Diagnostic Accuracy Study

Medicine®

OPEN

A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of breast masses on ultrasound

Added value for the inexperienced breast radiologist

Hee Jeong Park, MD[a], Sun Mi Kim, MD[a,*], Bo La Yun, MD[a], Mijung Jang, MD[a], Bohyoung Kim, PhD[b], Ja Yoon Jang, MD[c], Jong Yoon Lee, MD[d], Soo Hyun Lee, MD[e]

Abstract

To evaluate the value of the computer-aided diagnosis (CAD) program applied to diagnostic breast ultrasonography (US) based on operator experience.

US images of 100 breast masses from 91 women over 2 months (from May to June 2015) were collected and retrospectively analyzed. Three less experienced and 2 experienced breast imaging radiologists analyzed the US features of the breast masses without and with CAD according to the Breast Imaging Reporting and Data System (BI-RADS) lexicon and categories. We then compared the diagnostic performance between the experienced and less experienced radiologists and analyzed the interobserver agreement among the radiologists.

Of the 100 breast masses, 41 (41%) were malignant and 59 (59%) were benign. Compared with the experienced radiologists, the less experienced radiologists had significantly improved negative predictive value (86.7%–94.7% vs 53.3%–76.2%, respectively) and area under receiver operating characteristics curve (0.823–0.839 vs 0.623–0.759, respectively) with CAD assistance (all $P < .05$). In contrast, experienced radiologists had significantly improved specificity (52.5% and 54.2% vs 66.1% and 66.1%) and positive predictive value (55.6% and 58.5% vs 64.9% and 64.9%, respectively) with CAD assistance (all $P < .05$). Interobserver variability of US features and final assessment by categories were significantly improved and moderate agreement was seen in the final assessment after CAD combination regardless of the radiologist's experience.

CAD is a useful additional diagnostic tool for breast US in all radiologists, with benefits differing depending on the radiologist's level of experience. In this study, CAD improved the interobserver agreement and showed acceptable agreement in the characterization of breast masses.

Abbreviations: ACR = American College of Radiology, AUC = area under the receiver operating characteristic curve, BI-RADS = Breast Imaging Reporting and Data System, CAD = computer-aided diagnosis, NPV = negative predictive value, PPV = positive predictive value, ROI = region of interest, US = ultrasonography.

Keywords: breast, computer-aided, diagnosis, neoplasm, ultrasonography

1. Introduction

Along with mammography, breast ultrasonography (US) is regarded as the most effective diagnostic tool for evaluating breast abnormalities. Ultrasound is an easily available inexpensive imaging tool, without the accompanying risk of radiation, and therefore, there has been an expansion in indications for the use of breast US, including serving as an adjunctive screening tool to mammography, for preoperative staging, follow-up after cancer treatment, and interventional diagnosis.[1] However, US is a highly operator-dependent modality in terms of image acquisition and interpretation. Since a morphological analysis is essential for the diagnosis of benign and malignant lesions, the diagnostic accuracy is dependent on the skill and expertise of the operator. To overcome these problems, many studies have applied the computer-aided diagnosis (CAD) program to breast US. CAD is an image–analytic program that provides morphologic analysis of breast lesions seen on breast US. CAD systems were developed to overcome subjective disparity and observer variability of US,[2] and to improve the capability of radiologists in the analysis of US images and differentiation of tissue malignancy.[3–5] It has been reported that CAD enables efficient interpretation and improves the diagnostic accuracy in distinguishing benign and malignant breast lesions.[6,7]

Editor: Alberto Stephano Tagliafico.

This work was supported by grant no 02-2017-021 from the SNUBH Research Fund.

The authors report no conflicts of interest.

[a] Department of Radiology, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Gumi-dong, [b] Division of Biomedical Engineering, Hankuk University of Foreign Studies, Mohyeon-myeon, Cheoin-gu, Yongin-si, [c] Department of Radiology, Bundang Jesaeng Hospital, Bundang-gu, Seongnam-si, Gyeonggi-do, [d] Department of Radiology, Borame Medical Center 20, Boramae-ro 5-gil, Dongjak-gu, Seoul, [e] Department of Radiology, Chungbuk National University Hospital, Seowon-gu, Cheongju, South Korea.

[*] Correspondence: Sun Mi Kim, Department of Radiology, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Gumi-dong, Bundang-gu, Seongnam-si, Gyeonggi-do 436-707, South Korea (e-mail: kimsmlms@daum.net).

Medicine (2019) 98:3(e14146)

Received: 25 October 2018 / Received in final form: 22 December 2018 / Accepted: 26 December 2018

http://dx.doi.org/10.1097/MD.0000000000014146

S-Detect (Samsung Medison Co. Ltd., Seoul, South Korea) is a recently developed CAD program that provides computer-based analysis based on morphologic features, using a novel feature extraction technique and support vector machine classifier that provides final assessment data for breast masses in a dichotomized form (possibly benign or possibly malignant) based on the American College of Radiology Breast Imaging Reporting and Data System (ACR BI-RADS) ultrasonographic descriptors.[8,9] In a recent study analyzing the diagnostic performance of CAD and an experienced breast radiologist, CAD showed diagnostic performance equivalent to that of the radiologist.[10] Other studies have reported that CAD improved the diagnostic performance of breast US, regardless of the experience of the radiologist.[10,11] However, so far, there have been few studies evaluating its ability to improve the diagnostic performance of radiologists with different levels of experience by comparing diagnostic evaluation with and without assistance by CAD in experienced and inexperienced breast radiologists.

