. Am J Epidemiol 129:518-526, 1989

45. Vogel VG: High-risk populations as targets for breast cancer prevention trials. Prev Med 20:86-100, 1991

46. Feig SA: Breast masses. Mammographic and sonographic evaluation. Radiol Clin North Am 30:67-92, 1992

47. Fornage BD, Toubas O, Morel M: Clinical, mammographic, and sonographic determination of preoperative breast cancer size. Cancer 60:765-771, 1987

48. Liem SJ: Target ultrasonic mammography. An additional diagnostic tool for the detection of breast cancer. Diagn Imaging Clin Med 54:192-201, 1985

49. Stomper PC, van Voorhis BJ, Ravnikar VA, et al: Mammographic changes associated with postmenopausal hormone replacement therapy: A longitudinal study. Radiology 174:487-490, 1990

50. van Gils CH, Otten JD, Verbeek AL, et al: Short communication: breast parenchymal patterns and their changes with age. Br J Radiol 68:1133-1135, 1995

51. Feig SA: Hormonal reduction of mammographic densities: Potential effects on breast cancer risk and performance of diagnostic and screening mammography. J Natl Cancer Inst 86:408-409, 1994

52. Flook D, Gilhome RW, Harman J, et al: Changes in Wolfe mammographic patterns with aging. Br J Radiol 60:455-456, 1987