Diagnostic Accuracy of Digital Mammography in Patients with Dense Breasts Who Underwent Problem-solving Mammography: Effects of Image Processing and Lesion Type[1]

Elodia B. Cole, MS
Etta D. Pisano, MD
Emily O. Kistner, MS
Keith E. Muller, PhD
Marylee E. Brown, BA
Stephen A. Feig, MD
Roberta A. Jong, MD, FRCPC
Andrew D. A. Maidment, PhD
Melinda J. Staiger, MD
Cherie M. Kuzmiak, DO
Rita I. Freimanis, MD
Nadine Lesko, MD
Eric L. Rosen, MD
Ruth Walsh, MD
Margaret Williford, MD
M. Patricia Braeuning, MD

Index terms:
Breast radiography, comparative studies, 00.112, 00.1215
Diagnostic radiology, observer performance, 00.112, 00.1215
Images, processing, 00.112, 00.1215
Radiography, digital, 00.1215
Screens and films, 00.112

Published online before print
10.1148/radiol.2261012024
Radiology 2003; 226:153–160

Abbreviations:
AUC = area under the ROC curve
BI-RADS = Breast Imaging Reporting and Data System
CLAHE = contrast-limited adaptive histogram equalization
HIW = histogram-based intensity windowing
ROC = receiver operating characteristic
UNC = University of North Carolina

[1] From the Department of Radiology and Lineberger Comprehensive Cancer Center, University of North Carolina School of Medicine, 507 Old Infirmary, CB 7510, Chapel Hill, NC 27599-7510 (E.D.P.). The complete list of authors' affiliations and the author contributions are at the end of this article. Received December 10, 2001; revision requested February 18, 2002; revision received April 18; accepted May 24. Supported by grant 282-97-0078 from the Office of Women's Health, Department of Health and Human Services. **Address correspondence to** E.D.P. (e-mail: *etpisano@med.unc.edu*).

PURPOSE: To determine effects of lesion type (calcification vs mass) and image processing on radiologist's performance for area under the receiver operating characteristic curve (AUC), sensitivity, and specificity for detection of masses and calcifications with digital mammography in women with mammographically dense breasts.

MATERIALS AND METHODS: This study included 201 women who underwent digital mammography at seven U.S. and Canadian medical centers. Three image-processing algorithms were applied to the digital images, which were acquired with Fischer, General Electric, and Lorad digital mammography units. Eighteen readers participated in the reader study (six readers per algorithm). Baseline values for reader performance with screen-film mammograms were obtained through the additional interpretation of 179 screen-film mammograms. A repeated-measures analysis of covariance allowing unequal slopes was used in each of the nine analyses (AUC, sensitivity, and specificity for each of three machines). Bonferroni correction was used.

RESULTS: Although lesion type did not affect the AUC or sensitivity for Fischer digital images, it did affect specificity ($P = .0004$). For the General Electric digital images, AUC, sensitivity, and specificity were not affected by lesion type. For Lorad digital images, the results strongly suggested that lesion type affected AUC and sensitivity ($P < .0001$). None of the three image-processing methods tested affected the AUC, sensitivity, or specificity for the Fischer, General Electric, or Lorad digital images.

CONCLUSION: Findings in this study indicate that radiologist's interpretation accuracy in interpreting digital mammograms depends on lesion type. Interpretation accuracy was not influenced by the image-processing method.

An estimated 40% of the population of women undergoing mammography screening have dense breast tissue (1). Patients with dense breasts often require additional imaging for diagnosis beyond the standard four views (2), which results in additional examination time, cost, and radiation exposure to the patient and causes anxiety. Over the past 25 years, there have been substantial improvements in dedicated mammographic equipment, screen-film combinations, and processing units, and these improvements have resulted in improved images and reduced radiation dose (3). Because image acquisition factors are often interdependent, the optimization of one factor often comes at the expense of

another. For example, a decrease in dose may result in increased noise and decreased image sharpness.

Digital mammography allows the separation of image acquisition, processing, and display (4) and may therefore represent a solution to many of the inherent limitations of screen-film mammography (5,6). The digital detector has a linear response to x-ray intensity, in contrast to the sigmoidal response of screen-film systems. As a result, use of a digital detector provides a broader dynamic range of densities and higher contrast resolution (4). Through image processing, display parameters may be chosen independently from image acquisition factors. Small differences in attenuation between normal and abnormal breast tissue can be amplified. For example, contrast manipulation could improve lesion conspicuity. Findings in a previous study (7) suggested that different image-processing methods might be preferable for different tasks and different detector types. At digital mammography, there also is flexibility in image presentation. Images can be displayed in soft-copy format on high-resolution monitors or printed to film for display on a mammographic view box.

At the time of this study, none of the digital mammographic systems were approved by the Food and Drug Administration. Each of the systems required clinical trials showing, at minimum, equivalence to screen-film systems. The most critical determinant of equivalency is radiologist performance in the detection and characterization of lesions measured typically by means of receiver operating characteristic (ROC) analysis. While no consensus exists regarding the average standard performance for interpretation of screen-film mammographic images, researchers in several comparative studies have identified performance levels. Taplin et al (8) reported sensitivity of 79%, specificity of 81%, and area under the ROC curve (AUC) of 0.85 for a screening population with screen-film mammography. Jiang et al (9) reported sensitivity of 73.5%, specificity of 31.6%, and AUC of 0.61 for screen-film mammography in another screening population. Rosenberg et al (10) obtained sensitivity of 68% for a population with dense breasts. Van Gils et al (11) obtained a sensitivity of 59% for another population of women with dense breasts.