Therefore, the purpose of this study was to evaluate the added value of CAD applied to diagnostic breast US in operators with different levels of experience in breast imaging and to assess the interobserver agreement on lesion characterization and final assessment.

2. Materials and methods

This retrospective study was approved by the Institutional Review Board of our hospital, which waived the requirement for informed consent.

2.1. Patients

A total of 110 breast masses (69 benign and 41 malignant) from 101 women (mean age, 46.5 ± 12.3 years; range, 18–78 years) who were scheduled for breast US examination or US-guided biopsy between May and June 2015 were included in this retrospective study. Among them, 8 masses from 8 women in whom a follow-up US examination for benign mass was not conducted and 2 masses from 2 women with a size of over 4 cm, which could not be covered in the field of the CAD, were excluded. Finally, 100 breast masses in 91 women were included in the study. Seventy-six patients were asymptomatic, 15 had a palpable lump, and 1 presented with discharge.

2.2. US examination and biopsy

US examinations were performed using an RS80A US system (Samsung Medison Co., Seoul, South Korea) equipped with a linear high-frequency probe (frequency range, 3–12 MHz). One radiologist with 14 years of experience in breast imaging was involved in image acquisition. The radiologist was aware of the clinical and mammographic features and had access to the previous US images. Transverse and longitudinal static images were obtained for each lesion. For image analysis using CAD, video clips that included the entire mass and the surrounding normal breast parenchyma were recorded with the US machine during one-directional movement of the probe, starting at one end and ending at the other end of the mass. US-guided biopsy was performed after the US examination by 1 of the 2 radiologists who performed the scanning.

2.3. Image review and application of CAD

Three less experienced radiologists (first-year fellowship trainees in breast imaging) and 2 experienced radiologists (8 and 10 years of breast imaging experience) independently reviewed the obtained images. The reviewers were blinded to the clinical information and final pathologic outcomes. Two separate US image review sessions were carried out. The first session was for image review of grayscale breast US without CAD. All lesions were analyzed independently by the 5 radiologists. US features of each breast lesion captured by 2D sonography with video were analyzed, based on the fifth edition of the ACR BI-RADS lexicon and final assessment categories: shape, orientation, margin, echo pattern, and posterior features.[1] The reviewer chose and recorded the most appropriate term for each descriptor. Final assessments were made for each breast lesion using one of the assessment categories of BI-RADS: 3, probably benign; 4, suspicious finding; and 5, highly suggestive of malignancy. The cancer probability scale (0%–100%) was also recorded.

The second session, which was performed 1 month after the first session, used CAD on the video clip image for grayscale US feature analysis of each breast mass. When the operators placed a mark in the center of the target lesion by touching the screen, the program automatically drew a region-of-interest (ROI) along the border of the mass (Figs. 1 and 2). If the boundary drawn automatically by the CAD system was considered inaccurate, the operator manually readjusted the ROI. The CAD system analyzed the morphologic features of the mass according to the BI-RADS ultrasonographic descriptors, and analytic results including the BI-RADS descriptors and a final assessment, which were provided in a dichotomized form as "possibly benign" and "possibly malignant," were immediately visualized. After being informed of the analysis of the results of the CAD, each reviewer rescored the cancer probability scale and re-evaluated the sonographic features of the lesions using the BI-RADS lexicons. Due to the limited data analytical ability of CAD, calcifications were not analyzed.[12]

2.4. Treatment and follow-up

In cases with discordant pathologic and radiologic findings, if the pathological diagnosis showed malignancy or borderline result, such as atypical ductal hyperplasia, surgical excision was recommended; some benign lesions, such as phyllodes tumor also required surgical resection. For other cases where the imaging and histologic diagnoses were benign, follow-up US was recommended at 6-month intervals.

2.5. Data and statistical analysis

Clinical, radiological, and pathological data of the patients were collected for statistical evaluation.

For each reviewer, we compared the final results of the first and second reviews with the histologic diagnosis. Diagnostic performance of all included radiologists, without and with CAD assistance, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy were calculated according to the BI-RADS category (possibly benign or possibly malignant with the cut-off set at category 4) and then compared using the area under the receiver operating characteristic curve (AUC) for each review session based on the cancer probability scale using open-source statistical software R, version 3.3.2 (http://www.R-project.org).

Fleiss kappa statistics were used to analyze the interobserver agreement on each of the US breast lesion findings for the 2 image review sessions. Estimation of the overall kappa was based on the study of Landis and Koch: kappa value <0 indicated poor

Table 1

Histopathologic diagnosis of the 100 breast masses.

