Should We Ignore, Follow, or Biopsy? Impact of Artificial Intelligence Decision Support on Breast Ultrasound Lesion Assessment

Victoria L. Mango[1]
Mary Sun[2]
Ralph T. Wynn[2]
Richard Ha[2]

OBJECTIVE. The objective of this study was to assess the impact of artificial intelligence (AI)-based decision support (DS) on breast ultrasound (US) lesion assessment.

MATERIALS AND METHODS. A multicenter retrospective review of 900 breast lesions (470/900 [52.2%] benign; 430/900 [47.8%] malignant) on US by 15 physicians (11 radiologists, two surgeons, two obstetrician/gynecologists). An AI system (Koios DS for Breast, Koios Medical) evaluated images and assigned them to one of four categories: benign, probably benign, suspicious, and probably malignant. Each reader reviewed cases twice: 750 cases with US only or with US plus DS; 4 weeks later, cases were reviewed in the opposite format. One hundred fifty additional cases were presented identically in each session. DS and reader sensitivity, specificity, and positive likelihood ratios (PLRs) were calculated as well as reader AUCs with and without DS. The Kendall τ-b correlation coefficient was used to assess intra- and interreader variability.

RESULTS. Mean reader AUC for cases reviewed with US only was 0.83 (95% CI, 0.78–0.89); for cases reviewed with US plus DS, mean AUC was 0.87 (95% CI, 0.84–0.90). PLR for the DS system was 1.98 (95% CI, 1.78–2.18) and was higher than the PLR for all readers but one. Fourteen readers had better AUC with US plus DS than with US only. Mean Kendall τ-b for US-only interreader variability was 0.54 (95% CI, 0.53–0.55); for US plus DS, it was 0.68 (95% CI, 0.67–0.69). Intrareader variability improved with DS; class switching (defined as crossing from BI-RADS category 3 to BI-RADS category 4A or above) occurred in 13.6% of cases with US only versus 10.8% of cases with US plus DS ($p = 0.04$).

CONCLUSION. AI-based DS improves accuracy of sonographic breast lesion assessment while reducing inter- and intraobserver variability.

Keywords: artificial intelligence, breast cancer, breast ultrasound, computer-aided diagnosis, machine learning

doi.org/10.2214/AJR.19.21872

Received June 17, 2019; accepted after revision November 22, 2019.

Supported in part by the National Institutes for Health/National Cancer Institute Cancer Center Support Grant (P30 CA008748).

[1]Memorial Sloan Kettering Cancer Center, Breast and Imaging Center, 300 E 66th St, Ste 715, New York, NY 10065. Address correspondence to V. L. Mango (mangov@mskcc.org).

[2]Department of Radiology, Columbia University Medical Center, New York, NY.

AJR 2020; 214:1–8

ISSN-L 0361–803X/20/2146–1

The global cancer burden is increasing. By the end of this century, cancer is expected to rank as the leading cause of death in every country in the world [1]. In 2018, breast cancer was the most commonly diagnosed cancer in women other than nonmelanoma skin cancer, with over 2 million new cases diagnosed worldwide, and the leading cause of cancer death in female patients [1]. Given increasing oncologic needs of patients and complex oncologic imaging challenges, research increasingly focuses on artificial intelligence (AI) as a tool for detection, clinical decision making, diagnosis, characterization, and workflow support for radiologists [2, 3].

AI has been used in breast imaging for decades, with the use of computer-aided detection (CAD) reducing false-negative mammography interpretations and increasing detection of early-stage malignancies [3–5]. Breast imaging AI research has expanded broadly in recent years to include artificial neural networks, support vector machines, and deep learning applications in mammography, ultrasound (US), and MRI [3, 6–10]. Additionally, the U.S. Food and Drug Administration (FDA) recently granted approval for a CAD system used with screening automated breast US [11].

Breast US improves detection of small, invasive, node-negative cancers [12, 13]. In routine clinical practice, the classification and management of breast lesions depends on the radiologist's visual assessment, guided by the American College of Radiology's *BI-RADS Atlas* [14]. For lesions seen on US, this assessment is primarily based on lesion shape, orientation, margin, echo pattern, and posterior features [14]. Despite this guidance, breast US has low specificity and low positive predictive values (PPVs). Large

interobserver variability for lesion management has also been reported in clinical practice [12, 13, 15–18].

Initial applications of CAD to breast lesion analysis on US showed improved sensitivity for junior radiologists but decreased specificity for experienced radiologists [19]. A more recent study used a new AI-based decision support system and reported improved sensitivity and specificity with CAD, exceeding radiologist performance, with a 34–55% potential reduction in benign breast biopsies and increase in PPV of biopsies performed by 7–20% for the three radiologists studied [20]. Additionally, CAD has enabled minimally trained nonradiologist healthcare workers to triage palpable breast lumps with a low-cost portable US system with accuracy similar to that of specialist radiologist assessments [21].

Building on prior studies, we use the Koios Decision Support (DS) for Breast system, a software application designed to assist physicians in analyzing breast US images (Koios Medical). Using machine learning and AI, Koios DS for Breast automatically generates a probability of malignancy for a user-selected ROI that contains a breast lesion. This probability is then mapped into four categories (benign, probably benign, suspicious, probably malignant) via alignment to likelihood of malignancy (LOM). The four Koios output categories align with BI-RADS categories as follows: benign LOM denotes less than 0.5% BI-RADS 2, probably benign LOM denotes less than 2% BI-RADS 3, suspicious LOM denotes less than 50% BI-RADS 4A or 4B, and probably malignant LOM denotes more than 50% BI-RADS 4C or 5.

A continuous graphical confidence level indicator shows where a lesion falls within each LOM range. Previous studies used earlier versions of the core system presented in this article [21–23]. The system evaluated in this study has undergone training data, architecture, and user interface improvements that have significantly improved its clinical impact.

