Accuracy of Mammographic Breast Density Analysis: Results of Formal Operator Training[1]

**Sven Prevrhal, John A Shepherd,[2]
Rebecca Smith-Bindman, Steven R. Cummings,
and Karla Kerlikowske**

Departments of Radiology [S. P., J. A. S., R. S-B.], Medicine [S. R. C., K. K.],
Epidemiology and Biostatistics [S. R. C., K. K.], and General Internal
Medicine Section, Department of Veterans Affairs [K. K.], University of
California, San Francisco, California 94143-1250

Abstract

Mammographic density is a major risk factor for breast cancer. Breast density is not routinely quantified for research studies because present methods are time intensive and manual, and require expert training. We investigated whether individuals with and without backgrounds in radiology or medicine can achieve sufficient accuracy when compared with an expert (gold standard) reader of mammographic breast density. Nine readers (three radiologists, two non-radiology physicians, and four nonphysicians) assessed breast density on 144 digitized films (60 contralateral films of breast cancer cases and 84 controls) on a computer workstation with custom software. Readings were compared with a radiologist with training in mammography and density reading that read the same films. A correlation of $r = 0.9$ or higher with the gold standard reading was met by three of three radiologists, one of two nonradiology physicians, and one of four nonphysicians. Intrareader reproducibility measured as the residual sum of mean errors averaged 10% mammographic density for all readers and 9% for readers with a correlation of 0.9 or higher with the gold standard. The odds ratios associated with breast cancer when films with mammographic breast density of 50% or greater are considered "dense" ranged from 3.1 to 3.9 or a 1.9–2.4-per-population-SD increase in percentage density. Although it is advantageous to have a radiological background when quantifying mammographic density, it is not a prerequisite. Applying strict validation criteria to qualify readers to quantify mammographic breast density for research studies will enhance the chance of accurately assessing breast density and discriminating women at high and low risk of breast cancer.

Introduction

The mammographic appearance of the breast depends on its composition of radiolucent fatty tissue and more radiopaque epithelial or stromal tissue. Different radiological grading schemes like the density/pattern classification by Wolfe (1), or relative dense area assessment, show a strong association with subsequent development of breast cancer. In fact, breast density is perhaps the strongest but least recognized risk factor for breast cancer. Many studies have shown that women whose breast X-rays are composed of at least 50% "dense" area have a three to five times greater risk of breast cancer than women with less than 25% dense area (1–4).

Breast density can be crudely graded using a subjective scale that takes into account the quantitative (amount of density) and qualitative nature of the density (diffuse or associated with ductal structures; Refs. 1, 5, 6). Qualitative methods have a limited number of density categories and can detect only very large changes in density. Because of their subjectivity, qualitative methods have substantial intra- and interobserver variation (7). A more quantitative approach has been used to measure the area of dense breast as a proportion of the total projected breast area, or "mammographic density" (1, 8, 9). Mammographic density is expressed as PD[3] defined as PD = (radiographic dense area)/(total breast area) on a scale from 0 to 100% (10).

Mammographic density is not routinely quantified for research studies because current methods are time intensive, manual, and require expert training. On the other hand, a quantitative measurement appears to be superior to qualitative categorical methods such as Wolfe (5) and the American College of Radiology Breast Imaging and Reporting Data System (BI-RADS; Ref. 11). A recent tamoxifen trial measured breast density as a surrogate end point for breast cancer risk and found that the most significant annual changes in breast density were observed with a quantitative measurement (12).

Although, breast density has been shown to be a power indicator of cancer risk, there is no generalized method for training and validating individuals to perform the measure. We also ask the question: does it require a specialist in mammography to delineate the dense regions in the mammogram? If not, this may make the technique more clinically available by increasing the number of available trainees. In this study, we attempt to train people with a formal education in radiology,

[3] The abbreviations used are: PD, percentage (breast) density; RMSE, root mean square error; RAD (group), radiologists (with limited background in mammography); MD (group), physicians (with no background in radiology); NMD (group), nonphysicians.

other fields of medicine, and those with a nonmedical background to quantify mammographic density using a predefined training program. A secondary goal of the study was to demonstrate the association of quantitative PD measurement with the risk of breast cancer.

