European Journal of Radiology Open 12 (2024) 100545

Contents lists available at ScienceDirect

European Journal of Radiology Open

journal homepage: www.elsevier.com/locate/ejro

Diagnostic performance with and without artificial intelligence assistance in real-world screening mammography

Si Eun Lee[1], Hanpyo Hong[2], Eun-Kyung Kim[*,3]

Department of Radiology, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin, Korea

ARTICLE INFO

Keywords:
Breast cancer
Digital mammography
Diagnosis, Computer-assisted
Artificial intelligence

ABSTRACT

Purpose: To evaluate artificial intelligence-based computer-aided diagnosis (AI-CAD) for screening mammography, we analyzed the diagnostic performance of radiologists by providing and withholding AI-CAD results alternatively every month.

Methods: This retrospective study was approved by the institutional review board with a waiver for informed consent. Between August 2020 and May 2022, 1819 consecutive women (mean age 50.8 ± 9.4 years) with 2061 screening mammography and ultrasound performed on the same day in a single institution were included. Radiologists interpreted screening mammography in clinical practice with AI-CAD results being provided or withheld alternatively by month. The AI-CAD results were retrospectively obtained for analysis even when withheld from radiologists. The diagnostic performances of radiologists and stand-alone AI-CAD were compared and the performances of radiologists with and without AI-CAD assistance were also compared by cancer detection rate, recall rate, sensitivity, specificity, accuracy and area under the receiver-operating-characteristics curve (AUC).

Results: Twenty-nine breast cancer patients and 1790 women without cancers were included. Diagnostic performances of the radiologists did not significantly differ with and without AI-CAD assistance. Radiologists with AI-CAD assistance showed the same sensitivity (76.5%) and similar specificity (92.3% vs 93.8%), AUC (0.844 vs 0.851), and recall rates (8.8% vs. 7.4%) compared to standalone AI-CAD. Radiologists without AI-CAD assistance showed lower specificity (91.9% vs 94.6%) and accuracy (91.5% vs 94.1%) and higher recall rates (8.6% vs 5.9%, all $p < 0.05$) compared to stand-alone AI-CAD.

Conclusion: Radiologists showed no significant difference in diagnostic performance when both screening mammography and ultrasound were performed with or without AI-CAD assistance for mammography. However, without AI-CAD assistance, radiologists showed lower specificity and accuracy and higher recall rates compared to stand-alone AI-CAD.

1. Introduction

Mammography is the primary screening modality for breast cancer that improves survival rates [1,2]. However, fibroglandular breast tissue can obscure suspicious lesions which limits the sensitivity of mammography, and sometimes this tissue can mimic true lesions, resulting in false-positive recalls [3,4]. In the late 1990s, traditional computer-aided detection was introduced to improve the human detection of breast cancer, but subsequent large-scale validation studies concluded that it failed to improve human performance due to excessive recalls and an increase in unnecessary additional examinations [5–7].

Newly developed artificial intelligence-based computer-assisted

Abbreviations: **AI-CAD,**, Artificial intelligence-computer-assisted diagnosis; **PACS,**, Picture archiving and communication system; **PPV,**, Positive predictive value; **NPV,**, Negative predictive value; **AUC,**, Area under the receiver-operating-characteristics curve.

[*] Correspondence to: Department of Radiology, Yongin Severance Hospital, Yonsei University College of Medicine, 363, Dongbaekjukjeon-daero, Giheung-gul̦, Yongin-si, Gyeonggi-do, Korea.

E-mail address: ekkim@yuhs.ac (E.-K. Kim).

[1] 0000-0002-3225-5484

[2] 0000-0002-0573-4527

[3] 0000-0002-3368-5013

https://doi.org/10.1016/j.ejro.2023.100545
Received 15 October 2023; Received in revised form 27 December 2023; Accepted 29 December 2023
Available online 13 January 2024

diagnosis (AI-CAD) programs might improve the diagnostic performance of radiologists in reader studies [8–12], and have shown promise in reducing the missed cancer rate in a retrospective study [13]. Large-scale validation and simulation studies have proven that AI-CAD can act as a secondary reader for radiologists and triaging tool for normal exams [14–16]. In recent prospective trials, AI-CAD showed similar cancer detection rates without increasing recall rates when used to replace one of the radiologists in a double-reading system [17,18]. There is also an ongoing prospective trial on the role of AI assistance in a single reading system, but most current studies based on the retrospective application of AI-CAD suffer from bias [19]. Simulation studies cannot completely replicate what occurs in clinical practice. In real life, radiologists decide whether or not to accept the AI-CAD results after mammography, and their decision is also affected by prior studies, patient factors or even readily available AI-CAD results.