Our study was designed to determine the effects of lesion type (calcifications vs masses) and image processing on the radiologist's performance for AUC, sensitivity, and specificity for mass and calcification detection by using digital mammography in women with mammographically dense breasts.

MATERIALS AND METHODS

Image Production

The images used in the study were acquired at seven institutions: Massachusetts General Hospital, Boston; Hospital of the University of Pennsylvania, Philadelphia; University of Virginia, Charlottesville; Good Samaritan Hospital, West Islap, NY; University of Toronto, Ontario, Canada; Thomas Jefferson University Hospital, Philadelphia, Pa; and the University of North Carolina (UNC), Chapel Hill. The institutional review board at each participating site approved the study. Informed consent was obtained.

Two groups of women were enrolled between August 1998 and April 1999 into this study with different eligibility criteria. Group A included 167 consecutively enrolled women with mammographically dense breasts who had undergone diagnostic problem-solving mammography and were scheduled for either open (excisional) or percutaneous large core-needle breast biopsy within 12 weeks after the eligibility mammogram was obtained. Of the 167 women enrolled in group A, 165 had mammographically visible lesions at digital mammography. The remaining two patients had mammographically occult but palpable lesions. Group B consisted of a random sample of 34 women with mammographically dense breasts who underwent problem-solving mammography at the participating mammography clinics and were recommended for routine (1-year) follow-up rather than for biopsy. Some of the women in both groups had abnormal screening mammograms and had undergone problem-solving mammography for evaluation of findings detected at screen-film mammography. Data about the number of women who were entered into the study after they underwent screening mammography and had abnormal findings were not collected at all sites.

A total of 201 patients were included in the study. Mammographic images in 75 cases were obtained in women who underwent digital mammography at University of Virginia, Good Samaritan Hospital, and Thomas Jefferson University Hospital by using the Lorad digital mammography system (LoRad Digital Mammography System; Lorad, Danbury, Conn). Images in another 74 cases were obtained by using the Fischer digital mammography system (SenoScan; Fischer, Denver, Colo) located at University of Toronto and UNC. Images in the other 52 cases were obtained by using the General Electric digital mammography system (Senographe 2000D; GE Medical Systems, Milwaukee, Wis) located at the Hospital of the University of Pennsylvania and Massachusetts General Hospital.

At each institution, the recruiting radiologist determined eligibility among women who were undergoing screen-film problem-solving mammography. For those patients agreeing to participate, the recruiting radiologist, or designee, obtained informed consent at the time of enrollment into the study. Once consent was obtained, standard bilateral screen-film and digital mammograms were obtained during the same visit. "Standard" in this context implies that as many views as were necessary were obtained to include both breasts in their entirety.

The raw digital data were transmitted in an image format in use at the time by each manufacturer to UNC, where the images were converted to a standard image format for image processing and film printing.

All images were processed by using each of two different algorithms at UNC: histogram-based intensity windowing (HIW) and contrast-limited adaptive histogram equalization (CLAHE). In addition to HIW and CLAHE, each manufacturer provided its preferred algorithm for application to its machine-specific images. These three additional algorithms were collectively termed the manufacturer's recommended or default method. Each manufacturer was responsible for applying its default method to its machine-specific collection of images. The algorithms chosen were applied to the images following procedures outlined elsewhere (12).

The numbers of lesions classified according to machine type are shown in Table 1. This information was collected from biopsy reports and mammographic interpretation reports from clinical centers where the patients were recruited into the study. All biopsy-proved lesions were described by using the Breast Imaging Reporting and Data System (BI-RADS) lexicon on a standard reporting form devised for uniformity of information gathering across sites. The BI-RADS caner stage and size of pathologically proved cancers in the study population are listed in Tables 2 and 3, respectively.

Reader performance with conventional screen-film mammography was assessed to establish the baseline performance level of our group of readers in the same controlled

TABLE 1
Description of Diagnosis and Findings on 201 Digital and 179 Screen-Film Mammographic Images

Diagnosis and Findings	Digital Images (n = 201)			Screen-Film Images (n = 179)
	General Electric (n = 52)	Fischer (n = 74)	Lorad (n = 75)	
Group A				
Cancer				
Masses	10	14	13	33
Calcifications	9	8	8	18
Noncancer				
Masses	14	23	20	55
Calcifications	10	16	22	40
Group B*	9	13	12	33

Note.—Screen-film mammographic images were matched to digital images for lesion type, breast density, and cancer status.
* Group B included patients with normal mammograms and no findings.