Materials and Methods

Subjects. For 161 women ages 40 and older (64 with invasive breast cancer or ductal carcinoma *in situ* and 97 without breast cancer) who underwent screening mammography between April, 1985, and December, 1995, as part of the University of California-San Francisco (UCSF) Mobile Mammography Screening Program, we obtained one craniocaudal view of the right or left breast. For the 64 women that had been diagnosed with cancer, we selected the contralateral craniocaudal image without breast cancer taken at the time of diagnosis or a contralateral craniocaudal image from a screening examination that preceded the diagnosis of breast cancer. Women with bilateral breast cancer were not included in the study. Women without invasive breast cancer or ductal carcinoma *in situ* were randomly selected from the UCSF Mobile Mammography Screening Program database according to calendar year of the screening examination and age of the breast cancer cases. Institutional review board approval and patient informed consent were obtained.

Measurements. All of the films were initially assessed by a radiologist with training in mammography and density reading (gold standard, R.S-B.). Films were viewed directly using a standard radiology light box. A wax pencil was used to outline the breast area and breast densities. Films, with the wax pencil marks, were digitized on a Lumisys LumiScan 200 radiographic film digitizer (Kodak, Inc.) at a resolution of 200×200 μm^2 and the PD was determined by measuring the total area of the breast and number of pixels outlined in the dense regions using dedicated computer software. The software is based on the commercially available medical image-processing package MEDx (Version 3.31, Sensor Systems, Sterling, VA). Extensions to this package were written in the open-source scripting language Tcl/Tk (Version 8.3, www.tcltk.com).

Using the gold standard PD measurement, films were stratified into deciles of PD. For every density decile, 10 noncancer and 10 cancer films, when available, were selected to be included in a validation data set, resulting in 60 cancer and 84 noncancer images with PD ranging from 0 to 100%.

The wax pencil marks were erased from the validation set and films were redigitized without marks and patient identifiers. The digitized film files were transferred to CD-ROM for review by study readers.

The reading station program randomized the order of all films and consecutively displayed them with a default brightness/contrast setting on a high-resolution radiographic monitor. The reader was prompted to manually trace the breast contour using a polygonal drawing tool (clicking with the computer mouse inserts a polygon vertex at the cursor). After confirming the breast contour outline was correct, the reader proceeded to outline the dense areas of the breast using a "pencil" tool (the mouse cursor acts like the tip of a pencil). The number of dense regions was not limited and included zero for breasts that appeared to have no dense regions at all. PD was calculated as the ratio of the sum of all dense regions (overlaps are not counted twice) to the entire breast area. PD, and all drawn contours, were then stored in a database linked to the reader's study identification number, film identification number, and reading session number. Fig. 1 shows a screenshot of the main program window during an analysis session.

Standardized Training and Study Readers. We enrolled three radiologists with limited background in mammography (group RAD), two physicians with no background in radiology (group MD), and four nonphysicians (group NMD) with various backgrounds (one medical physicist, one research assistant with no experience in radiology, and two research assistants whose radiological experience was limited to bone densitometry.)

All of the readers were trained (by R.S-B.) in an hour-long training session in front of a light box (see Fig. 2). Mammography examinations ranging from very-low to very-high PD, and uniform to very-structured appearance were presented and discussed. After that, the readers were trained on the PD reading workstation (see Fig. 1). The readers could take as long as they wanted to read each film.

We selected a linear Pearson's correlation of $r < 0.9$ or lower between the gold standard and study readers as the criterion for additional training. If the study reader did not achieve an $r \geq 0.9$ after the first reading of the 144 selected films, the 10 cases with the biggest absolute difference from the gold standard were identified. The study readers compared their breast and dense tissue outlines with the gold standard outlines for those 10 cases and tried to identify any patterns that could account for the deviations from the gold standard to improve future readings. Scatter and Bland-Altman plots of study reader PD *versus* gold standard PD were also provided, and regression results were discussed with the study reader. After additional training, the study reader read the same 144 films a second time, again blinded to film identity. If a study reader's correlation with the gold standard was not ≥ 0.9, the study reader was not considered a certified reader for future studies.