Supplementary ultrasound is a good option for dense breasts, as it increases the cancer detection rate of screening mammography and has the potential to detect early invasive cancers despite decreasing the specificity of mammography [20–22]. In Asian countries, supplementary ultrasound is now commonly performed and has the potential to be more widely used with the introduction of automated breast ultrasound. When supplementary US is performed, the role of AI-CAD on mammography will differ from when only mammography is performed, and the advantages of AI-CAD may be partially diluted.

In our institution, we integrated a commercial AI-CAD software into the picture archiving and communication system (PACS), and radiologists could refer to the AI-CAD results during mammography interpretation. To evaluate artificial intelligence-based computer-aided diagnosis (AI-CAD) for screening mammography, we analyzed the diagnostic performance of radiologists by providing and withholding AI-CAD alternatively every month.

2. Materials and methods

2.1. Study protocol

Our hospital recommends that asymptomatic average-risk women who are 40 years and older, undergo mammography screenings every one to two years. Supplementary breast ultrasound is suggested for women with dense breast, however, the national health insurance program of Korea only covers screening mammography every two years, and additional examinations can be covered by personal insurance. Digital mammography was obtained using one mammography unit (Senographe Pristina Mammography System, GE Healthcare, Milwaukee, WI, United States) with cranio-caudal and medio-lateral-oblique views from each breast. AI-CAD was implemented on PACS and the AI-CAD results were shown on screen after the four mammographic views. Breast ultrasound was performed with either the conventional hand-held system (LOGIQ E10, GE Healthcare and EPIQ Elite, Philips Medical Systems, Bothell, WA) by radiologists or with automated breast ultrasound (Invenia ABUS, GE Healthcare) by a technician, depending on the clinician's or patient's preference.

A commercial AI-CAD program has been integrated into PACS in our institution since March 2020, and the first five months after its introduction was considered an adaptation period, during which radiologists freely referred to AI-CAD for interpretation. Since August 2020, we have alternatively provided and withheld the AI-CAD results monthly, with AI-CAD results being provided on even months. Four radiologists (dedicated to breast imaging for 1–28 years) interpreted mammography in clinical practice, and the radiologists could refer to the AI-CAD results every even month. Mammographic interpretations were based on the American College of Radiology Breast Imaging-Reporting and Data System (ACR BI-RADS). Mammographic density was assessed visually using the ACR BI-RADS 5th edition. As part of routine practice, any prior images of the breast were simultaneously reviewed if available. Every mammography examination was performed before the ultrasound. BI-RADS assessments were made for mammography and ultrasound separately, based on the imaging findings of each modality.

A retrospective image review was done for malignant cases to evaluate mammographic visible or occult cancers by a staff radiologist who had 4 years of experience in breast imaging.

2.2. Study population

Out of 3312 sets of screening mammography and ultrasounds performed on the same day from August 2020 to May 2022, cases that were confirmed as cancer (n = 29) or benign (n = 122) through histopathological biopsy or operative results within 11 months from the date of the initial examination were deemed as having a definitive diagnosis. Stable findings on follow-up mammography or ultrasound for at least 1 year (n = 1910) were considered as the standard reference for a negative result.

Finally, we included 1819 consecutive women with 2061 screening examinations. The mean age was 50.8 ± 9.4 years (range, 21 to 82 years) (Fig. 1).

AI-CAD.

We used a commercial AI-CAD program (Lunit insight MMG, version 1.1.7.1, available at https://insight.lunit.io) which was developed and validated through large-scale retrospective studies [12, 23, 24]. This program provides an abnormality score or malignancy risk in percentages of 0–100% per breast with a heatmap or grayscale map. Separate heatmaps are given for each craniocaudal and mediolateral oblique view, and the abnormality score is provided as the largest value per breast. When the abnormality score is less than 10%, the heatmap is not shown, and the malignancy risk is presented as "Low" with a corresponding score on PACS. AI-CAD results continued to be saved in the server even during the months that they were not automatically sent to PACS and withheld from radiologists, and we were able to use the stored data to evaluate the diagnostic performance of stand-alone AI-CAD.

2.3. Statistical analysis

Mean age, breast density, BI-RADS assessment, and ultrasound method (automated vs. hand-held) were compared between months with AI-CAD assistance and without using the t-test for age and the Chi-square test for other categorical variables.

For radiologists, BI-RADS 1 to 2 interpretations were regarded as test-negative and BI-RADS 0, 3, 4, and 5 as test-positive. For stand-alone AI-CAD, a score over 10 was regarded as test-positive. The diagnostic performances of the radiologists with or without AI-CAD assistance and of stand-alone AI-CAD were quantified in terms of sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, area under the receiver-operating-characteristics curve (AUC), cancer detection rate, and recall rate. The cancer detection rate was defined as the proportion of cancer detected within the follow-up period of a positive screening mammogram. The recall rate was defined as the proportion of screening mammograms leading to further work-up, calculated as true- and false-positive mammograms divided by all mammograms. The exact Clopper-Pearson confidence intervals were used to determine confidence intervals for sensitivity, specificity, and accuracy. To compare the proportions between the two groups, p-values were calculated using the two-proportion test.

