Original Article | Breast Imaging

eISSN 2005-8330
https://doi.org/10.3348/kjr.2021.0476
Korean J Radiol 2022;23(5):505-516

Korean Journal of Radiology
KJR

Improving the Performance of Radiologists Using Artificial Intelligence-Based Detection Support Software for Mammography: A Multi-Reader Study

Jeong Hoon Lee[1]*, Ki Hwan Kim[1]*, Eun Hye Lee[2], Jong Seok Ahn[1], Jung Kyu Ryu[3], Young Mi Park[4], Gi Won Shin[4], Young Joong Kim[5], Hye Young Choi[6]

[1]Lunit Inc., Seoul, Korea; [2]Department of Radiology, Soonchunhyang University Bucheon Hospital, Soonchunhyang University College of Medicine, Bucheon, Korea; [3]Department of Radiology, Kyung Hee University Hospital at Gangdong, Seoul, Korea; [4]Department of Radiology, Inje University Busan Paik Hospital, Inje University College of Medicine, Busan, Korea; [5]Department of Radiology, Konyang University Hospital, Konyang University College of Medicine, Daejeon, Korea; [6]Department of Radiology, Gyeongsang National University Hospital and College of Medicine, Gyeongsang National University, Jinju, Korea

Objective: To evaluate whether artificial intelligence (AI) for detecting breast cancer on mammography can improve the performance and time efficiency of radiologists reading mammograms.

Materials and Methods: A commercial deep learning-based software for mammography was validated using external data collected from 200 patients, 100 each with and without breast cancer (40 with benign lesions and 60 without lesions) from one hospital. Ten readers, including five breast specialist radiologists (BSRs) and five general radiologists (GRs), assessed all mammography images using a seven-point scale to rate the likelihood of malignancy in two sessions, with and without the aid of the AI-based software, and the reading time was automatically recorded using a web-based reporting system. Two reading sessions were conducted with a two-month washout period in between. Differences in the area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, and reading time between reading with and without AI were analyzed, accounting for data clustering by readers when indicated.

Results: The AUROC of the AI alone, BSR (average across five readers), and GR (average across five readers) groups was 0.915 (95% confidence interval, 0.876–0.954), 0.813 (0.756–0.870), and 0.684 (0.616–0.752), respectively. With AI assistance, the AUROC significantly increased to 0.884 (0.840–0.928) and 0.833 (0.779–0.887) in the BSR and GR groups, respectively ($p = 0.007$ and $p < 0.001$, respectively). Sensitivity was improved by AI assistance in both groups (74.6% vs. 88.6% in BSR, $p < 0.001$; 52.1% vs. 79.4% in GR, $p < 0.001$), but the specificity did not differ significantly (66.6% vs. 66.4% in BSR, $p = 0.238$; 70.8% vs. 70.0% in GR, $p = 0.689$). The average reading time pooled across readers was significantly decreased by AI assistance for BSRs (82.73 vs. 73.04 seconds, $p < 0.001$) but increased in GRs (35.44 vs. 42.52 seconds, $p < 0.001$).

Conclusion: AI-based software improved the performance of radiologists regardless of their experience and affected the reading time.

Keywords: *Breast cancer; Mammography; Screening; Deep-learning; Artificial intelligence; Reading time; Multi-reader study*

INTRODUCTION

Breast cancer is the second-leading cause of cancer-related death in female worldwide [1,2]. Mammography screening has been widely used to reduce breast cancer mortality; however, reading a large number of mammography screenings is laborious, time-consuming, and cost-intensive [3-5]. Computer-aided diagnosis (CAD) systems were

Received: March 10, 2021 **Revised:** January 4, 2022 **Accepted:** January 24, 2022
*These authors contributed equally to this work.
Corresponding author: Eun Hye Lee, MD, Department of Radiology, Soonchunhyang University Bucheon Hospital, Soonchunhyang University College of Medicine, 170 Jomaru-ro, Bucheon 14584, Korea.
• E-mail: grace@schmc.ac.kr

introduced as a support tool in the 1990s as a potential method for overcoming resource deficiency in mammography screening [6-8]. Recently, with the introduction of artificial intelligence (AI), breast experts have evaluated the performance of AI algorithms in mammography [9-11].

CAD systems were introduced into clinical practice based on their high sensitivity; however, subsequent large-scale studies have shown that these systems rarely change the decisions made by radiologists in breast cancer screening [12-14]. The main reason for the inability of CAD systems to improve radiologists' performance in practice was that CAD identified many redundant markers per image; therefore, radiologists did not trust the results provided [15]. Moreover, their high false-positive rates led to an increase in unnecessary examinations.

The use of AI in detecting breast cancer on mammograms is of great interest [16-19]. Although AI algorithms show great potential in breast cancer screening, their characteristics must be carefully evaluated and radiologists should fully understand these characteristics when they are used in practice [20,21]. This is because the effectiveness of the AI algorithms in diagnostic decision-making may vary with the radiologist's experience [10,16]. Therefore, the strengths and weaknesses of AI algorithms should be identified based on radiologist experience before they are introduced into clinical practice. This study investigated whether AI could improve radiologists' performance using a commercially available deep-learning-based software for mammography depending on their experience with external data [10]. The effects of AI on reading time efficiency were also evaluated.

MATERIALS AND METHODS

This retrospective study was approved by the Ethics Committee of the Institutional Review Board of Soonchunhyang University Hospital Bucheon (IRB No. SCHBC 2019-08-022-001), and the requirement for informed consent was waived. Lunit Inc. provided technical support for the AI system and statistical analysis of the study data. The first two authors, who are employees of Lunit Inc., contributed to the study design and analysis of the results of the image interpretation by the study readers; however, they did not participate in collecting and reading the mammograms in the reader study, which were collected from Soonchunhyang University Hospital Bucheon.

Patients and Datasets

A total of 200 external validation cases, including biopsy-proven malignant (n = 100), benign (n = 40), and negative (n = 60) cases, were collected from adult Asian female who underwent four-view mammography for breast cancer screening between March 2009 and September 2018 (Fig. 1, Table 1). Based on the assumption that the AUROC for mammography interpretation by radiologists is approximately 0.90, the AUROC difference between readings with and without AI assistance was greater than 0.15, and the variability between observers was moderate. The sample size of 200 was estimated to be sufficient for demonstrating the performance change elicited by the use of AI [22].