Statistical Analysis. Statistical analysis was programmed in SAS 8e (The SAS Institute, Cary, NC). Intrareader reproducibility was performed with ANOVA. Interreader agreement was calculated using linear regression and was expressed as Pearson product-moment correlations r. It is conceivable that reader reproducibility is dependent on breast density: "Extreme" cases might be easier to recognize and analyze than films in the midrange of PD values. Therefore, we categorized the average reading results (averaged over the two read passes) into four quarters (PD <25%; 25% $\leq$ PD < 50%; 50% $\leq$ PD < 75%; and PD $\geq$ 75%), and we calculated reproducibility separately for each of these four PD categories. For the same reasons as for intrareader reproducibility, we repeated the interreader regression analysis between readers and gold standard by PD category. Overall agreement between readers with similar background was tested with two-factor ANOVA, calculated separately for the reader groups RAD, MD, and NMD.

Odds ratios to determine the association between breast density and breast cancer status were calculated in two ways. First, a fixed threshold of PD = 50% was chosen to discriminate between cancer and noncancer cases, and the odds ratio was calculated from the resulting contingency table. Second, to avoid bias introduced by the arbitrary threshold, we executed unconditional logistic regression analysis with PD as factor and cancer status as outcome. To arrive at meaningful SDs, we calibrated reading results to the gold standard as:

$$PD_{calibrated} = PD * slope + intercept$$

with intercept and slope being regression of the particular reader to gold standard for the logistic regression. This is a

Fig. 1. Screenshot of mammographic breast density analysis program (outlines enhanced for this article).
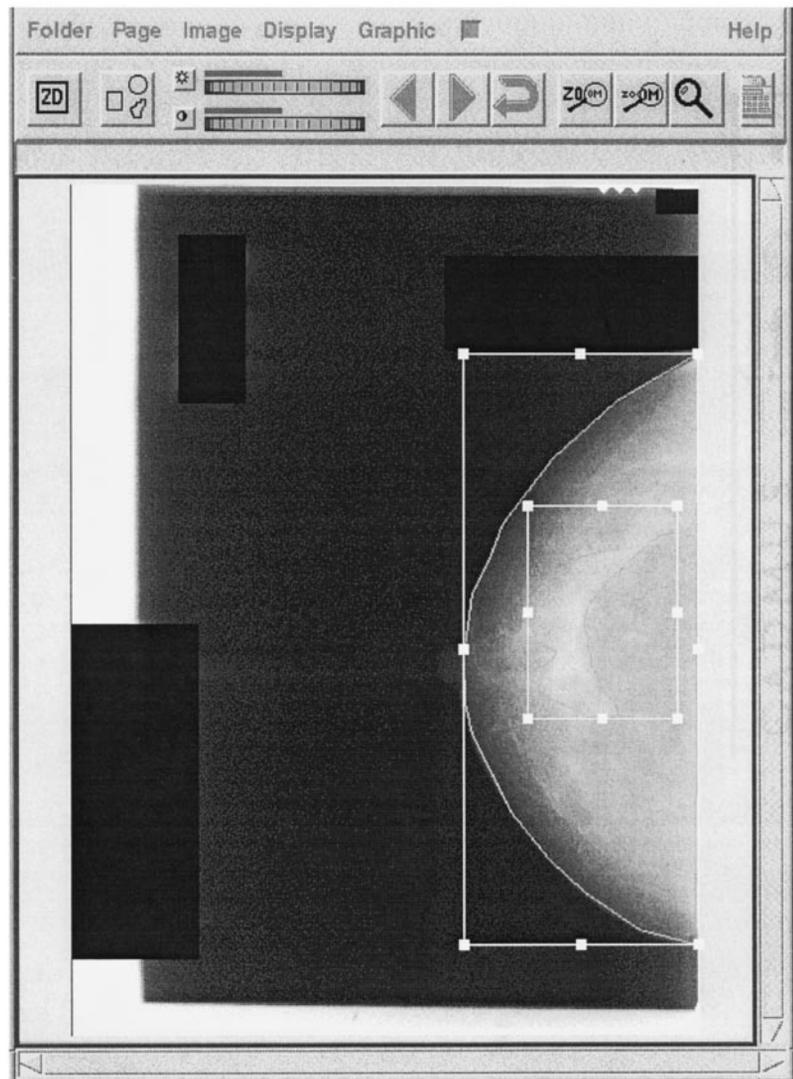
Results