Histopathologic diagnosis	No. (%)
Benign (n = 59)	
Fibroadenoma or complex fibroadenoma	32 (54.2)
Fibrocystic changes	7 (11.9)
Intraductal papilloma	5 (8.5)
Mammary duct ectasia	4 (6.7)
Benign phyllodes tumor	3 (5.1)
Nodular adenosis	2 (3.4)
Radial scar	1 (1.7)
Suture granuloma	1 (1.7)
Sclerosing adenosis	1 (1.7)
No diagnostic abnormality	1 (1.7)
Fibroadipose tissue	1 (1.7)
Fibroadenomatoid hyperplasia	1 (1.7)
Malignant (n = 41)	
Invasive ductal carcinoma	27 (65.9)
Ductal carcinoma in situ	10 (24.4)
Invasive lobular carcinoma	3 (7.3)
Mucinous carcinoma	1 (2.4)

agreement; ≤0.20, slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; and 0.81 to 1.00, almost perfect agreement.[12] We used a bootstrap method for comparing correlated kappa coefficients[13] using the statistical package STATA software version 14.0 (StataCorp, College Station, TX). A *P*-value of <0.05 was considered statistically significant.

3. Results

3.1. Characteristics of patients and lesions

Among the 100 breast masses included in this study, 41 (41%) were malignant, and 59 (59%) were benign. Surgery was performed for 39 malignant lesions. Two cases of atypical ductal hyperplasia and 2 of 3 phyllodes tumors were diagnosed on core needle biopsy. The histopathological results of 2 lesions were upgraded from atypical ductal hyperplasia to ductal carcinoma in situ. The other 57 (57%) lesions were diagnosed with US-guided core needle biopsy and these showed stability during US follow-up for a mean of 17 months (range, 6–35 months). The final histopathologic diagnoses are shown in Table 1. The mean size of the breast masses was 14 ± 7 mm (range, 4–39 mm). The mean size of the malignant masses was larger than that of the benign lesions, 14 ± 8 mm (range, 4–39 mm), and 12 ± 7 mm (range, 4–37 mm), respectively. However, the difference in size was not statistically significant ($P = .16$).

3.2. Diagnostic performance without and with CAD assistance

Diagnostic performance of the 5 reviewers for detecting breast malignancy with only grayscale US images and with CAD assistance were compared (Table 2). In the less experienced radiologists, all parameters of diagnostic performance (sensitivity, specificity, NPV, PPV, accuracy, and AUC) were improved when CAD was combined, except for specificity in reviewer 1. The NPV and AUC showed significant improvement in all 3 less experienced radiologists, whereas sensitivity, PPV, and accuracy showed statistically significant improvement in 2 of the 3 reviewers after CAD assistance ($P < .05$). In the 2 experienced radiologists, specificity and PPV were significantly improved, and in 1 of the 2 experienced radiologists, accuracy and AUC were significantly improved after CAD assistance ($P < .05$). The sensitivity and NPV were reduced when CAD was combined in 1 experienced reviewer, but it was not statistically significant (92.7%–90.2%, 91.4%–90.7%, respectively).

When the final assessments made by the individual radiologist were different from those of the CAD, less experienced radiologists changed the results of 18 to 25 of the 100 cases (18%–25%) to the conclusion made by CAD, and experienced radiologists changed the results of 10 to 14 of the 100 cases (10%–14%) (Table 3). Less experienced radiologists correctly downgraded 93% to 100% lesions from possibly malignant to possibly benign (Table 4, Fig. 1) and correctly upgraded 40% to 65% lesions from possibly benign to possibly malignant after CAD combination (Table 4, Fig. 2). Experienced radiologists correctly downgraded 89% to 100% lesions from possibly malignant to possibly benign and correctly upgraded 0% to 50% lesions from possibly benign to possibly malignant after CAD combination (Table 4).

3.3. Interobserver variability of US characteristics without and with CAD assistance

A summary of the interobserver variability in US features and final assessments among the reviewers with and without CAD assistance is presented in Table 5. Kappa statistics showed

Table 2

Diagnostic performance without and with computer-aided diagnosis.

	Less experienced									Experienced					
	Radiologist 1			Radiologist 2			Radiologist 3			Radiologist 4			Radiologist 5		
Diagnostic value	Without CAD	With CAD	*P* value	Without CAD	With CAD	*P* value	Without CAD	With CAD	*P* value	Without CAD	With CAD	*P* value	Without CAD	With CAD	*P* value
Sensitivity, %	65.9	97.6	<.001	75.6	85.4	.10	87.8	97.6	.05	85.4	90.2	.16	92.7	90.2	.327
Specificity, %	27.1	23.7	.56	50.8	66.1	.04	27.1	30.5	.59	52.5	66.1	.02	54.2	66.1	.02
NPV, %	53.3	93.3	<.001	75.0	86.7	.03	76.2	94.7	.03	83.8	90.7	.05	91.4	90.7	.759
PPV, %	38.6	47.1	.009	51.7	63.6	.01	45.6	49.4	.14	55.6	64.9	.008	58.5	64.9	.03
Accuracy, %	43.0	54.0	.03	61.0	74.0	.008	51.0	58.0	.15	66.0	74.0	.006	70.0	76.0	.05
AUC*	0.623 (0.501–0.746)	0.828 (0.745–0.912)	<.001	0.702 (0.596–0.808)	0.823 (0.742–0.904)	.001	0.759 (0.660–0.859)	0.839 (0.762–0.917)	.040	0.856 (0.776–0.936)	0.907 (0.848–0.967)	.02	0.889 (0.821–0.957)	0.904 (0.837–0.971)	.16

AUC = area under receiver operating characteristics curve; CAD = computer-aided diagnosis; NPV = negative predictive value; PPV = positive predictive value.
* AUC obtained using cancer probability scale, values in parentheses are 95% confidence intervals.

