Original Article | Breast Imaging

eISSN 2005-8330
https://doi.org/10.3348/kjr.2018.0193
Korean J Radiol 2019;20(2):218-224

Korean Journal of Radiology
KJR

Interpretive Performance and Inter-Observer Agreement on Digital Mammography Test Sets

Sung Hun Kim, MD[1], Eun Hye Lee, MD[2], Jae Kwan Jun, MD[3], You Me Kim, MD[4], Yun-Woo Chang, MD[5], Jin Hwa Lee, MD[6], Hye-Won Kim, MD[7], Eun Jung Choi, MD[8];
the Alliance for Breast Cancer Screening in Korea (ABCS-K)

[1]Department of Radiology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea; [2]Department of Radiology, Soonchunhyang University Hospital Bucheon, Soonchunhyang University College of Medicine, Bucheon, Korea; [3]National Cancer Control Institute, National Cancer Center, Goyang, Korea; [4]Department of Radiology, Dankook University Hospital, Dankook University College of Medicine, Cheonan, Korea; [5]Department of Radiology, Soonchunhyang University Hospital, Soonchunhyang University College of Medicine, Seoul, Korea; [6]Department of Radiology, Dong-A University Hospital, Busan, Korea; [7]Department of Radiology, Wonkwang University Hospital, Wonkwang University School of Medicine, Iksan, Korea; [8]Department of Radiology, Chonbuk National University Hospital, Jeonju, Korea

Objective: To evaluate the interpretive performance and inter-observer agreement on digital mammographs among radiologists and to investigate whether radiologist characteristics affect performance and agreement.

Materials and Methods: The test sets consisted of full-field digital mammograms and contained 12 cancer cases among 1000 total cases. Twelve radiologists independently interpreted all mammograms. Performance indicators included the recall rate, cancer detection rate (CDR), positive predictive value (PPV), sensitivity, specificity, false positive rate (FPR), and area under the receiver operating characteristic curve (AUC). Inter-radiologist agreement was measured. The reporting radiologist characteristics included number of years of experience interpreting mammography, fellowship training in breast imaging, and annual volume of mammography interpretation.

Results: The mean and range of interpretive performance were as follows: recall rate, 7.5% (3.3–10.2%); CDR, 10.6 (8.0–12.0 per 1000 examinations); PPV, 15.9% (8.8–33.3%); sensitivity, 88.2% (66.7–100%); specificity, 93.5% (90.6–97.8%); FPR, 6.5% (2.2–9.4%); and AUC, 0.93 (0.82–0.99). Radiologists who annually interpreted more than 3000 screening mammograms tended to exhibit higher CDRs and sensitivities than those who interpreted fewer than 3000 mammograms (p = 0.064). The inter-radiologist agreement showed a percent agreement of 77.2–88.8% and a kappa value of 0.27–0.34. Radiologist characteristics did not affect agreement.

Conclusion: The interpretative performance of the radiologists fulfilled the mammography screening goal of the American College of Radiology, although there was inter-observer variability. Radiologists who interpreted more than 3000 screening mammograms annually tended to perform better than radiologists who did not.

Keywords: *Screening; Medical audit; Radiologists; Observer variation; Sensitivity and specificity*

INTRODUCTION

The incidence of breast cancer has been rapidly increasing in Asian countries over the past two decades (1). Breast cancer in Asian countries has different characteristics compared with in developed Western countries. First, the incidence of breast cancer remains lower than in Western countries (1). Second, the age-specific incidence of female

Received March 28, 2018; accepted after revision October 9, 2018.
This study was supported by a grant from the National R&D Program for Cancer Control, Ministry of Health & Welfare, Republic of Korea (1520200).

Corresponding author: Eun Hye Lee, MD, Department of Radiology, Soonchunhyang University Hospital Bucheon, Soonchunhyang University College of Medicine, 170 Jomaru-ro, Bucheon 14584, Korea.
• Tel: (8232) 621-5851 • Fax: (8232) 621-5018 • E-mail: grace927@hanmail.net

breast cancer in Asia peaks at age 40–50 years, whereas in Western countries the peak occurs at age 60–70 years (2). Third, mammography accuracy is reduced in high-density breast tissue, and Asian women characteristically have higher-density breasts (3).

