nature communications

Article

https://doi.org/10.1038/s41467-025-57469-3

Artificial intelligence for breast cancer screening in mammography (AI-STREAM): preliminary analysis of a prospective multicenter cohort study

Received: 26 June 2024

Accepted: 22 January 2025

Published online: 06 March 2025

Check for updates

Yun-Woo Chang[1] ✉, Jung Kyu Ryu[2], Jin Kyung An[3], Nami Choi[4], Young Mi Park[5], Kyung Hee Ko[6,7] & Kyunghwa Han[8]

Artificial intelligence (AI) improves the accuracy of mammography screening, but prospective evidence, particularly in a single-read setting, remains limited. This study compares the diagnostic accuracy of breast radiologists with and without AI-based computer-aided detection (AI-CAD) for screening mammograms in a real-world, single-read setting. A prospective multicenter cohort study is conducted within South Korea's national breast cancer screening program for women. The primary outcomes are screen-detected breast cancer within one year, with a focus on cancer detection rates (CDRs) and recall rates (RRs) of radiologists. A total of 24,543 women are included in the final cohort, with 140 (0.57%) screen-detected breast cancers. The CDR is significantly higher by 13.8% for breast radiologists using AI-CAD ($n = 140$ [5.70‰]) compared to those without AI ($n = 123$ [5.01‰]; $p < 0.001$), with no significant difference in RRs ($p = 0.564$). These preliminary results show a significant improvement in CDRs without affecting RRs in a radiologist's standard single-reading setting (ClinicalTrials.gov: NCT05024591).

Worldwide, breast cancer is the most commonly diagnosed cancer in women and the leading cause of mortality among them. Randomized controlled trials, systemic reviews, and observational studies have demonstrated that including mammography in breast cancer screening can reduce breast cancer-related mortality rates by approximately 20–50%[1–3]. Accordingly, several countries have currently implemented screening programs that incorporate mammography for early detection and treatment of breast cancer, aiming to reduce mortality and morbidity. While mammography screening has proven effective in detecting early cancer, it can lead to undesirable false-positive recalls that may require additional imaging evaluation or biopsies. While early detection is the key to screening, breast cancer diagnosis can be alternatively delayed for inaccurate false negative interpretation of mammography; failure to recognize due to low sensitivity in dense breasts, misinterpretation due to lack of recognition of abnormal features, or a wrong interpretation due to no change compared to the previous study, previous biopsy site, or benign appearing lesion characteristics[4–6]. Roughly 20–33% of breast cancer cases are diagnosed based on symptoms between consecutive screening rounds (ie, interval cancers) or discovered

[1]Department of Radiology, Soonchunhyang University Seoul Hospital, Seoul, Korea. [2]Department of Radiology, Kyung Hee University Hospital at Gangdong, Seoul, Korea. [3]Department of Radiology, Nowon Eulgi University Hospital, Seoul, Korea. [4]Department of Radiology, Konkuk University Medical center, Seoul, Korea. [5]Department of Radiology, Inje University Busan Paik Hospital, Busan, Korea. [6]Department of Radiology, CHA Bundang Medical center, Seongnam, Korea. [7]Department of Radiology, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin, Korea. [8]Department of Radiology, Research Institute of Radiological Science and Center for Clinical Imaging Data Science, Severance Hospital, Yonsei University College of Medicine, Seoul, Korea. ✉e-mail: ywchang@schmc.ac.kr

through other imaging modalities[5]. To overcome these shortcomings, significant efforts and improvements have been made by performing double readings with two radiologists, increasing the frequency of screening examinations, or implementing additional supplementary imaging modalities[4,7]. Multiple studies have shown that digital breast tomosynthesis (DBT) combined with digital mammography (DM) has a lower recall rate (RR) and higher cancer detection rate (CDR), leading to lower false-positive rates than digital mammography (DM)[8–10].

As population-based breast cancer screening programs become more widespread, the demand for an increased number of daily breast cancer screening examinations is also progressively rising. However, medical resources remain limited, much short of this demand. Therefore, there is not only an inevitable need for assistance in screening interpretations to reduce interobserver variability but also a need to optimize and streamline current screening workflows[4,11,12]. With substantial promise in this space, artificial intelligence (AI) has been developed and validated with promising applications in mammography screening, which could support computer-assisted detection (CAD) to mark suspicious findings depending on the risk of malignancy. In support, multiple retrospective studies found that AI's diagnostic accuracy was similar to or better than that of breast radiologists (BRs)[9,11,13,14]. However, retrospective studies have inherent limitations, surfacing the need for prospective trials to assess the real-world effectiveness of AI-supported screening. A few prospective studies from Europe, which use a double-reading system, have assessed AI as either an independent reader or as triage tool. There are still knowledge gaps on how AI-CAD can affect radiologist's performance in a single-reading strategy. In this work, we evaluate the diagnostic accuracy of radiologists in interpreting screening mammograms with and without the assistance of AI-CAD within a single-reading strategy. This prospective, multicenter, cohort study, named AI-STREAM (Artificial Intelligence for Breast Cancer Screening in Mammography), demonstrates that the integration of AI-CAD can enhance diagnostic performance, providing evidence to support its potential utility in clinical breast cancer screening.

Results

Participant selection criteria and characteristics

Between February 1, 2021, and December 31, 2022, 25,008 women aged ≥40 years underwent regular mammography screening as part of the national breast cancer screening program in South Korea and were eligible for enrollment into the study. After applying the exclusion criteria of parenchymal change due to previous procedure, mammoplasty, or insertion of foreign substances ($n = 144$), individuals who withdraw consent ($n = 267$), and data errors on the cloud server ($n = 54$), 24,543 participants were included in the final cohort. Among participants who underwent additional assessment at the hospital where recall was conducted, the pathologically diagnosed breast cancer was analyzed 1 year after the last participant's enrollment (ensuring 1-year follow-up for all participants). There were 148 cases of pathologically confirmed cancers, of which there was a total of 140 (0.57%) screening-detected cancers, including 2 cases of bilateral breast cancers. Finally, a total of 24,545 mammograms from 24,543 participants, including 2 cases of bilateral breast cancer, were analyzed in this study (Fig. 1). The median age of the study cohort was 61 years (IQR, 51–68). Of all participants, 67.5% had dense breasts, and 80.7% of diagnosed breast cancer had dense breasts (Table 1).

Performance of BRs with and without AI-CAD

Overall, BR with AI-CAD detected 140 screening-detected cancers and 17 more cancers than BR without AI-CAD, which detected 123 breast cancers (Fig. 2). The PPV1 of BRs with AI-CAD was higher (12.6) compared to BRs without AI-CAD (11.2) ($p < 0.001$). The CDRs of BRs with AI-CAD ($n = 140$, 5.70 ‰ [95% CI 4.76, 6.65]) was significantly higher by 13.8% compared to BRs without AI ($n = 123$, 5.01 ‰ [95% CI 4.13, 5.89]) ($p < 0.001$), while no significant change in RRs was observed between BR with AI-CAD ($n = 1113$, 4.53% [4.27, 4.79]) and BR without AI-CAD ($n = 1100$, 4.48% [4.22, 4.74]) ($p = 0.564$) (Table 2).