Biopsy-proven malignant cases were consecutively collected from patients who were pathologically confirmed as having breast cancer within six months after mammography, and bilateral breast cancer was not included. Negative cases were collected from patients who were Breast Imaging Reporting and Data System (BI-RADS) final assessment category 1 on mammography and confirmed as negative over more than 2 years of follow-up imaging. Benign cases were confirmed via biopsy (biopsy-confirmed benign) or follow-up imaging for more than 2 years (clinically confirmed benign). During the data collection period, 20 cases were randomly collected for each benign category and 60 negative cases were randomly collected. Age and BI-RADS composition categories did not differ significantly among the three groups (ANOVA, $p > 0.05$). Of the malignant cases, 28 were ductal carcinoma *in situ* and 65 were invasive cancers. Of the latter, the majority were invasive ductal carcinomas (55/65, 84.6%). The remaining invasive cancers included lobular carcinoma (6/65, 9.2%), mucinous carcinoma (2/65, 3.1%), apocrine carcinoma (1/65, 1.5%), and tubular carcinoma (1/65, 1.5%). Seven cancer patients underwent surgery in other hospitals after receiving a core biopsy; therefore, their stage information was not available.

Mammography Examination

All digital mammography examinations included four views of full-field digital mammograms (i.e., right craniocaudal view, right mediolateral oblique view, left craniocaudal view, and left mediolateral oblique view), acquired using an equivalent full-field digital mammography system (Senographe 2000D; GE Healthcare).

https://doi.org/10.3348/kjr.2021.0476

Korean Journal of Radiology
KJR

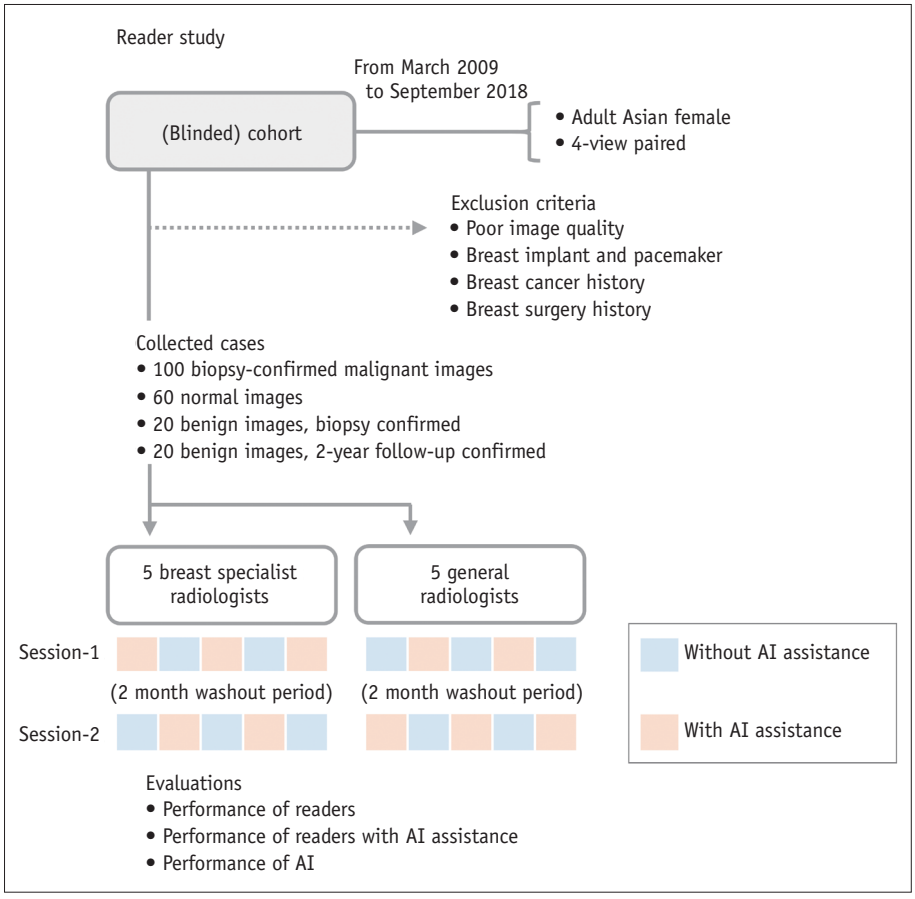

Fig. 1. Schematic of the workflow for the multi-reader study. The squares in sessions 1 and 2 represent readers. AI = artificial intelligence

AI Algorithm

Commercially available deep learning-based software for mammography (Lunit INSIGHT MMG, version 1.1.1.0; Lunit) was used in this study [10]. This AI algorithm was trained to detect breast cancer using data from 170230 mammograms, including data from over 30000 breast cancer cases. The AI algorithm provided an abnormality score based on four views between 0 and 100, indicating the possibility of breast cancer, and a heatmap was shown at the location of the abnormal region, indicating an abnormality score of 10 or more.

Reader Study

Ten radiologists with different levels of experience participated in the reader study. Of these, five were breast specialist radiologists (BSRs) with 4–19 years of breast imaging experience and five were general radiologists (GRs) with 1–10 years of post-training experience. BSRs read more than 3000 mammography examinations per year, whereas GRs did not. The workflow of this multicase multireader study is shown in Figure 1. Five readers (three BSRs and two GRs) read the mammograms with AI, and after two months of washout, they re-read the mammograms without AI. The remaining five readers (two BSRs and three GRs) read the mammograms with and without AI assistance. A reader-crossover design and two months of washout period were adopted to minimize the potential bias that the order of AI usage can affect the reading time and performance of the readers [23]. There was no change in the reading environment for each reader between reading sessions with and without AI. All readers performed this study using a web-based reader study system with the hardware configuration used in daily practice.