Three of three radiologists, one of two nonradiology physicians, and one of four nonphysicians had correlations ≥0.9 compared with the gold standard. Table 1 shows the summary statistics, correlation with the gold standard, and odds ratios for all readers. The mean PD for the gold standard was $PD_{mean} = 38\%$ for all of the subjects with a SD of 19%. In general, as the correlation with the gold standard reading increased, the odds ratio measuring breast densities association with cancer increased, but that was not true for all readers. Particularly, reader NMD2 had a very high odds ratio despite low correlation during her second reading. Reading time averaged between 2 and 3 min per film. Readers with correlations with the gold standard of greater than 0.9 tended to have slightly longer reading times (average, 2.1 min) than readers who failed (average, 1.5 min; *t* test, nonsignificant). RMSE values ranged around 10% PD for all readers (Table 1).

We determined reproducibility for four readers who read all of the films twice and had correlation values ≥0.9 on the first read, and who, therefore, did not receive additional training between the two readings. Readers RAD2 and RAD3 were not available to us for a second reading. Reproducibility ranged from RMSE = 6% PD to 11% PD and showed this range for all three of the reader groups. By comparison, the gold standard reader achieved RMSE = 6% PD ($r = 0.95$) on a subset of 100 duplicate readings. When categorized by PD quarters, we found that the reproducibility range was similar for each quarter. We found RSME = 8–13% PD for the first quarter (PD_{gold}, <25%), RSME = 8–12% PD for the second, and RSME = 5–13% PD for the third quarter. We did not have enough data points for meaningful regression analysis of the fourth (PD_{gold}, >75%) quarter.

Table 2 shows generalized ANOVA results for the comparison of overall reader-group performance. The RAD group exhibited highest intraclass correlation, followed by the MD group and the NMDs. The same was true when only validated readers were considered, but because only one NMD was validated, intraclass correlation could not be calculated for this group.
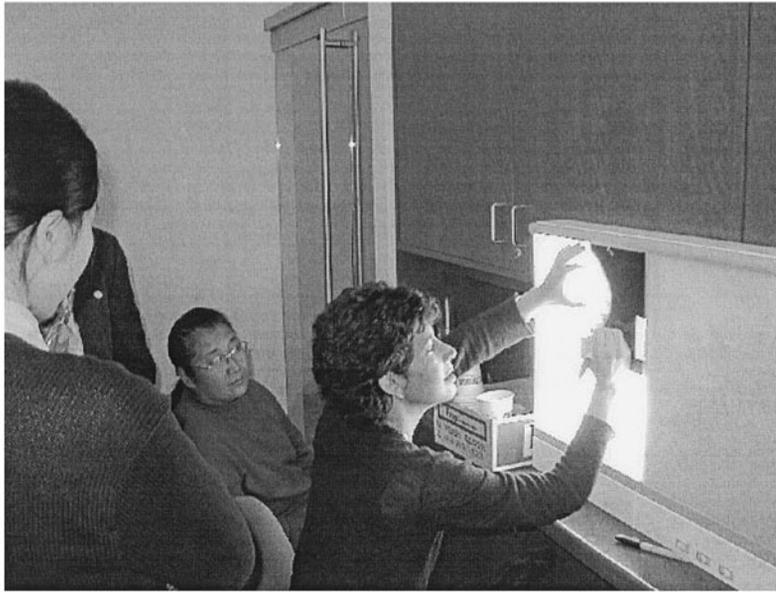
technique commonly used when comparing performance between readers or devices.