In cancer characteristic-specific subgroup analyses, the median tumor size was 16 mm (IQR 11–25). Of the total 140 cases, 46 (32.9%) had ductal cancer in situ (DCIS), and 94 (67.1%) had invasive cancers and among 84 cases examined for lymph node metastases, 70 (83.3%) showed negative lymph node metastasis. For molecular subtypes of

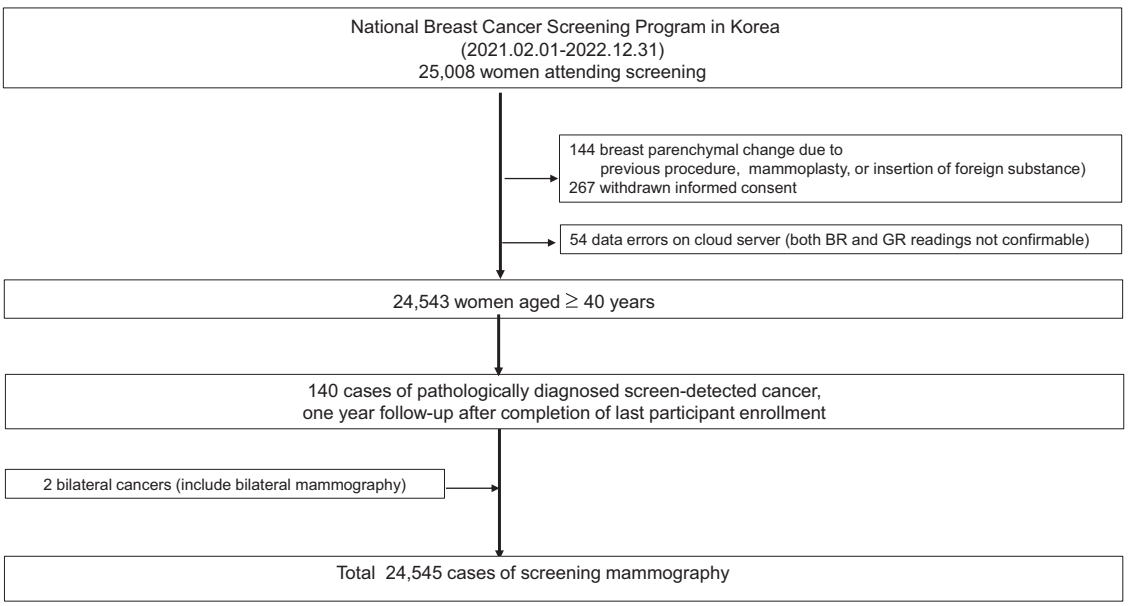


Fig. 1 | Study participant flow chart. Between February 1, 2021, and December 31, 2022, a total of 25,008 women aged over 40 underwent regular mammography screening as part of South Korea's national breast cancer screening program and were enrolled in the study. Among participants who underwent additional assessments at the hospital following a recall, pathologically diagnosed breast cancer cases were analyzed 1 year after the final participant's enrollment. In total, 24,545 mammograms from 24,543 participants, including two cases of bilateral breast cancer, were included in the analysis.

invasive ductal carcinoma (IDC), out of 68 cases, 50 (73.5%) were classified as luminal A, 18 (26.5%) as non-luminal A [8 (11.8%) as luminal B, 4 (5.9%) as HER2 overexpressing, and 6 (8.8 %) as basal]. Interpreting mammograms with AI-CAD detected 6 additional cases of DCIS and 11 additional cases of invasive cancer, leading to notable increase in detection for both DCIS ($p = 0.009$) and invasive cancers ($p < 0.001$).

Table 1 | Study participants' characteristics

	Overall population ($n = 24,543$)	Diagnosed with screen-detected breast cancer[a] ($n = 140$)
Age, median (IQR), years	61 (51–68)	62 (51–69)
Age, *N* (%)		
40–49	4890 (19.9)	30 (21.4)
50–59	6273 (25.6)	28 (20.0)
60–69	8196 (33.4)	50 (35.7)
70 and over	5184 (21.1)	32 (22.9)
Breast density, *N* (%)		
A	1079 (4.4)	2 (1.4)
B	6911 (28.1)	25 (17.9)
C	12,557 (51.2)	96 (68.6)
D	3996 (16.3)	17 (12.1)

IQR interquartile range.
[a]2 bilateral cancer included.

Furthermore, assistance with AI-CAD resulted in a significant increase in the detection of small-sized cancer less than 20 mm ($p = 0.002$), node-negative metastasis ($p = 0.001$), luminal A subtype ($p = 0.002$), and lower grade IDC NOS ($p = 0.009$) when compared without AI-CAD (Table 3).

Performance of GRs with and without AI-CAD

Results of a simulation study (exploratory analysis) found similar trends of diagnostic performance for GRs to that of BRs, but with a greater improvement, as CDRs for GRs with AI-CAD ($n = 120$, 4.89 ‰ [95% CI 4.02, 5.76]) was significantly higher by 26.4% than that of GRs without AI-CAD ($n = 95$, 3.87 ‰ [3.09, 4.65]; $p < 0.001$), resulting in 25 more detected cancers with AI-CAD. However, RRs between GRs with ($n = 1690$, 6.89% [6.57, 7.20]) vs without AI-CAD ($n = 1548$, 6.31%; [6.00, 6.61]) was also significantly increased ($p < 0.001$) (Fig. 3, Table 2). Comparing the characteristics of cancers detected in screening based on GR with and without AI-CAD, similar results as BR with and without AI-CAD (Table 4).

Performance of standalone AI and Radiologists with and without AI-CAD

An AI-CAD abnormal score ≥10 presents as "high", and is regarded as a test-positive result in this program. The areas marked with AI abnormal scores were checked to see if they matched the recalled areas and the areas confirmed as cancer, followed by a lesion-specific analysis. The CDR of AI standalone was 128 (5.21 ‰ [95% CI 4.31, 6.12]), which showed no significant difference vs BRs without AI-CAD ($p = 0.752$) or BRs with AI-CAD ($p = 0.462$). However, AI standalone showed a