All analyses were performed using MedCalc software version 20.0 (MedCalc Software Ltd., Ostend, Belgium). Two-sided P values and 95% confidence intervals were reported, with a statistical significance threshold of.05.

This retrospective study was approved by the institutional review board of ****, Yongin, Korea, with a waiver for informed consent.

3. Results

Out of 1819 women (mean age 50.8 ± 9.4 years, range 30 to 82 years) with 2061 screening examinations, we categorized the breast

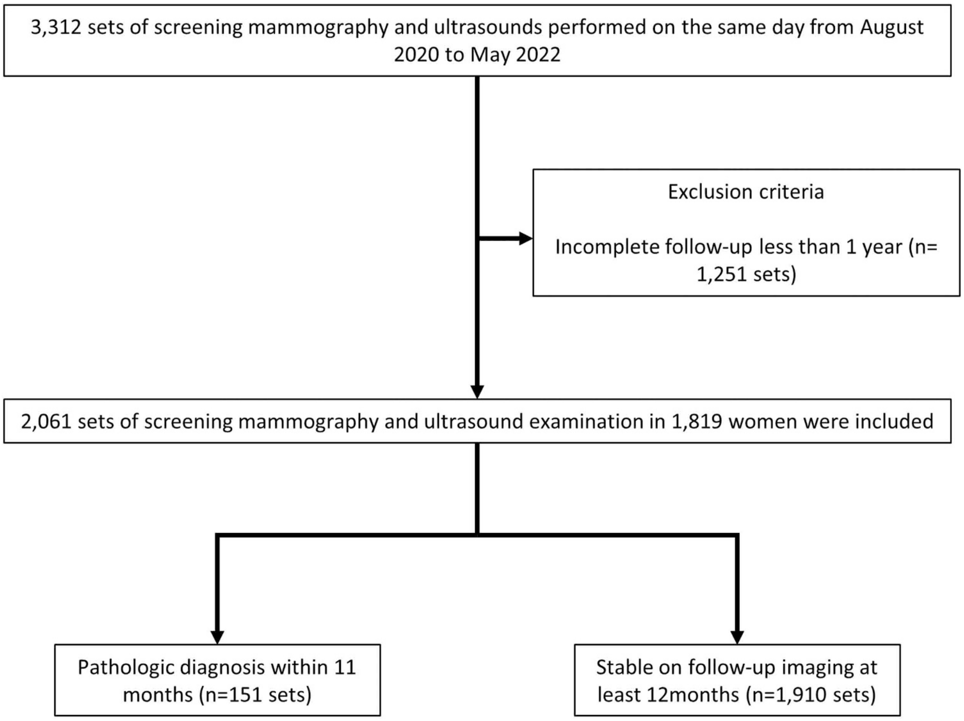
3,312 sets of screening mammography and ultrasounds performed on the same day from August 2020 to May 2022
Exclusion criteria
Incomplete follow-up less than 1 year (n= 1,251 sets)
2,061 sets of screening mammography and ultrasound examination in 1,819 women were included
Pathologic diagnosis within 11 months (n=151 sets)
Stable on follow-up imaging at least 12months (n=1,910 sets)

Fig. 1. Summary of patient selection.

density of 25 as fatty, 223 as scattered fibro-glandular, 1299 as heterogeneously dense, and 514 as extremely dense (Table 1). Among them, 63.0% (1298/2061) cases underwent automated breast ultrasound and the remaining 37.0% (763/2061) cases underwent hand-held ultrasound. Mean age, BI-RADS assessments, ultrasound methods, and standard reference did not differ between the months with AI-CAD assistance available and those without (Table 1). Breast density distribution was significantly different between the two periods (Table 1, p = 0.036).

Twenty-nine patients were diagnosed with breast cancer, consisting of 15 invasive cancers and 14 ductal carcinomas in situ (DCIS) on core biopsy. The overall cancer detection rate was 15.9 per 1000 women (29/1819). The average size of invasive cancers was 10.8 mm (range, 6–14 mm) on initial imaging, and 6.7% of patients (2/29) had metastatic lymph nodes in the ipsilateral axilla.