Fig. 2. Training session with the gold standard reader (R. S-B.).

Table 1 Reader validation results

The correlation coefficient (r) is with respect to the expert reader. RMSE and intercept are shown in units of PD. Regression is given as: reader value = gold value ∗ slope + intercept. Odds ratio (OR) to discriminate cancer status is based on a simple 50% PD threshold with 95% lower (LCL) and upper (UCL) confidence limits. All of the intercepts and slopes were significantly ($P < 0.05$) different from 0 and 1, respectively.

Reader	No. of reads to pass	r	RMSE (in PD)	Intercept (in PD)	Slope	OR (LCL, UCL)
RAD1-YC	1	0.91	9	7	1.04	3.09 (1.55, 6.18)
RAD2-XG	1	0.94	9	−7	1.17	3.89 (1.57, 9.63)
RAD3	1	0.92	10	3	1.16	3.29 (1.36, 7.93)
MD1-BF150	2	0.89	9	17	0.94	1.86 (0.93, 3.70)
MD2	2	0.93	8	−6	1.15	3.50 (1.65, 7.41)
NMD1	1	0.93	9	4	1.14	2.87 (1.44, 5.71)
NMD2	2[a]	0.70	14	31	0.73	4.25 (1.79, 10.09)
NMD3	3[a]	0.81	11	5	0.75	3.32 (1.32, 8.34)
NMD4	1[a]	0.81	13	−1	0.90	3.39 (1.48, 7.76)
GOLD	1					4.88 (2.31, 10.28)

[a] Last read was performed but still did not pass the validation criteria.

Table 3 lists odds ratios resulting from the unconditional logistic regression for a difference of PD = 1 SD, calculated for each study reader (OR_{SD}). The OR_{SD} values were significant and ranged from 1.7 to 2.1, being higher than that of the gold standard reader ($OR_{SD} = 1.9$).

Interreader correlations for all of the study readers with a high correlation ($r > 0.9$) with the gold standard ranged from $r = 0.82$ to 0.94, showing that their readings correlated well with each other.

Discussion

A principle finding of this study is that nonradiologists were able to be validated to read mammographic density. We also found that readers who had PD readings that were highly correlated with a gold standard could accurately discriminate between noncancer and cancer cases with odds ratios of 1.9 to

2.4 per population SD increase in PD. These results are similar to those reported by Boyd *et al.* (2)*,* who used a six-category scale for semiquantitative assessment and found per-category risk increments of 1.4. Other groups have compared the extreme ends of the PD range. In a case-control study of 160 paired images, Wolfe *et al.* (1) compared the extreme ends of the PD range and found odds ratios of 4.3 (95% confidence interval, 1.8–10.4). In one of the definitive studies of breast density as an independent risk factor, with about 2000 controls and as many cases, Byrne *et al.* (3) found a steady increase of odds ratios associated with PD category when the readers categorized films into either the 0%-PD category or into one of successive 25%-PD categories. Specifically, comparing the odds of breast cancer for PD = 0% *versus* the odds at PD > 75% yielded a significant odds ratio of 4.5.

In other published studies, inter- and intrareader variability values range from 0.86 to 0.96, similar to the results reported here (13, 14). Jong *et al.* (13) found an overall correlation of 0.89 between two readers and noted that the type of dense tissue distribution (homogeneous, nodular, linear) had a strong influence. We did not see such an effect in this study. Although we did not evaluate reproducibility by type of dense tissue, we retrospectively categorized the films by their density. We did not observe differences in reproducibility between categories. This might be of particular importance for study populations with a skewed or preselected mammographic density distribution.