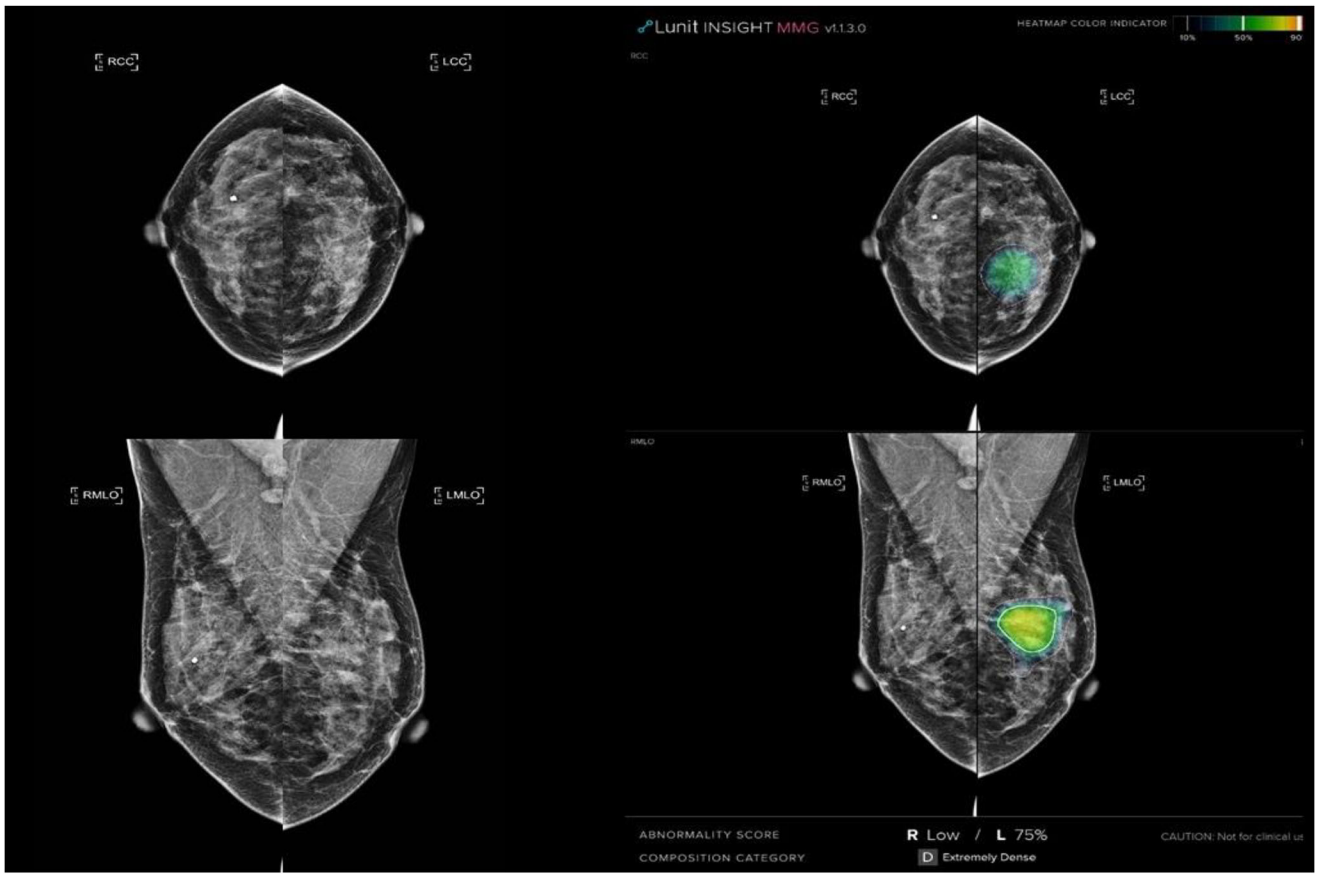


Fig. 2 | Example of cancer detection with the assistance of AI-CAD. In test 1, BR without AI interpreted as malignant scale 2 as non-recall, followed by automatic presentation of AI-CAD result, which marked an abnormal score representing 75%. In test 2, BR with AI-CAD interpreted mammography of focal asymmetry in the left upper inner breast as malignant scale 4 and decided on a recall. Ultrasonography and guided biopsy were performed and pathology revealed DCIS with microinvasion. The patient performed a nipple-sparing mastectomy.

https://doi.org/10.1038/s41467-025-57469-3

Table 2 | Comparison and pairwise test for the cancer detection rate and recall rate

Reading	Cancer number	CDR (‰)	Rate (95% CI)	
BRs without AI	**123**	**5.01**	5.01 (4.13, 5.89)	
BRs with AI	**140**	**5.70**	5.70 (4.76, 6.65)	
AI standalone	**128**	**5.21**	5.21 (4.31, 6.12)	
GRs without AI	**95**	**3.87**	3.87 (3.09, 4.65)	
GRs with AI	**120**	**4.89**	4.89 (4.02, 5.76)	
	Pairwise comparison		Est	***p*-value**
CDR (‰)	BR[a] without AI (5.01) vs BR[a] with AI (5.70)		−0.67 (−1.02, −0.36)	<0.001
	AI standalone (5.21) vs BR[a] without AI (5.01)		0.20 (−1.06, 1.47)	0.752
	AI standalone (5.21) vs BR[a] with AI (5.70)		−0.49 (−1.79, 0.81)	0.462
	GR[b] without AI (3.87) vs GR[b] with AI (4.89)		−1.02 (−1.46, −0.57)	<0.001
	AI standalone (5.21) vs GR[b] without AI (3.87)		1.34 (0.15, 2.53)	0.027
	AI standalone (5.21) vs GR[b] with AI (4.89)		0.33 (−0.93, 1.58)	0.611
Reading	Cancer number	RR (%)	Rate (95% CI)	
BRs without AI	1100	4.48	4.48 (4.22, 4.74)	
BRs with AI	1113	4.53	4.53 (4.27, 4.79)	
AL standalone	1535	6.25	6.25 (5.95, 6.56)	
GRs without AI	1548	6.31	6.31 (6, 6.61)	
GRs with AI	1690	6.89	6.89 (6.57, 7.2)	
	Pairwise comparison		Est	***p*-value**
RR (%)	BR[a] without AI (4.48) vs BR[a] with AI (4.53)		−0.05 (−0.23, 0.13)	0.564
	AI standalone (6.25) vs BR[a] without AI (4.48)		1.77 (1.37, 2.17)	<0.001
	AI standalone (6.25) vs BR[a] with AI (4.53)		1.72 (1.32, 2.12)	<0.001
	GR[b] without AI (6.31) vs GR[b] with AI (6.89)		−0.58 (−0.77, −0.39)	<0.001
	AI standalone (6.25) vs GR[b] without AI (6.31)		−0.05 (−0.48, 0.38)	0.809
	AI standalone (6.25) vs GR[b] with AI (6.89)		−0.63 (−1.07, −0.17)	0.005

All *p*-values were two-sided, and *p*-value < 0.05 indicates statistical significance.
AI artificial intelligence, *BRs* breast radiologists, *CDR* cancer detection rate, *GRs* general radiologists, *RR* recall rate.
[a]Experts in breast imaging with more than >10 years of experience.
[b]Radiologists not specializing in breast imaging.

significantly higher RR ($n = 1535$, 6.25% [5.95, 6.56]) vs BRs with AI-CAD and without AI-CAD (both, $p < 0.001$). When compared to GR without AI-CAD, the CDR of standalone AI was significantly higher than GR without AI-CAD ($p = 0.027$), without affecting RR ($p = 0.809$). The CDR of GR with AI-CAD showed improvement, showing no significant difference compared to AI alone ($p = 0.611$), but the RR significantly increased compared to AI alone ($p = 0.005$) (Fig. 3, Table 2).

Discussion

The preliminary results of the AI-STREAM study, which was a population-based prospective study provide real-world evidence that using AI-CAD for BRs' interpretation of screening mammograms significantly increased CDRs (5.70 per 1000 examinations) compared to BRs not using AI-CAD (5.01 per 1000 examinations). The assistance of AI-CAD that led to improved CDRs did not affect RRs, providing reassurance to radiologists when using AI-CAD in their routine practice in a single-reading setting.

The interpretation process for mammography in breast cancer screening has diverse strategies in each country and thus, is tailored to local needs, practice, and cancer prevalence and/or incidence. Recent advancements in AI have shown substantial promise in reading the results of screening mammography, including in double-reading systems, AI-aided, and AI standalone methods, as indicated by multiple retrospective studies[9,15–17]. While limited to double-read European settings, few prospective studies have been and are being conducted to assess the clinical effectiveness of using AI-CAD[18–20]. There are, however, limitations in directly comparing our study to these studies due to an inherent difference in the screening practice (single- versus double-reading strategy). The population-based, prospective Screen-Trust CAD study found that dual reading by one radiologist plus AI resulted in a 4% increase in screen-detected cancers compared to standard double-reading by radiologists, which was non-inferior in cancer detection. Results favoring the radiologist plus AI arm were also observed for RRs, with RRs reduced by 4% when compared with standard human double-reading during follow-up consensus discussions reviewing mammography. Consensus discussion, using medical history and AI information, has proven effective in preventing an increase in abnormal interpretation during double-reading by AI and radiologists[18]. Another prospective study worth noting, despite differences, is the Mammography Screening with Artificial Intelligence trial (MASAI) study, where the AI score was used to triage mammograms to either single or double-reading strategies; for instance, for low AI scores, independent single-reading was done. The implementation of AI-supported screen reading yielded a 20% increase in cancer detection without a rise in false-positive rates, as observed in the AI interventional triage group compared to the control standard double-reading group. Furthermore, there was a 44% reduction in the workload associated with screening-reading[19].

