

# Technology Assessment: Observer study directly compares screen/film to CR mammography

Lynn Fletcher-Heath\*, Anne Richards, Susan Ryan-Kron  
Health Group, Eastman Kodak Company  
Rochester, NY, USA 14650-2033

## ABSTRACT

A new study supports and expands upon a previous reporting that computed radiography (CR) mammography offers as good, or better, image quality than state-of-the-art screen/film mammography. The suitability of CR mammography is explored through qualitative and quantitative study components: feature comparison and cancer detection rates of each modality. Images were collected from 150 normal and 50 biopsy-confirmed subjects representing a range of breast and pathology types. Comparison views were collected without releasing compression, using automatic exposure control on Kodak MIN-R films, followed by CR. Digital images were displayed as both softcopy (S/C) and hardcopy (H/C) for the feature comparison, and S/C for the cancer detection task. The qualitative assessment used preference scores from five board-certified radiologists obtained while viewing 100 screen/film-CR pairs from the cancer subjects for S/C and H/C CR output. Fifteen general image-quality features were rated, and up to 12 additional features were rated for each pair, based on the pathology present. Results demonstrate that CR is equivalent or preferred to conventional mammography for overall image quality (89% S/C, 95% H/C), image contrast (95% S/C, 98% H/C), sharpness (86% S/C, 93% H/C), and noise (94% S/C, 91% H/C). The quantitative objective was satisfied by asking 10 board-certified radiologists to provide a BI-RADS™ score and probability of malignancy per breast for each modality of the 200 cases. At least 28 days passed between observations of the same case. Average sensitivity and specificity was 0.89 and 0.82 for CR and 0.91 and 0.82 for screen/film, respectively.

Keywords: CR mammography, observer study, feature comparison, screen/film mammography, ROC analysis, sensitivity, specificity

## 1. INTRODUCTION

Although analogue X-ray mammography remains the leading radiographic modality to detect breast cancers at an early stage, the transition to digital appears imminent, as there are many advantages associated with digital imaging, including storage, retrieval and shared access. This study was conducted to support the use of computed radiography (CR) for mammography in terms of perceived image quality and the ability of expert mammographers to detect cancer.

### 1.1 Digital debate

Computed radiography has been critically assessed based on intrinsic resolution, noise, and signal-to-noise transfer characteristics and is in worldwide use for general radiography.<sup>1</sup> CR mammography systems have been largely based on general radiography CR systems with various improvements. For example, the Kodak DirectView CR mammography feature is an upgrade that enables digital mammography imaging on any one of Kodak's DirectView CR 850, CR 950 or CR 975 systems. These systems and others contain laser scanners capable of reading the latent images formed on a storage phosphor imaging plate and producing a digital image for projection radiography applications.<sup>2</sup>

A common problem in mammography is that penetration of dense glandular tissue is difficult. This issue is further complicated in screen/film mammography by the limitations of film latitude and its inability to achieve adequate contrast at high and low dense regions simultaneously.<sup>3</sup> Digital image processing offers a solution to this problem by

---

\* [lynn.fletcher-heath@kodak.com](mailto:lynn.fletcher-heath@kodak.com); phone 1 585 588-5936; fax 1 585 722-4771; kodak.com  
Pre-print of Proc. SPIE, 2007; 6515-35

compressing the dynamic range of the image while enhancing the details necessary to make a diagnosis. This may not only impact the ability to diagnose within dense breast tissue but may also reduce the number of exposure retakes.<sup>4</sup>

## **2. MATERIALS AND METHODS: CLINICAL PROTOCOL**

Guidance for the statistical design of this study was obtained through respected literature.<sup>5,6,7</sup>

### **2.1 Data acquisition**

The following sections describe the data acquired for use in both parts of the two-armed study.

#### **2.1.1 Subject data**

All images were acquired from subjects under Institutional Review Board (IRB) approval, from one of six clinical sites (five U.S. and one Canadian site). Images were collected from 275 subjects representing a mixture of screening and diagnostic cases and a range of breast tissue types. The diagnostic patient populations consisted of 50 cases with biopsy-proven cancers.

##### **2.1.1.1 Enrollment inclusion criteria – diagnostic population**

The diagnostic population consisted of women, age 40 or older, who had a BI-RADS assessment of category 4 (suspicious abnormality) or 5 (highly suggestive of malignancy), and who were recommended for breast biopsy, able to have mediolateral oblique (MLO) and craniocaudal (CC) views taken, were otherwise in good general health, and were willing and able to provide a written Informed Consent form.

##### **2.1.1.2 Enrollment inclusion criteria – screening population**

The screening population consisted of women, age 40 or older, which were entering a facility for a routine screening mammogram, were in good general health, and were willing and able to provide a written Informed Consent form.

##### **2.1.1.3 Enrollment exclusion criteria – all subjects**

Women were excluded from the study if they: were under 40 years old, had any possibility of being pregnant, had breast implants, had breasts too large to be positioned on a 24 x 30 cassette, had a personal history of breast cancer treated with a lumpectomy, or were unable or unwilling to provide a written Informed Consent.

##### **2.1.1.4 Image acquisition**

All subjects had the standard CC and MLO views in screen/film mammography as it was typically performed at the facility. The default screen-film systems at these clinical sites were either Kodak Min-R 2000 or Min-R EV. Standard techniques for each site were used based on the calibration of the on-site automatic exposure control and the screen-film system in use. On average, the CR images were generated with less exposure than screen/film images.

Digital CR mammography images were acquired on the Kodak DirectView CR 850 system with the Kodak DirectView CR mammography feature enabled. Two views (CC, MLO) were taken of each breast for each subject. To minimize variations, each CR image was taken immediately following the corresponding screen/film image while the breast was still under compression.

##### **2.1.1.5 Clinical image acceptance criteria**

All CR and screen/film image data was required to pass a set of specified quality tests.

###### **2.1.1.5.1 Positioning**

All CR and screen/film images were required to have: breasts positioned on correct size of film and CR plate to ensure no breast tissue is omitted, one CC view and one MLO view for each breast per subject, nipples in profile on CC and MLO views to ensure that no tissue is superimposed and proper compression applied to ensure breast tissue is separated.

#### **2.1.1.5.2 Exposure and artifacts**

The following required items are based on best practices for mammography as prescribed by the American College of Radiology (ACR). No collimation of the breast could be visible, glandular tissue was a minimum of 0.90 optical density when measured on the screen/film mammogram (every 20<sup>th</sup> subject was measured), no plus density artifacts could be seen in the breast tissue (i.e. static, scratches), no more than 30 minus density artifacts caused by dust and/or dirt could be present in breast tissue and no minus density scratches were to be in any part of the image.

#### **2.1.1.5.3 CR clinical image acceptance criteria**

1. Screen/film clinical images are acceptable for positioning, exposure and compression.
2. CR clinical image was taken with no release of compression.
3. No visible artifacts in breast tissue caused by CR or printer.
4. The CR Reader met QC criteria
  - a. Sites ran an RMI 156 phantom daily prior to imaging subjects.
  - b. Phantom images were reviewed using softcopy. Image review followed the ACR Technologist scoring.<sup>8</sup> Pass Criteria: Visualize 4 fibers, 3 speck groups and 3 masses.
  - c. Phantoms failing to meet the above criteria, forced subject data to be excluded from the reader study, until the next acceptable phantom.