Table 3

Final assessments by Breast Imaging Reporting and Data System (BI-RADS) categories for each radiologist without and with computer-aided diagnosis.

			Without CAD			With CAD		
			Pathologic results			Pathologic results		
Radiologist		Interpretation*	Benign	Malignant	Total	Benign	Malignant	Total
Less experienced	1	Category 3	16 (53.3%)	14 (46.7%)	30 (100%)	14 (93.3%)	1 (6.7%)	15 (100%)
		Category 4 or 5	43 (61.4%)	27 (38.6%)	70 (100%)	45 (52.9%)	40 (47.1%)	85 (100%)
		Total	59	41	100	59	41	100
	2	Category 3	30 (75%)	10 (25%)	40 (100%)	39 (86.7%)	6 (13.3%)	45 (100%)
		Category 4 or 5	29 (48.3%)	31 (51.7%)	60 (100%)	20 (36.4%)	35 (63.6%)	55 (100%)
		Total	59	41	100	59	41	100
	3	Category 3	16 (76.2%)	5 (23.8%)	21 (100%)	18 (94.7%)	1 (5.3%)	19 (100%)
		Category 4 or 5	43 (54.4%)	36 (45.6%)	79 (100%)	41 (50.6%)	40 (49.4%)	81 (100%)
		Total	59	41	100	59	41	100
Experienced	4	Category 3	31 (83.8%)	6 (16.2%)	37 (100%)	39 (90.7%)	4 (9.3%)	43 (100%)
		Category 4 or 5	28 (44.4%)	35 (55.6%)	63 (100%)	20 (35.1%)	37 (64.9%)	57 (100%)
		Total	59	41	100	59	41	100
	5	Category 3	32 (91.4%)	3 (8.6%)	35 (100%)	39 (90.7%)	4 (9.3%)	43 (100%)
		Category 4 or 5	27 (41.5%)	38 (58.5%)	65 (100%)	20 (35.1%)	37 (64.9%)	57 (100%)
		Total	59	41	100	59	41	100

CAD = computer-aided diagnosis.

* BI-RADS category 3 considered possibly benign, category 4 to 5 considered possibly malignant.

significant improvement in the interobserver variability among all the radiologists, among less experienced radiologists, and between experienced radiologists for all US features ($P < .001$), except orientation in the experienced group ($P = .156$). Agreements among all reviewers without CAD assistance were moderate for shape (0.478); fair for orientation (0.322), posterior acoustic features (0.266), and echo pattern (0.302); and slight for margin (0.196). Agreements among all reviewers with CAD assistance were moderate for shape (0.544) and orientation (0.546); and fair for echo pattern (0.401), posterior acoustic features (0.350), and margin (0.285). Agreements in US features among less experienced radiologists without CAD assistance ranged from slight to moderate, but agreements improved from fair to substantial with CAD assistance ($\kappa = 0.139$–0.541 and 0.258–0.655, respectively). Agreement between experienced radiologists without CAD assistance in US features ranged from fair to moderate, but improved from fair to substantial ($\kappa = 0.219$–0.584 and 0.395–0.656, respectively) with CAD assistance. When the final assessment classified 3 categories by BI-RADS (3, 4, and 5), it also showed significant improvement in the agreement among all the reviewers, among less experienced radiologists, and between experienced radiologists; the agreement was fair without CAD to fair with CAD among all reviewers ($\kappa = 0.221$ and 0.320, respectively); fair without CAD to moderate with CAD among less experienced radiologists ($\kappa = 0.186$ and 0.412, respectively); and fair without CAD to moderate with CAD among experienced radiologists ($\kappa = 0.260$ and 0.510, respectively) (all $P < .001$).

4. Discussion

In this study, we evaluated the role of the CAD program, S-Detect, in diagnostic performance for the differential diagnosis of breast masses using US, with a focus on its value in less experienced and experienced breast radiologists. A few earlier studies have evaluated the diagnostic performance of CAD. Kim et al[8] reported that CAD showed significantly higher specificity, PPV, accuracy, and AUC compared with an experienced breast

Table 4

Change of reviewer's final assessment after computer-aided diagnosis combination and comparison with pathology.

			Pathology		
Radiologists		Change of final assessment	Benign, n (%)	Malignant, n (%)	Total, n
Less experienced	1	Upgrade	7 (35.0%)	13 (65.0%)	20
		Downgrade	5 (100.0%)	0 (0%)	5
	2	Upgrade	5 (50.0%)	5 (50.0%)	10
		Downgrade	14 (93.3%)	1 (6.7%)	15
	3	Upgrade	6 (60.0%)	4 (40.0%)	10
		Downgrade	8 (100.0%)	0 (0%)	8
Experienced	4	Upgrade	2 (50.0%)	2 (50.0%)	4
		Downgrade	10 (100.0%)	0 (0%)	10
	5	Upgrade	1 (100.0%)	0 (0%)	1
		Downgrade	8 (88.9%)	1 (11.1%)	9

Upgrade: radiologist changed the final assessments from possibly benign to possibly malignant after CAD combination; downgrade: radiologist changed the final assessments from possibly malignant to possibly benign after CAD combination.

CAD = computer-aided diagnosis.