The American College of Radiology (ACR) announced desirable goals for screening mammography outcomes (4). The performance recommendations of the ACR include recall rates of 5–12%, a cancer detection rate (CDR) of more than 2.5 per 1000 examinations, sensitivity greater than 75%, specificity of 88–96%, and a positive predictive value (PPV, abnormal interpretation) of 3–8%.

The Republic of Korea adopted the National Cancer Screening Program (NCSP) based on the results of randomized controlled trials conducted in developed countries since 1999. However, the diagnostic accuracy of the NCSP was suboptimal (5). Mammography interpretation is highly challenging and is not completely objective. Variability among radiologists in mammography interpretation is extensive, and radiologist characteristics affect screening accuracy (6-12). Fellowship training in breast imaging improved the sensitivity and the overall accuracy (11). Greater interpretive volume improved the sensitivity, but decreased specificity (12).

The Alliance for Breast Cancer Screening in Korea began the Mammography and Ultrasonography Study for Breast Cancer Screening Effectiveness (MUST-BE) trial in 2016, which compared the diagnostic performance and the cost effectiveness of combined mammography and ultrasonography screenings versus conventional digital mammography screening alone for women of 40–59 years of age. In order for this trial to be successful and to achieve reliable results, quality management should be preceded by periodic monitoring of the diagnostic performances of the participating radiologists.

Our study had two purposes: to evaluate the interpretive performance and inter-radiologist agreement among radiologists who participated in the trial, and to investigate whether these performance and agreement levels differed according to radiologist characteristics.

MATERIALS AND METHODS

Study Subjects

This study was approved by the Institutional Review Boards of three institutions (approval number SCHBC 2017-10-002-002, CMC 2017-6203-0001, DKUH 2017-11-011).

The test sets consisted of 12 cancer cases and 988 non-recall cases. The cases were selected from women aged 40–69 years who received screening between 2010 and 2011 at one of three institutions in Seoul, Bucheon, and Cheonan. Three radiologists with 10, 13, and 15 years of experience interpreting mammography, but who did not otherwise participate in this study, each collected 340 non-recall cases and 10 cancer cases. One of the three reviewed all of the images and finally selected 1000 cases, including 12 cancer cases. The cancer cases were all detected upon screening and mammographically occult cancers were excluded (Table 1). Non-recall cases were included when mammography and follow-up images greater than 12 months showed negative or benign findings (range, 12–29 months; median, 18.5 months). In these cases, cancer was not found with either mammography or ultrasonography during follow-up. Mammographically dense breasts constituted 57.3%

Table 1. Characteristics of Screen-Detected Cancers

Number	Age (Yrs)	Density	Lesion Type	Size* (mm)	Pathology	Percentage of Correct Answer
Cancer 1	58	b	Focal asymmetry	6	IDC grade 1, node (-)	100
Cancer 2	55	c	Calcifications	NA	DCIS	75
Cancer 3	49	d	Focal asymmetry	15	IDC grade 1, node (-)	100
Cancer 4	74	a	Focal asymmetry	15	IDC grade 2, node (-)	100
Cancer 5	78	b	Mass	8	IDC grade 1, node (-)	92
Cancer 6	64	c	Calcifications	20	IDC grade 3, node (-)	83
Cancer 7	64	a	Focal asymmetry	7	IDC grade 2, node (-)	100
Cancer 8	69	a	Focal asymmetry	2	IDC grade 1, node (-)	58
Cancer 9	62	b	Mass	1	Microinvasive, node (-)	100
Cancer 10	50	d	Mass	15	IDC grade 2, node (-)	75
Cancer 11	49	c	Mass	33	ILC grade 2 node (+1/21)	100
Cancer 12	52	c	Calcifications	NA	DCIS	75

*Size of invasive cancer. DCIS = ductal carcinoma *in situ*, IDC = invasive ductal carcinoma, ILC = invasive lobular carcinoma, NA = not applicable, Yrs = years

https://doi.org/10.3348/kjr.2018.0193

of the 1000 cases, which was similar to the 54.8% of the Korean population who present with this characteristic (13).