TABLE 2
Cancer Stage at Diagnosis in 62 Patients with Digital and 51 Patients with Screen-Film Mammographic Images

Imaging Method and Finding	BI-RADS Cancer Stage*					
	0 (n = 0)	I (n = 48)	II (n = 30)	III (n = 16)	IV (n = 2)	Missing (n = 17)
Screen-film						
Calcifications (n = 18)	0	15	2	1	0	0
Masses (n = 33)	0	14	14	3	1	1
Total	0	29	16	4	1	1
Digital						
Lorad						
Calcifications (n = 8)	0	5	1	2	0	0
Masses (n = 13)	0	3	5	4	0	1
General Electric						
Calcifications (n = 9)	0	3	1	0	0	5
Masses (n = 10)	0	0	0	1	0	9
Fischer						
Calcifications (n = 8)	0	6	1	1	0	0
Masses (n = 14)	0	2	6	4	1	1
Total	0	19	14	12	1	16

* Cancer stages are based on information provided by the National Cancer Registry.

environment as would be used for digital interpretation during the course of the reader interpretation study. The resulting baseline performance level was used to control for differences in reader experience. A total of 179 screen-film cases were selected from UNC image archives by a research technologist by using database descriptors of lesion type, cancer status, and breast density. Statisticians confirmed the cases that were matched through comparisons of these descriptors.

Of the 179 cases, 146 were matched to group A patients; 144 of 146 of the group A–matched screen-film images had mammographically visible lesions. The remaining two images had palpable lesions that were mammographically occult. These screen-film images were taken from a different group of patients but were matched for breast density, lesion type, and cancer status to the digital images. Table 1 also shows the distribution of findings in the matched population of women whose screen-film mammograms were included in this study. The 179 matched screen-film mammograms were used to obtain a baseline measure of mammographic interpretation performance for the group of 18 readers who participated in the study.

Reader Interpretation Study

A total of 18 radiologists interpreted the digital images. Fourteen readers were from teaching faculties or private practices with an average of 7.6 years (range, 2–18 years) of mammographic interpretation experience. The remaining four readers were breast imaging fellows with an average of 3 months (range, 0–12 months) of mammographic interpretation experience. Seventeen of the readers were certified by the American Board of Radiology; the remaining reader was eligible for certification but not yet certified by the American Board of Radiology. Readers were randomly assigned to interpret hard-copy digital mammograms that had been processed with one of the three methods (CLAHE, HIW, and manufacturer's recommendation). Six readers interpreted the 201 digital images processed with the manufacturer's recommended or default method. Another six readers interpreted the 201 digital images processed with HIW. A third group of six readers interpreted the 201 digital images processed with CLAHE.

Before beginning the study, readers were trained in viewing digital mammograms printed with the specific image-processing method assigned to them, with all machine types represented. The training set consisted of 28 mammograms, which were not included in the actual reader study, containing pathologically proved benign or malignant lesions or lesions documented to be benign by means of clinical and mammographic follow-up. During this training, the first 14 digital images were randomly ordered and presented alongside the corresponding screen-film mammograms for direct comparison so that the readers could evaluate the differences in the digital images in terms of lesion characteristics and image processing. The final 14 images presented were shown in digital format only. The readers were provided with written information identifying clinically relevant lesions and diagnoses for all training cases. The training cases also allowed the readers to become familiar with the data collection forms and malignancy ratings scales. Upon completion of training, the readers began the actual reader study.

The 201 digital mammograms, printed with the assigned image-processing method, and a randomly selected subset of 100 of

the 179 screen-film images were presented to each reader on a multiviewer (Panoramascope; RADX, Houston, Tex) appropriately masked for the presentation of bilateral mammograms. Nine readers first interpreted the 201 digital images in their assigned display method in random order and then interpreted 100 randomized screen-film mammograms. All 301 mammograms per reader were read over a 2-day period. The remaining nine readers first interpreted 100 randomized screen-film mammograms and then interpreted 201 randomly ordered digital mammograms.

Both digital and screen-film mammographic interpretations were determined without prior mammographic studies for comparison or pertinent patient history. The images were prehung in sets of 50 on a multiviewer by a research assistant. Five-minute breaks were required every 50 minutes and were otherwise permitted as necessary. Readers reported findings by using a standard set of interpretation forms. A research assistant electronically collected the study data by using data entry software (Epi-INFO, version 6, DOS; Center for Disease Control and Prevention, Atlanta, Ga) and a laptop computer (Satellite Pro 460CDT; Toshiba, Irvine, Calif). The data collected were reported by using the standard American College of Radiology BI-RADS lexicon and included descriptions of the overall breast parenchymal density and the location, number, and type of findings, that is, mass, calcification, architectural distortion, and asymmetric density, if any.

Each occurrence of a clinically relevant feature was individually characterized by using BI-RADS descriptors for a particular lesion type. A probability of malignancy rating was made for each perceived clinically relevant lesion on the basis of a five-point scale as follows: score 1, definitely not malignant; score 2, probably not malignant; score 3, possibly malignant; score 4, probably malignant; and score 5, definitely malignant. The radiologists were asked to assign a probability of malignancy rating for every lesion. When no lesion was detected, the mammogram was assigned a score of 0. The radiologists were also asked to provide a follow-up recommendation for all identified abnormalities, namely, routine follow-up, six-month follow-up mammography, immediate additional imaging (eg, ultrasonographic [US] images, magnification views, spot compression views), or biopsy.