All radiologists read the mammograms rated the likelihood of malignancy (LOM) in each case on a seven-point scale in both readings with and without AI: 1 = definitely normal, 2 = benign, 3 = probably benign, 4 = low suspicion of malignancy, 5 = moderate suspicion of malignancy, 6 = high suspicion of malignancy, and 7 = highly suggestive of malignancy. The radiologist was requested to record an LOM score of at least 3 when a suspected breast cancer lesion was detected. In the reading with AI assistance, the

 https://doi.org/10.3348/kjr.2021.0476

Table 1. Population Characteristics and Mammographic Features

	Cancer	Benign	Negative
Number of samples	100	40	60
Age, years			
Mean	53.03	50.15	51.97
Median	51	48	51
Range	36–78	36–75	39–78
Interquartile range	47.75–57.00	44.00–55.00	45.75–57.00
BI-RADS composition categories			
a	5 (5.0)	0 (0.0)	3 (5.0)
b	22 (22.0)	11 (27.5)	22 (36.7)
c	38 (38.0)	23 (57.5)	18 (30.0)
d	35 (35.0)	6 (15.0)	17 (28.3)
Mammographic features			
Calcification	49	27	
Mass	42	14	
Asymmetry	23	8	
Distortion	5	1	
T stage			
0	28		
I	54		
II	11		
Unknown	7		
N stage			
0	75		
1	18		
2	0		
Unknown	7		

Data are number of cases with % in parentheses, unless specified otherwise. BI-RADS = Breast Imaging Reporting and Data System

heatmap was overlaid on the original mammograms, and the heatmap could be turned on and off easily using the shortcut key. The readers considered both the AI results and the original mammogram findings and made their own final judgments using a seven-point scale. Additionally, for localization ROC (LROC) analysis, each radiologist was requested to mark one location of the most suspicious region and then provide the LOM using the same seven-point scale. For each case, the reading time from the initial image view by the reader to the final decision was automatically recorded using a web-based reporting system, and the readers were informed of this time measurement in advance.

Reference Standard

To conduct the LROC analysis, an expert with 20 years of experience in breast imaging annotated the location of malignant lesions with a free-form contour line referring to the original mammography and pathology reports, and this annotation was considered the ground truth. The expert also evaluated the BI-RADS composition categories and dominant mammographic features. For mammographic features, one of the three most prominent features (mass, asymmetry, and architectural distortion) was selected, and the presence of microcalcifications suspected of implying malignancy was evaluated separately.

Statistical Analysis

Diagnostic performance was primarily analyzed using conventional receiver operating characteristic (ROC) curve analysis. For AI, the area under the ROC (AUROC) was measured using the abnormality score (range: 0–100). The radiologists' AUROC values were obtained based on the LOM. The difference in AUROCs between the readings with and without AI was compared using the DeLong test in the pROC R package (version 3.6.3; R Project for Statistical Computing, https://www.r-project.org) [24]. The average AUROC for each group was obtained from the arithmetic mean of the AUC [25]. The average AUROC was compared using the iMRMC method [26]. To analyze the performance considering the correct localization of the lesion, LROC analysis was performed. The area under the LROC (AULROC) was measured using the non-parametric trapezoidal method [27]. $Sensitivity_{LROC}$ was calculated using true positives, excluding the incorrect location cases. The localization of the AI was considered correct if the location of the maximum pixel-level abnormality scores was inside the closed free-form line of the ground truth drawn by the expert radiologist. The correctness of the localization of the readers was determined on the basis of whether the reader's point mark was inside the ground truth [28]. The difference in AULROCs between the readings with and without AI was evaluated using Hillis' modification of the Dorfman-Berbaum-Metz (DBMH) method [29,30]. The DBMH method was conducted using the StSignificanceTesting function of the RJafroc R package v1.3.2 [31,32].

For both analyses, a representative set of sensitivity and specificity values was obtained and compared between the reading modes. The sensitivity and specificity of AI were determined using a threshold of 10, as described in a previous study [10]. For the radiologists' interpretation, an LOM of 1–2 was considered negative, and an LOM ≥ 3 was considered positive to calculate sensitivity and specificity. The change in sensitivity and specificity of each reader was compared using McNemar's test, and the change in

https://doi.org/10.3348/kjr.2021.0476

Korean Journal of Radiology
KJR

sensitivity and specificity of each reading group (BSR and GR) was evaluated using logistic regression with the generalized estimating equation (GEE) method, and 95% confidence intervals (CIs) were calculated for each group using the DTComPair and the geepack R package [33,34]. In cases of cancer, the sensitivities of the readers with and without AI were compared in subgroups according to T and N stages using the GEE method. Additionally, the Wilcoxon signed-rank test was used to compare the reading times pooled across readers with and without AI assistance.

Statistical significance was set at $p < 0.05$. Statistical analyses were performed using R statistical software (version 3.6.3).

RESULTS

Comparative Performance of AI-Unassisted and AI-Assisted Reading

Figure 2 shows the performance results of the AI algorithm and BSRs and GRs with and without AI assistance. The AUROC values of the AI, BSRs (average across five readers), and GRs (average across five readers) were 0.915, 0.813, and 0.684, respectively (Table 2). With AI assistance, the AUROC values significantly increased to 0.884 ($p = 0.007$) and 0.833 ($p < 0.001$) in the BSR and GR groups, respectively. Sensitivity was also significantly increased by AI assistance in both the BSR (74.6% vs. 88.6%, $p < 0.001$) and GR groups (52.1% vs. 79.4%, $p < 0.001$), but the specificity did not differ significantly (BSR: 66.6% vs. 66.4%, $p = 0.238$; GR: 70.8% vs. 70.0%, $p = 0.689$). When the GR group used AI assistance, the AUROC was not significantly different from that of the BSR group without AI ($p = 0.138$). Table 2 lists the performance of each reader. With AI, the AUROCs changed from 0.794–0.839 to 0.835–0.920 in the BSR group and from 0.608–0.727 to 0.794–0.859 in the GR group, with statistically significant differences, except for one BSR. The sensitivity also significantly increased for two of the BSRs and all GRs. However, there was no significant difference in the specificity of the nine readers (three BSRs and five GRs) between AI-assisted and unassisted readings.