The purpose of this study was to assess the impact of AI-based DS on breast US lesion assessment.

Materials and Methods

Overall Study Design

This study was approved by the Western Institutional Review Board. It was a HIPAA-compliant, multicenter retrospective review of 900 breast lesions. These lesions were visualized on US by 15 physicians who provided informed consent to participate in this study; patient informed consent was waived. Physician informed consent was obtained on the basis of recommendations from the Center for Devices and Radiologic Health at the FDA.

Patient Population and Breast Lesion Characteristics

Nine hundred women (mean age, 53.6 years; range, 17–96 years) with breast lesions on US images acquired between June 2004 and June 2016 were included. Lesions were identified from screening mammography recalls and scheduled biopsies from over 20 U.S. institutions with identifying information, including institution, removed during anonymization. Target patient demographics including race or ethnicity and tumor size distribution were informed by the Breast Cancer Surveillance Consortium (2006–2009) to ensure the study population was representative of national rates [22]. Race or ethnicity distribution was 592/900 (65.8%) white, 77/900 (8.6%) black, 73/900 (8.1% Hispanic), 133/900 (14.8%) Asian and 25/900 (2.8%) other.

Lesion size on US was as follows: 332/900 (36.9%), smaller than 10 mm; 229/900 (25.4%), 10–14 mm; 132/900 (14.7%), 15–19 mm; and 200/900 (22.2%), larger than 20 mm. Size was unavailable for seven lesions (0.8%) at the time of inclusion because the original images did not show calipers. ROIs were subsequently provided, and those cases were included in further analysis. Lesions were 470/900 (52.2%) benign and 430/900 (47.8%) malignant, as confirmed by pathology (BI-RADS 4 or 5 lesions) or stability on follow-up US for at least 1 year (BI-RADS 2 or 3 lesions). Of the 470 benign lesions, 249 (53.0%) underwent biopsy that yielded benign results, whereas 221 (47.0%) had at least 1 year of stability on follow-up US. Of the 430 malignant lesions, 369 (85.8%) were invasive cancer (320 invasive ductal carcinomas, 37 invasive lobular carcinomas, 12 mixed ductal and lobular features) and 61 (14.2%) were ductal carcinoma in situ (DCIS). Invasive cancers had a median largest dimension of 1.4 cm (0.3–6.4 cm).

The initial BI-RADS assessment for lesions, as determined by the reading clinical radiologist, was as follows: 68/900 (7.6%) BI-RADS 2, 173/900 (19.2%) BI-RADS 3, 562/900 (62.4%) BI-RADS 4, and 97/900 (10.8%) BI-RADS 5.

Ultrasound Imaging

Each breast lesion had static orthogonal US images available. Images were acquired with a minimum 12-MHz acquisition frequency. Cases were categorized as obtained at low (< 15 MHz) and high (≥ 15 MHz) frequency quality categories with equal distribution of these categories represented. Images were acquired on US units from a variety of vendors: 697/900 (77.4%) GE Healthcare, 170/900 (18.9%) Philips Healthcare, 26/900 (2.9%) Siemens Healthineers, 6/900 (0.7%) Toshiba, and 1/900 (0.1%) Supersonic.

Reader Characteristics and Training

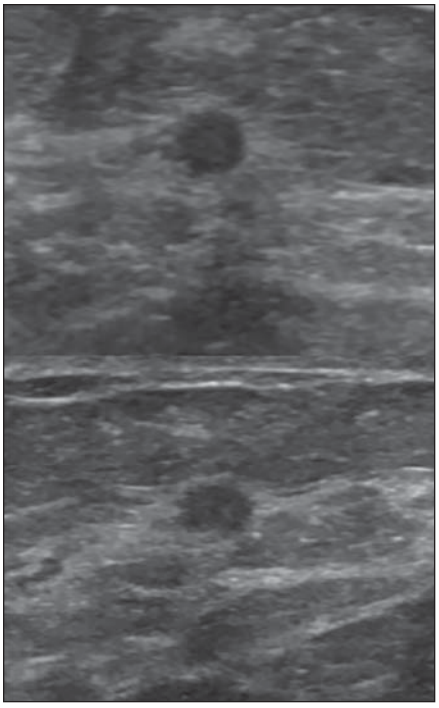
All 15 readers were physicians: 11 diagnostic radiologists who read breast imaging as part of their clinical practice, two breast surgeons, and two obstetrician/gynecologists. The surgeon and obstetrician/gynecologist readers were included because they are potential users of the DS system in their clinical practice, as permitted by the FDA. Readers had 0–39 years' experience reading US images. Four of the 11 diagnostic radiologists had completed breast fellowship training. Attending radiologists had a mean of 11.4 years of experience after fellowship (range, 2–31 years). Physicians were from 13 different centers; nine are in private practice, and six are in academic practices.

Each reader initially participated in a 30-minute online training session to ensure understanding of the software system and case format presentation. To avoid potential selection bias, readers were not tested for competency; however, all readers had direct access to the DS manufacturer support if questions arose about the platform or the study.

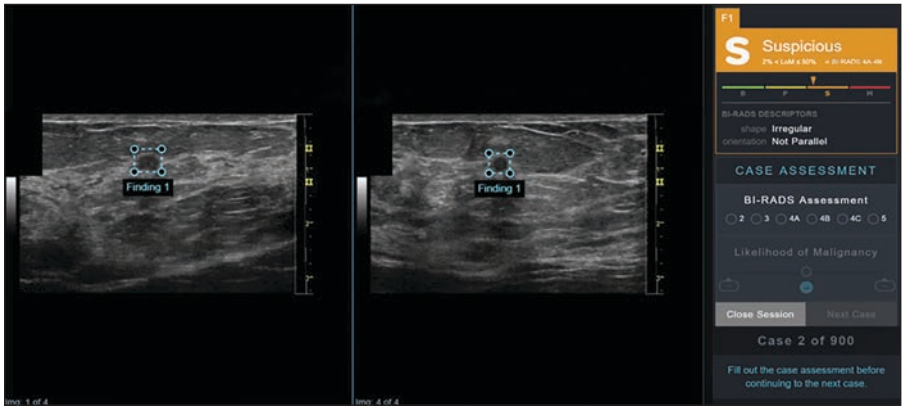
Artificial Intelligence Decision Support System

Two orthogonal views of each lesion with an ROI, as placed by the original interpreting physician on the original clinical images, were evaluated by the machine learning system, which generated a BI-RADS–aligned assessment [20]. The system processes these images via an ensemble of algorithms using either pathology or imaging follow-up as ground truth in its training process. The training data (over 400,000 clinical examples) were gathered from over 25 machines and 25 different healthcare systems and sites. The 900 cases used in this validation study were completely excluded from the testing and development of the AI system.