TABLE 3
Pathologic Cancer Size in Largest Dimension Diagnosed in 62 Patients with Digital and 51 with Screen-Film Mammographic Images

Imaging Method and Finding	Pathologic Cancer Size					
	<5 mm (n = 12)	5–10 mm (n = 21)	11–20 mm (n = 24)	21–30 mm (n = 17)	>30 mm (n = 8)	Missing (n = 31)
Screen-film						
Calcifications (n = 18)	4	4	5	1	2	2
Masses (n = 33)	2	8	8	9	5	1
Total	6	12	13	10	7	3
Digital						
Lorad						
Calcifications (n = 8)	0	2	1	0	0	5
Masses (n = 13)	0	4	4	0	0	5
General Electric						
Calcifications (n = 9)	2	0	1	0	0	6
Masses (n = 10)	0	0	0	1	0	9
Fischer						
Calcifications (n = 8)	3	0	1	2	0	2
Masses (n = 14)	1	3	4	4	1	1
Total	6	9	11	7	1	28

Statistical Methods

Very detailed rules were developed to assess the accuracy of a reader's ability to localize a lesion. To correctly specify location, a reader had to match side (left, right), depth, and clock referent location (eg, the "3-o'clock" position) of the lesion. All locations of lesions identified from pathology reports were verified by checking that the corresponding digital or baseline screen-film mammogram included the lesion on which biopsy was performed at the same location. When a lesion was seen on both views, the clock location had to be within three numbers from the true clock location that was determined by means of the biopsy report for a particular case. When a lesion was seen on only one view by a reader, the breast view and plane combination (ie, left craniocaudal) would have to include the actual clock location determined at biopsy to be judged a correct localization.

For example, if a mass with a 2-o'clock location as determined by means of biopsy was seen by a given reader only on the left craniocaudal view, laterally, the reader would be considered correct in localizing the lesion. In this instance, the reader would be considered correct since the lateral plane of the left breast included the 2-o'clock location. A binary variable, target, was created to denote whether a particular remark corresponded to an area for which pathologic findings were available. If the reader failed to rate the pathologically proved area, then an additional rating was created for the reader, with target = 1 and the rating = 0. This led to a false-negative finding for a malignant lesion. Next, all readings (for a particular reader and case) were reduced to at most two readings. For those cases with a lesion, the maximum rating among those judged to describe the lesion was computed. Similarly, the maximum rating was computed over all regions without a lesion. The approach corresponds to the alternative free-response ROC method (13).

Three primary outcomes were analyzed: *(a)* AUC from a nonparametric alternative free-response ROC analysis of cancer/no cancer; *(b)* sensitivity, with malignancy scores 0–2 defined as "benign" and malignancy scores 3–5 defined as "malignant"; and *(c)* specificity, with malignancy scores 0–2 defined as "benign" and malignancy scores 3–5 defined as "malignant." The alternative free-response ROC method, according to Chakraborty and Winter (13), was applied separately to each reader's data for each combination of machine and lesion type. To help control for multiple testing, $\alpha = .05/3 = .016$ was used for each outcome. For the purposes of simplicity, architectural distortions and asymmetric densities were classified as masses for the analyses. Baseline performance in interpreting screen-film mammographic images was used as a covariate in the analyses to control for differences in reader

TABLE 4
Estimated Mean AUC, Sensitivity, and Specificity for Fischer Unit

	AUC		Sensitivity		Specificity	
Algorithm	Calcifications	Masses	Calcifications	Masses	Calcifications	Masses
Default	0.682 ± 0.036	0.663 ± 0.026	0.663 ± 0.052	0.487 ± 0.040	0.656 ± 0.037	0.815 ± 0.042
CLAHE	0.581 ± 0.035	0.649 ± 0.026	0.580 ± 0.052	0.609 ± 0.040	0.597 ± 0.037	0.693 ± 0.042
HIW	0.635 ± 0.035	0.678 ± 0.026	0.554 ± 0.052	0.526 ± 0.040	0.685 ± 0.037	0.798 ± 0.042

Note.—All values are expressed as mean ± standard error.

TABLE 5
Estimated Mean AUC, Sensitivity, and Specificity for General Electric Unit

	AUC		Sensitivity		Specificity	
Algorithm	Calcifications	Masses	Calcifications	Masses	Calcifications	Masses
Default	0.577 ± 0.047	0.649 ± 0.034	0.498 ± 0.055	0.591 ± 0.059	0.635 ± 0.047	0.718 ± 0.034
CLAHE	0.604 ± 0.047	0.618 ± 0.034	0.590 ± 0.055	0.590 ± 0.059	0.635 ± 0.047	0.674 ± 0.034
HIW	0.556 ± 0.047	0.659 ± 0.034	0.375 ± 0.055	0.559 ± 0.059	0.693 ± 0.047	0.736 ± 0.034

Note.—All values are expressed as mean ± standard error.

TABLE 6
Estimated Mean AUC, Sensitivity, and Specificity for Lorad Unit

	AUC		Sensitivity		Specificity	
Algorithm	Calcifications	Masses	Calcifications	Masses	Calcifications	Masses
Default	0.624 ± 0.035	0.846 ± 0.026	0.539 ± 0.051	0.869 ± 0.033	0.695 ± 0.033	0.707 ± 0.031
CLAHE	0.597 ± 0.034	0.789 ± 0.025	0.601 ± 0.051	0.818 ± 0.033	0.602 ± 0.033	0.621 ± 0.031
HIW	0.666 ± 0.035	0.860 ± 0.025	0.631 ± 0.051	0.825 ± 0.033	0.669 ± 0.033	0.713 ± 0.030

Note.—All values are expressed as mean ± standard error.

experience or ability. Details of the analysis plan and model fitting are specified in the Appendix. Bonferroni correction was used.