3.1. Diagnostic performance of radiologists with or without AI-CAD assistance compared to stand-alone AI-CAD

Table 2 and Fig. 2 summarize the diagnostic performances of radiologists during two periods (months with and without AI-CAD assistance). Radiologists with AI-CAD assistance showed equal sensitivity (76.5%, 13/17) and cancer detection rate (13/1032) and similar specificity (92.3%, 937/1015 vs 93.8%, 952/1015), PPV (14.3%, 13/91 vs 17.1%, 13/76), NPV (99.6%, 937/941 vs 99.6%, 952/996), AUC (0.844 vs 0.851) and recall rate (8.8%, 91/1032 vs 7.4%, 76/1032) compared to stand-alone AI-CAD without statistical difference.

Radiologists without AI-CAD assistance showed equal sensitivity (50.0%, 6/12), NPV (99.4%, 935/941), and cancer detection rate (6/1029) compared to stand-alone AI-CAD. However, PPV(6.8%, 6/88 vs 9.8%, 6/61), specificity (91.9%, 935/1017 vs 94.6%, 962/1017), and accuracy (91.5%, 941/1029 vs 94.1%, 968/1029) were significantly lower without AI-CAD assistance, and the recall rate(8.6%, 88/1029) vs 5.9%, 61/1029) was higher for radiologists.

When we compared diagnostic performances between radiologists with or without AI-CAD assistance, the use of AI-CAD led to higher values in sensitivity (76.5%, 13/17 vs. 50.0%, 6/12, p = 0.135), specificity (92.3%, 937/1015 vs. 91.9%, 935/1017, p = 0.752), PPV (14.3%, 13/91 vs. 6.8%, 6/88, p = 0.101), accuracy (92.1%, 950/1032 vs. 91.4%, 941/1029, p = 0.617), AUC (0.84 vs. 0.71, p = 0.146), and cancer detection rate (13/1032 vs. 6/1029, p = 0.108). However, none of these differences were statistically significant. Also, the recall rate was slightly lower when AI-CAD results were available to the radiologists, but again without statistical significance (8.6%, 88/1029 vs. 8.8%, 91/1032, p = 0.830).

3.2. Performance of stand-alone AI-CAD according to breast density

For the entire study period, stand-alone AI-CAD showed a sensitivity of 65.5% (19/29), specificity of 94.2% (1914/2032), PPV of 13.9% (19/137), NPV of 99.5% (1914/1924), AUC of 0.8 and recall rate of 6.6% (137/2061).

When we classified patients by breast density into two groups (fatty and dense), in patients with fatty breasts, stand-alone AI-CAD showed significantly higher specificity (95.9%, 236/246 vs. 90.7%, 223/246, p = 0.021) and accuracy (95.6%, 237/248 vs. 90.7%, 225/248, p = 0.031) and lower recall rate (4.4%, 11/248 vs. 10.1%, 25/248, p = 0.015) compared to radiologists. In patients with dense breasts, stand-alone AI-CAD showed significantly higher specificity (94.0%, 1678/1786 vs 92.3%, 1649/1786, p = 0.044) compared to radiologists. The NPV of stand-alone AI-CAD was sustained nearly perfectly regardless of breast density (99.5%, 1678/1687 vs. 99.4%, 1649/1659, p = 0.696) (Table 3).

3.3. Cancers missed by AI-CAD

Among 29 breast cancers, 10 cancers (34%, 10/29) were missed by AI-CAD, consisting of 4 DCIS and 6 invasive cancers (5 invasive ductal carcinoma, 1 lobular carcinoma). The average size of the missed invasive cancers on initial imaging was 9.3 mm (range 6–14), and one had axillary lymph node metastasis.

Except for 1 case presenting as grouped microcalcifications, the other 9 lesions were not detected by radiologists and were diagnosed as BI-RADS 1 or 2 on screening mammography. All of these 9 patients had

Table 1
Patient demographics and outcome characteristics.

	Total	Months without AI-CAD	Months with AI-CAD	p-value
Women/Exams	1819/ 2061	887/1029	932/1032	
Mean age at screening*	50.8 ± 9.4	51.0 ± 9.4	50.7 ± 9.5	0.485
Density				0.036
Fatty	25(1.2)	10(1.0)	15(1.5)	
Scattered fibro-glandular	223 (10.8)	96(9.3)	127(12.3)	
Heterogeneously dense	1299 (63.0)	677(65.8)	622(60.3)	
Extremely dense	514 (24.9)	246(23.9)	268(26.0)	
Initial BI-RADS assessment of mammography				0.523
Incomplete	106 (5.1)	56(5.4)	50(4.8)	
Negative	1316 (63.9)	670(65.1)	646(62.6)	
Benign	566 (27.5)	271(26.3)	295(28.6)	
Probably benign	57(2.8)	24(2.3)	33(3.2)	
Suspicious	16(0.8)	8(0.8)	8(0.8)	
Ultrasound methods				0.238
Automated breast ultrasound	1298 (63.0)	661(64.2)	637(61.7)	
Hand-held breast ultrasound	763 (37.0)	368(35.8)	395(38.3)	
Standard reference				0.488
Cancer by biopsy and/or surgery	29(1.4)	12(1.2)	17(1.6)	0.354
-Invasive	15(0.7)	6(0.6)	9(0.9)	0.440
-Ductal carcinoma in situ	14 (0.7)	6(0.6)	8(0.8)	0.595
Benign by biopsy and/or surgery	122 (5.9)	57(5.5)	65(6.3)	0.465
Benign by follow-up imaging	1910 (92.7)	960(93.3)	950(92.1)	0.280
Retrospective image review				0.158
Mammography-visible cancer	23 (79.3)	8(66.7)	15(88.2)	
Mammography-occult cancer	6(20.7)	4(33.3)	2(11.8)	