In the LROC analysis, the AULROC values were significantly increased by AI assistance in the BSR (0.635 vs. 0.767, $p = 0.034$) and GR (0.420 vs. 0.694, $p < 0.001$) groups. The corresponding sensitivity ($sensitivity_{LROC}$) values were also significantly increased using AI in the BSR (68.2% vs. 80.6%, $p < 0.001$) and GR (45.4% vs. 74.8%, $p < 0.001$). A representative example of mammography examinations in which the number of correct recall decisions across readers changed between two reading sessions is shown in Figure 3.

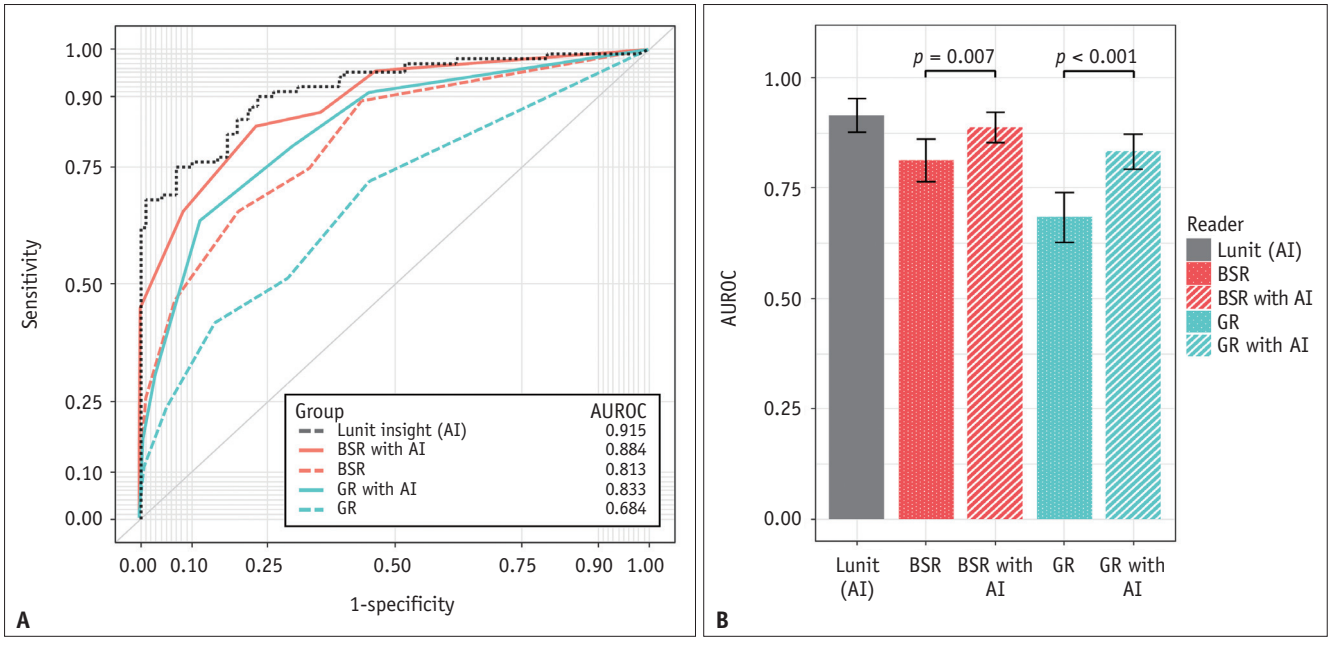

Fig. 2. Performance of the AI alone and BSR and GR groups with and without AI.
A. ROC curves of radiologists with and without AI. **B.** Difference of AUROC between AI and radiologists. AI = artificial intelligence, AUROC = area under the ROC curve, BSR = breast specialist radiologist, GR = general radiologist, ROC = receiver operating characteristic

https://doi.org/10.3348/kjr.2021.0476

Table 2. Diagnostic Performance of Five BSRs and Five GRs with and without AI Assistance for 100 Patients with and 100 Patients without Breast Cancer

	Conventional ROC Analysis									LROC Analysis					
	AUROC			Sensitivity			Specificity			AULROC			Sensitivity$_{LROC}$		
	AI-Unassisted	AI-Assisted	*P*	AI-Unassisted, %	AI-Assisted, %	*P*	AI-Unassisted, %	AI-Assisted, %	*P*	AI-Unassisted	AI-Assisted	*P*	AI-Unassisted, %	AI-Assisted, %	*P*
BSR															
1	0.839	0.903	0.008	75.0	86.0	0.005	69.0	68.0	0.819	0.673	0.787	0.003	71.0	81.0	0.012
2	0.826	0.901	0.002	83.0	90.0	0.052	58.0	66.0	0.033	0.714	0.820	0.004	77.0	85.0	0.045
3	0.808	0.835	0.293	79.0	83.0	0.317	68.0	61.0	0.108	0.646	0.689	0.242	70.0	75.0	0.225
4	0.797	0.863	0.004	77.0	82.0	0.166	63.0	67.0	0.317	0.615	0.713	0.003	68.0	76.0	0.021
5	0.794	0.920	< 0.001	59.0	92.0	< 0.001	75.0	60.0	< 0.001	0.526	0.824	< 0.001	55.0	86.0	< 0.001
Average (95% CI)	0.813 (0.756–0.870)	0.884 (0.840–0.928)	0.007	74.6 (70.8–78.4)	88.6 (83.6–89.6)	< 0.001	66.6 (62.5–70.7)	64.4 (60.2–68.6)	0.238	0.635 (0.564–0.706)	0.767 (0.700–0.833)	0.034	68.2 (64.1–72.3)	80.6 (77.1–84.1)	< 0.001
GR															
1	0.727	0.855	< 0.001	75.0	87.0	0.011	52.0	56.0	0.433	0.499	0.738	< 0.001	57.0	80.0	< 0.001
2	0.716	0.822	0.001	46.0	81.0	< 0.001	75.0	67.0	0.102	0.399	0.713	< 0.001	43.0	78.0	< 0.001
3	0.712	0.833	< 0.001	58.0	81.0	< 0.001	68.0	68.0	1.000	0.496	0.710	< 0.001	53.0	76.0	< 0.001
4	0.657	0.859	< 0.001	41.0	77.0	< 0.001	81.0	83.0	0.564	0.383	0.693	< 0.001	40.0	73.0	< 0.001
5	0.608	0.794	< 0.001	36.0	71.0	< 0.001	78.0	76.0	0.593	0.323	0.616	< 0.001	34.0	67.0	< 0.001
Average (95% CI)	0.684 (0.616–0.752)	0.833 (0.779–0.887)	< 0.001	51.2 (46.8–55.6)	79.4 (75.9–82.9)	< 0.001	70.8 (66.8–74.8)	70.0 (66.0–74.0)	0.689	0.420 (0.356–0.484)	0.694 (0.625–0.763)	< 0.001	45.4 (41.0–49.8)	74.8 (0.710–0.786)	< 0.001
All radiologists (95% CI)	0.748 (0.686–0.811)	0.858 (0.809–0.907)	< 0.001	62.9 (59.9–65.9)	83.0 (80.7–85.3)	< 0.001	68.7 (65.8–71.6)	67.2 (64.3–70.1)	0.273	0.527 (0.464–0.590)	0.730 (0.666–0.794)	< 0.001	56.8 (53.7–59.9)	77.7 (75.1–80.3)	< 0.001
AI (95% CI)	0.915 (0.876–0.954)			87.0 (80.4–93.6)			79.0 (71.0–87.0)			0.769 (0.689–0.849)			79.0 (78.8–93.0)		