#### **2.1.1.6 Image selection**

Images for use in the reader study were excluded if the CR or screen/film images did not meet the acceptance criteria outlined in section 2.1.1.5. Cases used for the cancer detection task were selected in chronologically to meet the required numbers for breast composition and tumor size distribution. Since partial cases could be used in the comparative feature analysis, cases meeting the criteria were selected chronologically provided a distribution of breast and pathology types were represented. Screenings were chosen randomly.

#### **2.1.1.7 Image preparation**

CR images were harvested and processed based on breast type for Kodak DirectView PTS, and EVP software. CR images were then displayed on a Kodak Carestream mammography workstation with 5MP flat-panel monitors for both arms of the reader study. Additionally, CR images were printed using the Kodak DryView 8900 mammography laser imaging system and Kodak DryView DVM+ mammographic laser imaging film for the comparative feature analysis arm of the study. The printed film density range was from 0.21 to 3.6. All H/C images were displayed on a Control Research Rolloscope M with an average luminance of 10,000 cd/m<sup>2</sup>.

#### **2.1.1.8 Case truth data**

Cases recommended for biopsy, were considered diagnostic cases, including cases that were originally part of the screening population. All positive biopsy reports were truth for malignancy and could be used for both arms of the study. All negative biopsy reports were considered truth for non-malignancy; however, the study design included normal screening cases and positive biopsy cases only. All screening cases, by definition were not biopsied and were therefore considered non-malignant by the clinical standard practices used at each participating data collection site.

## **2.2 Clinical evaluation**

Fifteen board-certified radiologists participated in the two-armed study. All have had extensive experience in mammography with varying levels of digital experience. Ten of the 15 radiologists participated in the cancer detection task and the remaining five participated in the comparative feature portion of the study. Viewing conditions for this

study utilized dedicated mammography view boxes, appropriate masking, and low-ambient lighting. Stray light was controlled and ambient conditions were maintained below 10 cd/m<sup>2</sup>.

## **2.2.1 Reader study design**

### **2.2.1.1 Cancer detection task**

In this arm of the study, 10 radiologists were asked to review 200 cases on each modality, comprising a cancer-to-normal ratio that was unknown to them. All CR images were repeat captures of the same breast compression that was acquired on screen/film. All same-subject, second-modality viewings were separated by a minimum of 28 days. CR images were viewed on a Kodak Carestream mammography workstation for this task. Bias was minimized in the reader study by utilizing many study-design options. One option was by setting up half of the cases to be viewed as film/screen images prior to the CR repeat captures, while the other half were viewed as CR captures before the film/screen images. Other attempts to reduce bias were (1) to have each radiologist begin review at a different insertion point into the master case list in order to reduce fatigue as a plauging reason for any one missed cancer, and (2) to randomly order the master case list, permitting an unpredictable spacing between cancer and normal cases.

Each radiologist was asked to provide the probability of malignancy and BI-RADS scores for each breast. BI-RADS scores were used primarily to assess the sensitivity and specificity of each system. The probability of malignancy was used primarily to assess the receiver operating characteristic (ROC) analysis (ability to detect cancer in terms of false and true indications) for each system.

### **2.2.1.2 Comparative feature test**

Each of the five radiologists reviewed each of the 50 subject pairings of the screen-film and CR images (same breast, same view, from the same patient). All observer data for this study was obtained given a comparative feature analysis of both screen/film to S/C CR and screen/film to H/C CR matching views for one view of the subject with cancer in the chosen view. The film image was placed on the left and the CR image on the right. Left/right randomization was not done because, the images were easy to differentiate, and it would have been very difficult to swap the S/C device from left to right.

Scores were provided in a total of 15 categories for each pairing, according to their impressions of overall noise, contrast, sharpness and image quality, and for the noise, contrast and sharpness in each of the breast regions (periphery and skin-line, parenchyma and fatty tissue). A rating scale from -4 to +4 was used, with a score of -4 corresponding to a definite preference for the screen/film image and a score of +4 indicating a definite preference for the CR image. A score of 0 indicated the CR image was equivalent to the conventional analog radiographic film (Table 1). There were 50 scores for each of 15 image-quality attributes. In addition, the radiologist was required to score each case based on the presence of suspicious abnormalities. Four attributes were scored if architectural distortions were present, five attributes were scored for each subject that exhibited a suspicious mass, and three attributes were scored for each case that exhibited suspicious microcalcifications (MCCs). A total of 62 suspicious findings comprising 50 masses: overlapping characterizations specified 9 as circumscribed, 12 as architectural distortions, 4 as having focal asymmetry, 14 as ill-defined, and 22 as spiculated masses, and 12 MCCs (9 pleomorphic or heterogeneous, and one each of amorphous, benign, and branching). Scores for suspicious findings totaled 431, categorizing: margin sharpness, lesion contrast, density and ability to diagnose masses; shape, count, sharpness, distribution and ability to diagnose MCCs; density, parenchymal edge and ability to diagnose architectural distortions.

All abnormalities categorized as MCCs, masses and architectural distortions were reviewed and scored using the same feature-analysis rating scale.

Table 1. Feature-analysis rating scale.

RATING	EXPLANATION
- 4	Film image is markedly better
- 3	Film image is moderately better
- 2	Film image is mildly better
- 1	Film image is minimally better
0	No difference between film image and CR image
+ 1	CR image is minimally better
+ 2	CR image is mildly better
+ 3	CR image is moderately better
+ 4	CR image is markedly better

### 2.2.2 Analysis methods

Several analysis methods were used to compare the performance of the CR mammography system to the screen/film mammography system. ROC analysis and sensitivity/specificity measurements were used in the quantitative study and frequency histograms were used to plot data trends of categorical variables for the qualitative analysis.

#### 2.2.2.1 ROC analysis

ROC curve analysis is a well-established method to compare the accuracy of two or more imaging modalities in which observers serve as detectors and decision makers.<sup>9,10</sup> The multiple-reader, multiple case (MRMC) study design was used, in which each case undergoes each of the diagnostic tests, every reader reads every case, and random effects models are used to account for both case and reader variability. The results can then be generalized to the population of cases and the population of readers.

In this study, ROC curve analysis was conducted based on the probability of malignancy, using the MRMC software LABMRMC provided by the University of Chicago, for each reader in the study and for the combined results of all 10 readers. This software is based on the Dorfman-Berbaum-Metz (DBM) approach, the most frequently used method for analyzing multireader ROC data,<sup>11</sup> assuming a mixed-effects analysis of variance (ANOVA) model.<sup>12</sup> A breast level analysis, in which scores for each breast were analyzed, is presented in this paper. The correlation between the probabilities of cancer between the breasts was examined and adjusted by the model. The areas under the ROC curves for the CR and the screen/film systems were calculated and compared using this model.