RESULTS

The multivariate analysis of covariance model was chosen for this exploratory analysis of AUC, sensitivity, and specificity. The baseline performance of the group of 18 readers for interpretation with screen-film images was 0.691 for AUC, 62.0% for sensitivity, and 69.4% for specificity.

There were no interactions for image-processing method according to lesion type, and there were no small *P* values for any of the three machine types. Tables 4 (Fischer), 5 (General Electric), and 6 (Lorad) include mean values for AUC, sensitivity, and specificity for all combinations of digital image-processing method and lesion type.

Fischer Unit

For AUC, interpretation of masses (mean, 0.663; standard error, 0.015) was better than interpretation of calcifications (mean, 0.633; standard error, 0.020) across all three image-processing algorithms, but this resulted in no small *P* values (Fig 1, left). AUC was best for default-processed images (mean, 0.673; standard error, 0.020), followed by HIW-processed images (mean, 0.657; standard error, 0.020), and CLAHE-processed images (mean, 0.615; standard error, 0.020) across both lesion types. There were no small *P* values for tests of differences between the means for the three image-processing methods.

Sensitivity was better for interpretation of calcifications (mean, 0.599; standard error, 0.030) than for interpretation of masses (mean, 0.541; standard error, 0.023) across all three image-processing algorithms, but this did not result in small *P* values. CLAHE-processed images (mean, 0.594; standard error, 0.032) had a higher sensitivity than default-processed images (mean, 0.575; standard error, 0.032) or HIW-processed images (mean, 0.540; standard error, 0.032) across the two lesion types. There were no small *P* values for tests of the differences between the means for the image-processing methods.

There was a higher specificity for interpretation of masses (mean, 0.769; standard error, 0.024) than for interpretation of calcifications (mean, 0.646; standard error, 0.021) across the three image-processing algorithms. This difference in mean resulted in a small *P* value ($P = .0004$). HIW was best for specificity (mean, 0.742; standard error, 0.032), followed by Default (mean, 0.735; standard error, 0.032), and CLAHE (mean, 0.645; standard error, 0.032) across the two lesion types. Tests of differences between the means of the three image-processing methods did not result in small *P* values.

General Electric Unit

For AUC, interpretation of masses (mean, 0.642; standard error, 0.020) was

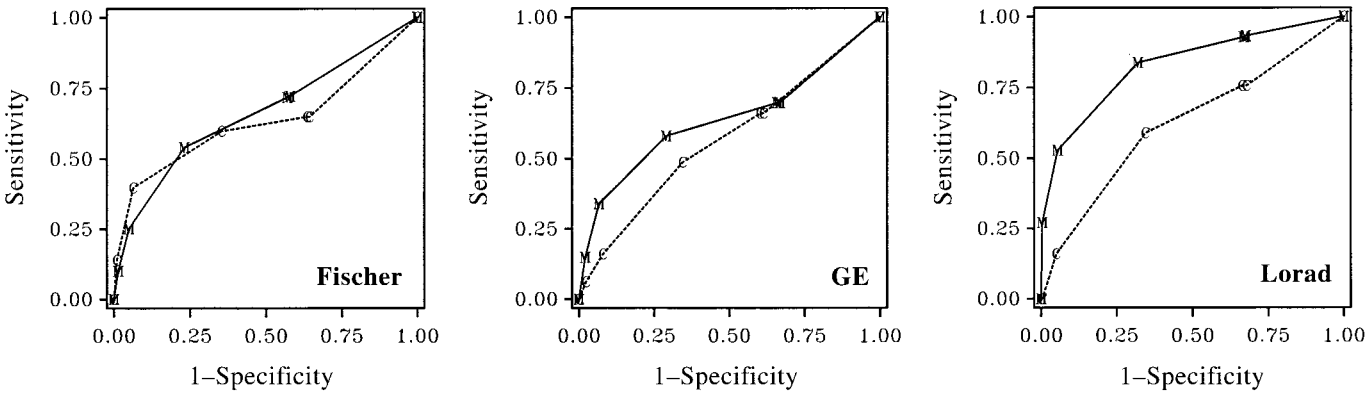

Figure 1. ROC curves for images obtained with Fischer, General Electric, and Lorad units. Each graph includes the average performance of 18 readers for interpretation of masses *(M)* and calcifications *(C)* across all image-processing algorithms.

better than interpretation of calcifications (mean, 0.579; standard error, 0.027) across the three image-processing algorithms, but this did not result in a small *P* value (Fig 1, center). AUC was best for default-processed images (mean, 0.613; standard error, 0.030), followed by CLAHE-processed images (mean, 0.611; standard error, 0.030) and HIW-processed images (mean, 0.608; standard error, 0.030) across the two lesion types. Tests of differences between the means of the three image-processing methods did not result in small *P* values.

Sensitivity was better for interpretation of masses (mean, 0.580; standard error, 0.034) than for interpretation of calcifications (mean, 0.488; standard error, 0.032) across the three image-processing algorithms. The difference in means did not result in a small *P* value. For processing methods, CLAHE-processed images (mean, 0.590; standard error, 0.042) had a higher sensitivity than default-processed images (mean, 0.544; standard error, 0.042) or HIW-processed images (mean, 0.467; standard error, 0.042) across the two lesion types. Tests of differences between the means of the three image-processing methods did not result in small *P* values.