However, many regions outside of Europe and some private practices in Europe adopt single reading as their many readings strategy, raising the need for prospective evidence on the effect of AI in a real-world, single-read setting. The preliminary analysis of the current AI-STREAM study would be expected to specifically address this knowledge gap by not only utilizing real-world screening data collected prospectively from multiple centers but also increased CDRs and unaffected RRs in BRs with AI-CAD. In the single-reading setting of this AI-STREAM study, the radiologist's decision to recall or not recall a participant was made based on a comprehensive assessment of paired results with and without AI-CAD. Despite utilizing a single-reading strategy in this study, comprehensive decisions, including the comparison of prior mammograms, appear to have contributed to reducing false-positive recalls by BRs while maintaining the strengths of consensus reading from a dual-reading strategy. Additionally, radiologists can compare current mammograms to previous ones, whereas the AI algorithm does not have a temporal analysis feature for previous images. That gives radiologists with AI-CAD interpretation an advantage over compared to standalone AI-CAD.

In a real-world clinical environment, the setting of AI thresholds is an important factor in mammography readings using AI, where this threshold is set and calibrated differently for each individual study[9]. In the AI-STREAM study, specifically, the AI score was considered positive if it was above the predefined cutoff of 10 based on the abnormality score per breast; if the AI score was 10 or higher, the abnormality score and CAD mark were displayed on the screen readout for mammography, where radiologists could fully detect these markings. Thresholds could be varied depending on the purpose where for instance, an AI threshold with high sensitivity should be used if the final decision is made by radiologists[9], whereas higher thresholds for specificity may be essential when AI is used as an independent reader. Regardless, repeated calibration of the AI threshold will likely be necessary in actual clinical use to maintain the desired operating point. Moreover, determining AI thresholds based on retrospective data alone may not

https://doi.org/10.1038/s41467-025-57469-3

Table 3 | Characteristics for screening-detected cancer according to breast radiologist with or without the use of AI

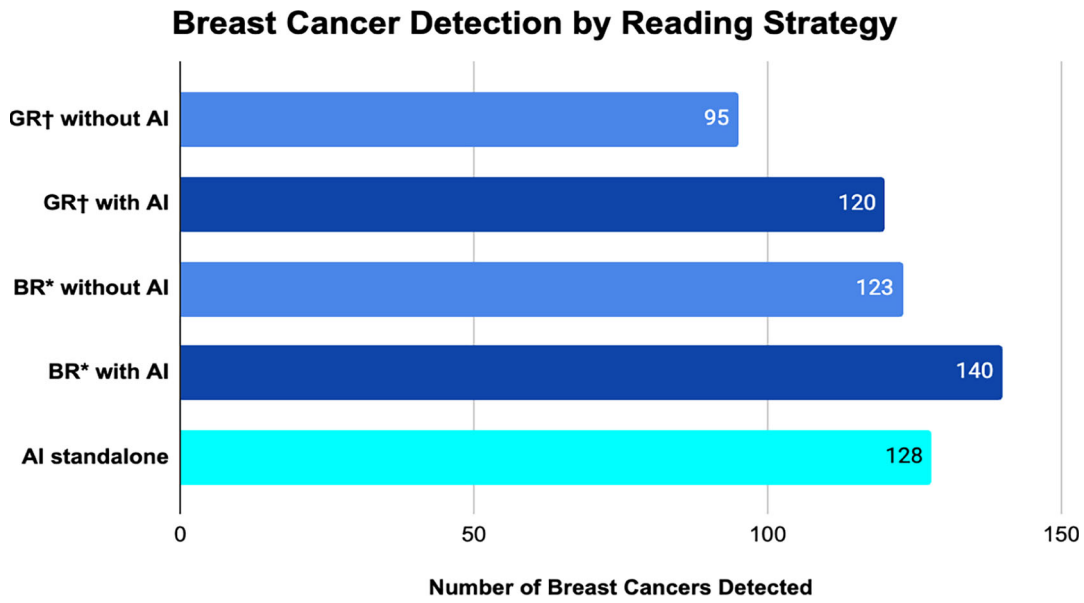
No	BRs with AI (A)	Proportion (95% CI)	BRs (B)	Proportion (95% CI)	AI standalone (C)	Proportion (95% CI)	Difference (95% CI)			*p*-value		
							A vs B	A vs C	B vs C	A vs B	A vs C	B vs C
Cancer, detection	140		123		128							
DCIS	46	97.9 (93.7, 102)	40	85.1 (74.9, 95.3)	41	87.2 (77.7, 96.8)	−12.8 (−22.3, −3.2)	−10.6 (−19.5, −1.8)	2.1 (−11.7, 15.9)	0.009	0.018	0.763
Invasive cancer	94	93.1 (88.1, 98)	83	82.2 (74.7, 89.6)	87	86.1 (79.4, 92.9)	−10.9 (−17, −4.8)	−6.9 (−11.9, −2)	4 (−4.2, 12.2)	<0.001	0.006	0.344
Size[a]	93		81		84							
<20 mm	57	95 (89.5, 100.5)	49	81.7 (71.9, 91.5)	52	86.7 (78.1, 95.3)	−13.3 (−21.9, −4.7)	−8.3 (−15.3, −1.3)	5 (−6.7, 16.7)	0.002	0.02	0.403
≥20 mm	36	90 (80.7, 99.3)	32	80 (67.6, 92.4)	32	80 (67.6, 92.4)	−10 (−19.3, −0.7)	−10 (−19.3, −0.7)	0 (−13.9, 13.9)	0.035	0.035	>0.99
Lymph node[a]	84		74		75							
Negative	70	93.3 (87.7, 99)	60	80 (70.9, 89.1)	61	81.3 (72.5, 90.2)	−13.3 (−21, −5.6)	−12 (−19.4, −4.6)	1.3 (−10.1, 12.7)	0.001	0.001	0.818
Positive	14	93.3 (80.7, 106)	14	93.3 (80.7, 106)	14	93.3 (80.7, 106)	0 (0, 0)	0 (0, 0)	0 (0, 0)	>0.99	>0.99	>0.99
IHC type of IDC[a]	68		60	100 (100, 100)	63							
Luminal A	50	89.3 (81.2, 97.4)	42	75 (63.7, 86.3)	45	80.4 (70, 90.8)	−14.3 (−23.5, −5.1)	−8.9 (−16.4, −1.5)	5.4 (−7.2, 17.9)	0.002	0.019	0.402
Non-luminal A	18	100 (100, 100)	18	100 (100, 100)	18	100 (100, 100)	0 (0, 0)	0 (0, 0)	0 (0, 0)	NA	NA	NA
Subtype of IDC[a]	94		83		87							
IDC NOS	66		58		60							
Grade I	20	87 (74.3, 99.6)	15	65.2 (46, 84.4)	16	69.6 (48, 91.1)	−21.7 (−38, −5.4)	−17.4 (−35, 0.2)	4.3 (−22.4, 31.1)	0.009	0.053	0.75
Grade II	30	88.2 (77.4, 99.1)	27	79.4 (65.8, 93)	29	85.3 (73.4, 97.2)	−8.8 (−18.4, 0.7)	−2.9 (−8.6, 2.7)	5.9 (−5.5, 17.2)	0.07	0.31	0.31
Grade III	16	100 (100, 100)	16	100 (100, 100)	15	93.8 (82.3, 105.2)	0 (0, 0)	−6.2 (−17.7, 5.2)	−6.2 (−17.7, 5.2)	>0.99	0.284	0.284
Unkown grade/ILC/ Special type	18/7/3		16/6/3		18/7/2							

All *p*-values were two-sided, and *p*-value < 0.05 indicates statistical significance.

Non-luminal A: luminal B, HER2 (Human epidermal growth factor receptor-2) enrich, triple negative.

BRs breast radiologists, *AI* artificial intelligence, *DCIS* ductal carcinoma in situ, *IDC* invasive ductal carcinoma, *NOS* not otherwise specified, *IHC* immunohistochemistry.

[a]Numbers for available results.