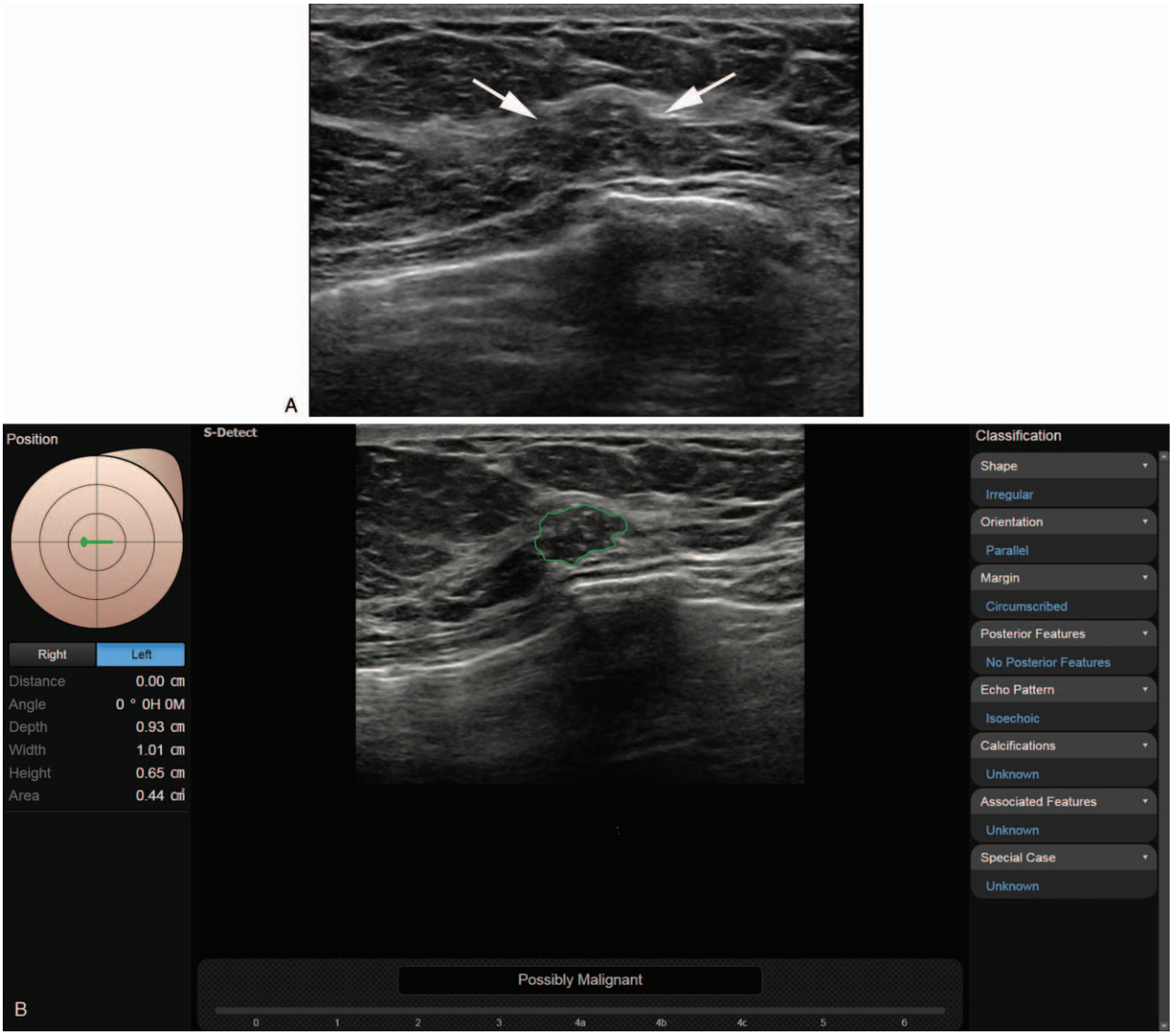

Figure 1. A. The grayscale ultrasound image in a 57-year-old woman with incidentally detected breast mass on screening examination shows an indistinct irregular heterogeneous hypoechoic mass (arrows) at the 9 o'clock position in the left breast that was diagnosed as breast imaging reporting and data system (BI-RADS) category 3, 4a, and 3, respectively, by less experienced radiologists and 4a and 4b, respectively, by experienced radiologists. B. After review of the CAD application, (where the conclusion was "possibly malignant"), each reviewer recategorized the mass as 4a, 4a, 4b, 4b, and 4c, respectively; core biopsy confirmed the lesion as ductal carcinoma in situ. CAD = computer-aided diagnosis.

radiologist, with moderate agreement of US characteristics between the CAD system and the radiologist. However, another study showed that the usefulness of CAD for breast US could differ according to the degree of experience of the radiologist. Choi et al[10] reported that combining CAD with breast US led to improved specificity and AUC for both experienced (5 years of breast imaging experience) and less experienced radiologists (1st-year residents with 1 week of training in breast imaging), while sensitivity was improved only for less experienced radiologists. Wang et al[14] evaluated the benefit of CAD when used by 8 radiologists with varied experience in breast US (ranging from 1 to 16 years; 4 residents and 4 senior radiologists). The AUCs of all residents were significantly improved when interpretation was performed using CAD, whereas AUCs of only half of the senior radiologists showed significant improvement when CAD was integrated. These studies suggest that the benefits of CAD may be greater for less experienced radiologists.