Two views (mediolateral-oblique and craniocaudal) of full-field digital images were provided to radiologists as Digital Imaging and Communications in Medicine files.

Radiologist Characteristics

Radiologist characteristics were obtained from self-administered questionnaires. Questionnaires included the following data: years of experience interpreting mammography, fellowship training in breast imaging of more than one year, annual volumes, and percentage of examinations that were screening mammograms.

Data Review Process

Twelve radiologists independently interpreted all of the test set mammograms and did not review the follow up mammograms. These radiologists were blinded to the original mammographic interpretations and cancer status. Readers rated the mammograms using two scales: the four-point NCSP scale and the seven-point malignant scale. The NCSP scale was modified using the ACR Breast Imaging Reporting and Data System (BI-RADS) categories: 1, negative (BI-RADS category 1); 2, benign (category 2); 3, incomplete, additional evaluation needed (categories 3 and 0); 4, breast cancer doubt (categories 4 and 5).

A seven-point malignant scale was used to obtain suitable receiver-operating-characteristic (ROCs) curves for analysis (14): 1, definitely not malignant; 2, almost definitely not malignant; 3, probably not malignant; 4, possibly malignant; 5, probably malignant; 6, almost definitely malignant; 7, definitely malignant. We collapsed these assessments into two categories for recall (yes or no): recall, NCSP scale 3–4 and malignant scale 4–7; no recall, NCSP scale 1–2 and malignant scale 1–3. The percentage of correct answers was expressed as the percentage of radiologists who recalled the cancer cases.

To evaluate intra-radiologist agreement, 150 mammograms were interpreted a second time by the participating radiologists. The 150 cases were selected using random numbers out of non-recall cases (n = 988). There was an interval of three months between the readings.

Statistical Analysis

We calculated performance indicators as a function of each radiologist's performance and characteristics. Performance indicators included the recall rate, CDR, PPV, sensitivity, specificity, and false positive rate (FPR). The recall rate was calculated as the percentage of women screened who were recalled for further evaluation. The CDR was calculated as the number of breast cancer cases detected per 1000 examinations. The overall mammography accuracy according to radiologist characteristics was assessed using a ROC curve that plotted the true positive rate against the FPR. The significance of the differences among individual characteristics was estimated using the Wilcoxon rank sum test.

We measured intra- and inter-radiologist agreement using percent agreement and the kappa statistic. Percent agreement was a "row measure" that provided the percentage of interpretations for which both radiologists agreed. Cohen's kappa and its 95% confidence interval (CI) were calculated to measure intra- and inter-radiologist variability for both assessments. Because both variables used an ordinal scale, we also used the weighted kappa statistic. The method used to estimate an overall kappa in the case of multiple radiologists and multiple categories was based on the work of Hayes and Krippendorff (15): Hayes' interpretation of Krippendorff's alpha, which is equivalent to an overall weighted kappa, was used as a measure of the overall agreement among the twelve radiologists. Kappa values were interpreted as follows: poor agreement, kappa less than 0.0; slight agreement, 0.0–0.2; fair agreement, 0.2–0.4; moderate agreement, 0.4–0.6; substantial agreement, 0.6–0.8; almost perfect agreement, 0.8–1.0 (16). All statistical analyses were conducted using SAS software, version 9.2 (SAS Institute Inc., Cary, NC, USA) and p values of less than 0.05 were considered statistically significant.

RESULTS

Pathologic and mammographic characteristics of cancer cases are summarized in Table 1. Cancers were ductal carcinoma *in situ* (n = 2), microinvasive cancer (n = 1), and invasive cancers (n = 9). Mammographic abnormalities were focal asymmetry (n = 5), mass (n = 4), and calcifications (n = 3). The percentage of correct answers was 75–100% for mass and calcifications and 58–100% for focal asymmetry.