To provide its clinical assessment, the system generates a probability of malignancy that is then used to assign one of four categoric outputs: benign, probably benign, suspicious, or probably malignant. These categories were designed to align to the LOM values for BI-RADS 2, 3, 4A or 4B, and 4C or 5 assessments. In addition, the system provides a continuous output that represents the confidence of the assessment within each category. These score ranges and categories are inherent to the system and were not designed or altered for this study. These scores were presented to study readers in a graphical form as an electronic case report form and constituted the AI DS (Fig. 1).



A



B

Fig. 1—75-year-old woman with invasive ductal carcinoma.
A, Orthogonal ultrasound transverse (*top*) and sagittal (*bottom*) images of 0.4-cm breast mass that could be categorized as oval and parallel and interpreted as benign or probably benign by reader.
B, Artificial intelligence decision support (DS) output scores were presented to study readers in graphical form as electronic case report form in conjunction with orthogonal ultrasound images of lesion for that case. Right panel shows categoric assessment, in this case "suspicious," with triangle marker indicating confidence of assessment within that category. In this example, DS support correctly classifies this lesion as suspicious; malignancy (invasive ductal carcinoma) was confirmed by ultrasound-guided biopsy. LoM = likelihood of malignancy, B = benign, P = probably benign, S = suspicious, M = probably malignant.

Reader Workflow

Cases were presented to and scored by readers in a single software environment (Koios DS Study Tool, version 2.0.0.1, Koios Medical) that showed two orthogonal US images of a breast lesion with ROIs. ROIs were based on measurement calipers as used in usual practice. During the course of the study, readers had the ability to toggle the overlay presence to ensure the lesion could be well visualized without overlap of the ROI or label on the lesion itself. This ROI overlap would be especially problematic for small lesions and was addressed during user training. Cases with DS displayed the AI output in the electronic case report form. The study platform instructed the reader to input a BI-RADS category (BI-RADS 2, 3, 4A, 4B, 4C, or 5) and LOM as a percentage.

All 900 cases were reviewed twice, in two sessions (900 cases per session) separated by a 4-week washout period. Each session consisted of 750 US cases randomized to US images alone (US only) or US images plus the AI DS (US plus DS). In the second session, the same US images were provided in the opposite format (i.e., cases that were initially US plus DS were displayed as US only and vice versa).

In addition, in each session, 150 cases (75 US only and 75 US plus DS) were presented to readers identically without switching the reading condition. These cases were included to assess intraoperator variability.

System Evaluation

To evaluate the system's robustness to variation in the ROI boundaries, two assessments were performed.

First, we assessed DS output robustness to ROI boundary variation by evaluating the 900 cases 40 times, randomly varying the ROI boundary each time. Specifically, each corner of the region was shifted at random by up to 20% from the predetermined optimal cropping. ROC curves and AUC distributions were calculated. Second, we evaluated the level of DS output switching between probably benign and suspicious caused by ROI variability using the 40 ROI boundary variations generated in the previous analysis, and we counted the number of times a category was switched, compared with the initial DS output based on the ROI drawn on the original US image.

In addition, the potential impact of US transducer frequency (low versus high) on DS output was examined. Each frequency subset was bootstrapped such that it contained a statistically equivalent BI-RADS distribution. These groups of low- and high-frequency cases were analyzed via ROC curves, and their respective AUCs were computed.

System and Reader Performance Evaluation

System and reader performance were evaluated by comparing the positive likelihood ratio (PLR) of the DS output to each reader. PLR is defined as sensitivity / (1 – specificity) and indicates the likelihood that a biopsy recommendation is indicative of malignancy. Because of the system's four potential categoric outputs, we considered both a benign and a probably benign output from the system as a recommendation to not biopsy; suspicious and probably malignant outputs were considered recommendations to biopsy. This grouping is in alignment with the system's BI-RADS risk alignment, in which the first two categories align to BI-RADS 2 and 3, and the latter two align with BI-RADS 4A–5.

The AUC for each radiologist was calculated and compared between reading paradigms (US only vs US plus DS). The estimate of the change in AUC with and without DS and 95% CIs were made using the Dorfman-Berbaum-Metz method of multireader multicase analysis [24].

Intra- and interoperator variability were assessed via Kendall τ-b correlation coefficient [25]. This assessment was done in a pairwise fashion across each pair of readers before and after being provided with the DS output.

Finally, to assess the specific cases in which readers agreed or disagreed with the DS system, total disagreement counts were computed and probed to determine whether the system or the reader was correct. These data were further evaluated to ascertain the net effects of whether the system or the reader was correct based on pathologic results, when available, or at least 1-year sonographic stability.

