J Korean Soc Breast Screening 2023;20:19-29

Original Article

One-on-One Comparison Between Conventional CAD and AI-CAD Applied to Screening Mammography

Si Eun Lee[1], Jung Hyun Yoon[2], Hanpyo Hong[1], Nak-Hoon Son[3], Eun-Kyung Kim[1]

[1]*Department of Radiology, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin, Gyeonggi-do, Korea*
[2]*Department of Radiology, Severance Hospital, Yonsei University College of Medicine, Seoul, Korea*
[3]*Department of Statistics, Keimyung University, Daegu, Korea*

Background: Artificial intelligence-based computer assisted diagnosis (AI-CAD) has been shown to improve the diagnostic performance of breast cancer diagnosis on mammography. We evaluated the diagnostic performances of AI-CAD and the conventional CAD through a one-on-one comparison.

Materials and Methods: From January to December 2017, of 997 women who visited a health examination center to undergo screening mammography, 978 had normal or benign results with stable follow-up for two years and 19 had cancer diagnosed within the two years of follow-up. Conventional CAD was applied when performing mammography and AI-CAD was retrospectively applied. We compared the diagnostic performances of the two CADs used and did a case-level comparison for immediately and delayed diagnosed cancers.

Results: Standalone AI-CAD presented significantly higher specificity (92.7% vs. 48.7%, P <0.001), PPV (14.5%, 2.3%, P <0.001), and accuracy (92.2% vs. 48.9%, P <0.001) than conventional CAD. For 978 women without breast cancers, conventional CAD presented at least one mark for 502 women (51.3%), which was significantly higher than AI-CAD for 71 women (7.2%). AI-CAD correctly localized three cancers among eight delayed diagnosed cancers, and conventional CAD detected two of them.

Conclusion: AI-CAD showed better diagnostic performance than conventional CAD, by lowering the number of false-positive results with higher specificity.

Index words: Digital Mammography; Breast cancer; Diagnosis, Computer-Assisted; Artificial Intelligence

Correspondence to: Eun-Kyung Kim, MD, PhD
Department of Radiology, Yongin Severance Hospital
Yonsei University College of Medicine
363, Dongbaekjukjeon-daero, Giheung-gu,
Yongin-si, Gyeonggi-do 16995, Korea
Tel: 82-31-5189-8321, Fax: 82-2-2227-8337
e-mail:ekkim@yuhs.ac

Introduction

Mammography is the standard modality for screening breast cancer with proven survival benefits since its introduction 30 years ago (1-

3). To overcome variability in sensitivity, which is probably due to the intrinsic limitation of screening mammography being a 2-dimentional projection and reader dependent, conventional computer-aided detection/diagnosis (CAD) was developed and widely applied to reduce the frequency of lesions overlooked by the interpreting radiologist(4). Conventional CAD has been met with positive reviews in initial studies as cancer detection rates have increased with its use (5-7). However, in subsequent large-scale studies, conventional CAD did not significantly improve diagnostic performance, although it lowered specificity and positive predictive values (8-10). A recent study found sensitivity to be lower in the subset of radiologists using CAD (11). Still, a major pitfall of conventional CAD is the high rate of false-positive markings which can overload the interpreting radiologist and cause unnecessary additional exams to be ordered for patients (12).

The primary distinguishing feature of conventional CAD is that it uses hand-crafted features suggested by professional radiologists to identify suspicious lesions, while recently developed artificial intelligence-based CAD (AI-CAD) is trained with image features that are achieved by the algorithm itself based on a large amount of digital mammography data (4, 13). In retrospective studies, several AI-CAD algorithms have improved the diagnostic performance of radiologists without increasing the number of false-positive recalls (14-16). However, there have not been enough studies that directly compare the performances of conventional CAD and AI-CAD for related findings to be conclusive (17, 18).

In this study, we evaluated the stand-alone diagnostic performance of AI-CAD and conventional CAD through a one-on-one comparison in a screening cohort. We also classified the diagnosis of breast cancer according to detection time and compared the results of both CADs.