There was a higher specificity for interpretation of masses (mean, 0.709; standard error, 0.019) than for interpretation of calcifications (mean, 0.654; standard error, 0.027) across the three image-processing algorithms. The difference in means did not result in a small *P* value. HIW-processed images were best for specificity (mean, 0.714; standard error, 0.028), followed by default-processed images (mean, 0.676; standard error, 0.028) and CLAHE-processed images (mean, 0.655; standard error, 0.028) across the two lesion types. Tests of differences between the means of the three image-processing methods did not result in small *P* values.

Lorad Unit

For AUC, interpretation of masses (mean, 0.832; standard error, 0.015) was better than interpretation of calcifications (mean, 0.629; standard error, 0.020) across the three image-processing algorithms (Fig 1, right). The difference in means resulted in a small *P* value ($P < .0001$). AUC was best for HIW-processed images (mean, 0.763; standard error, 0.021), followed by default-processed images (mean, 0.735; standard error, 0.021) and CLAHE-processed images (mean, 0.693; standard error, 0.021) across the two lesion types. Tests of differences between the means of the three image-processing methods did not result in small *P* values.

Sensitivity was better for interpretation of masses (mean, 0.838; standard error, 0.019) than for interpretation of calcifications (mean, 0.590; standard error, 0.029) across the three image-processing algorithms. The difference in means resulted in a small *P* value ($P < .0001$). For processing methods, HIW-processed images (mean, 0.728; standard error, 0.034) had a higher sensitivity than CLAHE-processed images (mean, 0.709; standard error, 0.034) or default-processed images (mean, 0.704; standard error, 0.034) across the two lesion types. Tests of differences between the means of the three image-processing methods did not result in small *P* values.

There was a higher specificity for interpretation of masses (mean, 0.680; standard error, 0.018) than for interpretation of calcifications (mean, 0.656; standard error, 0.019) across the three image-processing algorithms. The difference did not result in a small *P* value. Default-processed images were best for specificity (mean, 0.701; standard error, 0.025), followed by HIW-processed images (mean, 0.691; standard error, 0.025) and CLAHE-processed images (mean, 0.612; standard error, 0.025) across the two lesion types. Tests of differences between the means of the three image-processing methods did not result in small *P* values.

Assessment of the likelihood of diagnosing cancer from findings on a mammogram is based on specific characteristics of the lesion: the distribution, morphology, number of particles in a cluster, and cluster size for calcifications and the margins, size, shape, and density for masses. BI-RADS provides standard terminology to describe the specific features of lesions. These features, some of which may be easily discernable on standard mammograms, are generally verified by using some type of diagnostic imaging (eg, US images, spot compression views) to determine a probability of malignancy. Since the readers of our study only had the standard views, the results presented here are based on their impressions of the visual characteristics that could be seen on the standard four-view mammograms. Some lesions were very obvious, and diagnosis was easy (Fig 2). Other lesions were very subtle, and diagnosis was difficult (Fig 3). In addition, there were lesions that were visible, but the features of the lesions as seen on the mammograms were not characteristic of the actual pathologic diagnoses (Fig 4).

DISCUSSION

With digital mammography, the fact that there is a difference in the radiologist's

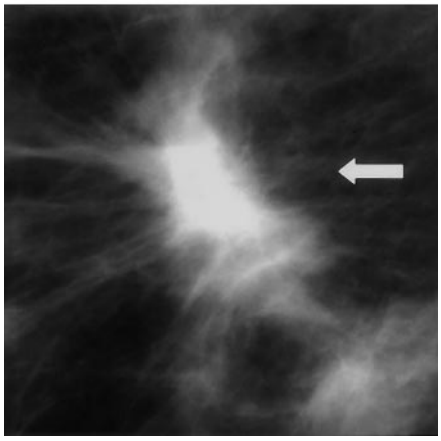

Figure 2. Image obtained with Lorad unit and processed with CLAHE algorithm. Cropped left mediolateral oblique view of a spiculated mass (arrow). This 30 × 15 × 25-mm mass was of higher density than the surrounding parenchyma. The lesion was pathologically proved to be invasive ductal carcinoma. Fourteen of 18 readers rated this lesion with a score of 5, or definitely malignant. The remaining four readers rated this lesion with a score of 4, or probably malignant. Note how it lies in the fatty part of this patient's heterogeneously dense breasts; detection and characterization are easy.

Figure 3. Image obtained with General Electric unit and processed with HIW algorithm. Cropped left craniocaudal view of an architectural distortion (mass) (arrows) 15 × 15 × 20 mm in size. The density of the lesion was equivalent to that of the surrounding parenchyma and was pathologically proved to be invasive ductal carcinoma. This lesion was not seen or was not considered clinically relevant by 17 of 18 readers.

performance in interpretation of masses and calcifications is expected. As individual calcifications tend to be considerably smaller than masses, a reader's ability to define the morphology of calcifications by using the standard four views, given

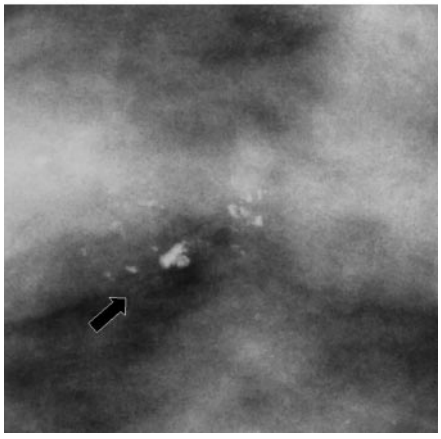

Figure 4. Image obtained with Fischer unit and processed with the manufacturer's default algorithm. Cropped right mediolateral oblique view of a cluster of pleomorphic calcifications (arrow) 10 × 6 × 4 mm in size. The lesion was pathologically proven to be ductal carcinoma in situ. This lesion was not seen or was not considered clinically relevant by 13 of 18 readers. The remaining five readers saw the lesion, but all rated it with a score of 2, or probably benign.