Note.—Unless otherwise specified, data are numbers of women, with percentages in parentheses. Percentages do not always sum to 100%, due to rounding.
* Data are means ± standard deviations
P-value, months without AI-CAD assistance vs. months with AI-CAD assistance

dense breasts, and subsequent ultrasound examinations revealed suspicious lesions, which were then diagnosed through ultrasound-guided biopsies.

4. Discussion

When screening mammography was performed with supplementary ultrasound, the radiologists who interpreted screening mammography with AI-CAD assistance showed better sensitivity, specificity, AUC, cancer detection rates and recall rates compared to the radiologists who did not refer to AI-CAD, although this improvement was not statistically significant. Interestingly, the radiologists with AI-CAD assistance showed similar performances with stand-alone AI-CAD. However, the radiologists without AI-CAD assistance showed significantly lower specificity, accuracy and higher recall rate compared to stand-alone AI-CAD.

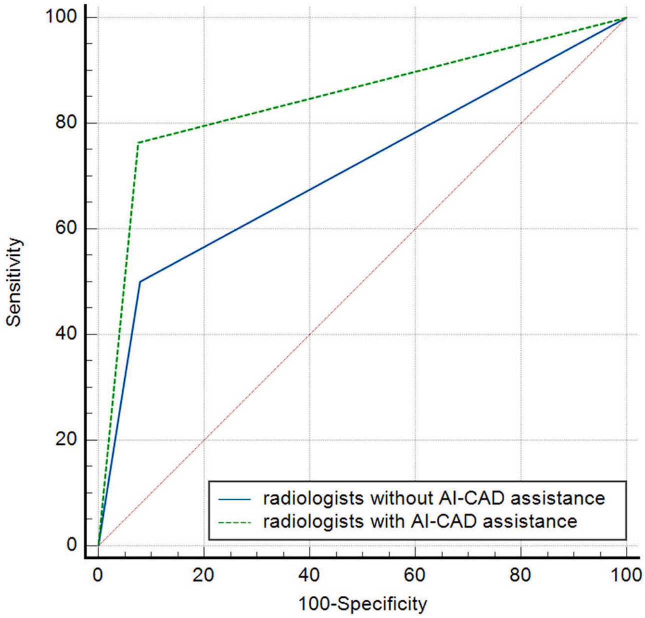

Fig. 2. Receiver operating characteristic curve analysis of radiologists with and without AI-CAD assistance. The blue line indicates radiologists with AI-CAD assistance, and the green dotted line indicates radiologists without AI-CAD assistance (AUC 0.84 vs 0.71, p = 0.146).

Table 2
Comparison of diagnostic performance between radiologists without the assistance of AI-CAD and stand-alone AI-CAD, and between radiologists with AI-CAD and stand-alone AI-CAD.

	Months without AI			Months with AI			*P*-value†
	Radiologists without AI-CAD	Stand-alone AI	p-value	Radiologists with AI-CAD	Stand-alone AI	p-value	
Sensitivity	50.0 (6/12) (21.1, 78.9)	50.0 (6/12) (21.1, 78.9)	>0.999	76.5 (13/17) (50.1, 93.2)	76.5 (13/17) (50.1, 93.2)	>0.999	0.135
Specificity	91.9 (935/1017) (90.1, 93.5)	94.6 (962/1017) (93.0, 95.9)	0.017	92.3 (937/1015) (90.5, 93.9)	93.8 (952/1015) (92.1, 95.2)	0.190	0.752
PPV	6.8 (6/88) (3.9, 11.8)	9.8 (6/61) (5.5, 16.9)	0.518	14.3 (13/91) (10.6, 19.0)	17.1 (13/76) (12.6, 22.8)	0.619	0.101
NPV	99.4 (935/941) (98.9, 99.6)	99.4 (962/968) (98.9, 99.6)	0.961	99.6 (937/941)(99.0, 99.8)	99.6 (952/956) (99.0, 99.8)	0.982	0.526
AUC	0.710 (0.68, 0.74)	0.723 (0.70, 0.75)	0.903	0.844 (0.82, 0.87)	0.851 (0.83, 0.87)	0.926	0.146
Accuracy	91.5 (941/1029) (89.6, 93.1)	94.1 (968/1029) (92.5, 95.4)	0.021	92.1 (950/1032) (90.2, 93.6)	93.5 (965/1032) (91.8, 94.9)	0.202	0.617
Detection rate	6/1029	6/1029	>0.999	13/1032	13/1032	>0.999	0.108
Recall rate	8.6 (88/1029) (6.9, 10.4)	5.9 (61/1029) (4.6, 7.6)	0.021	8.8 (91/1032) (7.2, 10.7)	7.4 (76/1032) (5.8, 9.1)	0.226	0.830