AI = artificial intelligence, AUROC = area under the receiver operating characteristics curve, BSR = breast specialist radiologist, CI = confidence interval, GR = general radiologist, LROC = localization receiver operating characteristic

Korean Journal of Radiology
KJR

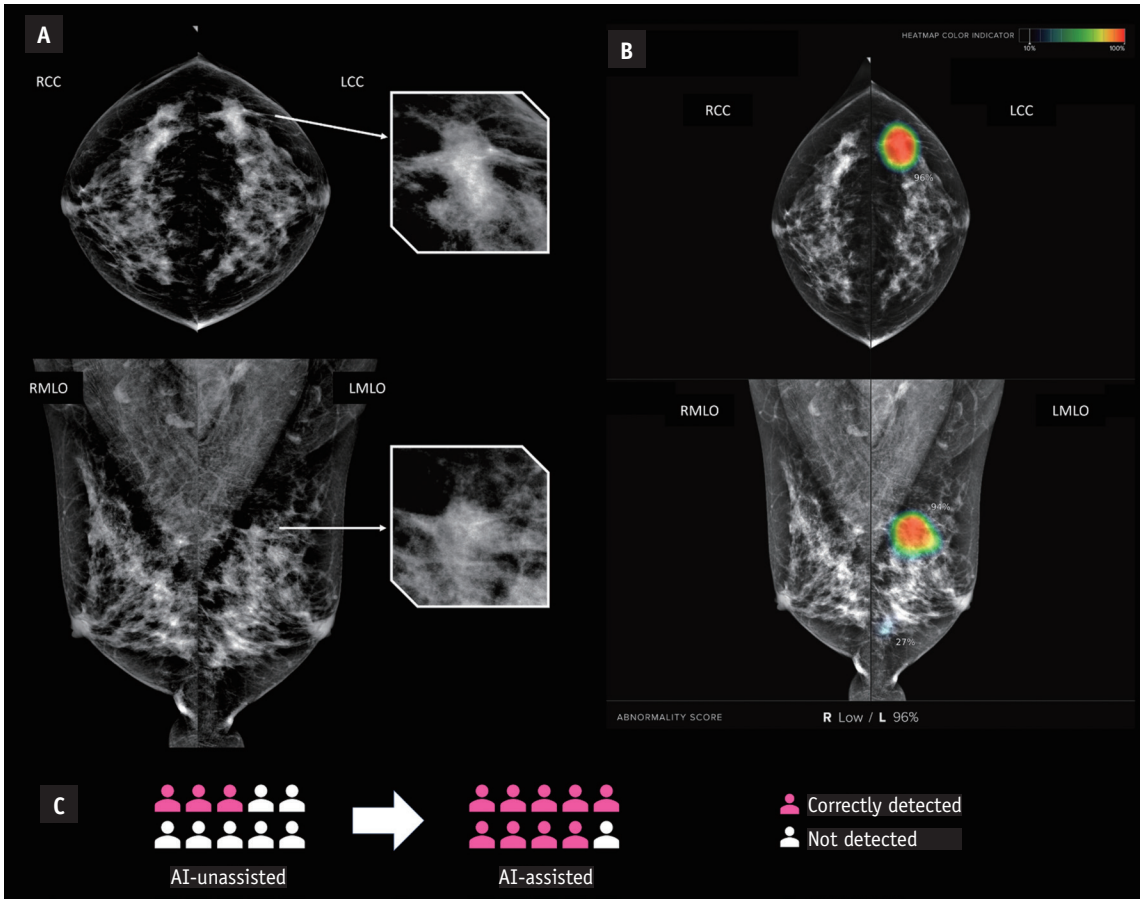

Fig. 3. Examples of breast cancer detected with the aid of AI.
A. Mammograms in 47-year-old female with invasive ductal carcinoma. **B.** Heatmap and abnormality score are shown as in the viewer of the AI-based software. **C.** The patient was recalled by three of 10 radiologists when reading without AI assistance and by nine of 10 radiologists using AI-based software for support. AI = artificial intelligence

Effects of AI Assistance in Breast Cancer Detection according to T and N Stages

The sensitivities of the AI and readers were evaluated in patients (n = 93) with T-stage and N-stage information. Table 3 lists the changes in sensitivity and statistical significance according to T-stage and N-stage, representing tumor size and lymph node metastasis status, respectively. In both the BSR and GR groups, sensitivity significantly increased with AI assistance, regardless of the T-stage and N-stage of the patients.