Materials and Methods

This retrospective study was approved by the institutional review board (IRB) of yongin severance hospital, with a waiver for informed consent.

Study population

From January to December 2017, 6,575 women visited a health examination center that was an affiliated center of a tertiary institution to screen for breast cancer. Among these women, we enrolled 19 patients who were diagnosed with breast cancer within 2 years since their most recent screening examination (the cancer group). For the normal/benign group, we enrolled 994 women who visited the health examination center between January and June 2017 and who were followed for at least 2 years. Next, we excluded 10 patients without available CAD results, and 6 patients with a prior history of breast cancer. Therefore, 964 patients with stable follow-up results for at least 2 years, and 14 patients with benign pathology after 14-gauge core needle biopsy or vacuum-assisted biopsy were included. For the cancer group, 11 patients who were diagnosed with breast cancer immediately after initial mammography in 2017, and 8 patients who were diagnosed with breast cancer later (delayed diagnosis) within 2 years were included.

Finally, a total of 997 women were included in our study; 19 cancer patients and 978 women with normal breasts or breasts with benign disease (Fig. 1).

Mammography acquisition and interpretation

Screening mammography was obtained with the bilateral mediolateral oblique (MLO) and craniocaudal (CC) view using two dedicated digital mammography units (Lorad Selenia, Hologic Inc., Danbury, CT, USA). Two radiologists (1 with fellowship training in breast imaging, 1 general radiologist) each with 9 and 3 years of experience