The radiologist characteristics are summarized in Table 2. Most radiologists had less than 10 years of experience interpreting mammography (mean, 9.22; range, 3–16 years) (58.3%), reported no fellowship training in breast imaging (58.3%), had a mean annual diagnostic

Table 2. Characteristics of Radiologists Participating in Study

Characteristic	No. of Radiologists (%)
Total	12 (100.0)
Years' experience interpreting mammography	
< 10	7 (58.3)
≥ 10	5 (41.7)
Fellowship training in breast imaging	
Yes	5 (41.7)
No	7 (58.3)
Mean annual diagnostic volume (no. of mammograms)	
< 3000	9 (75.0)
≥ 3000	3 (25.0)
Mean annual screening volume (no. of mammograms)	
< 3000	7 (58.3)
≥ 3000	5 (41.7)
Percentage of all examinations that were screening mammograms*	
< 50	5 (41.7)
≥ 50	7 (58.3)

*Average per year over previous 1 year.

mammography volume of < 3000 (75%), had a mean annual screening mammography volume of < 3000 (58.3%), and a mammography screening percentage of ≥ 50% (58.3%).

Interpretive Performance

The mean and range of performance indicators were as follows: recall rate, 7.5% and 3.3–10.2%; number of cancer detections, 10.6 and 8.0–12.0 per 1000 examinations; PPV, 15.9% and 8.8–33.3%; sensitivity, 88.2% and 66.7–100%; specificity, 93.5% and 90.6–97.8%; FPR, 6.5% and 2.2–9.4%; area under the curve, 0.93 and 0.82–0.99, respectively (Table 3, Fig. 1). Radiologists interpreting more than 3000 screening mammograms annually tended to have higher CDRs and sensitivities than those interpreting less than 3000 mammograms; however, there was no statistical significance (*p* = 0.064). Years of experience, fellowship training in breast imaging, volume of diagnostic mammograms and percentage of screening mammograms did not affect interpretive performance.

Observer Variability

All twelve radiologists completed the first assessment of 1000 cases and the repeat assessment of 150 cases. Table 4 shows intra- and inter-radiologist agreement.

All radiologists had more than 76% intra-individual agreement, ranging from 85.4% (95% CI, 82.1–88.6%) to 94.6% (95% CI, 92.6–96.5%); kappa values of 0.410 (95% CI, 0.264–0.557) to 0.572 (95% CI, 0.514–0.628) indicated moderate agreement. There was no difference in

Table 3. Interpretive Performances and Radiologist Characteristics

Characteristics	Recall Rate, %	No. Cancers Detected*	PPV†, %	Sensitivity, %	Specificity, %	False Positive Rate, %	AUC (Range)
	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	
Total	7.5 (3.3–10.2)	10.6 (8.0–12.0)	15.9 (8.8–33.3)	88.2 (66.7–100.0)	93.5 (90.6–97.8)	6.5 (2.2–9.4)	0.93 (0.82–0.99)
Years experience in interpreting mammography							
< 10	7.2 (3.6–9.6)	10.4 (8.0–12.0)	15.1 (11.5–22.2)	86.9 (66.7–100.0)	93.7 (91.6–97.2)	6.3 (2.8–8.5)	0.92 (0.82–0.99)
≥ 10	7.7 (3.3–10.2)	10.8 (9.0–12.0)	16.9 (8.8–33.3)	90.0 (75.0–100.0)	93.3 (90.6–97.8)	6.7 (2.2–9.4)	0.95 (0.90–0.99)
p value	0.725	0.652	0.293	0.651	0.741	0.753	0.516
Fellowship training in breast imaging							
Yes	8.0 (6.6–10.2)	10.0 (9.0–11.0)	12.9 (8.8–15.2)	83.3 (75.0–91.7)	92.9 (90.6–94.3)	7.1 (5.7–9.4)	0.92 (0.88–0.99)
No	7.0 (3.3–9.6)	11.0 (8.0–12.0)	18.1 (10.4–33.3)	91.7 (66.7–100.0)	94.0 (91.3–97.8)	6.0 (2.2–8.7)	0.95 (0.82–0.99)
p value	0.487	0.208	0.192	0.206	0.425	0.432	0.389
Mean annual screening volume (no. of mammograms)							
< 3000	6.9 (3.3–10.2)	10.0 (8.0–12.0)	17.3 (8.8–33.3)	83.3 (66.7–100.0)	94.0 (90.6–97.8)	6.0 (2.2–9.4)	0.91 (0.82–0.99)
≥ 3000	8.2 (7.3–9.6)	11.4 (10.0–12.0)	14.0 (12.5–15.8)	95.0 (83.3–100.0)	92.8 (91.6–93.7)	7.2 (6.3–8.5)	0.96 (0.88–0.99)
p value	0.332	0.064	0.424	0.064	0.384	0.377	0.096
Percentage of all examination that were screening mammograms‡							
< 50	6.5 (3.3–10.2)	10.4 (8.0–12.0)	19.3 (8.8–33.3)	86.7 (66.7–100.0)	94.4 (90.6–97.8)	5.6 (2.2–9.4)	0.93 (0.82–0.99)
≥ 50	8.1 (6.6–9.6)	10.7 (10.0–12.0)	13.5 (10.4–15.8)	89.3 (83.3–100.0)	92.9 (91.3–94.3)	7.1 (5.7–8.7)	0.94 (0.88–0.99)
p value	0.258	0.707	0.138	0.706	0.263	0.258	0.884