Subgroup analysis was performed on AUROC, sensitivity, and specificity by dividing the BI-RADS composition category into two groups: a-b and c-d (Supplementary Table 1). In both groups, the AUROC and sensitivity increased significantly with AI. In the BSR group, the specificity significantly increased in the BI-RADS composition category a-b group (67.8% vs. 74.4%, p = 0.014) and decreased in the c-d group (65.9% vs. 58.8%, p = 0.003). There was no significant difference in specificity in the GR group.

To observe how the decision of readers was changed by AI assistance, the numbers of detected and missed cases according to the AI and radiologists are shown in Table 4. The majority of radiologists made the decision. For cases detected by AI, the number of cases detected correctly by radiologists increased from 54 to 75 with AI assistance. For cases that the AI alone could not detect, radiologists additionally detected three cases using AI.

Effects of AI Assistance on Reading Time

With AI assistance, the overall average reading time pooled across readers significantly changed between before and after using AI in the BSR (decrease from 82.73 seconds to 73.04 seconds, p < 0.001) and GR (increase from 35.44 seconds to 42.52 seconds, p < 0.001) groups. The average reading times of the BSR group were significantly decreased

https://doi.org/10.3348/kjr.2021.0476

Table 3. Effects of AI Assistance on Sensitivity for Radiologist Groups according to T-Stage and N-Stage

	T-0 (n = 28)			T-1 (n = 54)			T-2 (n = 11)			N-0 (n = 75)			N-1 (n = 18)		
	AI-Unassisted, %	AI-Assisted, %	*P*	AI-Unassisted, %	AI-Assisted, %	*P*	AI-Unassisted, %	AI-Assisted, %	*P*	AI-Unassisted, %	AI-Assisted, %	*P*	AI-Unassisted, %	AI-Assisted, %	*P*
BSR	73.6 (66.3–80.9)	82.9 (76.6–89.1)	0.001	72.6 (67.3–77.9)	86.7 (82.6–90.7)	0.004	72.7 (61.0–84.5)	89.1 (80.9–97.3)	0.048	72.3 (67.7–76.8)	84.5 (80.9–88.2)	0.001	75.6 (66.7–84.4)	91.1 (85.2–97.0)	0.011
GR	56.4 (48.2–64.6)	75.7 (68.6–82.8)	< 0.001	47.4 (41.5–53.4)	80.4 (75.6–85.1)	< 0.001	41.8 (28.8–54.9)	74.5 (63.0–86.1)	0.004	50.9 (45.9–56.0)	77.6 (73.4–81.8)	< 0.001	43.3 (33.1–53.6)	81.1 (73.0–89.2)	< 0.001
AI	82.1 (68.0–96.3)			87.0 (78.1–96.0)			90.9 (73.9–107.9)			84.0 (75.7–92.3)			94.4 (83.9–105.0)		

p values of unaided versus with AI from generalized estimating equation analysis. Values in parentheses are 95% confidence interval. AI = artificial intelligence, BSR = breast specialist radiologist, GR = general radiologist

regardless of the three case types: cancer, benign and negative (83.05 seconds vs. 74.79 seconds, $p = 0.004$; 86.41 seconds vs. 74.20 seconds, $p = 0.025$; and 79.72 seconds vs. 69.37 seconds, $p = 0.040$, respectively) (Fig. 4). However, in the GR group, there was a significant increase in cancer and benign cases (37.99 seconds vs. 46.94 seconds, $p < 0.001$ and 33.62 seconds vs. 47.45 seconds, $p < 0.001$, respectively).

DISCUSSION

We investigated whether deep-learning-based AI assistance could improve the detection performance of radiologists. This study analyzed the change in detection performance and reading time using AI assistance according to the reader's experience and characteristics of the samples. AI assistance improved detection performance regardless of the experience of the radiologists and the characteristics of the samples. AI increased the efficiency of the BSR, which had a relatively long reading time, and helped the GR focus the lesions in detail, thereby increasing the reading time. Before introducing AI algorithms into clinical practice, this study will be helpful in anticipating how AI affects the detection performance of radiologists in the real world.

Recently, many AI algorithms have been developed to assist breast cancer detection, and they have focused on increasing sensitivity without reducing specificity, which can lead to unnecessary recalls and higher medical costs [15]. Consequently, many studies have shown that AI algorithms have superior performance compared to traditional CAD [9,10,16,18]. A study by Kooi et al. [35] reported that a deep learning model outperformed the traditional CAD system at a high sensitivity. AI algorithms can accurately identify malignant lesions and ignore benign lesions by training using a large amount of mammography data. In a previous study with 546 cases (110 cancers), the AUROC for reading with AI was significantly higher than that without AI (0.886 vs. 0.866, respectively) [16]. The sensitivity (83.0%–86.0%) and specificity (77.1%–79.3%) were improved using AI. Another study also demonstrated that AI could improve the AUROC value (0.801–0.881) in a reader study [10]. In this study, the AUROCs of the BSR and GR groups were significantly improved (0.813 vs. 0.884, $p = 0.007$ and 0.684 vs. 0.833, $p < 0.001$, respectively), which is consistent with previous studies.

In terms of specificity, AI was superior to radiologists

https://doi.org/10.3348/kjr.2021.0476

Korean Journal of Radiology
KJR

Table 4. Number of Breast Cancer Detected or Missed by AI and Radiologists

	Detected by Both	Detected by Radiologists Alone	Detected by AI Alone	Missed by Both
Unaided	54	3	25	18
With AI	75	6	4	15