#### 2.2.2.2 Sensitivity and specificity

Sensitivity was based on the BI-RADS rating provided by each observer for each breast. The rating system is listed in Table 2. The presence of disease was defined as a breast with biopsy pathology positive for breast cancer. Sensitivity to breast cancer of CR and screen/film was compared within each individual reader, then for all readers combined. The proportions of cases correctly classified by a reader on a specific modality (true positives among cancer cases by that modality divided by all known cancers) were compared.

Specificity of breast cancer detection was calculated within each observer, for each modality, and reported for each breast. All subjects not recommended for biopsy for either breast were considered disease negative. The proportions of disease-negative cases correctly classified by a reader on a specific modality (true negatives by that modality divided by all the disease-negative cases) were compared. Combined results for all 10 readers for sensitivity and specificity measurements are presented using a generalized linear mixed model of the covariance structure of correlated data, allowing the correlation among the repeated measurements in the same subject to be incorporated into the estimates of parameters and standard errors.

#### 2.2.2.3 Comparative feature analysis

Comparative ratings for the 15 general image quality attributes were averaged for all five expert observers. In addition, suspicious masses, microcalcifications, and architectural distortions were evaluated for pairs that contained them, for up

to 12 additional ratings per pair. For each attribute evaluated, a score was calculated for each image pair as the mean reading of all observers.

Table 2. BI-RADS categories and descriptions

BI-RADS Category	Description
0	Needs Additional Imaging Evaluation: Considered positive mammogram result
1	Negative mammogram: Considered negative mammogram result
2	Benign finding: Considered negative mammogram result
3	Probable benign finding: Considered negative mammogram result
4	Suspicious abnormality: Considered positive mammogram result
5	Highly suggestive of malignancy: Considered positive mammogram result

### 3. RESULTS

#### 3.1 ROC analysis

The ROC curve for each modality was calculated using the probabilities of malignancy (0 – 100%) given by each observer after independent interpretation of each four-view case. The DBM method was used in the analysis.

Table 3 includes a comparison of areas under the ROC curves for each reader on each modality. Six readers had a higher performance rating on screen/film than CR, three readers were higher on CR than screen/film and one reader had a matched performance on each modality. For two readers the 95% confidence bound of the difference between the modalities was greater than the 0.10 limit. The MRMC analysis for all cases and all readers showed an ROC curve area of 0.92 for the CR system. The 95% upper confidence bound of the difference was 0.7, which is within the 0.10 limit. The p-value of non-inferiority was 0.004.

Table 3. ROC curve areas for individual readers and for all readers, combined using multiple-reader, multiple-case software (LABMRMC). P-value based on one-sided, two-sample t-test for non-inferiority with an equivalence limit of 0.1.

n=398 Reader	Screen/Film System		CR System		Mean Difference	95% Confidence Bound of Difference	
	ROC Area	95% CI	ROC Area	95% CI			
1	0.93	(0.87, 0.99)	0.92	(0.86, 0.97)	0.01	0.08	
2	0.95	(0.90, 0.99)	0.96	(0.94, 0.99)	-0.01	0.04	
3	0.94	(0.90, 0.98)	0.96	(0.93, 0.99)	-0.02	0.02	
4	0.89	(0.82, 0.97)	0.89	(0.81, 0.96)	0.00	0.09	
5	0.95	(0.90, 0.99)	0.92	(0.87, 0.98)	0.03	0.09	
6	0.93	(0.88, 0.98)	0.95	(0.91, 0.99)	-0.02	0.03	
7	0.93	(0.87, 0.99)	0.80	(0.68, 0.92)	0.13	0.24	
8	0.96	(0.92, 0.99)	0.93	(0.89, 0.98)	0.03	0.08	
9	0.93	(0.89, 0.98)	0.90	(0.84, 0.96)	0.03	0.09	
10	0.96	(0.93, 0.99)	0.95	(0.91, 0.99)	0.01	0.05	
All	0.94	(0.90, 0.98)	0.92	(0.88, 0.97)	0.02	0.07	p-value 0.004

Overall ROC curve plots were generated by averaging the fitted true positive fractions at the fixed false-positive fractions of the individual reader’s ROC curve for each modality.

The ROC curve represents a system’s complete sensitivity (true positive rate) and specificity (1–false-positive rate) range and offers an accuracy comparison between two imaging systems without the influence of a specific threshold cut-off value. To prove efficacy of the CR mammography system, the ROC area must not be significantly lower than the

ROC area of the film system (more than 10%). This study measured the ROC areas to be within 2%, indicating that the CR system and screen/film system performed similarly.

**Overall ROC Curves by Probability of Malignancy: Breast Level Analysis**

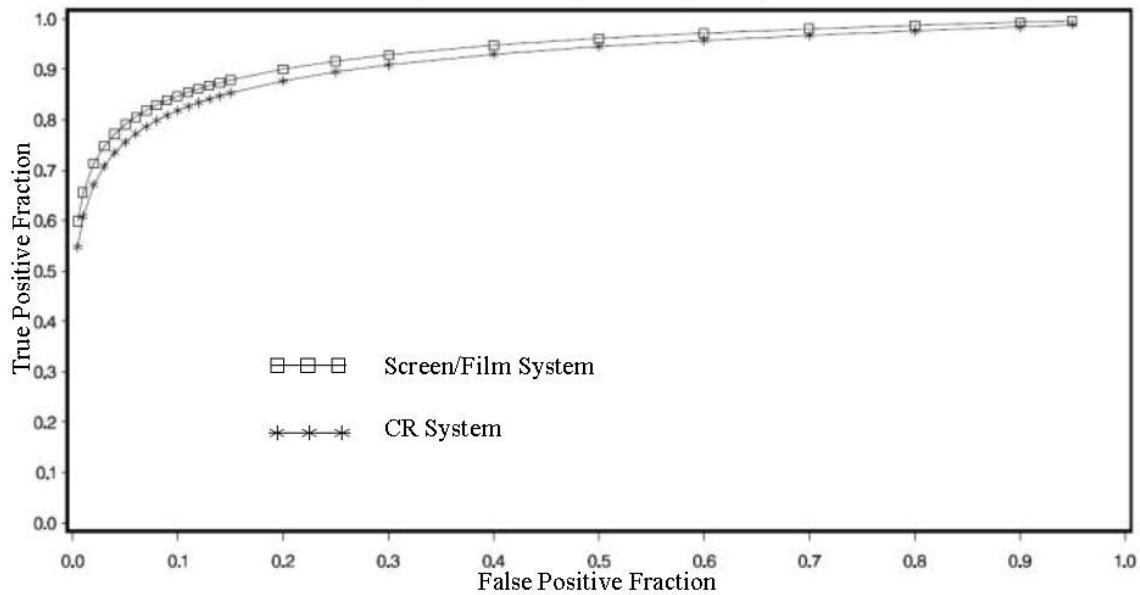


Fig. 1. ROC curves for 10 readers combined for each of the 398 breasts for CR and screen/film.