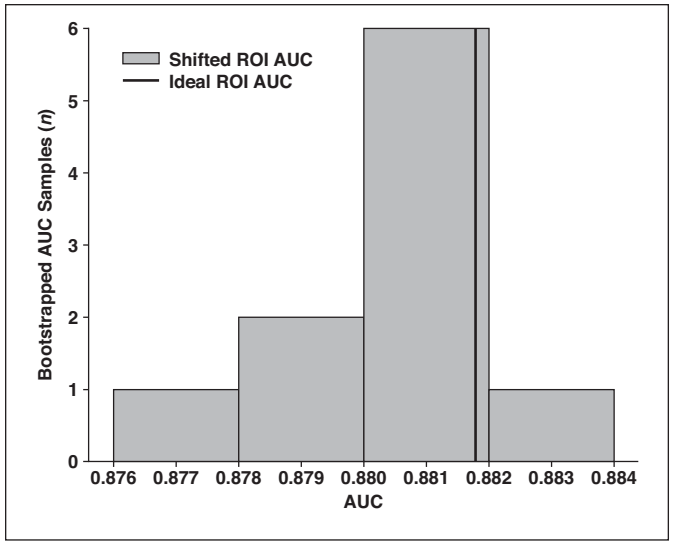


Fig. 2—Bar graph shows decision support output variation due to ROI boundary variation (varied by up to 20% in each dimension) in all 900 cases, which was assessed by randomly varying ROI boundary 40 times from ROI drawn on initial clinical ultrasound images (line labeled "Ideal ROI AUC" indicates ROI generated by original interpreting physician). Class switching from probably benign to suspicious was 1.9% and suspicious to probably benign was 1.7%.

Results

System Evaluation

Variability in ROI boundary produced no significant change in either the shape of the ROC curve or the AUC values (Fig. 2). Similarly, the ROI boundary variation showed minimal class switching between probably benign to suspicious (1.9%) and suspicious to probably benign (1.7%).

The transducer frequency dependence analysis generated ROC curves with the AUC for the low- and high-frequency cases (0.876 [95% CI, 0.838–0.916] and 0.893 [95% CI, 0.848–0.926], respectively). These results show that there is no statistically significant difference in performance of the DS system between the low- and high-frequency US transducers ($p = 0.56$).

System and Reader Lesion Evaluation

Results for each of the categoric assessments supplied by the DS system (using the ROI drawn by the initial clinical radiologist) were compared with the BI-RADS assessments provided by the 15 readers without DS (US only). DS sensitivity was 0.98 (95% CI, 0.96–0.99) and mean reader sensitivity was 0.94 (95% CI, 0.91–0.96). DS specificity was 0.50 (95% CI, 0.45–0.55) and mean reader specificity was 0.40 (95% CI, 0.36–0.45). Individual reader operating points (defined as the sensitivity and specificity at the BI-RADS 3–BI-RADS 4A boundary) for US only and US plus DS can be seen in Figure 3. The DS system misclassified seven cancers as benign or probably benign: two (28.6%) invasive ductal carcinoma (IDC), four (57.1%) DCIS, and one (14.3%) mixed IDC and DCIS. Misclassified invasive cancers included one grade 1 tumor and two grade 2 tumors. Misclassified DCIS cases included two grade 3 tumors and two grade 1 tumors. Additionally, 190 benign lesions were marked as suspicious, and 42 benign lesions were marked as probably malignant by the system.

To evaluate the alignment on metrics beyond sensitivity, PLR, negative likelihood ratio, PPV, and negative predictive values were calculated for DS and all participating readers (Table 1)

Further analysis yielded an AUC for the DS system alone of 0.88 (95% CI, 0.86–0.91). Mean US-only AUC was 0.83 (95% CI, 0.78–0.89), whereas mean US-plus-DS AUC was 0.87 (95% CI, 0.84–0.90). To assess if the difference of these values was significant, a two-tailed *t* test was performed that yielded a statistically significant difference ($p < 0.0001$). Each reader's performance with and without DS is shown in Figure 3, along with a tabular representation of the differences and standard error in Table 2. Mean ROC curves and their corresponding AUC values can be seen in Figure 4.

Reader 1 was the only reader below the equivalence line, although the difference was not statistically significant. On further investigation, this reader's AUC shift was due to a reduction in sensitivity (from 0.92 with US only to 0.91 with US plus DS) and increase in specificity (0.56 with US only to 0.57 with US plus DS). To establish whether the DS system was responsible for a reduction in sensitivity, confusion tables were built for the malignant lesions and DS recommendations (Table 3). Table 3 shows that the drop in sensitivity for reader 1 was due exclusively to intrareader variability; all cases in which reader 1 switched from a biopsy recommendation (BI-RADS category of 4 or higher) to a nonbiopsy recommendation had a DS output that recommended biopsy.

To characterize the effect of the DS (US plus DS) system on interreader variability, we computed the Kendall τ-b correlation coefficient in a pairwise manner for all readers. The mean Kendall τ-b was 0.54 (95% CI, 0.53–0.55) for US only and 0.68 (95% CI, 0.67–0.69) for US plus DS, showing a significant shift in this metric ($\alpha = 0.05$).

To assess for intrareader variability, the 150 cases for which the format did not change between reading sessions (75 US only and 75 US plus DS) were analyzed.