*Total of 12 cancer cases detected per 1000 screening mammograms, †PPV, ‡Average per year over previous year. AUC = area under curve, PPV = positive predictive value

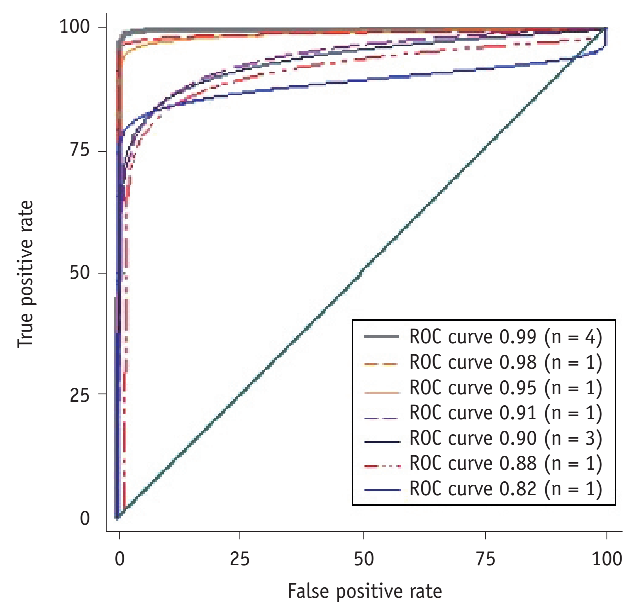

Fig. 1. Areas under curve of twelve radiologists ranged from 0.82 to 0.99 with mean value of 0.93. ROC = receiver-operating-characteristic

the intra-radiologist agreement according to the radiologist characteristics.

The inter-radiologist agreement for twelve radiologists was as follows. Pairwise percent agreements ranged from 77.2% (95% CI, 75.4–79.1%) to 88.8% (95% CI, 87.3–90.2%); pairwise kappa values were fair: from 0.27 (95% CI, 0.15–0.63) to 0.34 (95% CI, 0.21–0.46). There was no difference in observer variability according to radiologist characteristics.

DISCUSSION

In the present study, radiologists who participated in the MUST-BE trial exhibited good interpretive performance for digital screening mammography, and surpassed most performance recommendations of the ACR (4) and performance measures of the Breast Cancer Surveillance Consortium (BCSC) for screening digital mammography examinations (17). The BCSC assessed the trends in screening mammography performance in the United States and published screening mammography performance benchmarks (17). The mean recall rate of the present study was lower than that of the BCSC (7.5% vs. 11.6%, respectively). In addition, the mean PPV and specificity of the present study were higher than those of the BCSC (15.9% vs. 4.4% and 93.5% vs. 88.9%, respectively), and the mean sensitivities were similar (88.2% vs. 86.9%). The results of the