PPV: Positive predictive value, NPV: Negative predictive value, AUC: Area under the receiver operating characteristic curve The number in parentheses indicates the 95% confidence interval. † P-value, Radiologists without AI-CAD assistance (odd months) vs. Radiologists with AI-CAD assistance (even months)

Table 3

Comparison of diagnostic performance of mammography between radiologists with or without AI-CAD assistance and stand-alone AI-CAD during the whole period according to breast density.

	Fatty breasts			Dense breasts			*P*-value[†]
	Radiologists	Stand-alone AI	p-value	Radiologists	Stand-alone AI	p-value	
Sensitivity	100.0 (2/2) (15.8100.0)	50.0 (1/2) (1.3, 98.7)	0.317	63.0 (17/27) (42.4, 80.6)	66.7 (18/27) (46.0, 83.5)	0.778	0.638
Specificity	90.7 (223/246) (86.3,94.0)	95.9 (236/246) (92.7, 98.0)	0.021	92.3 (1649/1786) (91.0, 93.5)	94.0 (1678/1786) (92.7, 95.0)	0.044	0.231
PPV	8.0 (2/25) (5.6,11.4)	9.1 (1/11) (2.2, 31.2)	0.914	11.0 (17/154) (8.2, 14.7)	14.3 (18/126) (10.8, 18.7)	0.407	0.634
NPV	100.0 (223/223) (98.4100.0)	99.6 (236/237) (98.3, 99.9)	0.345	99.4 (1649/1659) (99.0, 99.6)	99.5 (1678/1687) (99.1, 99.7)	0.696	0.836
AUC	0.953 (0.92,0.98)	0.730 (0.67, 0.78)	0.373	0.776 (0.76, 0.80)	0.803 (0.78, 0.82)	0.684	0.773
Accuracy	90.7 (225/248) (86.4, 94.0)	95.6 (237/248) (92.2, 97.8)	0.031	91.9 (1666/1813) (90.5, 93.1)	93.5 (1696/1813) (92.3, 94.6)	0.064	0.200
Detection rate	2/248	1/248	0.562	17/1813	18/1813	0.865	0.205
Recall rate	10.1 (25/248) (6.6, 14.5)	4.4 (11/248) (2.2, 7.8)	0.015	8.5 (154/1813) (7.3, 9.9)	6.9 (126/1813) (5.8, 8.2)	0.071	0.137

PPV: Positive predictive value, NPV: Negative predictive value, AUC: Area under the receiver operating characteristic curve

The number in parentheses indicates the 95% confidence interval. [†] *P*-value, comparison of stand-alone AI performance between fatty and dense breasts

The largest scale validation study of AI-CAD so far using more than 160,000 mammograms and multiple AI algorithms found that the combination of a radiologist and AI-CAD outperformed a single radiologist[15]; however, retrospectively applied AI-CAD or simulation studies are not the same as clinical practice where radiologists and clinicians refer to AI-CAD in real-time. Recent prospective trials reported that the selected replacement of one of two radiologists in a double-reading system by AI-CAD showed similar to higher cancer detection without increasing recall rates [17,18]; however, this was not validated in a single-reading system with AI-CAD assistance. In our institution, AI-CAD has become part of all mammography examinations with its integration into PACS since the hospital first opened in March 2020. With the ubiquitous application of AI-CAD possible, we used a monthly alternate setting to provide and withhold the AI-CAD results from PACS during the study period, and we thought this method could minimize interruptions that might occur with radiologists, radiographers, mammography equipment and other clinical circumstances.

In our study, AI-CAD assistance seemed to lead to better diagnostic performance for screening mammography, but it did not add clinical value with statistical significance. We still consider our findings meaningful because they were deduced with mammography screening performed on the same day as ultrasound. While we recommend supplementary ultrasound for dense breasts, women with non-dense breasts also receive ultrasound for reasons other than medical conditions such as the availability of personal insurance, anxiety or a history of benign breast lesions. This could be the reason behind the relatively high cancer detection rate in our study population, and it is consistent with the findings of a recent article that reported a cancer detection rate of 9.3 per 1000 women in Korea, who underwent mammography plus supplementary screening ultrasound [25].