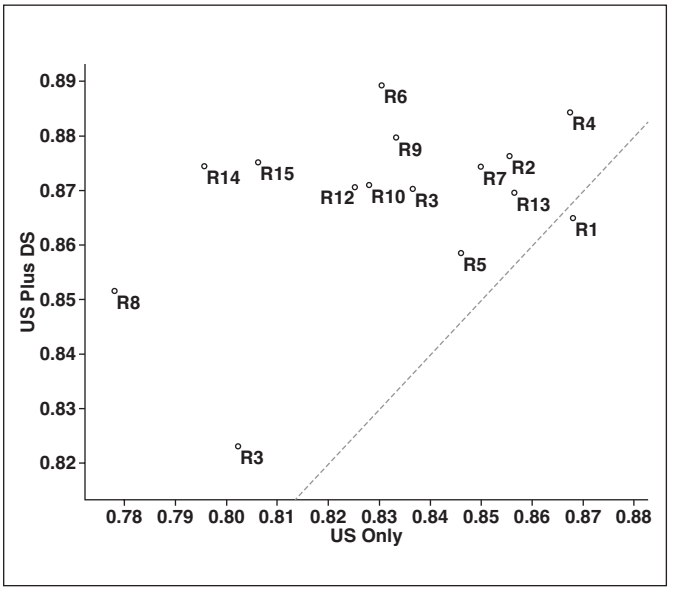


Fig. 3—Scatterplot shows comparison of AUCs for assessments with ultrasound (US) only versus those with US plus decision support (US plus DS) for 15 readers (R1–R4, breast fellowship–trained radiologists; R5–R11, radiologists without breast fellowship training; R12 and R13, breast surgeons; R14 and R15, obstetrician/gynecologists). Only R1 showed decrease in performance with US plus DS compared with US only; difference was not statistically significant. Stand-alone DS performance is not shown but had AUC of 0.88 (95% CI, 0.86–0.91). Dashed line denotes equivalency.

TABLE 1: PLR, NLR, PPV, and NPV for DS and 15 Readers Evaluating Cases Without DS (Ultrasound Only)

Reader	PLR	NLR	PPV	NPV
DS output	1.98 (1.78–2.18)	0.03 (0.03–0.04)	0.64 (0.60–0.68)	0.97 (0.94–0.99)
Reader 1	2.08 (1.88–2.29)	0.14 (0.13–0.15)	0.66 (0.61–0.70)	0.89 (0.84–0.92)
Reader 2	1.18 (0.89–1.47)	0.08 (0.06–0.10)	0.52 (0.48–0.56)	0.93 (0.84–0.98)
Reader 3	1.95 (1.75–2.15)	0.28 (0.25–0.30)	0.64 (0.60–0.68)	0.80 (0.75–0.84)
Reader 4	1.77 (1.57–1.98)	0.07 (0.07–0.08)	0.62 (0.58–0.66)	0.94 (0.89–0.97)
Reader 5	1.35 (1.11–1.58)	0.06 (0.05–0.07)	0.55 (0.51–0.59)	0.95 (0.89–0.98)
Reader 6	1.46 (1.25–1.68)	0.11 (0.10–0.13)	0.57 (0.53–0.61)	0.91 (0.85–0.95)
Reader 7	1.88 (1.68–2.08)	0.10 (0.09–0.11)	0.63 (0.59–0.67)	0.92 (0.87–0.95)
Reader 8	1.56 (1.36–1.76)	0.20 (0.17–0.22)	0.59 (0.55–0.63)	0.85 (0.79–0.90)
Reader 9	1.74 (1.54–1.94)	0.14 (0.13–0.16)	0.61 (0.57–0.66)	0.88 (0.83–0.92)
Reader 10	1.72 (1.52–1.92)	0.10 (0.09–0.11)	0.61 (0.57–0.65)	0.92 (0.87–0.95)
Reader 11	1.21 (0.93–1.49)	0.08 (0.06–0.09)	0.53 (0.49–0.56)	0.94 (0.85–0.98)
Reader 12	1.55 (1.35–1.75)	0.21 (0.18–0.24)	0.59 (0.54–0.63)	0.84 (0.78–0.89)
Reader 13	1.94 (1.74–2.14)	0.15 (0.14–0.17)	0.64 (0.60–0.68)	0.88 (0.83–0.92)
Reader 14	1.72 (1.52–1.92)	0.18 (0.16–0.20)	0.61 (0.57–0.65)	0.86 (0.81–0.90)
Reader 15	1.36 (1.13–1.59)	0.10 (0.08–0.12)	0.55 (0.51–0.59)	0.92 (0.85–0.96)

Note—Readers 1–4 were breast fellowship–trained radiologists, readers 5–11 were radiologists without breast fellowship training, readers 12 and 13 were breast surgeons, and readers 14 and 15 were obstetrician/gynecologists. Values in parentheses are 95% CIs. PLR = positive likelihood ratio, NLR = negative likelihood ratio, PPV = positive predictive value, NPV = negative predictive value, DS = decision support.

TABLE 2: Difference in AUC by Reader

Reader	AUC Difference[a]	Propagated SE
1	–0.003	0.013
2	0.079	0.015
3	0.045	0.014
4	0.021	0.016
5	0.059	0.014
6	0.021	0.013
7	0.069	0.015
8	0.013	0.014
9	0.047	0.014
10	0.043	0.014
11	0.013	0.014
12	0.017	0.013
13	0.074	0.016
14	0.034	0.014
15	0.024	0.014

Note—Readers 1–4 were breast fellowship–trained radiologists, readers 5–11 were radiologists without breast fellowship training, readers 12 and 13 were breast surgeons, and readers 14 and 15 were obstetrician/gynecologists. SE = standard error.
[a]Calculated by subtracting AUC for ultrasound-only readings from AUC for ultrasound with decision support.

From these cases, the amount of class switching from lower than BI-RADS 4A to BI-RADS 4A or higher and vice versa was measured. These rates were plotted with respect to each switching rate and assessed for statistical differences using a paired *t* test (Fig. 5). US-only class switching rate, defined as crossing the BI-RADS 3–BI-RADS 4A boundary, was 13.6%, and the US-plus-DS class switching rate was 10.8% ($p = 0.04$). The highest rates of class switching were for reader 4 (20% for US plus DS and 25% for US only.