### 3.2 Sensitivity and specificity

A summary of the sensitivity and specificity calculations are listed in Table 4. The sensitivity of the CR system proved slightly lower than the screen/film system. The 95% confidence bound of the difference between the two systems (0.02) was 0.11, which is very near the tolerable margin of 0.10. The p-value representing non-inferiority was 0.078. The specificity of the CR system proved to be the same as for the screen/film system. The p-value of non-inferiority was less than 0.001.

Table 4. Overall sensitivity and specificity for two mammographic modalities. BI-RADS score for every breast was included. Of the 50 subjects with cancer, two had bi-lateral cancer.

Estimated Mean Sensitivity							
Number of Breasts	Screen/Film System	95% CI	CR System	95% CI	Difference	95% Confidence Bound	p-value
52	0.91	(0.88, 0.93)	0.89	(0.86, 0.92)	0.02	0.11	0.078

Estimated Mean Specificity							
Number of Breasts	Screen/Film System	95% CI	CR System	95% CI	Difference	95% Confidence Bound	p-value
346	0.82	(0.81, 0.83)	0.82	(0.81, 0.84)	0	0.04	< 0.001

### 3.3 Comparative feature analysis

Statistical treatment of the observer data demonstrates a few interesting points, which will be correlated with trend analysis.<sup>13,14</sup> Table 5 lists average values for each attribute for 50 image pairs taken from the biopsy positive subjects. Five radiologists viewed and rated aspects of each pair, screen/film to CR, where CR was presented in both and S/C and H/C formats, for a total of 100 pairs. General image-quality attributes were rated for each image pair by each

radiologist. Each radiologist who considered a lesion to be suspicious rated attributes related to that pathology. Recall values above zero indicate a preference towards the CR image, values less than zero indicate a preference toward the screen/film image, and zero indicates equivalence.

Highlighting key attributes:

- Both CR H/C and S/C were preferred over film for Overall Image Quality. The mean scores and (std. dev.) were S/C 0.9 (0.38) and H/C 1.2 (0.4). A graphical depiction of this attribute is shown in Fig. 2.
- Volume of tissue at the chest wall was comparable between CR and screen/film.
- Tissue visibility at the skin-line yielded a strong preference toward the CR images: S/C 1.8 (0.45) and H/C 2.2 (0.46).
- Mean contrast scores for both S/C and H/C showed a preference for CR over screen/film. Mean (std. dev.) scores for: S/C were 0.7 (0.32), 0.6 (0.26), and 0.5 (0.35); for H/C were 1.2 (0.38), 1.0 (0.38) and 0.8 (0.45) for the breast periphery, fatty tissue and glandular tissue respectively.
- Mean detail/sharpness scores for both S/C and H/C also showed a preference for CR over screen/film. Mean (std. dev.) scores for: S/C were 0.8 (0.35), 0.7 (0.26), and 0.6 (0.32); for H/C were 1.3 (0.34), 1.1 (0.33) and 0.8 (0.40) for the breast periphery, fatty tissue and glandular tissue respectively.
- Mean noise scores for both S/C and H/C also showed a weaker preference for CR over screen/film. Mean (std. dev.) scores for: S/C were 0.3 (0.21), 0.4 (0.17), and 0.3 (0.22); for H/C were 0.6 (0.25), 0.6 (0.25) and 0.5 (0.23) for the breast periphery, fatty tissue and glandular tissue respectively.
- All means of attributes describing the appearance of pathology were also above zero, indicating an average preference toward CR for both S/C and H/C.
- H/C CR performed better overall than S/C CR, which in turn, on average, out-performed screen/film.

Graphical representations of some of the attributes listed are shown in Figs. 2 – 12. One of the most general attributes was overall image quality. This was one of the 15 attributes rated for every image pair for both H/C and S/C CR, when compared to the screen/film image. Fig. 2 demonstrates how the overall image quality was preferred on the CR image 64% of the time when CR was displayed on S/C and 76% when CR was displayed on H/C. The screen/film image was preferred 8% of the time with the S/C comparison and 3.6% with the H/C comparison. The pairs were rated equivalently 28% and 20.4% of the time, in the S/C and H/C comparisons, respectively.

Further breakdown by breast type, of the data in Fig. 2, can be seen in Fig. 3. Ratings are collapsed to preference for screen/film (< 0), no preference (= 0) and preference for CR (> 0). Although the percent of the number of pairs for each category is displayed on the y-axis, there are varying numbers of pairs within each breast type. Above each bar is the total number of pairs within each rating and breast type category. Note that there is a general preference for CR, regardless of breast type, with a slight preference to the H/C CR image for fatty and scattered fibroglandular breasts.

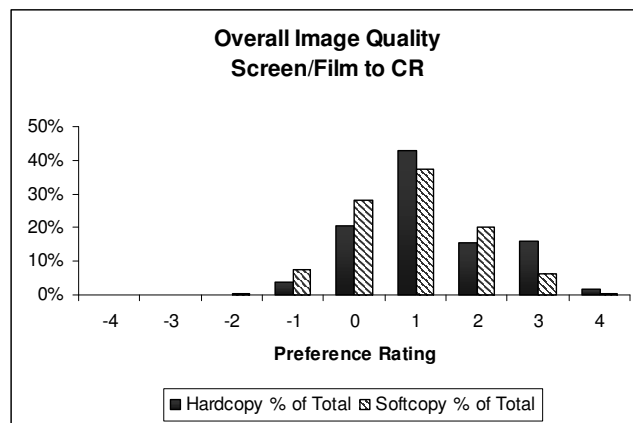


Fig. 2. Distribution of 500 diagnostic preference ratings (250 H/C and 250 S/C) based on overall image quality for 50 paired comparisons of cancer cases.

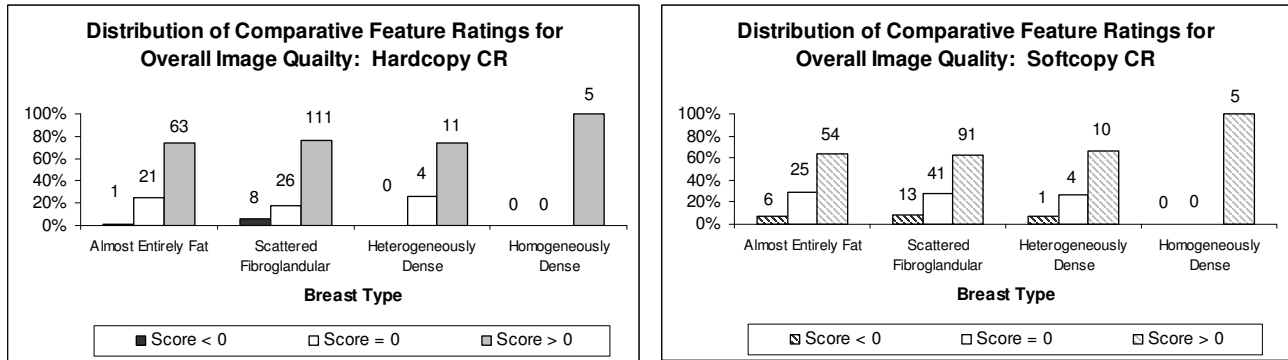


Fig. 3. Overall image quality subdivided by breast type for H/C CR (left) and S/C CR (right).