A limitation of this study is that not all of the readers were available for a second reading. Therefore, we could not present a complete picture of intra- and interreader variability. Although those readers could have exhibited a performance drop during the second read, three of four readers who did finish a second reading maintained or even improved their performance. In addition, as noted above, the intraclass correlations are similar to those reported by others. Our reproducibility analysis did not consider redigitization. Although we do not expect a large effect from digitization because routine digitizer quality assurance showed that the device is stable and linearly maps film absorbance to pixel gray-scale value, this step needs to be verified. Lastly, because we attempted to provide a full range of density values for all of the decades, this most probably inflated

Table 2 Intraclass Correlations, r^2, by reader group[a]

	RAD	MD	NMD
All readers	0.98 (4)	0.88 (12)	0.81 (12)
Validated readers	0.98 (4)	0.88 (11)	

[a] The RMSE in units of PD are shown in parentheses.

Table 3 Logistic regression results[a]

Odds ratios (ORs) and their 95% confidence limits [lower (LCL) and upper (UCL)] of present cancer status for a change in PD of 1 SD for all validated readers (SD calibrated to gold standard) and the gold standard. SD is reader specific.

Reader	Unit (SD, % PD)	OR (LCL, UCL)
RAD1	24	1.99 (1.35, 2.94)
RAD2	22	2.05 (1.28, 3.28)
RAD3	22	2.13 (1.33, 3.40)
MD1	21	1.68 (1.18, 2.40)
MD2	22	1.77 (1.24, 2.52)
NMD1	20	2.00 (1.39, 2.90)
Gold standard (R. S-B.)	19	1.87 (1.33, 2.64)

the PD variance and, thus, improved the correlation coefficient. However, it is our view that using this approach (stratified PD values in every decade) will allow others to reproduce our results.

Reader Quality Control. Reader RAD1 in our study had PD readings highly correlated with the gold standard and high odds ratios predicting cancer risk on the first reading. When this reader was retested, he had a markedly lower correlation with the gold standard and slightly different slope and intercept. This suggests that PD readers for research studies may need to be monitored for consistent reading quality. Quantifying reproducibility is highly important because it influences the least significant difference that the breast density measure can detect with confidence. Continuous monitoring may be achieved by including films from the validation set with the study data so that, if necessary, a reader can be retrained, or removed if skills wane over time.

Of note, several readers achieved slightly, albeit nonsignificantly, higher odds ratios than the gold standard reader. This merely reflects that the gold standard itself is subjective but also suggests that the correlation criterion might need to be supplemented with a more objective one such as an odds ratio threshold. Automation would obviously alleviate the subjectivity problem. Efforts in this direction have recently been undertaken by a number of groups (15–18) but are also based on maximization of the correlation to a human gold standard.

Conclusions. The goal of this study was to establish whether or not a background in mammography or radiology is necessary to quantify mammographic breast density and to validate readers for future mammographic density studies based on the degree of correlation to a gold standard reader. We found that, although it seems beneficial to have a radiological background, it is not a prerequisite. Of nine study readers, five (all three radiologists, one of two physicians of other disciplines having some radiological background, and one of four nonphysicians) had readings that sufficiently correlated with the gold standard that they could be considered breast density readers for future

research studies. All of the readers with breast density readings highly correlated with the gold standard reading achieved odds to predict breast cancer of similar magnitude (on the order of three with PD dichotomized as less than 50% or 50% or greater), which is comparable with values in the literature. Strict validation criteria must be applied to qualify readers for mammographic breast density quantification. For research studies, this will enhance the chance of accurately assessing breast density and discriminating women at high and low risk of breast cancer.

References

1. Wolfe, J. N., Saftlas, A. F., and Salane, M. Mammographic parenchymal patterns and quantitative evaluation of mammographic densities: a case-control study. Am. J. Roentgenol., *148:* 1087–1092, 1987.

2. Boyd, N. F., Byng, J. W., Jong, R. A., *et al.* Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening study. J. Natl. Cancer Inst. (Bethesda), *87:* 670–675, 1995.

3. Byrne, C., Schairer, C., Wolfe, J. N., *et al.* Mammographic features and breast cancer risk: effects with time, age, and menopause status. J. Natl. Cancer Inst. (Bethesda), *87:* 1622–1629, 1995.