spatial resolution constraints, may make it more difficult to assess them well. Clearly, the assessment of calcification morphology is improved with focal magnification techniques, which were not available to our readers (16). Evaluation of the validity of any study about digital mammographic interpretation in regard to AUC, sensitivity, and specificity results requires knowledge of the lesion types and their frequency distribution within the study's case set, as well as breast density distribution.

This study was set up to represent a population of patients in whom the mammograms that were obtained were relatively difficult to interpret because of the patients' increased breast density. Researchers in some studies have investigated the effect of breast density on sensitivity, specificity, and AUC. Sensitivity, specificity, and AUC performance for populations of patients with dense breasts have been documented as being lower than the sensitivity, specificity, and AUC performance measures for a more general population of patients (10,11), because the lesions in dense breasts are often less conspicuous, and mammographic interpretation in these cases is more difficult.

The sensitivity, specificity, and AUC values obtained in this study are somewhat lower than values published in the literature for screen-film mammography (8,9) and digital mammography (17) for screening populations. Beam et al (18) showed sensitivity of 79.3%, specificity of 88.5%, and AUC of 0.845 for a population of patients who underwent screen-film mammography, neither additional views nor other imaging studies were required for diagnosis. The only digital mammography unit with published results is General Electric, and in this regard, Lewin et al (17) indicated a sensitivity of 60% for digital mammography and 63% for screen-film mammography for a screening population. A standard screening population would include women whose breasts are mostly or completely fatty in composition, and thus the frequency of visible lesions would be increased in these patients.

In a previous study (12), the application of image processing led to differences in perceived effectiveness by readers. Similarly in this study, there were slight differences in AUC, sensitivity, and specificity, depending on the image-processing method applied to the images from each machine, but none of the tests for mean differences resulted in small *P* values. This could be due to the lack of power in the study for this effect or, just as likely, there simply is no difference in performance in interpretation of the processed digital mammograms with the three tested algorithms. There is also the question of subtlety of the lesions within the case sets, which may function as a predictor of case difficulty. The subtlety, or visibility of the detail of a lesion, is dependent on a number of variables and these include overall breast density, density of the lesion in relationship to the surrounding breast tissue, and lesion size.

The application of different image-processing algorithms may lead to different results; therefore, this study will be repeated with an additional three image-processing algorithms applied to the same 201 digital mammograms. The screen-film images that will be read in that study are the screen-film images for the same 201 patients in whom digital images were obtained. Because careful control of lesion size and cancer stage was not maintained between the screen-film images used to establish baseline performance and the digital images, a direct comparison between performance with digital images and with screen-film images was not possible for the study documented here. Use of the screen-film mammograms obtained at the time of enrollment of patients in this study (ie, the screen-film images were obtained at the same time as the digital images) will

eliminate this limitation in the replication study currently underway.

It is important to note and explain why we chose not to use the widely familiar BI-RADS lexicon in the reader study we conducted. In fact, while radiologists are well trained in the use of BI-RADS terms, BI-RADS terms do not measure anything in the required and statistically continuous fashion that would be suitable for ROC analysis. Our scale of malignancy did allow the reader to record an impression of the probability of malignancy on a five-point scale. This allows the creation of smooth ROC curves. In addition, radiologists are not consistent in their use of BI-RADS terms (19). The creation of a new scale that each reader was trained to use circumvented the problems that would be created by the use of a familiar but inconsistently used scale.

The American College of Radiology Imaging Network "Digital Mammographic Imaging Screening Trial" is a study that will be conducted to assess the effectiveness of digital mammography in a screening setting. The study will enroll 49,500 women at 18 institutions and include four different digital mammographic systems. Both digital mammograms and corresponding screen-film mammograms will be acquired for subsequent use in reader studies. The cases will be drawn from a screening population for a more inclusive and statistically conclusive comparison of digital and screen-film mammography. In addition, it is possible that algorithms useful for diagnosis may not be useful in the screening setting (7). Additional imaging algorithm evaluation in that population will probably be needed.

APPENDIX: PLANNED ANALYSIS

The original analysis (not reported here) involved fitting a model of difference scores (digital-analog). Predictors that varied between readers included processing method, analog (baseline) performance, as well as the interaction of method and baseline. Machine and lesion type varied within reader and, therefore, were treated as repeated measures. Concern over not using the same cases for analog and digital modalities led to changing analyses to the following. Three primary outcomes (AUC, sensitivity, and specificity) were analyzed separately for each machine (Fischer, General Electric, and Lorad units) by using the same approach. A general linear multivariate model was used, with repeated-measures tests based on the Geisser-Greenhouse test (14). The response matrix contained 18 rows and two columns, and each row represented the performance of a radiologist for interpretation of calcifications and masses. The between-subject design was an analysis of covariance (14), and this design included a three-level categoric-predictor-of-processing method (CLAHE, default, HIW) and baseline analog performance as a continuous predictor. Calcifications and masses were treated separately for digital responses but were pooled for the analog or screen-film responses.