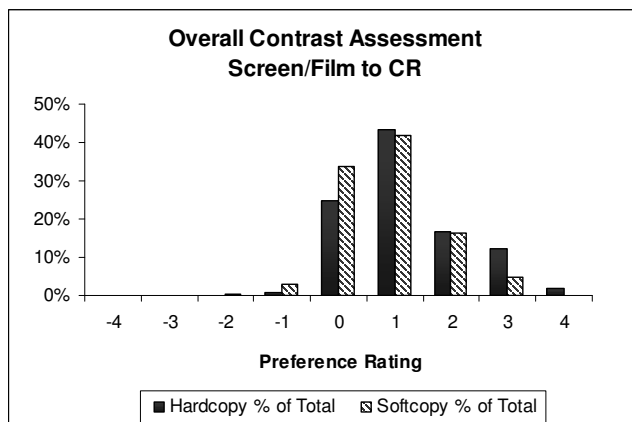


Fig. 4. Distribution of diagnostic preference ratings (250 H/C and 250 S/C) based on overall contrast assessment for 50 paired comparisons of cancer cases.

Figure 4 demonstrates how the overall contrast was preferred on the CR image 63.2% of the time when CR was displayed on S/C and 74.4% of the time when CR was displayed on H/C. The screen/film image was preferred 3.2% of the time with the S/C comparison and 0.8% with the H/C comparison. The pairs were rated equivalently 33.6% and 24.8% of the time, in the S/C and H/C comparisons, respectively.

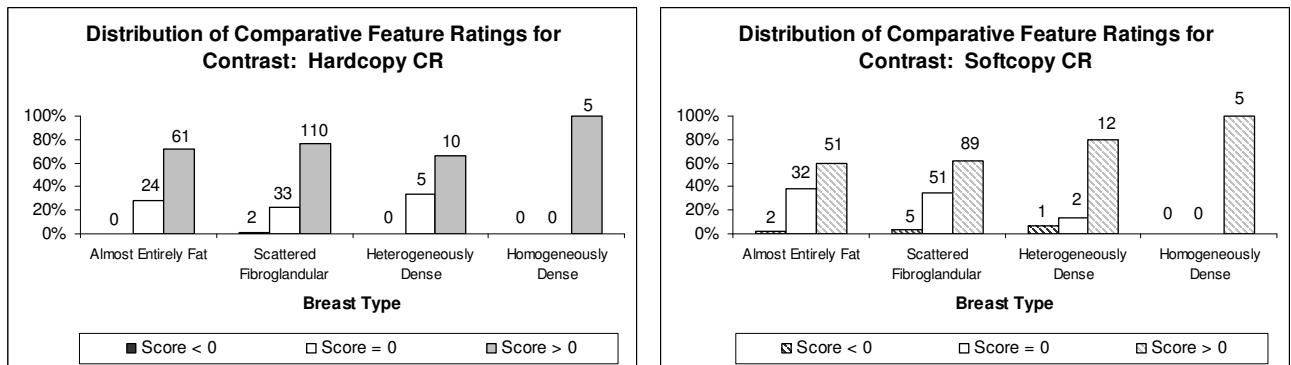


Fig. 5. Overall contrast assessment subdivided by breast type for H/C CR (left) and S/C CR (right).

Table 5. Summary of mean scores for comparative feature evaluation.

Attributes	CR Softcopy		CR Hardcopy	
	n	Mean (Std Dev) (a,b)	n	Mean (Std Dev) (a,b)
Overall Image Quality	50	0.9 (0.38)	50	1.2 (0.40)
Volume of Tissue at Chest Wall	50	0.2 (0.29)	50	0.4 (0.31)
Tissue Visibility of Skin Line	50	1.8 (0.45)	50	2.2 (0.46)
<b>Contrast</b>				
Breast Periphery	50 (c)	0.7 (0.32)	50	1.2 (0.38)
Fatty Tissue	50	0.6 (0.26)	50	1.0 (0.35)
Glandular Tissue	50	0.5 (0.35)	50	0.8 (0.45)
Overall Image Contrast	50	0.8 (0.35)	50	1.2 (0.39)
<b>Detail/Sharpness</b>				
Breast Periphery	50	0.8 (0.35)	50	1.3 (0.34)
Fatty Tissue	50	0.7 (0.26)	50	1.1 (0.33)
Glandular Tissue	50	0.6 (0.32)	50	0.8 (0.40)
Overall Image Detail/Sharpness	50	0.8 (0.31)	50	1.1 (0.41)
<b>Noise</b>				
Breast Periphery	50	0.3 (0.21)	50	0.6 (0.25)
Fatty Tissue	50	0.4 (0.17)	50	0.6 (0.25)
Glandular Tissue	50	0.3 (0.22)	50	0.5 (0.23)
Overall Noise	50	0.3 (0.16)	50	0.6 (0.23)
<b>Masses</b>				
Ability to Assess Size of Mass	44 (d)	0.2 (0.32)	46	0.2 (0.43)
Ability to Assess Shape of Mass	44	0.2 (0.30)	46	0.3 (0.49)
Ability to Assess Margin of Mass	44	0.4 (0.40)	46	0.4 (0.54)
Overall Conspicuity	44	0.2 (0.34)	46	0.3 (0.57)
Ability to Make Diagnosis	44	0.2 (0.27)	46	0.2 (0.39)
<b>Microcalcifications</b>				
Ability to Assess Shape of MCC	45 (d)	0.3 (1.09)	41	0.4 (0.77)
Conspicuity	45	0.4 (0.88)	41	0.5 (0.81)
Ability to Make Diagnosis	45	0.3 (0.80)	41	0.4 (0.70)
<b>Architectural Distortions</b>				
Parenchymal Edge Distortion	38 (d)	0.5 (0.60)	39	0.4 (1.10)
Conspicuity	38	0.2 (0.40)	39	0.5 (1.07)
Ability to Make Diagnosis	38	0.2 (0.39)	39	0.4 (0.95)
<p>Note: (a) Scores based upon a scale from -4 to 4 where -4 represents screen-film as being markedly superior and 4 represents CR as being markedly superior.</p> <p>(b) Score for each image pair was the average of all readings for this image pair across all readers.</p> <p>(c) Average of five radiologists' ratings for all general image quality attributes.</p> <p>(d) Total number of pairs with between one and five ratings each.</p>				

Contrast assessment is subdivided by breast type in Fig. 5. Again, there is a general preference for CR, regardless of breast type, with a slight preference to the H/C CR image for Fatty and Scattered Fibroglandular breasts. Preference toward S/C CR for Heterogeneously Dense breasts was not statistically significant in this data.