While radiologists were asked to interpret mammography independently from ultrasound, we could not completely exclude the impact that ultrasound results could have on the diagnostic performance of mammography, such as overestimated performance values for mammography or dilution of the role of AI-CAD. In another study in our institution which applied AI-CAD retrospectively to patients who underwent both screening mammography and ultrasound, AI-CAD failed to improve diagnostic performance [26]. Besides, available prior examinations were also reviewed for the best interpretation in routine practice, which could increase the baseline diagnostic performance of mammography. While the use of supplementary ultrasound might be a confounding factor, findings that accept its routine use reflect what happens in actual clinical practice.

Most mammograms (2015/2061, 97.8%) were interpreted by experienced breast radiologists (5–28 years) in our study. We know from previous studies that inexperienced readers are the ones who show more improvement in diagnostic performance with AI-CAD [12]. Radiologists were completely free to follow or disregard the results of AI-CAD. However, stand-alone AI-CAD showed significantly higher specificity, accuracy, and AUC and lower recall rates than radiologists throughout the whole period. Other parameters including sensitivity, PPV, NPV and cancer detection rate were not statistically different. We analyzed real-world interpretation results, and unlike in reader studies, radiologists in actual clinical settings tend to confirm lesions that appear likely benign, with additional magnification views, ultrasound, or short-term follow-up to minimize the risk of missed cancers. This tendency can lower specificity and, consequently, reduce overall accuracy. A recent meta-analysis found that the pooled AUC of stand-alone AI-CAD was similar to better than radiologists[27].

NPV of AI-CAD remained very high, regardless of breast density (dense breast, 99.6%, 1678/1687 vs fatty breast, 99.6%, 236/237). However, the false-negative rate of AI-CAD, also known as the missed cancer rate, was 34% (10/29), which was higher than previously published studies using the same platform at 11–19.4% [12,28]. In other words, there is still room for screening ultrasound to find additional cancers, especially for dense breasts, which generally make up the more complex cases for radiologists. A recent study found that supplementary automated ultrasound increased cancer detection rates regardless of breast density [25]; however, as expected, it decreased the specificity of digital mammography. Since stand-alone AI-CAD showed higher specificity and NPV, it may compensate for the decreased specificity of supplementary ultrasound. As the number of patients in our study was not enough to analyze how AI-CAD affected the performance of radiologists according to breast density, future research will need to elucidate this with a higher number of patients. Generally, for women who perform screening mammography only without ultrasound, results are expected to differ, and this method needs to be further validated in prospective studies.

There were several limitations to this study. First, a large portion (38%) of the study sample was excluded due to incomplete follow-up with many visits being a one-time check-up, especially in women with negative or benign results. This likely contributed to an atypically high cancer detection rate in the study. Second, the study was performed in a single institution with a single AI-CAD platform. Third, our results might not apply to institutions where screening populations are examined by mammography only without ultrasound.

In conclusion, radiologists showed no significant difference in diagnostic performance when both screening mammography and ultrasound were performed with or without AI-CAD assistance for mammography. However, without AI-CAD assistance, radiologists showed lower specificity and accuracy and higher recall rates compared to stand-alone AI-CAD.

Ethical approval details

This retrospective study was approved by the institutional review board (Yongin Severance Hospital, 9–2022-0059) with a waiver for informed consent.

Funding sources

This work was supported by Medical AI Clinic Program through the National IT Industry Promotion Agency (NIPA), funded by the Ministry of Science and ICT (MSIT).

CRediT authorship contribution statement

Kim Eun-Kyung: Writing – review & editing, Supervision, Data curation, Conceptualization. **Hong Hanpyo:** Writing – review & editing, Visualization, Methodology. **Lee Si Eun:** Writing – review & editing, Writing – original draft, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Tabar, B. Vitak, H.H. Chen, M.F. Yen, S.W. Duffy, R.A. Smith, Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality, Cancer 91 (2001) 1724–1731.

[2] L. Tabar, M.F. Yen, B. Vitak, H.H. Chen, R.A. Smith, S.W. Duffy, Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening, Lancet 361 (2003) 1405–1410.

[3] S. Ciatto, N. Houssami, D. Bernardi, et al., Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study, Lancet Oncol. 14 (2013) 583–589.

[4] N. Ni Mhuircheartaigh, L. Coffey, H. Fleming, O.D. A, S. McNally, With the advent of tomosynthesis in the workup of mammographic abnormality, is spot compression mammography now obsolete? an initial clinical experience, Breast J. 23 (2017) 509–518.