On average, intrareader variability resulted in less class switching with US plus DS than with US only. Although we found a statistically significant trend toward lower intrareader variability with US plus DS (nine readers showed decreased class switching with DS), one reader showed equivalent class switching and five showed more class switching with DS. These results indicate that the

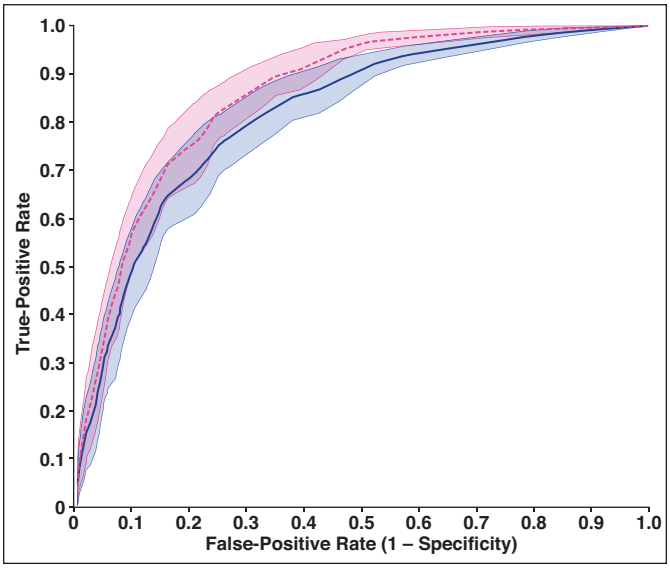


Fig. 4—ROC curves show mean performance of assessment with ultrasound (US) only (*solid blue line;* 95% CI, 0.78–0.89 [dashed red line]) and with US plus decision support (DS) (*red shading;* 95% CI, 0.84–0.90 [*blue shading*]). Stand-alone DS performance is not shown but had AUC of 0.88 (95% CI, 0.86–0.91).

TABLE 3: Impact of the DS System on Malignant Lesion Assessment by Readers Evidenced by Reader Decision Switches After DS System Recommendations

Reader No., DS Recommendation[a]	Change in Reader Decision	
	Correct to Incorrect	Incorrect to Correct
1		
No biopsy	0	0
Biopsy	10	7
2		
No biopsy	2	0
Biopsy	0	25
3		
No biopsy	1	0
Biopsy	2	20
4		
No biopsy	2	1
Biopsy	21	32
5		
No biopsy	4	0
Biopsy	0	13

(Table 3 continues on next page)

TABLE 3: Impact of the DS System on Malignant Lesion Assessment by Readers Evidenced by Reader Decision Switches After DS System Recommendations (continued)

Reader No., DS Recommendation[a]	Change in Reader Decision	
	Correct to Incorrect	Incorrect to Correct
6		
No biopsy	0	0
Biopsy	1	2
7		
No biopsy	4	0
Biopsy	1	5
8		
No biopsy	0	0
Biopsy	1	22
9		
No biopsy	2	2
Biopsy	3	17
10		
No biopsy	4	0
Biopsy	0	14
11		
No biopsy	1	0
Biopsy	1	4
12		
No biopsy	1	0
Biopsy	1	7
13		
No biopsy	4	0
Biopsy	2	24
14		
No biopsy	1	1
Biopsy	0	4
15		
No biopsy	2	0
Biopsy	1	12

Note—The DS system marked 190 benign lesions as suspicious and 42 benign lesions as probably malignant. DS = decision support.

[a]A "no biopsy" recommendation was defined as a "benign" or "probably benign" output from the DS; a "biopsy" recommendation was defined as a "suspicious" or "probably malignant" output. The resulting biopsy recommendation by the reader changed from a correct to incorrect or incorrect to correct assessment when DS output was provided.

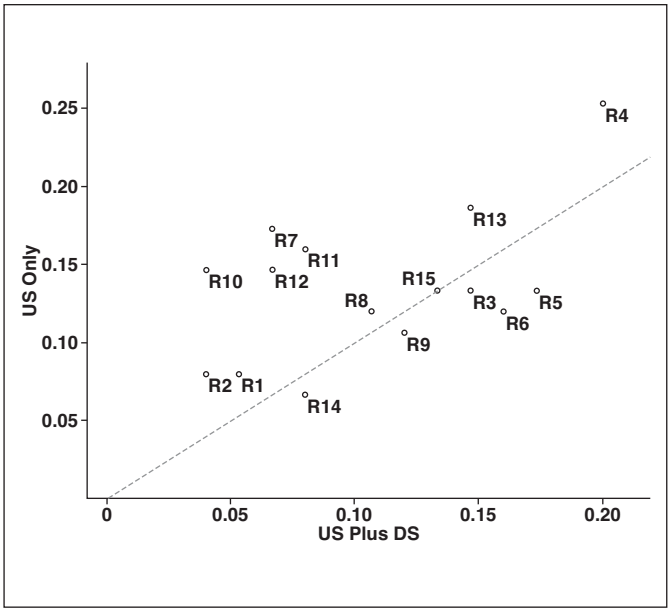


Fig. 5—Scatterplot shows intrareader variability represented by class switching from probably benign to suspicious or from suspicious to probably benign between first reading and second reading 4 weeks later for 15 readers (R1–R4, breast fellowship–trained radiologists; R5–R11, radiologists without breast fellowship training; R12 and R13, breast surgeons; R14 and R15, obstetrician/gynecologists). There is statistically significant trend to lower intrareader variability with ultrasound (US) plus decision support (DS) compared with US only. Dashed line denotes equivalency.

overall improvement in intrareader variability did not extend to all readers.

The impact of DS on each reader's sensitivity and specificity was also analyzed (Fig. 6). To assess any systematic bias in the DS output itself, a subsequent analysis was performed in which the disagreement between the reader's US-only reading was compared with the DS output (Table 4). To establish a single operating point, a DS output of suspicious or probably malignant was treated as a determination of malignancy.