*Experts in breast imaging with more than >10 years of experience

†Radiologists not specializing in breast imaging

Note: AI, artificial intelligence; CAD, computer-aided detection

Fig. 3 | Number of screen-detected breast cancers in breast or general radiologist with and without AI-CAD, AI standalone, and biopsy-proven true positives. Breast radiologists (BR) using AI-CAD detected 140 screening-detected cancers, which was 17 more than the 123 cancers detected by BRs without AI-CAD. Results from a simulation study (exploratory analysis) showed similar trends in diagnostic performance for general radiologists (GR) compared to BRs, with an even greater improvement. GRs using AI-CAD detected 120 cancers, 25 more than the 95 cancers detected by GRs without AI-CAD. Standalone AI detected 128 cancers.

always be sufficient, raising the need for repeated calibrations in a prospective manner. In support, the ScreenTrust CAD study aimed for a 2% increase in the true positive fraction, which, however, resulted in an actual 4% increase[18]. Currently, the lack of quality assurance protocols to detect and correct data drift, which impacts the performance of AI systems, stands as a major barrier to the practical implementation of AI, Therefore, further evaluation is required regarding AI thresholds to better understand how best to address these unresolved issues.

One prior study of differences in screening mammography interpretation performance by radiologist experience found that RRs for specialist radiologists were significantly lower than that of GRs whereas the biopsy performed CDRs were significantly higher for specialist radiologists[21]. These differences in performance, in which specialist radiologists make more true positive and fewer false-positive interpretations of screening mammography, may be related to increased amounts of initial and continuing education in mammography, as well as accumulated experience. Additionally, consideration may be given to reducing the differences by using multiple rather than single-reading systems[21]. The AI-STREAM study was additionally designed to evaluate the impact of using AI-CAD when differences exist in radiologists' experience in mammography interpretation (eg, BRs and GRs). Similar positive impacts of using AI-CAD on screening mammography interpretation by BRs were also observed with GRs, but to a much larger extent, as CDRs significantly increased by 26.4%. GRs have a chance to see only a few cases of breast cancer in an entire year which may cause limited self-confidence in the interpretation of mammography. This showed a modest increase in RRs, unlike BRs which had no significant differences in RRs. Although comparison with prior mammograms was possible in this simulation analysis similar to a real clinical setting of BR reading, these results are consistent with GRs relying more on AI-CAD results, given their relatively lower self-confidence in interpreting mammography compared to BRs, which seems to have induced increased false-positive RRs. When comparing the results of standalone AI vs GR without AI-CAD, the standalone AI had a significantly higher CDR and no difference in RR, suggesting that AI could surpass the reading capabilities of inexperienced radiologists, or GRs. Based on these results, it is clearly expected that AI-CAD as an aid in mammography interpretation can particularly benefit those with less experience in interpreting mammograms, and produce positive results demonstrating the effect of multiple readings.

Limited studies have explored the use of AI for DBT screening in real clinical settings. In a retrospective analysis of a 12-month period where AI was used in DBT screening examinations at a subspecialized academic breast cancer, the use of AI modestly increased CDR and positive predictive values for screenings with abnormal interpretations without an adverse effect on abnormal interpretation rates or positive predictive values for biopsies performed[22].

Results of standalone AI showed no significant differences in CDRs compared to both BRs with and without AI-CAD, indicating comparable CDRs of standalone AI versus experienced, expert radiologists. Thus, these findings are also in line with previous meta-analysis, which reported that standalone AI in DM is either equivalent to or superior to the interpretation by radiologists[9]. However, RRs from standalone AI were significantly higher compared to those from BRs with and without AI-CAD. This could be owed to the uniformly applied AI-CAD abnormality score thresholds to determine AI's recall or no recall, as well as the fact that AI-CAD could not consider and compare with prior mammograms. Furthermore, the use of standalone AI as a mammography reader, without any human involvement, presents many challenges of current ethical and medicolegal uncertainties.

https://doi.org/10.1038/s41467-025-57469-3

Table 4 | Characteristics for screening-detected cancer according to general radiologist with or without the use of AI

No	GRs with AI (A)	Proportion (95% CI)	GRs (B)	Proportion (95% CI)	AI standalone (C)	Proportion (95% CI)	Difference (95% CI)			*p*-value		
							A vs B	A vs C	B vs C	A vs B	A vs C	B vs C
Cancer, detection	120		95		128							
DCIS	39	83 (72.2, 93.7)	33	70.2 (57.1, 83.3)	41	87.2 (77.7, 96.8)	−12.8 (−22.3, −3.2)	4.3 (−4, 12.5)	17 (4.8, 29.3)	0.009	0.312	0.006
Invasive cancer	81	80.2 (72.4, 88)	62	61.4 (51.9, 70.9)	87	86.1 (79.4, 92.9)	−18.8 (−27.8, −9.8)	5.9 (0.6, 11.3)	24.8 (15.1, 34.4)	<0.001	0.030	<0.001
Size[a]	79		63		84							
<20 mm	46	76.7 (66, 87.4)	37	61.7 (49.4, 74)	52	86.7 (78.1, 95.3)	−15 (−26.1, −3.9)	10 (2.4, 17.6)	25 (13.1, 36.9)	0.008	0.010	<0.001
≥20 mm	33	82.5 (70.7, 94.3)	26	65 (50.2, 79.8)	32	80 (67.6, 92.4)	−17.5 (−31.2, −3.8)	−2.5 (−7.3, 2.3)	15 (0.2, 29.8)	0.012	0.311	0.047
Lymph node[a]	72		56		75							
Negative	58	77.3 (67.9, 86.8)	44	58.7 (47.5, 69.8)	61	81.3 (72.5, 90.2)	−18.7 (−29.6, −7.8)	4 (−1.8, 9.8)	22.7 (11.2, 34.1)	0.001	0.174	<0.001
Positive	14	93.3 (80.7, 100)	12	80 (59.8, 100)	14	93.3 (80.7, 100)	−13.3 (−30.5, 3.9)	0	13.3 (−3.9, 30.5)	0.129	NA	0.129
IHC type of IDC[a]	58		44		63							
Luminal A	40	71.4 (59.6, 83.3)	30	53.6 (40.5, 66.6)	45	80.4 (70, 90.8)	−17.9 (−31.1, −4.7)	8.9 (1.5, 16.4)	26.8 (13.2, 40.3)	0.008	0.019	<0.001
Non-luminal A	18	100 (100, 100)	14	77.8 (58.6, 97)	18	100 (100, 100)	−22.2 (−41.4, −3)	0	22.2 (3, 41.4)	0.023	NA	0.023
Subtype of IDC[a]	94		83		87							
IDC NOS	66		43		60							
Grade I	14	60.9 (40.9, 80.8)	8	34.8 (15.3, 54.2)	16	69.6 (50.8, 88.4)	−26.1 (−50.8, −1.3)	8.7 (−2.8, 20.2)	34.8 (11.9, 57.7)	0.039	0.139	0.003
Grade II	28	82.4 (69.5, 95.2)	23	67.6 (51.9, 83.4)	29	85.3 (73.4, 97.2)	−14.7 (−29.1, −0.3)	2.9 (−2.7. 8.6)	17.6 (2.5, 32.8)	0.046	0.310	0.023
Grade III	15	93.8 (81.9, 100)	12	75 (53.8, 96.2)	15	93.8 (81.9, 100)	−18.8 (−37.9, 0.4)	0	18.8 (−0.4, 37.9)	0.055	>0.99	0.055
Unkown grade/ILC/ Special type	16/6/2		11/6/2		18/7/2							

All *p*-values were two-sided, and *p*-value < 0.05 indicates statistical significance.

Non-luminal A: luminal B, HER2 (Human epidermal growth factor receptor-2) enrich, triple negative.

GRs general radiologists, *AI* artificial intelligence, *DCIS* ductal carcinoma in situ, *IDC* invasive ductal carcinoma, *NOS* not otherwise specified, *IHC* immunohistochemistry.

[a]Numbers for available results.