Our study also showed that diagnostic performance was significantly improved in all the radiologists when CAD was combined with US. However, there were differences in the benefits of using CAD between the less experienced and experienced radiologists. In less experienced radiologists, the NPV and AUC were significantly improved after CAD assistance compared with the performance of radiologists without CAD. However in experienced radiologists, the specificity and PPV were improved significantly after CAD assistance ($P < .05$). Based on our results, CAD assistance could be useful as a second opinion for making the final decision and in improving diagnostic performance of radiologists, regardless of the level of experience. However, the type of benefit would be different between the less experienced and experienced radiologists.

BI-RADS categorization is extremely important as it directly impacts the management plan.[1] Previous studies have shown that the use of CAD can lead to a change in the final BI-RADS

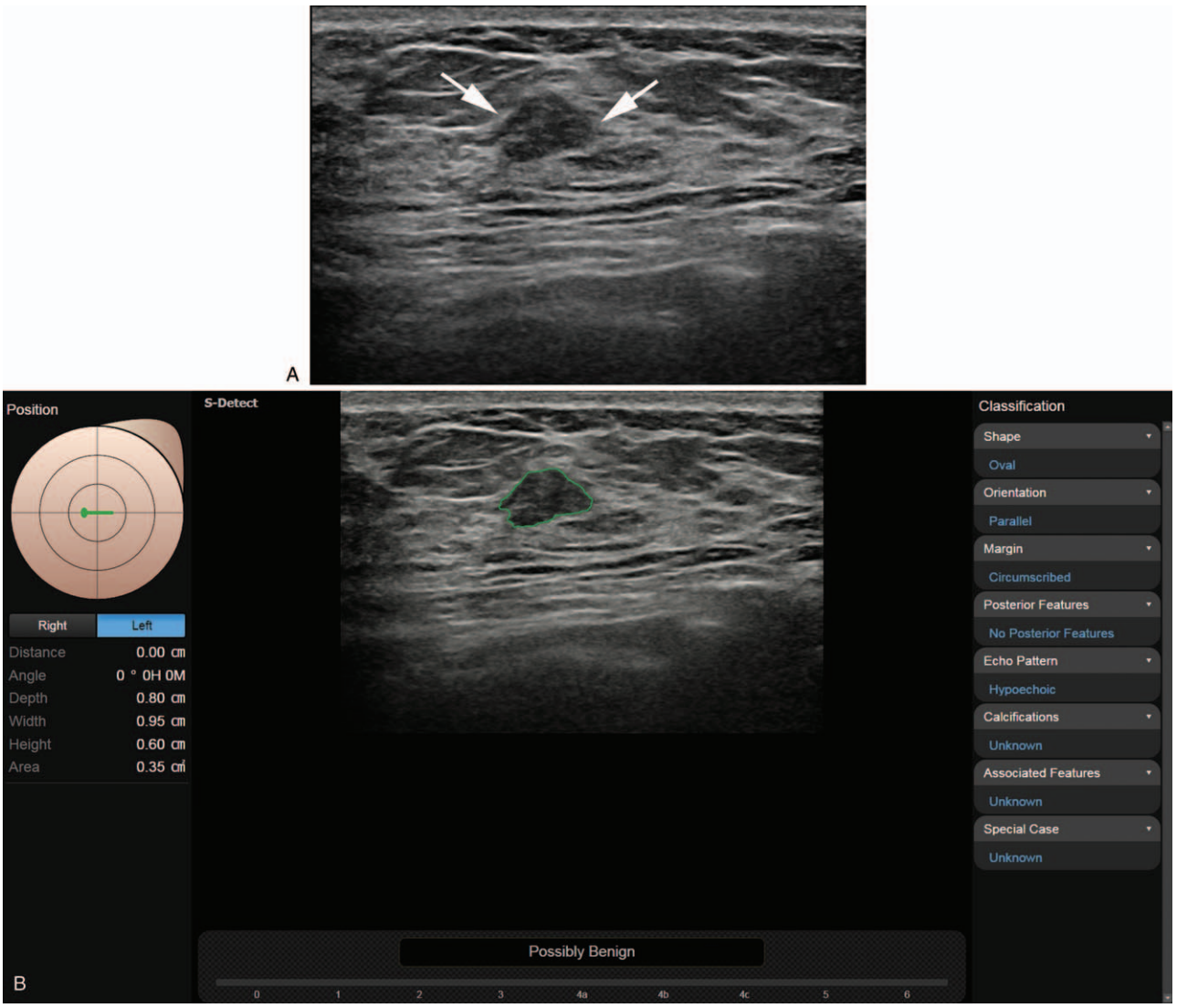

Figure 2. A. The grayscale ultrasound image in a 35-year-old woman with a palpable lesion in her right breast, shows an indistinct oval hypoechoic mass (arrow) at the 9 o'clock position in the left breast that was diagnosed as breast imaging reporting and data system (BI-RADS) category 4a, 4a, and 4a, respectively, by 3 fellowship-trained radiologists, and as category 4a and 3, respectively, by experienced radiologists. B. After review of the CAD application, (where the conclusion was "possibly benign"), each reviewer recategorized the mass as 4a, 3, 3, 3, and 3, respectively; core biopsy finally confirmed the mass as fibroadenoma. CAD = computer-aided diagnosis.