Table 4. Intra- and Inter-Radiologist Agreement and Radiologist Characteristics

Characteristics	Intra-Radiologist Agreement (n = 150 Cases)					Inter-Radiologist Agreement (n = 1000 Cases)				
	Percent Agreement		Kappa		*P*	Pairwise Percent Agreement		Pairwise Kappa		*P*
	Range	Mean (95% CI)	Range	Mean (95% CI)		Range	Mean (95% CI)	Range	Mean (95% CI)	
NCSP scale					0.111					0.272
2 categories	84.7–98.0	92.5 (90.1–94.8)	0.116–0.733	0.457 (0.326–0.588)		84.6–95.7	88.7 (87.3–90.1)	0.118–0.433	0.266 (0.213–0.325)	
4 categories	78.7–95.3	85.4 (82.1–88.6)	0.444–0.745	0.572 (0.514–0.628)		57.1–88.2	76.3 (74.6–77.9)	0.108–0.507	0.342 (0.216–0.463)	
Malignancy scale					0.169					0.631
2 categories	88.7–98.7	94.6 (92.6–96.5)	0.088–0.686	0.410 (0.264–0.557)		84.6–95.7	88.8 (87.3–90.2)	0.117–0.433	0.265 (0.159–0.638)	
7 categories	76.0–95.3	85.8 (82.0–89.6)	0.364–0.709	0.522 (0.452–0.592)		55.5–92.1	77.2 (75.4–79.1)	0.129–0.432	0.327 (0.246–0.407)	
Years experience in interpreting mammography					0.239					0.809
< 10	78.7–95.3	86.6 (82.6–90.6)	0.488–0.745	0.602 (0.500–0.700)		57.3–87.9	66.6 (64.4–68.8)	0.108–0.426	0.323 (0.190–0.447)	
≥ 10	78.7–88.0	83.6 (80.5–86.7)	0.444–0.666	0.529 (0.462–0.596)		62.6–83.5	77.4 (75.8–79.0)	0.200–0.507	0.344 (0.230–0.453)	
Fellowship training in breast imaging					0.177					0.646
Yes	78.7–90.0	82.6 (76.6–88.6)	0.444–0.611	0.529 (0.449–0.609)		58.1–84.3	80.6 (79.1–82.0)	0.174–0.415	0.345 (0.225–0.461)	
No	80.7–95.3	87.3 (83.0–91.7)	0.482–0.745	0.602 (0.512–0.691)		62.6–87.0	71.7 (69.7–73.7)	0.135–0.507	0.322 (0.193–0.443)	
Mean annual screening volume (no. of mammograms)					0.598					0.062
< 3000	78.7–95.3	85.4 (81.1–89.8)	0.444–0.666	0.560 (0.510–0.610)		57.1–88.2	75.3 (73.6–76.9)	0.108–0.507	0.285 (0.162–0.400)	
≥ 3000	80.7–90.7	85.2 (82.4–88.1)	0.482–0.745	0.588 (0.496–0.679)		75.2–83.3	83.1 (81.5–84.7)	0.302–0.438	0.448 (0.320–0.566)	
Percentage of all examination that were screening mammograms					0.380					0.475
< 50	78.7–95.3	85.1 (80.0–90.2)	0.444–0.666	0.547 (0.478–0.616)		62.6–83.5	77.4 (75.8–79.1)	0.135–0.507	0.298 (0.179–0.417)	
≥ 50	78.7–90.7	85.5 (82.4–88.7)	0.488–0.745	0.589 (0.525–0.653)		57.1–88.2	66.6 (64.3–68.8)	0.170–0.463	0.361 (0.234–0.485)	

Agreement among radiologist characteristics was calculated using weighted kappa values based on four NCSP categories. CI = confidence interval, NCSP = National Cancer Screening Program present study are evidences of quality control of the MUST-BE trial, which was to investigate the diagnostic performance and the cost effectiveness of combined mammography and ultrasonography screenings in comparison with those of conventional digital mammography screening for women in their forties and fifties.