[5] C.D. Lehman, R.D. Wellman, D.S.M. Buist, et al., Diagnostic accuracy of digital screening mammography with and without computer-aided detection, JAMA Intern. Med. 175 (2015) 1828–1837.

[6] J.J. Fenton, L. Abraham, S.H. Taplin, et al., Effectiveness of computer-aided detection in community mammography practice, J. Natl. Cancer Inst. 103 (2011) 1152–1161.

[7] J.J. Fenton, S.H. Taplin, P.A. Carney, et al., Influence of computer-aided detection on performance of screening mammography, N. Engl. J. Med 356 (2007) 1399–1409.

[8] J.H. Lee, K.H. Kim, E.H. Lee, et al., Improving the performance of radiologists using artificial intelligence-based detection support software for mammography: a multi-reader study, Korean J. Radio. 23 (2022) 505–516.

[9] S.M. McKinney, M. Sieniek, V. Godbole, et al., International evaluation of an AI system for breast cancer screening, Nature 577 (2020) 89–94.

[10] J.H. Yoon, E.-K. Kim, Deep learning-based artificial intelligence for mammography, Korean J. Radio. 22 (2021) 1225–1239.

[11] Y.S. Kim, M.-j Jang, S.H. Lee, et al., Use of artificial intelligence for reducing unnecessary recalls at screening mammography: a simulation study, Korean J. Radio. 23 (2022) 1241–1250.

[12] H.E. Kim, H.H. Kim, B.K. Han, et al., Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study, Lancet Digit Health 2 (2020) e138–e148.

[13] K. Lång, S. Hofvind, A. Rodríguez-Ruiz, I. Andersson, Can artificial intelligence reduce the interval cancer rate in mammography screening? Eur. Radiol. 31 (2021) 5940–5947.

[14] K. Lång, M. Dustler, V. Dahlblom, A. Åkesson, I. Andersson, S. Zackrisson, Identifying normal mammograms in a large screening population using artificial intelligence, Eur. Radiol. 31 (2021) 1687–1692.

[15] T. Schaffter, D.S.M. Buist, C.I. Lee, et al., Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms, JAMA Netw. Open 3 (2020) e200265-e200265.

[16] C. Leibig, M. Brehmer, S. Bunk, D. Byng, K. Pinker, L. Umutlu, Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis, Lancet Digit. Health 4 (2022) e507–e519.

[17] K. Lång, V. Josefsson, A.M. Larsson, et al., Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study, Lancet Oncol. 24 (2023) 936–944.

[18] K. Dembrower, A. Crippa, E. Colón, M. Eklund, F. Strand, Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study, Lancet Digit Health 5 (2023) e703–e711.

[19] K. Freeman, J. Geppert, C. Stinton, et al., Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy, Bmj 374 (2021) n1872.

[20] W.A. Berg, J.D. Blume, J.B. Cormack, et al., Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer, Jama 299 (2008) 2151–2163.

[21] N. Harada-Shoji, A. Suzuki, T. Ishida, et al., Evaluation of adjunctive ultrasonography for breast cancer detection among women aged 40-49 years with varying breast density undergoing screening mammography: a secondary analysis of a randomized clinical trial, JAMA Netw. Open 4 (2021) e2121505-e2121505.

[22] W.H. Yuan, H.C. Hsu, Y.Y. Chen, C.H. Wu, Supplemental breast cancer-screening ultrasonography in women with dense breasts: a systematic review and meta-analysis, Br. J. Cancer 123 (2020) 673–688.

[23] K. Dembrower, E. Wåhlin, Y. Liu, et al., Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study, Lancet Digit. Health 2 (2020) e468–e474.

[24] M. Salim, E. Wahlin, K. Dembrower, et al., External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms, JAMA Oncol. 6 (2020) 1581–1588.

[25] M.-r Kwon, J.S. Choi, M.Y. Lee, et al., Screening outcomes of supplemental automated breast US in Asian women with dense and nondense breasts, Radiology 307 (2023) e222435.

[26] S.E. Lee, J.H. Yoon, N.H. Son, K. Han, H.J. Moon, Screening in patients with dense breasts: comparison of mammography, artificial intelligence, and supplementary ultrasound, AJR Am. J. Roentgenol. (2023).

[27] Yoon J.H., Strand F., Baltzer P.A.T., et al. Standalone AI for Breast Cancer Detection at Screening Digital Mammography and Digital Breast Tomosynthesis: A Systematic Review and Meta-Analysis. Radiology. 0:222639.

[28] S.E. Lee, K. Han, J.H. Yoon, J.H. Youk, E.K. Kim, Depiction of breast cancers on digital mammograms by artificial intelligence-based computer-assisted diagnosis according to cancer characteristics, Eur. Radio. 32 (2022) 7400–7408.