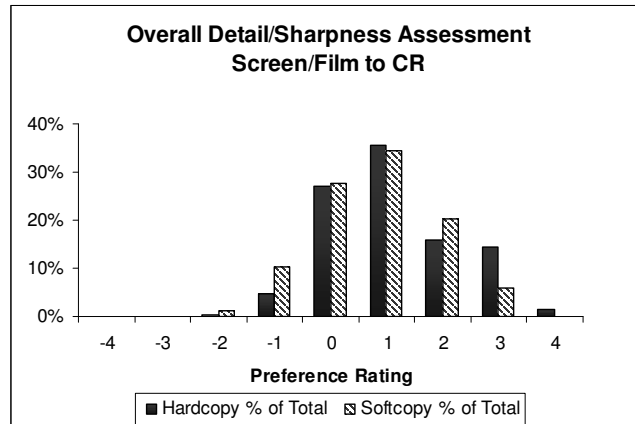


Fig. 6. Distribution of 500 diagnostic preference ratings (250 H/C and 250 S/C) based on detail/sharpness assessment for 50 paired comparisons of cancer cases.

Figure 6 demonstrates another clear preference for the CR image. The assessment of details and sharpness was preferred on the CR image 60.8% of the time when CR was displayed on S/C and 67.6% of the time when CR was displayed on H/C. The screen/film image was preferred 11.6% of the time with the S/C comparison and 5.2% with the H/C comparison. The pairs were rated equivalently 27.6% and 27.2% of the time, in the S/C and H/C comparisons respectively. Further division of the data into breast type categories is shown in Fig. 7 where there continues to be a preference towards H/C CR, then to S/C CR then to the screen/film image.

Preference towards CR for noise assessment is less profound with 20.8% and 38% preference for S/C and H/C CR, respectively (Fig. 8). The majority of the ratings showed equivalency, 74.4% and 55.2% for the S/C and H/C assessments, respectively. Screen/film was preferred 4.8% and 6.8% of the ratings. Figure 9 further demonstrates that the noise was generally observed as equivalent between the modalities, although there is a trend for the H/C CR image to outperform the S/C CR image.

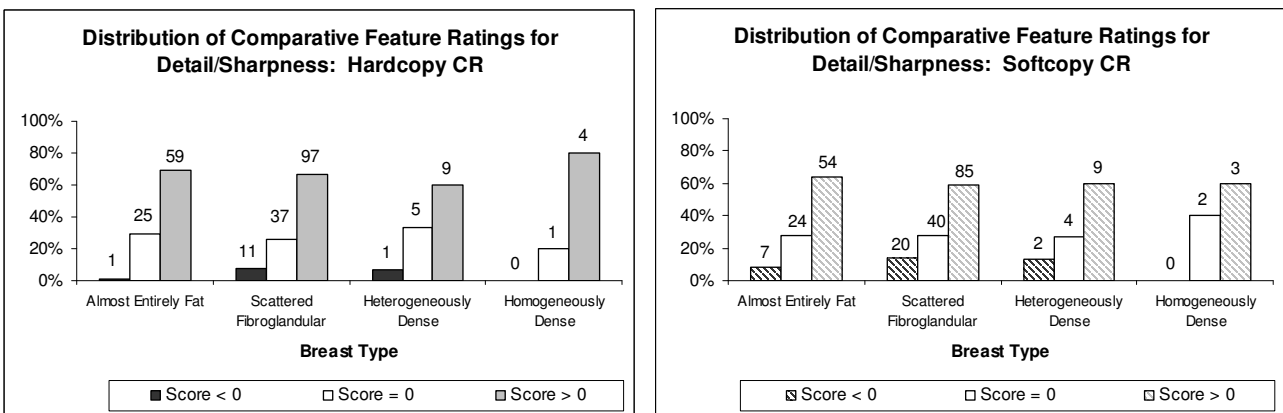


Fig. 7. Overall detail/sharpness assessment subdivided by breast type for H/C CR (left) and S/C CR (right).

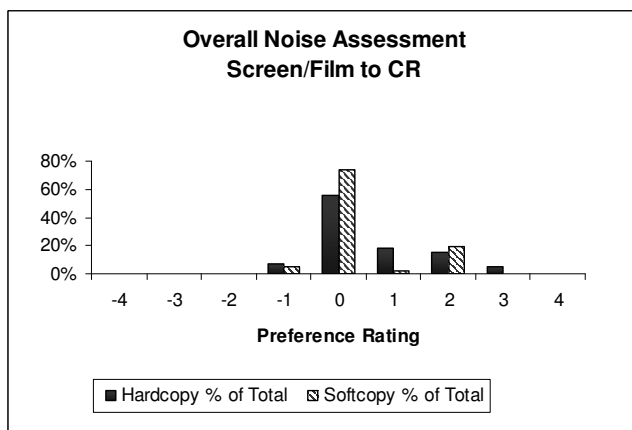


Fig. 8. Distribution of 500 diagnostic preference ratings (250 H/C and 250 S/C) based on noise assessment for 50 paired comparisons of cancer cases.

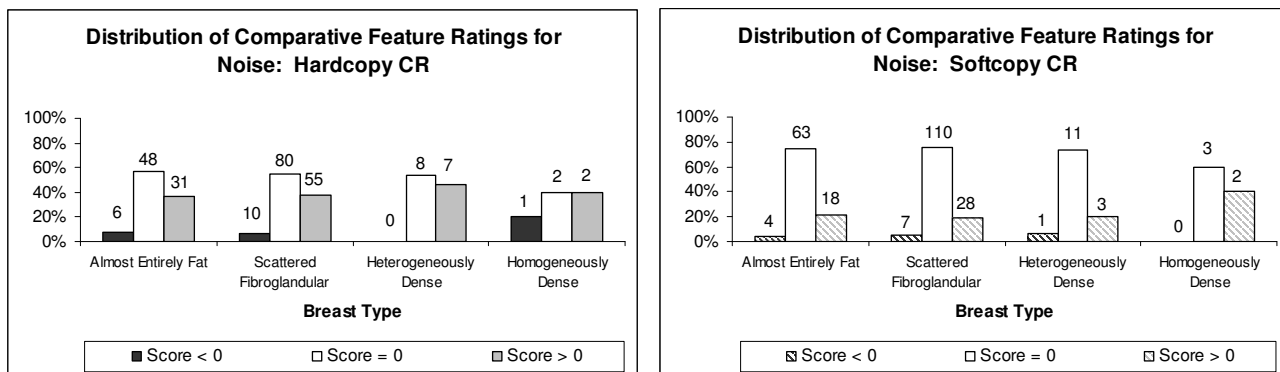


Fig. 9. Noise assessment subdivided by breast type for H/C CR (left) and S/C CR (right).

A further separation of only the pathologies presents another important aspect of the required diagnostic quality, shown in the next set of figures. Eleven additional attributes, or a subset of the eleven, were rated for each image pair. Observers were asked to rate suspicious findings, and each determined what was suspicious, so there were an average of 35 masses, 31 MCCs and 19 architectural distortions. Each suspicious finding was subjected to a series of relevant questions corresponding to the characteristics of each type of finding. Trends may be observed for each attribute, each type of finding, and each modality. Figures 10, 11 and 12 demonstrate a sample of sub-categorical attributes of the finding.