There were several reports that characteristics of radiologists affected the diagnostic performance of mammographic screening (6-12), but this study did not show any significant difference. The mean annual screening volume 3000 or more alone showed a tendency to increase the CDR and sensitivity. This result supports the previous results that an increased volume of screening mammography improved the CDR (18). These results agreed with the basic rationale that radiologists who participate in funded screening programs must comply with the national accreditation standards and requirements. The annual minimum number of mammograms that radiologists read is 480 in the United States, 5000 in the United Kingdom and Germany, and 2000 in Canada (18, 19). The Korean National Cancer Screening quality guidelines recommend that radiologists read an annual minimum number of 1000 mammograms (20), but it is necessary to systematically monitor the quantity of screening volume and diagnostic performance of radiologists to manage quality of breast cancer screening, as in Western Europe.

There was a study that reported that the volume of diagnostic mammography was associated with improved sensitivity and decreased FPR (21), but there were no differences between the diagnostic volume and performance of radiologists in the present study.

We expected that radiologists who completed fellowship training in breast imaging would have a better diagnostic performance, as with the previous result (11), but the present study did not show a significant effect. This may be due to more intensive fellowship training of diagnosis for symptomatic patients and preoperative evaluation for breast cancer patients than breast cancer screening. A curriculum that gives a greater weight to the education of breast cancer screening is needed.

Variability in radiologist performance due to differences in both lesion detection and interpretation has been observed (22). Several studies reported considerable variability in the assignment of BI-RADS categories (6-10), resulting in a wide range of recall and FPRs. Such false positives can cause increased anxiety and breast cancer-specific worry, as well as financial loss to patients (23). The present study showed a fair degree of agreement among the twelve radiologists; however, the recall rate was within the acceptable range of the ACR recommendations (4). The FPR was similar to previous results that reported the range of FPRs as 3.5–7.9% after adjusting for patient, radiologist, and testing factors (22, 24, 25). Proper training and continued practice might improve variability and performance.

There are some limitations to this study. First, the mammography test sets were composed of non-recall cases (true negativity) and recall cases (true positivity). False positive and false negative cases were not included and the test sets may not have adequately represented actual clinical practices in community-based screening settings. In addition, the test sets included more cancer cases compared to the incidence of breast cancer in the real world. So, it is questionable as to whether the diagnostic performance measured using the test sets would reflect the diagnostic performance in actual clinical practice. As such, the comparison of the results between them is limited. Second, breast cancer epidemics show different characteristics in Western and Asian countries and, therefore, the populations' desirable performance goals are different. Although we compared our results to published Western data, this was not the most appropriate comparison. Performance goals appropriate for Korean women need to be developed. Third, 150 non-recall cases were selected to evaluate the intra-radiology agreement. Cancer cases were excluded, resulting in selection bias. Nevertheless, this is the first study that used test sets to evaluate digital mammography performance in breast cancer screening in Korea.

In conclusion, the interpretative performances of radiologists participating in the MUST-BE trial fulfilled the ACR goal of screening mammography, although inter-observer variability persisted. Sufficient volume of screening mammography and specialized training for radiologists are needed to perform the national breast cancer screening successfully.

Conflicts of Interest

The authors have no financial conflicts of interest.

Acknowledgments

We would like to thank Seung Hoon Song and Hye-Mi Jo for data collection and management.

ORCID
Eun Hye Lee
https://orcid.org/0000-0002-8773-700X
Sung Hun Kim
https://orcid.org/0000-0003-4478-9720