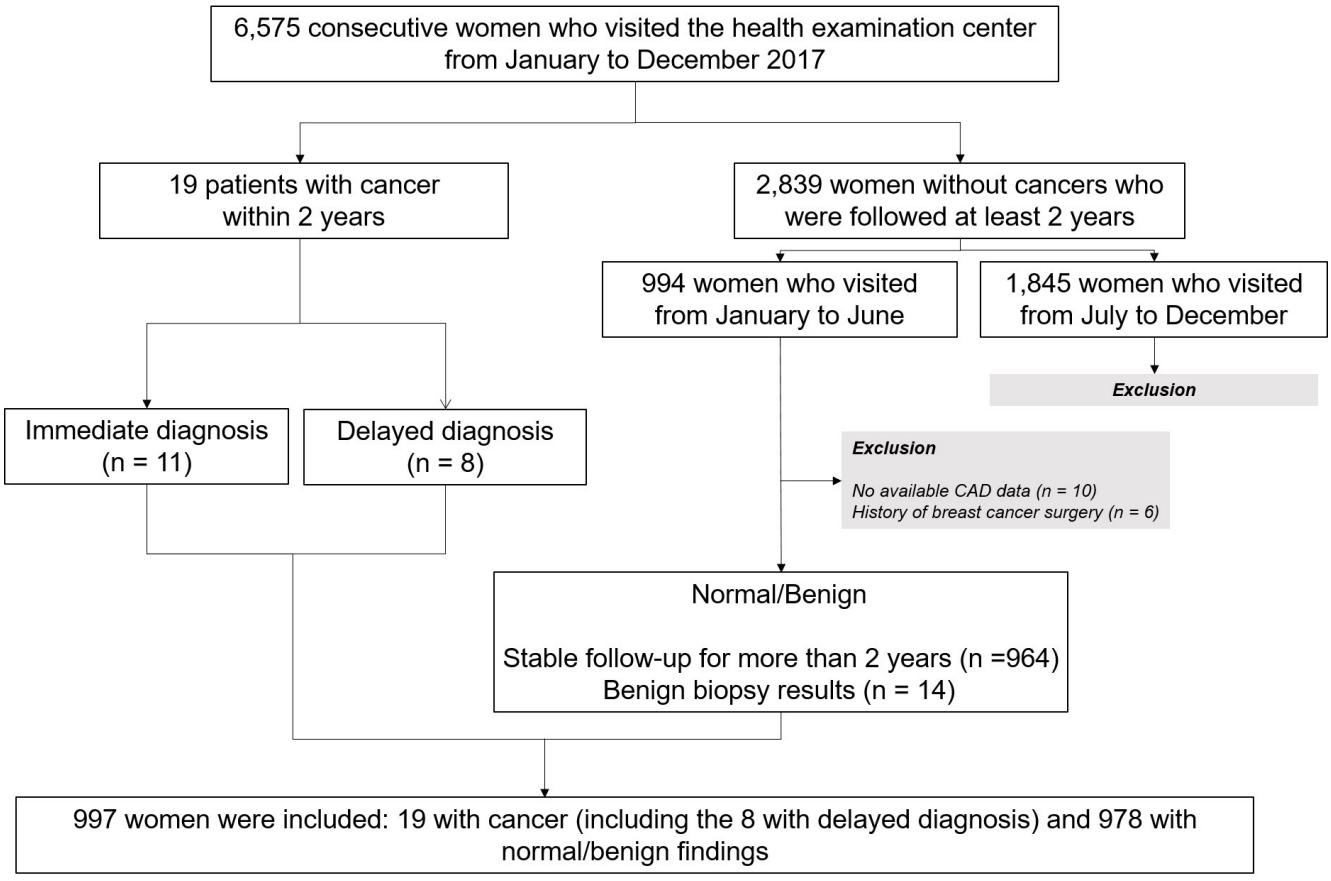


Fig. 1. Flow diagram for the study population.

in breast imaging independently interpreted the screening mammograms based on the American College of Radiology Breast Imaging-Reporting And Data System (ACR BI-RADS) (19). Radiologists could refer to the conventional CAD results on demand when interpreting mammography. Breast density was assessed according to a four-grade system; grade A: almost entirely fat, grade B: scattered areas of fibroglandular density, grade C: heterogeneously dense, grade D: extremely dense breast.

CAD systems for mammography

After the mammography examination, conventional CAD (R2, version 8.3 and version 9.4, Hologic, California, USA) was automatically applied through post-processing. Two different versions of conventional CADs were embedded in each mammography unit respectively. Conventional CAD presented three kinds of markers according to imaging findings on each mammographm; ▲ indicating microcalcifications, * indicating mass/asymmetry, and + indicating combined features.

We retrospectively applied a AI-based diagnostic support software dedicated to breast cancer detection on digital mammography (Lunit INSIGHT for Mammography, version 1.1.0.1, Lunit Inc., Seoul, Korea) (14), so AI-CAD was not used during the initial mammographic interpretations. AI-CAD provided circular marks on all suspicious findings with an abnormality score of 10 or higher, and a representative maximum score per breast. The continuous abnormality score ranging between 0 to 100% represents the level of suspicion for breast cancer being detected on that specific image.

Statistical analysis

Ground truth was dichotomized into 'breast cancer' and 'normal/benign' based on histopathologic diagnosis via biopsy/surgery and stable follow-up mammography for more than 2 years. Another dedicated breast radiologist (S.E.L., 4 years of experience) who did not take part in the original interpretation retrospectively reviewed the serial mammographic images of breast cancer patients with the AI-CAD and conventional CAD results to correlate the known cancer site and sites selected by the two CADs. Those images with any mark were considered to have 'positive' CAD findings. When the markings by the two CADs correctly localized the cancer, it was regraded as a true-positive finding.

We also counted the number of markings drawn by the two CAD systems. An abnormality score of 10% was used as the cutoff threshold for AI-CAD (13), and between abnormality scores calculated for the right and left breast, the higher score was chosen to represent each patient. For the false-positive analysis, we counted the number of AI-CAD and CAD marks per each mammographic view, and per patient among women without breast cancers. Since two different versions of conventional CAD were used in our study, we compared the number of false-positive marks between them using the Mann-Whitney U test.