classification with a significant rate of correct re-classification.[15,16] Bartolotta et al[15] showed a re-classification rate of 21.3% by 2 experienced radiologists after the use of CAD and 81% cases were correctly re-classified. Among the re-classified lesions, there was a correct change in clinical management in 42.2% cases and incorrect change in clinical management in 18.7% cases. In our study, the final assessments, according to BI-RADS categories, were reclassified 10% to 25% by less experienced radiologists and 10% to 14% by experienced radiologists after CAD combination. Less experienced radiologists correctly upgraded 40% to 65% cases that were initially assigned to BI-RADS category 3 benign lesion to BI-RADS category 4 or 5 malignant lesion, and experienced radiologists correctly upgraded 0% to 50% cases. The percentage of correct downgrading of cases initially assigned to BI-RADS category 4 or 5 malignant lesion to BI-RADS category 3 benign lesion in all reviewers was 89% to 100%. As biopsy is recommended for breast masses classified as category 4 or higher, these high correct downgrade rates with CAD assistance can reduce misdiagnosis and unnecessary breast biopsies.

Interobserver variability in all radiologists, with various levels of experience, measured using kappa values, significantly improved with CAD assistance compared with grayscale US. Previous studies have reported that kappa values for each BI-RADS US descriptor had fair to substantial agreement (0.33–0.69), and fair to moderate agreement (0.28–0.53) for final BI-RADS categories between radiologists with a variable range of experience.[17–19] In our study, slight to moderate agreement was observed for US descriptors (0.196–0.478) among all radiologists; this is a relatively low rate of agreement compared with the previous studies. However, when CAD was used, kappa values for the US descriptors improved significantly to fair to moderate agreement (0.285–0.546). The 2 level upgrade (slight to moderate) was observed for orientation and posterior acoustic features in less experienced radiologists, whereas, only 1 level upgrade or no change was observed in the experienced

Table 5

Interobserver variability of US characteristics and final assessment categories among the radiologists without and with the computer-aided diagnosis.

US features	All			Less experienced			Experienced		
	Kappa (95% CI)			Kappa (95% CI)			Kappa (95% CI)		
	Without CAD	With CAD	*P* value	Without CAD	With CAD	*P* value	Without CAD	With CAD	*P* value
Shape	0.478 (0.422–0.535)	0.544 (0.487–0.600)	<.001	0.541 (0.440–0.643)	0.655 (0.553–0.758)	<.001	0.478 (0.422–0.535)	0.656 (0.473–0.838)	<.001
Orientation	0.322 (0.260–0.384)	0.546 (0.485–0.606)	<.001	0.139 (0.026–0.252)	0.560 (0.447–0.673)	<.001	0.584 (0.388–0.780)	0.584 (0.402–0.767)	.156
Margin	0.196 (0.156–0.236)	0.285 (0.242–0.328)	<.001	0.152 (0.076–0.229)	0.258 (0.177–0.339)	<.001	0.219 (0.097–0.341)	0.395 (0.267–0.523)	<.001
Posterior acoustic feature	0.266 (0.221–0.310)	0.350 (0.304–0.396)	<.001	0.164 (0.084–0.244)	0.455 (0.363–0.548)	<.001	0.374 (0.230–0.520)	0.499 (0.363–0.634)	<.001
Echo pattern	0.302 (0.258–0.347)	0.401 (0.350–0.451)	<.001	0.408 (0.330–0.487)	0.570 (0.469–0.671)	<.001	0.249 (0.100–0.398)	0.431 (0.281–0.581)	<.001
Final assessment categories	0.221 (0.168–0.275)	0.320 (0.268–0.371)	<.001	0.186 (0.087–0.285)	0.412 (0.324–0.499)	<.001	0.260 (0.096–0.425)	0.510 (0.314–0.706)	<.001

CAD = computer-aided diagnosis; CI = confidence interval; US = ultrasonography.

radiologists. This could be because inexperienced radiologists might find analyzing the intrinsic characteristics of breast masses or selecting the terms of US features difficult. However, we believe that by providing uniform knowledge through the CAD and with consistent diagnostic performance, the interobserver variability can be reduced. In addition, agreement between all the radiologists with regard to final assessment improved from 0.221 to 0.320 after CAD assistance, but the improvements were only up to fair agreement. Although improved, the levels of agreement were still relatively low. We believe the reason for this to be the fact that CAD provides the final assessment in a dichotomized form (possibly benign and possibly malignant), whereas, we classified the lesions into BIRADS-category 3, 4, and 5. However in the less experienced group, the 2 levels (slight: 0.186 to moderate: 0.412) upgrade was observed in the final assessment. In the experienced radiologists, the upgrade was fair (0.260) to moderate (0.510).

This study has several limitations. First, data were obtained by 1 experienced radiologist initially and analysis was then performed by the less experienced and experienced radiologists retrospectively. However, to make the images as real-time grayscale as possible, we used cine images and the same acquired grayscale US and CAD data were analyzed by the 5 readers; we believe that this would have reduced the operator dependency in image acquisition, and that is a major advantage of this study. Second, selection bias probably exists in this study, as only biopsy-confirmed lesions were included. Therefore, BI-RADS category 2 lesions were not included. Third, the value of CAD was not evaluated for calcifications and non-mass lesions, since analysis was limited to lesions within the present breast US CAD system. Further technical developments are required in this area. Finally, the data were derived from a relatively small number of patients. Hence, large-scale studies will be required in the future to generalize the results.