REFERENCES

1. Youlden DR, Cramb SM, Yip CH, Baade PD. Incidence and mortality of female breast cancer in the Asia-Pacific region. *Cancer Biol Med* 2014;11:101-115
2. Leong SP, Shen ZZ, Liu TJ, Agarwal G, Tajima T, Paik NS, et al. Is breast cancer the same disease in Asian and Western countries? *World J Surg* 2010;34:2308-2324
3. Ohuchi N, Suzuki A, Sobue T, Kawai M, Yamamoto S, Zheng YF, et al.; J-START investigator groups. Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the Japan Strategic Anti-cancer Randomized Trial (J-START): a randomised controlled trial. *Lancet* 2016;387:341-348
4. American College of Radiology. *ACR BI-RADS Atlas®*, 5th ed. Reston, VA: American College of Radiology, 2013
5. Lee EH, Kim KW, Kim YJ, Shin DR, Park YM, Lim HS, et al. Performance of screening mammography: a report of the alliance for breast cancer screening in Korea. *Korean J Radiol* 2016;17:489-496
6. Baker JA, Kornguth PJ, Floyd CE Jr. Breast Imaging Reporting and Data System standardized mammography lexicon: observer variability in lesion description. *AJR Am J Roentgenol* 1996;166:773-778
7. Lazarus E, Mainiero MB, Schepps B, Koelliker SL, Livingston LS. BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. *Radiology* 2006;239:385-391
8. Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast Imaging Reporting and Data System: inter- and intraobserver variability in feature analysis and final assessment. *AJR Am J Roentgenol* 2000;174:1769-1777
9. Timmers JM, van Doorne-Nagtegaal HJ, Zonderland HM, van Tinteren H, Visser O, Verbeek AL, et al. The Breast Imaging Reporting and Data System (BI-RADS) in the Dutch breast cancer screening programme: its role as an assessment and stratification tool. *Eur Radiol* 2012;22:1717-1723
10. Duijm LE, Louwman MW, Groenewoud JH, van de Poll-Franse LV, Fracheboud J, Coebergh JW. Inter-observer variability in mammography screening and effect of type and number of readers on screening outcome. *Br J Cancer* 2009;100:901-907
11. Elmore JG, Jackson SL, Abraham L, Miglioretti DL, Carney PA, Geller BM, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology* 2009;253:641-651
12. Barlow WE, Chi C, Carney PA, Taplin SH, D'Orsi C, Cutter G, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. *J Natl Cancer Inst* 2004;96:1840-1850
13. Kim YJ, Lee EH, Jun JK, Shin DR, Park YM, Kim HW, et al. Analysis of participant factors that affect the diagnostic performance of screening mammography: a report of the Alliance for Breast Cancer Screening in Korea. *Korean J Radiol* 2017;18:624-631
14. Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, et al.; Digital Mammographic Imaging Screening Trial (DMIST) Investigators Group. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med* 2005;353:1773-1783
15. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 2007;1:77-89
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-174
17. Lehman CD, Arao RF, Sprague BL, Lee JM, Buist DS, Kerlikowske K, et al. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. *Radiology* 2017;283:49-58
18. Rickard M, Taylor R, Page A, Estoesta J. Cancer detection and mammogram volume of radiologists in a population-based screening programme. *Breast* 2006;15:39-43
19. Albert US, Altland H, Duda V, Engel J, Geraedts M, Heywang-Köbrunner S, et al. 2008 update of the guideline: early detection of breast cancer in Germany. *J Cancer Res Clin Oncol* 2009;135:339-354
20. National Cancer Center. Ministry of Health & Welfare. *Quality guidelines of breast cancer screening*, 2nd ed. Goyang: National Cancer Center, 2018:43
21. Haneuse S, Buist DS, Miglioretti DL, Anderson ML, Carney PA, Onega T, et al. Mammographic interpretive volume and diagnostic mammogram interpretation performance in community practice. *Radiology* 2012;262:69-79
22. Berg WA, D'Orsi CJ, Jackson VP, Bassett LW, Beam CA, Lewis RS, et al. Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography? *Radiology* 2002;224:871-880
23. Nelson HD, Pappas M, Cantor A, Griffin J, Daeges M, Humphrey L. Harms of breast cancer screening: systematic review to update the 2009 U.S. Preventive Services Task Force recommendation. *Ann Intern Med* 2016;164:256-267
24. Lee EH, Jun JK, Jung SE, Kim YM, Choi N. The efficacy of mammography boot camp to improve the performance of radiologists. *Korean J Radiol* 2014;15:578-585
25. Elmore JG, Miglioretti DL, Reisch LM, Barton MB, Kreuter W, Christiansen CL, et al. Screening mammograms by community radiologists: variability in false-positive rates. *J Natl Cancer Inst* 2002;94:1373-1380