4. Saftlas, A. F., Wolfe, J. N., Hoover, R. N., *et al.* Mammographic parenchymal patterns as indicators of breast cancer risk. Am. J. Epidemiol., *129:* 518–526, 1989.

5. Wolfe, J. N. Risk for breast cancer development determined by mammographic parenchymal pattern. Cancer (Phila.), *37:* 2486–2492, 1976.

6. Saftlas, A. F., and Szklo, M. Mammographic parenchymal patterns and breast cancer risk. Epidemiol. Rev., *9:* 146–174, 1987.

7. Kerlikowske, K., Grady, D., Barclay, J., *et al.* Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J. Natl. Cancer Inst. (Bethesda), *90:* 1801–1809, 1998.

8. Brisson, J., Merletti, F., Sadowsky, N. L., Twaddle, J. A., Morrison, A. S., and Cole, P. Mammographic features of the breast and breast cancer risk. Am. J. Epidemiol., *115:* 428–437, 1982.

9. Boyd, N. F., O'Sullivan, B., Campbell, J. E., *et al.* Mammographic signs as risk factors for breast cancer. Br. J. Cancer, *45:* 185–193, 1982.

10. Byng, J. W., Boyd, N. F., Fishell, E., John, R. A., and Yaffe, M. J. The quantitative analysis of mammographic densities. Phys. Med. Biol., *39:* 1629–1638, 1994.

11. Kopans, D. B., D'Orsi, C. J., Adler, D. D., *et al.* American College of Radiology Breast Imaging and Reporting Data System (BI-RADS), Ed. 3. Reston, VA: American College of Radiology, 1998.

12. Chow, C. K., Venzon, D., Jones, E. C., Premkumar, A., O'Shaughnessy, J., and Zujewski, J. Effect of tamoxifen on mammographic density. Cancer Epidemiol. Biomark. Prev., *9:* 917–921, 2000.

13. Jong, R., Fishell, E., Little, L., Lockwood, G., and Boyd, N. F. Mammographic signs of potential relevance to breast cancer risk: The agreement of radiologists' classification. Eur. J. Cancer Prev, *5:* 281–286, 1996.

14. Ursin, G., Astrahan, M. A., Salane, M., *et al.* The detection of changes in mammographic densities. Cancer Epidemiol. Biomark. Prev., *7:* 43–47, 1998.

15. Byng, J. W., Yaffe, M. J., Lockwood, G. A., Little, L. E., Tritchler, D. L., and Boyd, N. F. Automated analysis of mammographic densities and breast carcinoma risk. Cancer (Phila.), *80:* 66–74, 1997.

16. Boone, J. M., Lindfors, K. K., Beatty, C. S., and Seibert, J. A. A breast density index for digital mammograms based on radiologists' ranking. J. Digit. Imaging, *11:* 101–115, 1998.

17. Heine, J. J., and Velthuizen, R. P. A statistical methodology for mammographic density detection. Med. Phys., *27:* 2644–2651, 2000.

18. Lou, S. L., and Fan, Y. Automatic evaluation of breast density for full field mammography. *In:* SPIE Medical Imaging 2000. San Diego, CA: SPIE, 2000.

Accuracy of Mammographic Breast Density Analysis: Results of Formal Operator Training

Sven Prevrhal, John A Shepherd, Rebecca Smith-Bindman, et al.

Updated version	Access the most recent version of this article at: http://cebp.aacrjournals.org/content/11/11/1389

Cited articles	This article cites 16 articles, 2 of which you can access for free at: http://cebp.aacrjournals.org/content/11/11/1389.full#ref-list-1
Citing articles	This article has been cited by 7 HighWire-hosted articles. Access the articles at: http://cebp.aacrjournals.org/content/11/11/1389.full#related-urls

E-mail alerts	Sign up to receive free email-alerts related to this article or journal.
Reprints and Subscriptions	To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.
Permissions	To request permission to re-use all or part of this article, use this link http://cebp.aacrjournals.org/content/11/11/1389. Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.