AI = artificial intelligence

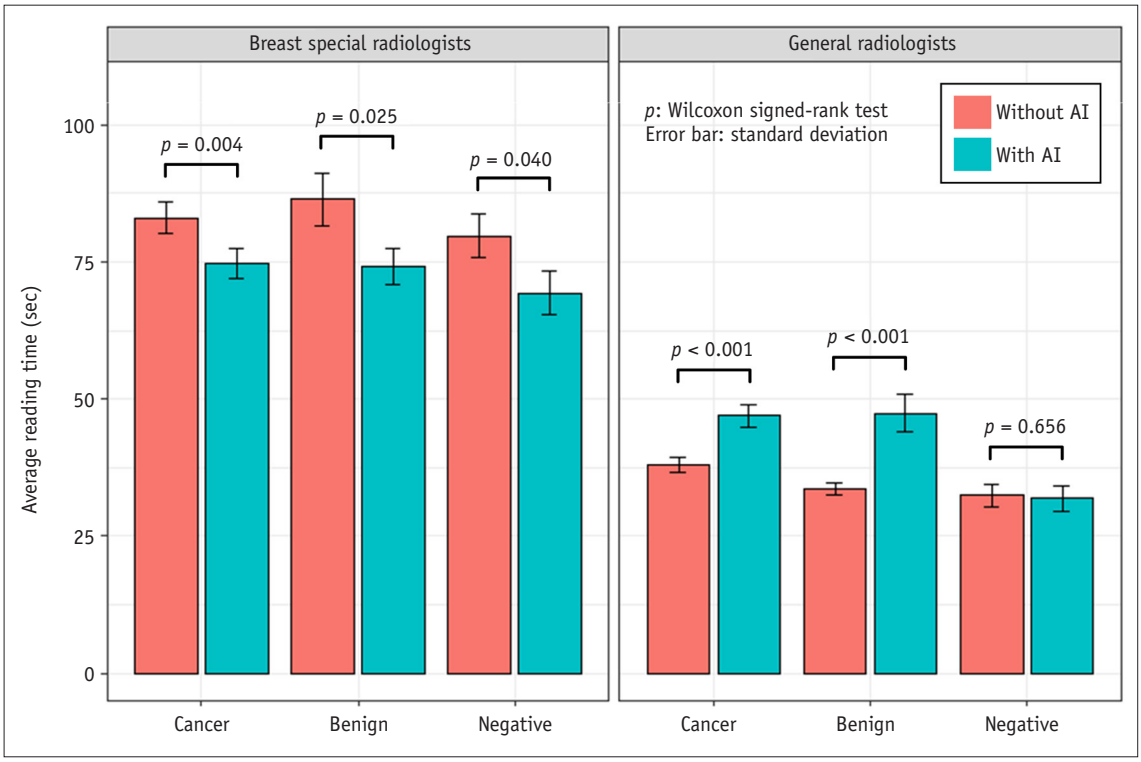

Fig. 4. Reading time pooled across radiologists with and without AI assistance. AI = artificial intelligence

in distinguishing clinically confirmed benign cases from cancer, and AI did not differ from BSRs in biopsy-confirmed benign cases. Although false-positive cases detected by the AI tended to increase the false positives identified by radiologists, the overall specificity was not statistically different because the false positive rate of radiologists also decreased in cases where the AI analysis was negative. This high AI specificity could help improve radiologists' performance without losing specificity in reader studies. In terms of sensitivity, AI can assist radiologists based on the fact that 25 cancers were only detected using AI (Table 4). With AI assistance, the number of cancers detected only using AI was significantly reduced to 4.

Although it has been reported that reading time is significantly lower for digital breast tomosynthesis (DBT) reading with AI assistance [36], there may be a limit to reducing the reading time in 2D mammography because the number of images is significantly smaller than that of DBT. In another study, Rodriguez-Ruiz et al. [18] found that reading time tended to decrease in low-suspicion examinations and increase in high-suspicion examinations but did not differentiate radiologists' experience. However, we found that the use of AI could also reduce the average reading time in the BSR group, probably because AI helps radiologists to detect lesions more quickly. This reduction in reading time is expected to improve radiologists' efficiency in clinical practice. Moreover, the reading time for GRs increased slightly with the use of AI, presumably because the average reading time of GRs was shorter than that of BSRs, and the sensitivity increased with the use of AI. This increase in sensitivity indicates that AI can encourage GRs to focus more on mammograms with suspicious findings. However, negative cases with AI had the shortest reading time regardless of the reader's experience, suggesting that the use of AI may increase the efficiency of breast cancer screening, where the negative is relatively dominant.

This study had some limitations. First, it included a small number of cancer cases, so it was difficult to perform

 https://doi.org/10.3348/kjr.2021.0476

a detailed analysis in terms of mammographic features and cancer stage. Second, there was a high proportion of cancer and biopsy-confirmed benign cases, which does not reflect breast cancer screening. However, in this study, we focused on changes in radiologists' decisions following AI assistance. Third, multiple lesions were not considered; therefore, comparisons at the lesion level were impossible. Fourth, the reader's study environment differs from real-world clinical settings. Furthermore, the readers were probably not fully committed to accepting AI results because of their lack of experience in using AI before participating in this study. This can explain why the AUROC of one reader did not change significantly between that before and after using the AI algorithm and why the specificity of the readers was lower than that of the AI. Further studies, including radiologists with sufficient experience in AI-based software and a larger dataset representing the real-world environment, such as breast screening data or use in a diagnostic setting, are needed.

In summary, this study demonstrated the effect of AI-based software in interpreting mammograms. AI improved cancer detection without sacrificing specificity, regardless of reader experience, and affected reading time.

Supplement

The Supplement is available with this article at https://doi.org/10.3348/kjr.2021.0476.

Availability of Data and Material

The datasets generated or analyzed during the study are available from the corresponding author on reasonable request.

Conflicts of Interest

J.H. Lee, K.H. Kim, and J.S. Ahn are Employees of Lunit. All other authors (E.H. Lee, J.K. Ryu, Y.M. Park, G.W. Shin, Y.J. Kim, H.Y. Choi) declare no competing interests.

Author Contributions

Conceptualization: Eun Hye Lee, Ki Hwan Kim. Data curation: Jung Kyu Ryu, Young Mi Park, Gi Won Shin, Young Joong Kim, Hye Young Choi. Formal analysis: Jeong Hoon Lee. Funding acquisition: Eun Hye Lee, Ki Hwan Kim. Investigation: Jeong Hoon Lee. Methodology: Jeong Hoon Lee. Project administration: Eun Hye Lee, Ki Hwan Kim. Resources: Eun Hye Lee, Ki Hwan Kim. Software: Jeong Hoon Lee, Jung Kyu Ryu, Young Mi Park, Gi Won Shin, Young Joong Kim, Hye Young Choi. Supervision: Eun Hye Lee. Validation: Ki Hwan Kim. Visualization: Jeong Hoon Lee. Writing—original draft: Jeong Hoon Lee. Writing—review & editing: Eun Hye Lee, Ki Hwan Kim, Jeong Hoon Lee, Jong Seok Ahn.