The mass attributes were the ability to assess shape, margin sharpness, and size, as well as conspicuity and the ability to diagnose. Figure 10 demonstrates two of these five features, conspicuity and ability to assess the margin of the mass. Statistics for all features are listed in Table 5. Note the largest percentage of scores demonstrated equivalency (conspicuity: S/C 82.1%, H/C 70.2%)(margin: S/C 57.8%, H/C 52.3%), with a preference towards CR (conspicuity: S/C 12.1%, H/C 22.5%)(margin: S/C 36.4%, H/C 34.8%). Fewer preferences were noted toward screen/film (conspicuity: S/C 5.8%, H/C 7.3%)(margin: S/C 5.8%, H/C 12.9%).

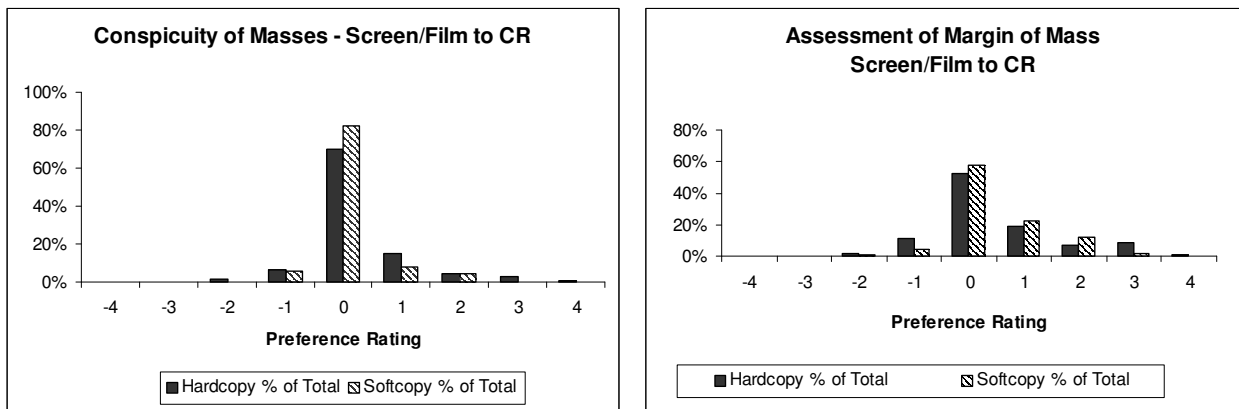


Fig. 10. Distribution of 350 diagnostic preference ratings for each mass attribute, based on conspicuity (left) and margin of masses (right) for the cancer cases containing masses.

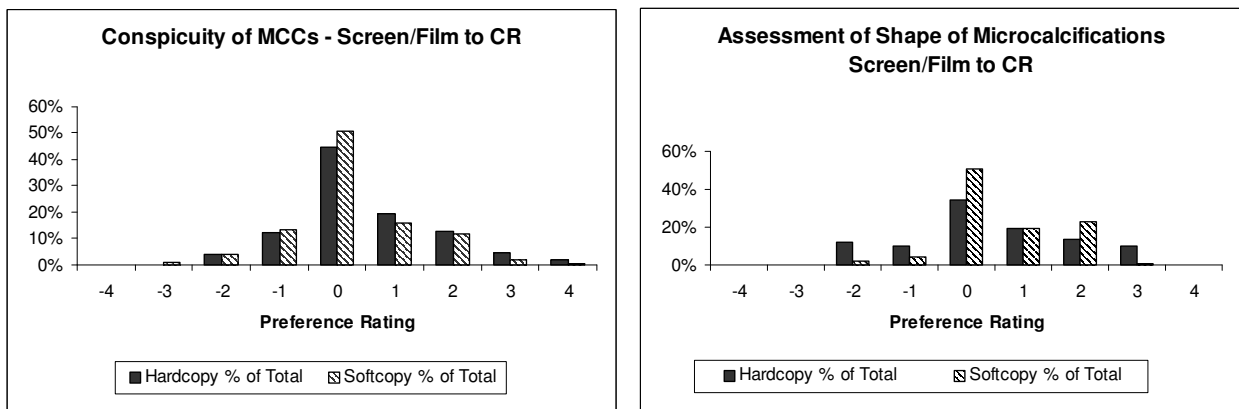


Fig. 11. Distribution of 310 diagnostic preference ratings for each MCC attribute based on conspicuity (left) and assessment of shape (right) for the cancer cases containing MCCs.

The MCC attributes were recorded as the ability to assess shape, conspicuity and the ability to diagnose. Figure 11 demonstrates the trend of conspicuity and the ability to assess shape. As with masses, the largest percentage of scores demonstrated equivalency (conspicuity: S/C 51.0%, H/C 44.6%)(shape: S/C 43.6.8%, H/C 41.2%), with a slight preference toward CR for conspicuity: S/C 30.0%, H/C 39.2%. Assessment of shape demonstrated a preference towards H/C CR (37.2%) but no clear preference for S/C CR or screen/film (28.8% S/F and 27.6% S/C CR). Fewer preferences were noted towards screen/film (conspicuity: S/C 19.0%, H/C 16.2%)(margin: H/C 21.6%).

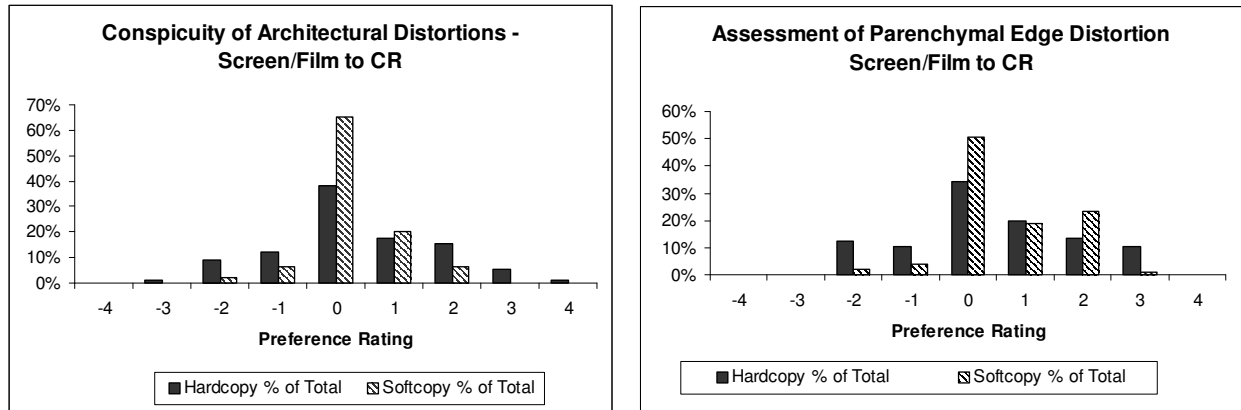


Fig. 12. Distribution of 192 diagnostic preference ratings for each architectural distortion attribute based on conspicuity (left) and edge distortion (right) of architectural distortions for the cancer cases containing them.