Discussion

Our study indicates that AI-based DS output sensitivity and specificity compare favorably with those of interpreting physicians from various subspecialties in the evaluation of static orthogonal breast US images. Interestingly, the system's stand-alone performance, as measured by AUC, was still higher than US plus DS. Given the performance of the stand-alone system, the DS output may have a larger impact if it is used more frequently. Similarly, a study assessing a CAD

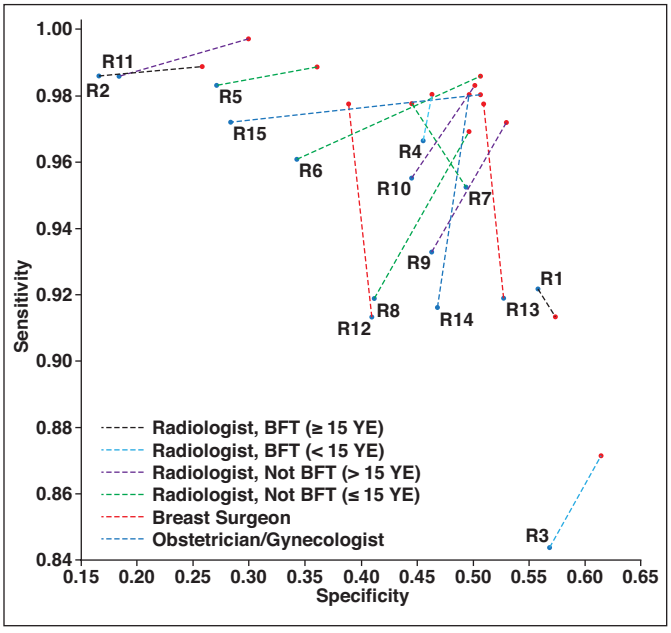


Fig. 6—Operating point shifts with original operating points in blue and final operating points in red for 15 readers (R1–R4, breast fellowship–trained [BFT] radiologists; R5–R11, radiologists not BFT; R12 and R13, breast surgeons; R14 and R15, obstetrician/gynecologists). All operating points are defined at transition between BI-RADS category 3 and 4A. YE = years of experience.

TABLE 4: Cases in Which DS Output and Reader Assessment (US Only) Disagreed

Reader	Disagreements Between Reader Assessment (US only) and DS Output for Any Case	Malignant Cases		
		No. of Correct Reader Assessments (But Incorrect DS Outputs)	No. of Correct DS Outputs (But Incorrect Reader Assessments)	Difference Between No. of Correct DS Outputs and No. of Correct Reader Assessments[a]
1	121	2	24	22
2	146	3	27	24
3	158	3	28	25
4	173	3	53	50
5	139	5	13	8
6	157	4	3	–1
7	142	4	8	4
8	130	1	24	23
9	128	2	20	18
10	109	4	14	10
11	143	6	6	0
12	107	2	8	6
13	137	4	27	23
14	158	5	4	–1
15	115	4	15	11

Note—Readers 1–4 were breast fellowship–trained radiologists, readers 5–11 were radiologists without breast fellowship training, readers 12 and 13 were breast surgeons, and readers 14 and 15 were obstetrician/gynecologists.

[a]Positive values indicate that the DS system made more correct assessments than the reader (US only) for malignant lesions.

system for detection in mammography found that sensitivity increased 9% and callback increased 12%, but radiologists ignored 71% of true-positive assessments from the system that were otherwise missed [26]. The PLR, the likelihood a biopsy recommendation indicates malignancy, for the DS system was 1.98, higher than all readers interpreting US images alone, except for one reader (reader 1) who had a higher PLR (2.08) with a 95% CI (1.88–2.29) that overlapped the DS PLR. The DS system performance in our study was similar to prior studies and was at least comparable with that of fellowship-trained breast imagers, indicating the system could be used as a valuable second opinion tool clinically [24].

Adding AI-based DS to US images improved correct BI-RADS classification of sonographic breast lesions by most physicians studied. The AUC improved for all physicians but one when they were provided with US images plus DS. Only one reader (reader 1) showed a decrease in performance with US plus DS compared with US only. The difference was not statistically significant, and further evaluation revealed that it was due to a slight reduction in sensitivity and an increase in specificity. Correlation with the confusion matrix indicated that this reader's decreased sensitivity was most likely caused by intrareader variability, given that the DS system recommended biopsy for all malignancies the reader changed from correct to incorrect or vice versa between reading sessions. When the impact of DS on each reader's sensitivity and specificity was analyzed, improvements appeared to depend on the reader's initial operating point. Subpopulations of readers tended to show improvements in areas where they were weak; for example, specific, but not sensitive, readers saw greater improvements in sensitivity. We did see significant deviations within physician specialty subgroups; the initial and final operating points across all groups except breast surgeons varied significantly. Although all saw an overall benefit, it would be inappropriate to average these populations because they appeared to be bimodal in nature (i.e., readers who were more sensitive and less specific and readers who were more specific and less sensitive). This variability limits the ability to group physicians on the basis of training.

As outlined in the confusion matrix, with the addition of DS, readers changed some assessments of malignant lesions from correct to incorrect and others from incorrect to correct. The magnitude of switching varied by reader and may not be completely attributable to the presence of the DS system given that readings with and without DS were 4 weeks apart; as a result, some switching may have been due to intrareader variability. Our study indicates that an AI-based DS system decreases inter- and intraobserver variability overall, which could facilitate more consistent clinical care for patients.

Given that lesion caliper placement on the US image by the clinical radiologist defines the ROI for the DS system, examining how variability in caliper placement may affect the DS results is essential. Our analysis shows ROI boundary variation produced no significant change in either the shape of the ROC curve or the AUC values. Clinically significant differences in DS output, such as deeming a lesion suspicious versus probably benign, represent an important clinical threshold between recommending biopsy rather than imaging follow-up. ROI variation produced minimal class switching from probably benign to suspicious (1.9%) and suspicious to probably benign (1.7%). Because the system's output is categoric, changes across this decision boundary can change clinical management.

Similarly, variation in US transducer frequency (low versus high) did not produce statistical difference in DS system performance. Our study also included images from

multiple institutions and multiple US vendors. Ensuring consistency of DS output despite variations in equipment and technique is vital for potential future clinical applications of this software and applicability of our results to other users.