Hence, the importance of leaving the ultimate clinical decision to the radiologists must be emphasized, not only to meet established medicolegal requirements but also to minimize false-positive results, and future discussions specifically addressing this topic are needed.

With AI assistance, BRs demonstrated an approximate 11 and 13% increase in cancer detection in invasive cancer and DCIS, respectively. Compared with AI assistance, cancer detected by AI assistance increased small-sized (<20 mm), node-negative metastasis, low grade of IDC, and better prognostic luminal A subtype. Although it is a preliminary analysis and the number of cancers is small, this suggests that AI assistance can improve the early detection of breast cancer with relevant prognostic features, with minimal unnecessary recalls. However, a 2-year follow-up is needed to evaluate the true impact of AI-CAD use on interval cancers detected after two years (biennial screening interval in South Korea) and whether there is an increase in interval cancers with poor prognosis. The final results of the AI-STREAM study reflecting these results will be announced after 2026 on the analysis of data to linked to the National Cancer Register.

Strengths of the AI-STREAM study are that, first, as part of a multicenter prospective study conducted on patients participating in national cancer screening, it used various mammography devices (GE and Hologic DM devices). Second, the radiologists who participated in the analysis were proven experts in the interpretation of mammography with many years of experience, and the analysis results were evaluated separately, according to experience. Third, according to the standard procedure of interpreting screening mammography by a single radiologist, to the best of our knowledge, this study is one of the few clinical trials conducted as a prospective multicenter study evaluating the diagnostic accuracy between radiologists with and without AI-CAD. Although only one AI system was used for mammography interpretation, it was verified as the best-performing algorithm compared with others used in previous research[11]. There were several limitations in our study. First, the study was an observational trial although a randomized controlled trial would be ideal for direct comparison between with AI and without AI in screening. The study was performed to evaluate the effect of AI assistance in a single-reading strategy by a single radiologist because it was not easy to design a protocol involving multiple centers as a randomized trial that applied AI in real clinical practice. However, it is hard to assess the direct AI-CAD effect due to which information affected the change in mind for recall of the radiologists in a single-reading strategy. Second, the interim analysis was indeed planned for 2026 upon linkage to and reviewing the National Cancer Registry data. However, this preliminary analysis focused on pathologically proven cases a year after the last participant's enrollment (screen-detected cancers). Although this is a preliminary analysis, all participant's data had been collected and cleaned and the database was locked. Further, despite the short follow-up to evaluate cumulative effectiveness, it does not affect the false-positive results and maintains the pre-planned Statistical Analysis Plan (SAP) (Supplementary Information); we had statistical power when assuming a cancer prevalence rate of 3.21 per 1000 tests, by recruiting and enrolling 24,000 participants, of which we detected over 90 cancers. In this prospective study, the prevalence is higher than the expected national standard. We speculate that the elevated prevalence rate in our study may be due to the screening group being conducted at academic hospitals, and screening mammograms being interpreted by breast imaging specialists. This preliminary analysis evaluated screening-detected cancers including the RR of radiologists with and without AI-CAD, which can be substituted as the result of an interim analysis of AI-STREAM. Third, the results of GR interpretation may not accurately reflect how the GRs would interpret mammography in a clinical setting due to the retrospective simulation nature of the readings, despite the use of prospectively enrolled data. It has been reported that retrospective laboratory experiments may not represent expected performance levels or inter-reader variability during clinical interpretations of the same set of mammograms in a real-world clinical environment[22]. Fourth, this study used DM to interpret the screening effect of AI-CAD. Although DM and DBT are used for routine breast screening, the use of DBT has been rapidly disseminated for breast cancer screening. Further study should assess the performance of AI-CAD in the interpretation of DBT. Fifth, the highest AI score of mammography was included in the analysis, except in only two cases of bilateral cancer, where scores from both sides were included. Therefore, not all instances where the AI score was bilateral-sided in non-cancer cases were evaluated. Lastly, the conclusive results for breast cancer were obtained through additional diagnostic workup following recall after mammography screening, as well as electronic medical and pathology reports from the same hospital where the surgery was conducted. The sample size was relatively small because we could only analyze data for cases with available results for lesion size, nodal metastasis, and molecular subtypes.

In conclusion, given the diverse mammography interpretation procedures across countries worldwide, there is a need to demonstrate the true positive impact of increased CDRs when AI is applied in various ways in real-world clinical environments. The preliminary results from this prospective AI-STREAM study demonstrated positive potential that AI assistance in radiologists' interpretation is indeed beneficial for both BRs and GRs in a single-reading strategy. With the assistance of AI-CAD, BRs improved CDR and increased early cancer detection without affecting RRs in a single-reading strategy.

Methods

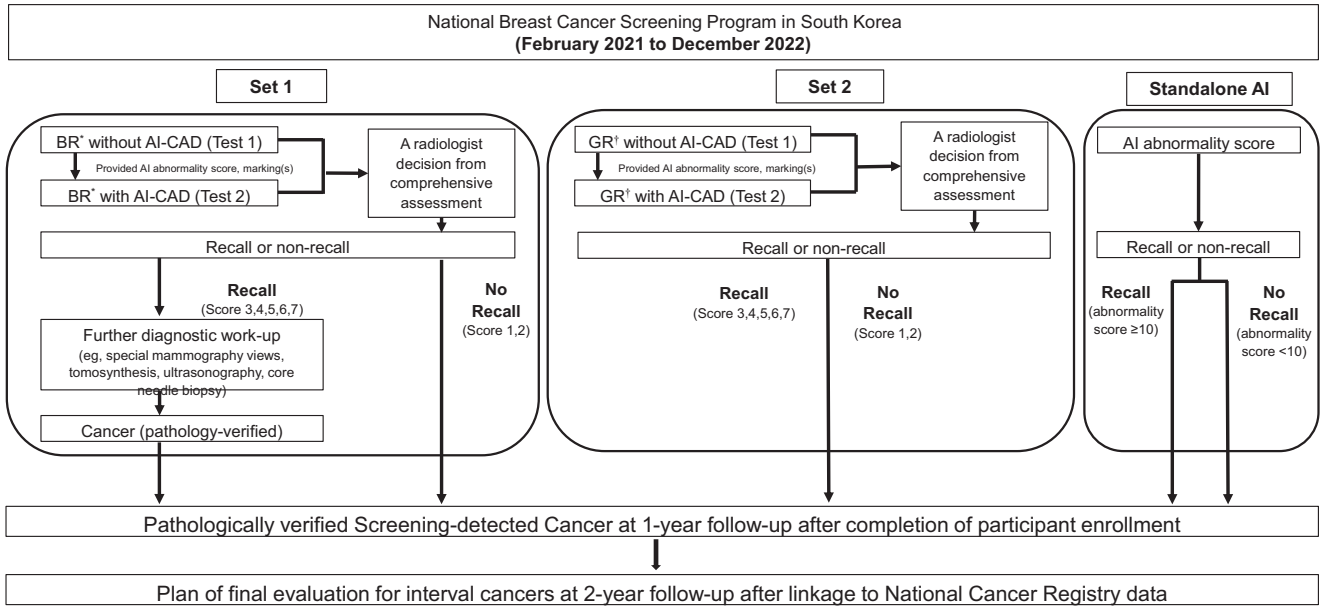
Participants

The AI-STREAM study, described in more detail elsewhere[23], is a prospective, population-based study aimed to compare the diagnostic accuracy of BRs when interpreting screening mammograms with or without AI-CAD for breast cancer (Fig. 4). The trial enrolled women aged ≥40 years from six academic hospitals that participated in national breast cancer screening program in South Korea. All women participating in the study provided their consent by completing an informed consent form about taking part in the study and reading a participant information sheet. Of all eligible women, those with a history of breast cancer or mammoplasty were excluded as the used AI software was not validated for these subgroups. During breast cancer screening, mammography was performed by technologists and interpreted by a single radiologist (ie, single-reading strategy) using two standard craniocaudal and mediolateral oblique views of each breast using DM, which is the standard procedure for interpreting mammograms in South Korea.