The ratings obtained regarding architectural distortions were in assessment of the parenchymal edge distortion, conspicuity and the ability to diagnose. Figure 12 demonstrates conspicuity and ability to assess the parenchymal edge distortion. Statistics for all features are listed in Table 5. Equivalency was demonstrated much of the time for conspicuity: S/C 65.3%, H/C 38.1% and for the edge distortion: S/C 50.5%, H/C 34.0%). A preference was noted for CR (conspicuity: S/C 26.3%, H/C 39.2%)(edge distortion: S/C 43.2%, H/C 43.3%). Again, fewer preferences were noted for screen/film overall (conspicuity: S/C 8.4%, H/C 22.7%)(edge distortion: S/C 6.3%, H/C 22.7%). One key difference existed in the assessment of architectural distortions rather than with masses, MCCs or with general image-quality attributes, and that was S/C CR being preferred over H/C CR.

#### 4. DISCUSSION AND CONCLUSIONS

A clinical study has been described that provides evidence of two key points: (1) diagnostic accuracy, based on the area under the ROC curves, sensitivity, and specificity was observed as similar between state-of-the-art mammographic film systems and CR mammography, and (2) a diagnostic-look preference for the digital CR mammography images was prevalent over screen/film images of the same breast compression for all image-quality attributes, while maintaining a minimum equivalent preference between the two systems for all attributes related to pathology.

The observed values of the area under the ROC curves for the enriched data set were within 0.02, with an upper 95% CI of 0.07. Sensitivity and specificity were also shown to be very similar to screen/film with a difference of sensitivity of 0.02 (upper 95% CI = 0.11) and specificity of 0.00 (upper 95% CI = 0.05). The area under ROC analyses demonstrates non-inferiority of the CR system to the screen/film system based on the negated null hypothesis that the ROC area of film was greater than 0.10 more than the ROC area of the CR digital system (p-value of non-inferiority = 0.004). Sensitivity and specificity calculations also support non-inferiority, as each null hypothesis was defined as the expectation that the film system would prove to be 0.10 better than the CR system. The p-values of non-inferiority were calculated as 0.078 for sensitivity and < 0.001 for specificity.

One study limitation requires discussion. Only 20 of 250 cases used in this study were of women with dense breasts. This makes it difficult to further the statement made by the Digital vs. Film Mammography in the Digital Mammographic Screening Trial (DMIST) that women with dense breasts benefit from digital imaging. Correlative observations may be made, however. Figures 3, 5, 7, and 9 do show that heterogeneously and homogeneously dense breasts demonstrate a consistent skew toward a CR preference. Grouping display types together (S/C + H/C), Fig. 3 shows that for overall image quality, only one comparison out of 30 for dense breasts, showed a preference towards screen/film. Similarly, for overall contrast (Fig. 5) only one comparison of 30 for dense breasts, showed a preference towards screen/film. Detail/sharpness (3 out of 30 preferred S/F) and noise (2 out of 30 preferred S/F) demonstrated similar skew, however, due to numbers of available cases in the dense category, it is difficult to assess if a benefit exists when using a digital-imaging modality.

Image processing plays a significant role in the presentation of optimal renderings. The on-going discussion about required acceptance and verification tests for digital mammography systems is warranted. Judging a system by its final output can be done easily for screen/film imaging systems where final assessment is done on the developed film. For digital systems, the path is not as clear because image-processing algorithms are generally designed to work optimally for anatomy, not phantoms. This study, performed with expert mammographers, provides statistics of diagnostic accuracy as a measure of system performance. Results from this study show diagnostic equivalence and visual preference trends in favor of the use of CR for mammography.

## ACKNOWLEDGMENTS

The authors sincerely thank the technologists, physicists and mammographers of the following clinics for their support during this study - Advocate Illinois Masonic Medical Center (Chicago, IL), Alpena Regional Medical Center (Alpena, MI), Memorial Health Care System (Chattanooga, TN), MeritCare Health System (Fargo, ND), Providence Hospital (Washington, DC) and St. Joseph's Health Care (London, ON, Canada). Special thanks for the time and effort of the radiologists who participated in the study, namely Drs. Anne Archer, Peggy Avagliano, Ermelinda Bonnacio, Stamatia Destounis, Bill Eklund, Fabius Fox, Kathleen Gordon, Lydia Liao, Michael Racenstein, Sughra Raza, Michelle Rossman, Barbara Savader, Rene Shumak, Ann Swinford, and Christine Watt.

## REFERENCES

- 
1. E. Samei and M. Flynn. Physical Measures of Image Quality in Photostimulable Phosphor Radiographic Systems. *Proc SPIE*, 1997; 3032:328-338.
  2. T. Bogucki, D. Trauernicht and T. Kocher. Characteristics of a Storage Phosphor System for Medical Imaging. Health Sciences Division, Eastman Kodak Company, Rochester, NY 14650, 1995. Technical and Scientific Monograph 6.
  3. E. Pisano, C. Gatsonis, E. Hendrick, M. Yaffe, J. Baum, et. al. Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. *The New England Journal of Medicine*, 2005; 353(17): 1773-1783.
  4. L. Fletcher-Heath, A. Richards and S. Ryan-Kron. Investigation of Diagnostic and Image Quality Attributes: Comparison of Screen-Film to CR Mammography. *Proc SPIE*, 2006; 6142: 1016-1026.
  5. N. Obuchowski, "Multireader receiver operating characteristic studies: a comparison of study designs," *Academic Radiology*, 2(8): 709-716, 1995.
  6. N. Obuchowski, S. Beiden, K. Berbaum, S. Hillis, H. Ishwaran, H. Song, R. Wagner. "Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods". *Academic Radiology*, 11(9): 980-995, 2004.
  7. H. Rockette, Campbell WL, Britton CA, Holbert JM, King JL, and Gur D. "Empiric assessment of parameters that affect the design of multireader receiver operating characteristic studies," *Academic Radiology*, 6: 723-729, 1999.
  8. The American College of Radiology Mammography Quality Control Manual 1999 edition, 167-171 (radiologic technologist section).
  9. A. Hanley and B. McNeil. The meaning and use of the area under the Receiver Operator Characteristic (ROC) curve. *Radiology*, 1982; 143: 29-36.
  10. A. Hanley and B. McNeil. A method of comparing the areas under Receiver Operating Characteristic curves derived from the same cases. *Radiology*, 1983; 148: 839-843.
  11. D. Dorfman, K. Berbaum and C. Metz. ROC Rating analysis generalization to the population of readers and cases with the jackknife method. *Invest. Radiol.* 1992; 27: 723.
  12. S. Hillis, N. Obuchowski, K. Schartz and K. Berbaum. Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification. *Academic Radiology* 2005; 12:1534-1541.
  13. D. Clason and T. Dormody. Analyzing data measure by individual Likert-Type items. *Journal of Agriculture Education*, 1994; 4:31-35.
  14. Jaccard, James and Choi K. Wan (1996). LISREL approaches to interaction effects in multiple regression. Thousand Oaks, CA: Sage Publications.