This study has several limitations. As a retrospective reader study of static US images, it does not replicate a true clinical environment. In clinical practice, lesions are often scanned in real time by the radiologist and evaluated in the context of patient symptoms, risk factors, and correlation with mammography, prior imaging, or both. Thus, a BI-RADS assessment based on two orthogonal static US images may not always correspond with a reader's assessment when all clinical information and images are available. US images provided already had an ROI provided by the original interpreting physician, so the ability of the reader to recognize the presence of a lesion and define lesion boundaries was not tested and is beyond the scope of this study. Only US images containing lesions were used, and no images of normal breast tissue were used. Also, benignity for some lesions was established by 1 year of stability on US, which does not meet satisfy the 2–3 years of US stability required by American College of Radiology criteria for a probably benign lesion.

Our study did not examine the integration of the DS system in real time and did not examine integration into clinical practice. The AI platform evaluated here can have its interface integrated into US scanners and PACS clients, enabling accessibility in the imaging area with the patient or at the PACS workstation while interpreting images. Further study would be needed to assess the impact of this system on clinical workflow.

Additionally, our readers had no familiarity with the DS system before the study. With increased use, readers may learn to trust or distrust the DS system output, which could affect their agreement or disagreement with the system and subsequently affect their AUC with continued use of DS. This difference in system trust may partially explain why some fellowship-trained breast imagers did not alter their responses on the basis of DS output as much as others. Future study directions include looking at the effect of ongoing DS use on radiologists' performance and potential benefits for bridging experience and training gaps.

Conclusion

An AI-based DS improved correct assessment of sonographic breast lesions by most physicians while reducing inter- and intraobserver variability. Future study of its prospective use in a true clinical environment is needed.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018; 68:394–424
2. Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA Cancer J Clin* 2019; 69:127–157
3. Mendelson EB. Artificial intelligence in breast imaging: potentials and limitations. *AJR* 2019; 212:293–299
4. Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology* 2000; 215:554–562
5. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology* 2001; 220:781–786
6. Cai H, Peng Y, Ou C, Chen M, Li L. Diagnosis of breast masses from dynamic contrast-enhanced and diffusion-weighted MR: a machine learning approach. *PLoS One* 2014; 9:e87387
7. Jerez JM, Franco L, Alba E, et al. Improvement of breast cancer relapse prediction in high risk intervals using artificial neural networks. *Breast Cancer Res Treat* 2005; 94:265–272
8. Qiu Y, Yan S, Gundreddy RR, et al. A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology. *J XRay Sci Technol* 2017; 25:751–763
9. Saritas I. Prediction of breast cancer using artificial neural networks. *J Med Syst* 2012; 36:2901–2907
10. van Zelst JCM, Tan T, Platel B, et al. Improved cancer detection in automated breast ultrasound by radiologists using computer aided detection. *Eur J Radiol* 2017; 89:54–59
11. U.S. Food and Drug Administration website. QVIEW Medical, Inc. premarket approval for QVCAD system: summary of safety and effectiveness data. www.accessdata.fda.gov/cdrh_docs/pdf15/P150043B.pdf. Approved November 9, 2016. Accessed March 10, 2020
12. Berg WA, Bandos AI, Mendelson EB, Lehrer D, Jong RA, Pisano ED. Ultrasound as the primary screening test for breast cancer: analysis from ACRIN 6666. *J Natl Cancer Inst* 2015; 108:djv367
13. Bae MS, Han W, Koo HR, et al. Characteristics of breast cancers detected by ultrasound screening in women with negative mammograms. *Cancer Sci* 2011; 102:1862–1867
14. D'Orsi CJ, Sickles EA, Mendelson EB, et al. *ACR BI-RADS Atlas, Breast Imaging Reporting and Data System*. Reston, VA: American College of Radiology, 2013
15. Berg WA, Blume JD, Cormack JB, Mendelson EB. Operator dependence of physician-performed whole-breast US: lesion detection and characterization. *Radiology* 2006; 241:355–365
16. Berg WA, Blume JD, Cormack JB, Mendelson EB. Training the ACRIN 6666 Investigators and effects of feedback on breast ultrasound interpretive performance and agreement in BI-RADS ultrasound feature analysis. *AJR* 2012; 199:224–235
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; 239:385–391
18. Abdullah N, Mesurolle B, El-Khoury M, Kao E. Breast imaging reporting and data system lexicon for US: interobserver agreement for assessment of breast masses. *Radiology* 2009; 252:665–672
19. Chabi ML, Borget I, Ardiles R, et al. Evaluation of the accuracy of a computer-aided diagnosis (CAD) system in breast ultrasound according to the radiologist's experience. *Acad Radiol* 2012; 19:311–319
20. Barinov L, Jairaj A, Paster L, et al. Decision quality support in diagnostic breast ultrasound through artificial intelligence. In: *Signal Processing in Medicine and Biology Symposium*. Piscataway, NJ: IEEE, 2016:SPMB-L3.03
21. Love SM, Berg WA, Podilchuk C, et al. Palpable breast lump triage by minimally trained operators in Mexico using computer-assisted diagnosis and low-cost ultrasound. *J Glob Oncol* 2018; 4:1–9
22. Breast Cancer Surveillance Consortium website. Dataset (HHSN261201100031C). tools.bcsc-scc.org/dataexplorer/. Accessed April 8, 2020
23. Barinov L, Jairaj A, Becker M, et al. Impact of data presentation on physician performance utilizing artificial intelligence-based computer-aided diagnosis and decision support systems. *J Digit Imaging* 2019; 32:408–416
24. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. *Invest Radiol* 1992; 27:723–731
25. Kendall M. A new measure of rank correlation. *Biometrika* 1938; 30:81–89
26. Nishikawa RM, Schmidt RA, Linver MN, Edwards AV, Papaioannou J, Stull MA. Clinically missed cancer: how effectively can radiologists use computer-aided detection? *AJR* 2012; 198:708–716