The first analysis step was the evaluation of residuals (15). All results were consistent with the assumptions needed. The next step was the testing of the interaction of lesion type according to processing method. The interaction was not considered to demonstrate a statistically significant difference, so the main effects of lesion type and processing method were tested. All tests were at the $\alpha = .05/3 = .016$ level for each machine.

Author affiliations: From the Departments of Radiology (E.B.C., E.D.P., M.E.B.) and Biostatistics (E.O.K., K.E.M.) and Lineberger Comprehensive Cancer Center (E.B.C., E.D.P., M.E.B., C.M.K.), University of North Carolina School of Medicine, Chapel Hill; Department of Radiology, Mount Sinai Hospital, New York, NY (S.A.F.); Department of Medical Imaging, Sunnybrook and Women's College Health Sciences Center, Toronto, Ontario, Canada (R.A.J.); Department of Radiology, Thomas Jefferson University Hospital, Philadelphia, Pa (A.D.A.M.); Department of Radiology, Good Samaritan Hospital Medical Center, West Islap, NY (M.J.S.); Department of Radiology, Wake Forest University Baptist Medical Center, Winston-Salem, NC (R.I.F., N.L.); Department of Radiology, Duke University Medical Center, Durham, NC (E.L.R., R.W., M.W.); and Department of Radiology, Christ Hospital, Cincinnati, Ohio (M.P.B.).

Author contributions: Guarantor of integrity of entire study, E.D.P.; study concepts and design, E.D.P., K.E.M., M.E.B.; literature research, E.B.C., S.A.F.; clinical studies, E.D.P., M.P.B., R.W., M.J.S., C.M.K., R.I.F., N.L., E.L.R., M.W.; data acquisition, E.B.C., E.D.P., S.A.F., R.A.J., A.D.A.M., M.J.S., C.M.K., R.I.F., N.L., E.L.R., R.W., M.W., M.P.B.; data analysis/interpretation, E.B.C., E.D.P., E.O.K., K.E.M.; statistical analysis, E.O.K., K.E.M.; manuscript preparation, E.B.C., E.D.P., E.O.K., K.E.M.; manuscript revision/review, E.B.C., E.D.P., E.O.K., K.E.M.; manuscript definition of intellectual content, editing, and final version approval, all authors.

References

1. Jackson VP, Hendrick RE, Feig SA, et al. Imaging the radiographically dense breast. Radiology 1993; 188:297–301.
2. Feig SA. The importance of supplementary views to diagnostic accuracy. AJR Am J Roentgenol 1988; 151:40–41.
3. Haus AG, Yaffe MJ. Screen-film and digital mamography image quality and radiation dose considerations. Radiol Clin North Am 2000; 38:871–898.
4. Feig SA, Yaffe MJ. Digital mammography. RadioGraphics 1998; 18:893–901.
5. Shtern F. Digital mammography and related technologies: a perspective from the National Cancer Institute. Radiology 1992; 183:629–630.
6. Pisano ED. Current status of full-field digital mammography. Radiology 2000; 214: 26–28.
7. Pisano ED, Cole EB, Major S, et al. Radiologist preferences for digital mammography display. Radiology 2000; 216:820–830.
8. Taplin SH, Rutter CM, Elmore JG, Seger D, White D, Brenner RJ. Accuracy of screening mammography using single versus independent double interpretation. AJR Am J Roentgenol 2000; 174:1257–1262.
9. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. Acad Radiol 1999; 6:22–33.
10. Rosenberg RD, Hunt William C, Williamson MR, et al. Effects of age, breast density, ethnicity, and estrogen replacement therapy on screening mammographic sensitivity and cancer stage at diagnosis: review of 183134 screening mammograms in Albuquerque, New Mexico. Radiology 1998; 209:511–518.
11. Van Gils CH, Otten JD, Verbeek AL, Hendriks JH, Holland R. Effect of mammographic breast density on breast cancer screening performance: a study in Nijmegen, the Netherlands. J Epidemiol Community Health 1998; 52:267–271.
12. Pisano ED, Cole EB, Hemminger BM, et al. Image processing algorithms for digital mammography: a pictorial essay. RadioGraphics 2000; 20:1479–1491.
13. Chakraborty DP, Winter HL. Free response methodology: alternative analysis and a new observer-performance experiment. Radiology 1990; 174:873–881.
14. Kirk RE. Experimental design: procedures for the behavioral sciences. 3rd ed. Pacific Grove, Calif: Brooks/Cole, 1995.
15. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied regression analysis and other multivariable methods. 3rd ed. Pacific Grove, Calif: Duxbury, 1998.
16. Sickles EA. Mammographic detectability of breast microcalcifications. AJR Am J Roentgenol 1982; 139:913–918.
17. Lewin JM, Hendrick RE, D'Orsi CJ, et al. Comparison of full-field digital mammography with screen-film mammography for cancer detection: results of 4,945 paired examinations. Radiology 2001; 218:873–880.
18. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample. Arch Intern Med 1996; 156:209–213.
19. Berg WA, Campassi C, Langenberg P. Breast Imaging Reporting and Data System: inter- and intraobserver variability in feature analysis and final assessment. AJR Am J Roentgenol 2000; 174:1769–1777.