Study design

The study received approval from the Institutional Review Board (IRB) of all participating centers (Kyung Hee University Hospital at Gangdong, Soonchunhyang University Seoul Hospital, Konkuk University Medical Center, Noweon Eulji Medical Center, CHA Bundang Medical Center, and Inje University Busan Paik Hospital), and written consent for data publication was obtained from all screening participants. In this study, mammography interpretation was performed by a breast-specialized radiologist at each of the six university hospitals in a real clinical setting (Supplementary Table 1). A radiologist specializing in breast imaging (i.e., BR) was defined as a radiologist with >10 years of experience at a university hospital, having specializing expertise in breast imaging. In the standard single-reading strategy, the subsequent clinical decision about whether the participant requires further diagnostic workup is determined by a single radiologist. Generally, the radiologist performs a comparative reading with a previous mammogram to determine recall, if previous mammography is available. In this study, mammograms were read by radiologists without, and then with AI-CAD. As part of the standard single screening procedure, even when AI was assisted following the radiologist's

https://doi.org/10.1038/s41467-025-57469-3



*Experts in breast imaging with more than >10 years of experience

†Radiologists not specializing in breast imaging

Note: AI, artificial intelligence; BI-RADS, Breast Imaging-Reporting and Data System; BR, breast radiologist; CAD, computer-aided detection; GR, general radiologist.

Fig. 4 | Overview of the study design. Set 1: As part of the standard single screening procedure, mammograms were read by radiologists without AI-CAD, and then read with AI-CAD assistance. A radiologist specializing in breast imaging (i.e., BR) was defined as someone with more than 10 years of experience at a university hospital and expertise in breast imaging. Even when AI-CAD assistance was provided after the radiologist's initial interpretation without AI-CAD, the final decision to recall a patient for further diagnostic evaluation was based on the radiologist's comprehensive decision. Set 2: As a secondary and exploratory objective, a separate simulation study was conducted involving general radiologists (i.e., GR), who did not specialize in breast imaging, to compare their performance with and without AI-CAD. Standalone AI: For exploratory purposes, results from standalone AI-CAD were also evaluated. AI results were considered positive when the abnormality score exceeded a predefined cutoff value of 10. The study compared the diagnostic performance of radiologists, with and without AI-CAD assistance, in identifying screening-detected cancers confirmed pathologically within a 1-year follow-up period.

reading without AI-CAD, the final recall for further diagnostic workup was made based on the radiologist's comprehensive decision (set 1 from Fig. 4).

For exploratory purposes, results from standalone AI-CAD were also collected and compared to the diagnostic performance of radiologists with and without the use of AI-CAD. Moreover, a separate simulation study was conducted (set 2 from Fig. 4) to compare the cancer detection rare (CDR) and recall rate (RR) of general radiologists (i.e., GRs), who did not specialize in breast imaging, with versus without AI-CAD. Note that these results from GRs did not impact any real-world clinical decision-making of study participants for further workup, given that mammograms are those examined from BRs. The main rationale behind this additional study was that GRs comprise the majority in South Korea's breast cancer screening program due to shortage of BRs. Thus, gaining an insight into the potential effect of AI-CAD on GR's performance would be highly reflective of and valuable to real-world screening practice. All radiologists who participated in the study, including both BRs and GRs, had no prior experience with the AI-CAD program, minimizing bias.

Procedures

For image acquisition and data management, a cloud-based imaging data management platform (IRM's BEST Image) was used (Supplementary Fig. 1, Supplementary Note 1). Mammograms were processed by a Snupi program, which performed various operations (eg, search, inquiry, de-identification, separation) and transmission functions on Digital Imaging and Communication in Medicine (DICOM) files. After de-identifying the participant information and assigning identifications, mammograms from participants were exported to the platform to record reading results. If participants had mammograms within the past four years, these were also exported to the study platform for comparison to replicate the same procedure as the real-world reading procedure of each site.

As part of the routine screening process, BRs interpreted the mammography without AI-CAD and recorded the findings (Test 1), followed by the automatic presentation of AI-CAD results (abnormality score and marks) for review and recording. The final records results were based on a comprehensive assessment that considered interpretations from both with and without AI-CAD (Test 2). Results of Test 1 (or without AI-CAD) could not be modified or adjusted once the AI-CAD results were reviewed. Likewise, the results of Test 2 (or with AI-CAD) could also not be corrected after reading (Supplementary Fig. 1).

For variables recorded per 'Test', the radiologist first recorded the breast density according to Breast Imaging Reporting and Data System (BI-RADS) 5th edition (A, B, C, D). Second, the radiologist assessed malignancy using a 7-point scale (1, definitely normal; 2, benign; 3, probably benign [0–2%]; 4, low suspicion for malignancy [2–10%]; 5, moderate suspicion of malignancy [10–50%]; 6, high suspicion for malignancy [50–95%]; 7, highly suggestive of malignancy [≥95%]). For cases not recalled, the radiologist could choose from a malignant assessment score of 1 or 2, whereas for cases recalled, the radiologist could choose a score between 3 and 7. Scores 1 or 2 were considered negative (BI-RADS 1 or 2), while scores of 3 and higher were considered positive (BI-RADS 3 to 5), indicating the need for a recall. In the case a recall decision was made, the location (left, right, both) of the recall was also recorded. The recall for further diagnostic workup of participants was a BR's comprehensive decision informed by the results from considering the paired reading resulting with and without AI-CAD.

If a participant was recalled and visited the same hospital where the screening mammography was performed, additional diagnostic workup (e.g., special mammography views, DBT, ultrasonography) was conducted. If needed, a biopsy was performed, and if a pathologist subsequently diagnosed breast cancer (screen-detected), the participant's information was recorded separately. If surgery was performed, the final pathology was confirmed. Participants diagnosed with breast cancer were reviewed for other breast imaging and pathologic features from electronic medical records and pathology reports (if available).

As a secondary and exploratory objective, a separate reading set (Set 2) was designed and conducted as a simulation study (Fig. 4). In set 2, five general radiologists (GRs) who did not specialize in breast imaging interpreted the same participant's mammography; GRs had variable experience as radiologists and in interpreting mammography. The participant's mammography was interpreted using the same research platform with and without AI-CAD, and the corresponding results were recorded on the same platform. All participating radiologists, including both BRs and GRs, had no prior experience with the AI-CAD program.

The study used a commercial AI-CAD system (Lunit INSIGHT MMG, available at https://insight.lunit.io, version 1.1.7.1), which has been validated through various studies[11,24] (Supplementary Note 3). In brief, the AI system improves radiologists' performance and has diagnostic performance equivalent to or superior to those of radiologists alone[24]. It also has shown superior performance compared with two other commercial AI-based software products[11]. The AI system provides abnormal scores ranging from 0 to 100, per breast based on mammograms. These scores can also be presented as a heatmap or grayscale map. AI results were considered positive if the abnormality score was above a predefined cutoff of 10. The highest lesion abnormality score was reflected, and examinations with scores of 10 or higher were considered positive.

Outcomes

The primary outcomes were CDRs and RRs of BRs with and without AI-CAD in mammography reading for screen-detected breast cancer, including invasive or ductal in situ (or both). The secondary outcome was to compare CDRs and RRs of mammography reading in the following comparisons: (1) BRs without AI-CAD vs AI standalone, (2) BRs with AI-CAD vs AI standalone, (3) GRs without AI-CAD vs GRs with AI-CAD, (4) GRs without AI-CAD vs AI standalone, and (5) GRs with AI-CAD and AI standalone. PPV1 was defined as the percentage of all positive screening exams with a pathologic cancer diagnosis within 1 year, and PPV1 of screening-detected cancer by BR with or without AI was obtained.