In conclusion, CAD is a useful additional diagnostic tool to improve the diagnostic performance in all radiologists with different benefits for radiologists with different levels of experience in breast imaging. It may be helpful in refining lesion descriptions and in making management decisions, such as the need for biopsy and determination of clinical strategy. In addition, CAD improved the interobserver agreement and showed acceptable agreement in the characterization of breast masses, especially in less experienced radiologists.

Acknowledgments

The authors thank the Division of Statistics of the Medical Research Collaborating Center at Seoul National University Bundang Hospital for the statistical analyses.

Author contributions

Conceptualization: Sun Mi Kim.
Data curation: Hee Jeong Park, Bo La Yun, Mijung Jang, Ja Yoon Jang, Jong Yoon Lee.
Formal analysis: Ja Yoon Jang, Jong Yoon Lee.
Investigation: Soo Hyun Lee.
Methodology: Sun Mi Kim, Bohyoung Kim.
Project administration: Sun Mi Kim.
Validation: Soo Hyun Lee.
Writing – original draft: Hee Jeong Park.
Writing – review & editing: Bo La Yun, Mijung Jang, Bohyoung Kim.

Hee Jeong Park orcid: 0000-0002-9378-7775.

Sun Mi Kim orcid: 0000-0003-0899-3580.
Bo La Yun orcid: 0000-0002-5457-7847.
Mijung Jang orcid: 0000-0001-9619-6877.
Bohyoung Kim orcid: 0000-0002-2183-5651.
Ja Yoon Jang orcid: 0000-0001-7945-7684.
Jong Yoon Lee orcid: 0000-0003-4972-2357.
Soo Hyun Lee orcid: 0000-0002-4178-2008.

References

[1] D'Orsi CJ. ACR BI-RADS Atlas: Breast Imaging Reporting and Data System. 5th ed.American College of Radiology, Reston:2013.
[2] Singh S, Maxwell J, Baker JA, Nicholas JL, Lo JY. Computer-aided classification of breast masses: performance and interobserver variability of expert radiologists versus residents. Radiology 2011;258:73–80.
[3] Doi K, MacMahon H, Katsuragawa S, Nishikawa RM, Jiang Y. Computer-aided diagnosis in radiology: potential and pitfalls. Eur J Radiol 1999;31:97–109.
[4] Vyborny CJ, Giger ML, Nishikawa RM. Computer-aided detection and diagnosis of breast cancer. Radiol Clin North Am 2000;38: 725–40.
[5] Giger ML, Karssemeijer N, Armato SG3rd. Computer-aided diagnosis in medical imaging. IEEE Trans Med Imaging 2001;20:1205–8.
[6] Horsch K, Giger ML, Vyborny CJ, Lan L, Mendelson EB, Hendrick RE. Classification of breast lesions with multimodality computer-aided diagnosis: observer study results on an independent clinical data set. Radiology 2006;240:357–68.
[7] Horsch K, Giger ML, Vyborny CJ, Venta LA. Performance of computer-aided diagnosis in the interpretation of lesions on breast sonography. Acad Radiol 2004;11:272–80.
[8] Kim K, Song MK, Kim EK, Yoon JH. Clinical application of S-Detect to breast masses on ultrasonography: a study evaluating the diagnostic performance and agreement with a dedicated breast radiologist. Ultrasonography 2017;36:3–9.
[9] Lee JH, Seong YK, Chang CH. Novak CL, Aylward S, et al. Computer-aided lesion diagnosis in B-mode ultrasound by border irregularity and multiple sonographic features. Medical Imaging 2013: Computer-Aided Diagnosis 2013;International Society for Optics and Photonics, 86701O.
[10] Choi J, Kang B, Baek J, Lee H, Kim S. Application of computer-aided diagnosis in breast ultrasound interpretation: improvements in diagnostic performance according to reader experience. Ultrasonography 2018;37:217–25.
[11] Cho E, Kim EK, Song MK, Yoon JH. Application of computer-aided diagnosis on breast ultrasonography: evaluation of diagnostic performances and agreement of radiologists according to different levels of experience. J Ultrasound Med 2018;37:209–16.
[12] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
[13] Vanbelle S, Albert A. A bootstrap method for comparing correlated kappa coefficients. J Stat Comput Simulat 2008;78:1009–15.
[14] Wang Y, Jiang S, Wang H, et al. CAD algorithms for solid breast masses discrimination: evaluation of the accuracy and interobserver variability. Ultrasound Med Biol 2010;36:1273–81.
[15] Bartolotta TV, Orlando A, Cantisani V, et al. Focal breast lesion characterization according to the BI-RADS US lexicon: role of a computer-aided decision-making support. Radiol Med 2018;123: 498–506.
[16] Buchbinder SS, Leichter IS, Lederman RB, et al. Computer-aided classification of BI-RADS category 3 breast lesions. Radiology 2004;230:820–3.
[17] 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;239:385–91.
[18] Lee HJ, Kim EK, Kim MJ, et al. Observer variability of breast imaging reporting and data system (BI-RADS) for breast ultrasound. Eur J Radiol 2008;65:293–8.
[19] Park CS, Lee JH, Yim HW, et al. Observer agreement using the ACR Breast Imaging Reporting and Data System (BI-RADS)-ultrasound, First Edition (2003). Korean J Radiol 2007;8:397–402.