ORCID iDs

Jeong Hoon Lee
https://orcid.org/0000-0002-1789-8270
Ki Hwan Kim
https://orcid.org/0000-0001-7684-235X
Eun Hye Lee
https://orcid.org/0000-0002-8773-700X
Jong Seok Ahn
https://orcid.org/0000-0003-1189-5981
Jung Kyu Ryu
https://orcid.org/0000-0001-8195-0785
Young Mi Park
https://orcid.org/0000-0001-7332-3853
Gi Won Shin
https://orcid.org/0000-0002-6202-1945
Young Joong Kim
https://orcid.org/0000-0002-7084-0289
Hye Young Choi
https://orcid.org/0000-0002-3714-5700

Funding Statement

This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: KMDF_PR_20200901_0300).

Acknowledgments

The English in this document has been checked by at least two professional editors, both native speakers of English. For a certificate, please see: http://www.textcheck.com/certificate/ILXoRO. The scientific guarantor of this publication is the corresponding author, Eun Hye Lee.

REFERENCES

1. DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, et al. Breast cancer statistics, 2019. *CA Cancer J Clin* 2019;69:438-451

2. 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
3. Oeffinger KC, Fontham ET, Etzioni R, Herzig A, Michaelson JS, Shih YC, et al. Breast cancer screening for women at average risk: 2015 guideline update from the American Cancer Society. *JAMA* 2015;314:1599-1614
4. Myers ER, Moorman P, Gierisch JM, Havrilesky LJ, Grimm LJ, Ghate S, et al. Benefits and harms of breast cancer screening: a systematic review. *JAMA* 2015;314:1615-1634
5. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. *Br J Cancer* 2013;108:2205-2240
6. Baker JA, Rosen EL, Lo JY, Gimenez EI, Walsh R, Soo MS. Computer-aided detection (CAD) in screening mammography: sensitivity of commercial CAD systems for detecting architectural distortion. *AJR Am J Roentgenol* 2003;181:1083-1088
7. Eberhard JW, Alyassin AM, Kapur A. Computer aided detection (CAD) for 3D digital mammography. Web site. https://patents.google.com/patent/US7218766B2/en. Accessed January 11, 2021
8. Warren Burhenne LJ, Wood SA, D'Orsi CJ, Feig SA, Kopans DB, O'Shaughnessy KF, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology* 2000;215:554-562
9. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. *Sci Rep* 2018;8:4165
10. Kim HE, Kim HH, Han BK, Kim KH, Han K, Nam H, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health* 2020;2:e138-e148
11. Yoon JH, Kim EK. Deep learning-based artificial intelligence for mammography. *Korean J Radiol* 2021;22:1225-1239
12. Cole EB, Zhang Z, Marques HS, Edward Hendrick R, Yaffe MJ, Pisano ED. Impact of computer-aided detection systems on radiologist accuracy with digital mammography. *AJR Am J Roentgenol* 2014;203:909-916
13. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL; Breast Cancer Surveillance Consortium. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015;175:1828-1837
14. Fenton JJ, Abraham L, Taplin SH, Geller BM, Carney PA, D'Orsi C, et al. Effectiveness of computer-aided detection in community mammography practice. *J Natl Cancer Inst* 2015;103:1152-1161
15. Mahoney MC, Meganathan K. False positive marks on unsuspicious screening mammography with computer-aided detection. *J Digit Imaging* 2015;24:772-777
16. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology* 2019;290:305-314
17. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577:89-94
18. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst* 2019;111:916-922
19. Schaffter T, Buist DSM, Lee CI, Nikulin Y, Ribli D, Guan Y, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Netw Open* 2020;3:e200265
20. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018;286:800-809
21. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA* 2017;318:517-518
22. Obuchowski NA. Sample size tables for receiver operating characteristic studies. *AJR Am J Roentgenol* 2000;175:603-608
23. Sung J, Park S, Lee SM, Bae W, Park B, Jung E, et al. Added value of deep learning-based detection system for multiple major findings on chest radiographs: a randomized crossover study. *Radiology* 2021;299:450-459
24. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;12:77
25. Chen W, Samuelson FW. The average receiver operating characteristic curve in multireader multicase imaging studies. *Br J Radiol* 2014;87:20140016
26. Gallas BD, Hillis SL. Generalized Roe and Metz receiver operating characteristic model: analytic link between simulated decision scores and empirical AUC variances and covariances. *J Med Imaging (Bellingham)* 2014;1:031006
27. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys* 1996;23:1709-1725
28. He X, Frey E. ROC, LROC, FROC, AFROC: an alphabet soup. *J Am Coll Radiol* 2009;6:652-655
29. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. *Stat Med* 2007;26:596-619
30. 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
31. Chakraborty DP. Analysis of location specific observer performance data: validated extensions of the jackknife free-response (JAFROC) method. *Acad Radiol* 2006;13:1187-1193
32. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis, and validation. *Med Phys* 2004;31:2313-2330

 https://doi.org/10.3348/kjr.2021.0476

33. Stock C, Hielscher T. DTComPair: comparison of binary diagnostic tests in a paired study design. R package version 1.0.3. Web site. http://CRAN.R-project.org/package=DTComPair. Published 2014. Accessed July 16, 2020
34. Højsgaard S, Halekoh U, Yan J. The R package geepack for generalized estimating equations. *J Stat Softw* 2006;15:1-11
35. Kooi T, Litjens G, van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, et al. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal* 2017;35:303-312
36. Conant EF, Toledano AY, Periaswamy S, Fotin SV, Go J, Boatsman JE, et al. Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. *Radiol Artif Intell* 2019;1:e180096

https://doi.org/10.3348/kjr.2021.0476