Statistical analysis

The sample size was estimated using McNemar's test to detect differences in CDRs between groups of radiologists with and without AI-CAD, with a two-sided test at a significance level of 0.05 and 80% power. The assumed cancer prevalence was 3.21 per 1000 examinations, determined from data in a previous retrospective study, and the target sample size was chosen based on this expected cancer prevalence[25]. The target sample size was 32,714 participants, corresponding to approximately 16,000 participants per year. The total number of expected participants was, however, adjusted from the initial study design due to the COVID-19 pandemic, but no effect on the primary study endpoint was observed. Assuming the same cancer prevalence as 3.21 per 1000 examinations, it was calculated that the sample size could be maintained if approximately 24,000 people were recruited while still maintaining 80% power and detecting more than 90 cases of cancers (Supplementary Note 4).

All statistical analyses, including CDRs and RR, were taken place based on a 1-year follow-up from all the study participants. All statistical analyses were performed using R version 4.3.3. (R Foundation for Statistical Computing, Vienna, Austria). Descriptive statistics were used for continuous and categorical variables, as appropriate. Logistic regression analysis using a generalized estimating equation to account for reader variability was used to estimate 95% CI and for comparative analysis. Pairwise comparisons were performed to compare BRs with AI-CAD, BRs without AI-CAD, AI standalone, GRs with AI-CAD, and GRs without AI-CAD. In addition, corrections based on multiple comparisons are necessary for confirmation, but no corrections were made considering the preliminary nature of the study.

Prespecified subgroup analyses were performed to examine results in different age groups (40–49, 50–59, 60–69, 70+ years), mammographic density (four categories of BI-RADS by the American College of Radiology), and malignant scale assessment using a 7-point scale [defined above]. In breast cancer, the following subgroups were analyzed for cancer characteristics including invasiveness, categories of tumor size (<20 mm, ≥20 mm), presence of axillary lymph node metastasis, and molecular subtypes (luminal A, non-luminal A [luminal B, human epidermal growth factor receptor-2 [HER2] enriched, or triple negative]) (Supplementary Note 5).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Individual patient data will be shared to the extent that anonymity can be maintained, that the recipient has ethical approval to conduct the research and with a data transfer agreement. A request to obtain study data can be discussed with the committee comprising researchers associated with the study hospital, to ensure compliance with General Data Protection Regulations and other legal agreements. A request to obtain study data can be discussed with the committee comprising researchers associated with the study hospital, to ensure compliance with General Data Protection Regulations and other legal agreements. The Figshare DOI is https://doi.org/10.6084/m9.figshare.26104606.

References

1. Myers, E. R. et al. Benefits and harms of breast cancer screening: a systematic review. *JAMA* **314**, 1615–1634 (2015).
2. Independent UK Panel on Breast cancer screening. The benefits and harms of breast cancer screening: an independent review. *Lancet* **380**, 1778–1786 (2012).
3. Gøtzsche, P. C. & Jørgensen, K. J. Screening for breast cancer with mammography. *Cochrane Database Syst. Rev.* **2013**, Cd001877 (2013).
4. Yoon, J. H. & Kim, E. K. Deep learning-based artificial intelligence for mammography. *Korean J. Radiol.* **22**, 1225–1239 (2021).
5. Hovda, T., Tsuruda, K., Hoff, S. R., Sahlberg, K. K. & Hofvind, S. Radiological review of prior screening mammograms of screen-detected breast cancer. *Eur. Radiol.* **31**, 2568–2579 (2021).
6. Lamb, L. R., Mohallem Fonseca, M., Verma, R. & Seely, J. M. Missed breast cancer: effects of subconscious bias and lesion characteristics. *Radiographics* **40**, 941–960 (2020).
7. Taylor-Phillips, S. & Stinton, C. Double reading in breast cancer screening: considerations for policy-making. *Br. J. Radiol.* **93**, 20190610 (2020).
8. Lowry, K. P. et al. Screening performance of digital breast tomosynthesis vs digital mammography in community practice by patient age, screening round, and breast density. *JAMA Netw. Open* **3**, e2011792 (2020).
9. Yoon, J. H. et al. Standalone AI for breast cancer detection at screening digital mammography and digital breast tomosynthesis: a systematic review and meta-analysis. *Radiology* **307**, e222639 (2023).

10. Sprague, B. L. et al. Digital breast tomosynthesis versus digital mammography screening performance on successive screening rounds from the Breast Cancer Surveillance Consortium. *Radiology* **307**, e223142 (2023).
11. Salim, M. et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. *JAMA Oncol.* **6**, 1581–1588 (2020).
12. Lee, C. S. et al. Radiologist characteristics associated with interpretive performance of screening mammography: a National Mammography Database (NMD) study. *Radiology* **300**, 518–528 (2021).
13. Rodriguez-Ruiz, A. et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J. Natl. Cancer Inst.* **111**, 916–922 (2019).
14. Freeman, K. et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *BMJ* **374**, n1872 (2021).
15. Hickman, S. E. et al. Machine learning for workflow applications in screening mammography: systematic review and meta-analysis. *Radiology* **302**, 88–104 (2022).
16. Larsen, M. et al. Artificial intelligence evaluation of 122 969 mammography examinations from a population-based screening program. *Radiology* **303**, 502–511 (2022).
17. Romero-Martín, S. et al. Stand-alone use of artificial intelligence for digital mammography and digital breast tomosynthesis screening: a retrospective evaluation. *Radiology* **302**, 535–542 (2022).
18. Dembrower, K., Crippa, A., Colón, E., Eklund, M. & Strand, F. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. *Lancet Digit. Health* **5**, e703–e711 (2023).
19. Lång, K. 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**, 936–944 (2023).
20. Ng, A. Y. et al. Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. *Nat. Med.* **29**, 3044–3049 (2023).
21. Sickles, E. A., Wolverton, D. E. & Dee, K. E. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology* **224**, 861–869 (2002).
22. Letter, H. et al. Use of artificial intelligence for digital breast tomosynthesis screening: a preliminary real-world experience. *J. Breast Imaging* **5**, 258–266 (2023).
23. Chang, Y. W. et al. Artificial intelligence for breast cancer screening in mammography (AI-STREAM): a prospective multicenter study design in Korea using AI-based CADe/x. *J. Breast Cancer* **25**, 57–68 (2022).
24. Kim, H. E. et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit. Health* **2**, e138–e148 (2020).
25. Hong, S. et al. Effect of digital mammography for breast cancer screening: a comparative study of more than 8 million Korean women. *Radiology* **294**, 247–255 (2020).

Acknowledgements

This study received a grant from the Korea Health Industry Development Institute with its third Korea Medical Device Development fund in 2020. We thank the trial participants, trial support nurses at each hospital, radiologists at the simulation mammography reading (K.W.R., T.H.N., J.Y.L., D.Y.Y.), and Lunit for their support. We would like to express special thanks to Dr. Ki Hwan Kim for management, information, and organizational contributions and to Dr. Han Eol Jeong for research support.

Author contributions

Y.-W.C. and K.H. conceptualized the design of the trial with input from Y.-W.C. K.H. did the statistical analysis. Y.-W.C., J.K.A., N.C., K.H.K., Y.M.P., and J.K.R. directly assessed and verified the underlying data reported in the manuscript. Y.-W.C. and K.H. interpreted the results of the validation study. Y.-W.C. wrote the first draft of the report with input from K.H. All authors subsequently edited the report. J.K.R. and Y.-W.C. supervised the project. All authors approved the final version of the manuscript and had final responsibility for the decision to submit it for publication.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-57469-3.

Correspondence and requests for materials should be addressed to Yun-Woo Chang.

Peer review information *Nature Communications* thanks Robert Nishikawa and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.