We compared sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and accuracy between AI-CAD and CAD using the generalized estimated equation. All analyses were conducted using SAS statistical software (version 9.2, SAS Inc., Cary, NC, USA). $P<0.05$ was considered to have statistical significance.

Result

A total of 997 Asian women were included (mean age, 52 ± 10 years). More than 80% of women had dense breasts. Original interpretation and final pathology are shown in Table 1.

Comparison between the two different versions of conventional CAD

Among 997 women, 484 (48.5%) were analyzed by the prior version of conventional CAD (Imagechecker, version 8.3, R2 technology) and 513 (51.5%) by the updated version (version 9.4). There was no statistically significant difference in the number of false-positive markings made between the two versions, however, the prior version of conventional CAD showed better specificity and accuracy: specificity 52.7% (250/474) vs. 44.8% (226/504), P=0.014; accuracy 53.1% (257/484) vs. 45.0% (231/504), P=0.01. Sensitivity, PPV and NPV were not statistically different between the two conventional CADs.

Comparison of diagnostic performance between AI-CAD and CAD

In 997 women consisting of 19 breast cancer

Table 1. Demographics for the study population

	Numbers
No. women	997
Mean age (years)	52 ± 10 years
Mammography density	
Entire fatty	13 (1.3%)
Scattered	157 (15.7%)
Heterogeneous	486 (48.7%)
Extreme dense	341 (34.2%)
Initial mammography interpretation	
BI-RADS 1, 2	960 (96.3%)
BI-RADS 3	5 (0.5%)
BI-RADS 0	32 (7.7%)
Final pathology	
Benign with follow-up	964 (96.7%)
Benign with biopsy	14 (1.4%)
DCIS	7 (0.7%)
Invasive cancer	12 (1.2%)

patients and 978 women without breast cancers, AI-CAD presented significantly higher specificity (92.7% vs. 48.7%, P <0.001), PPV (14.5%, 2.3%, P <0.001), and accuracy (92.2% vs. 48.9%, P <0.001) than conventional CAD.

In 978 women without breast cancers, conventional CAD presented at least one mark on 502 women (51.3%) with 1.12 marks per patient on average (Table 3). On the other hand, AI-CAD presented marks on images for only 71 women (7.2%) with 0.11 marks per patient on average, a number significantly lower than that for conventional CAD. In conventional CAD, the mean number of markers for microcalcifications (▲), mass/asymmetry (*) or their combination (+) per woman was 0.52, 0.58 and 0.03, respectively. Seventy-one women with false-positive results on AI-CAD presented a mean abnormality score of 28 (range 10-95). Fig. 2 shows the most contrasting case between the two CADs, with 15 markers on conventional CAD but no markings on AI-CAD.

Table 2. Diagnostic performance of AI-CAD and conventional CAD (CAD) for 997 patients

	CAD (%, CI)	AI-CAD (%, CI)	P-value
Sensitivity	63.2 (12/19, 38.4-83.7)	63.2 (12/19, 38.4-83.7)	1.0
Specificity	48.7 (476/978, 45.5-51.9)	92.7 (907/978, 90.9-94.3)	<0.001
PPV	2.3 (12/514, 1.7-3.3)	14.5 (12/83, 10.1-20.3)	<0.001
NPV	98.6 (476/483, 97.4-99.2)	99.2 (907/914, 98.6-99.6)	0.286
Accuracy	48.9 (488/997, 45.8-52.1)	92.2 (919/997, 90.3-93.8)	<0.001

CI = confidence interval, PPV = positive predictive value, NPV = negative predictive value

Table 3. Number of false-positive markings per patient in 978 women without breast cancers compared between AI-CAD and conventional CAD

No. markers	CAD (No. women)	AI-CAD (No. women)
0	476 (48.7%)	907 (92.7%)
1	243 (24.8%)	44 (4.5%)
2	125 (12.8%)	17 (1.7%)
More than 3	134 (13.7%)	10 (0.1%)
Average (per woman)	1.12	0.11

Breast cancer cases

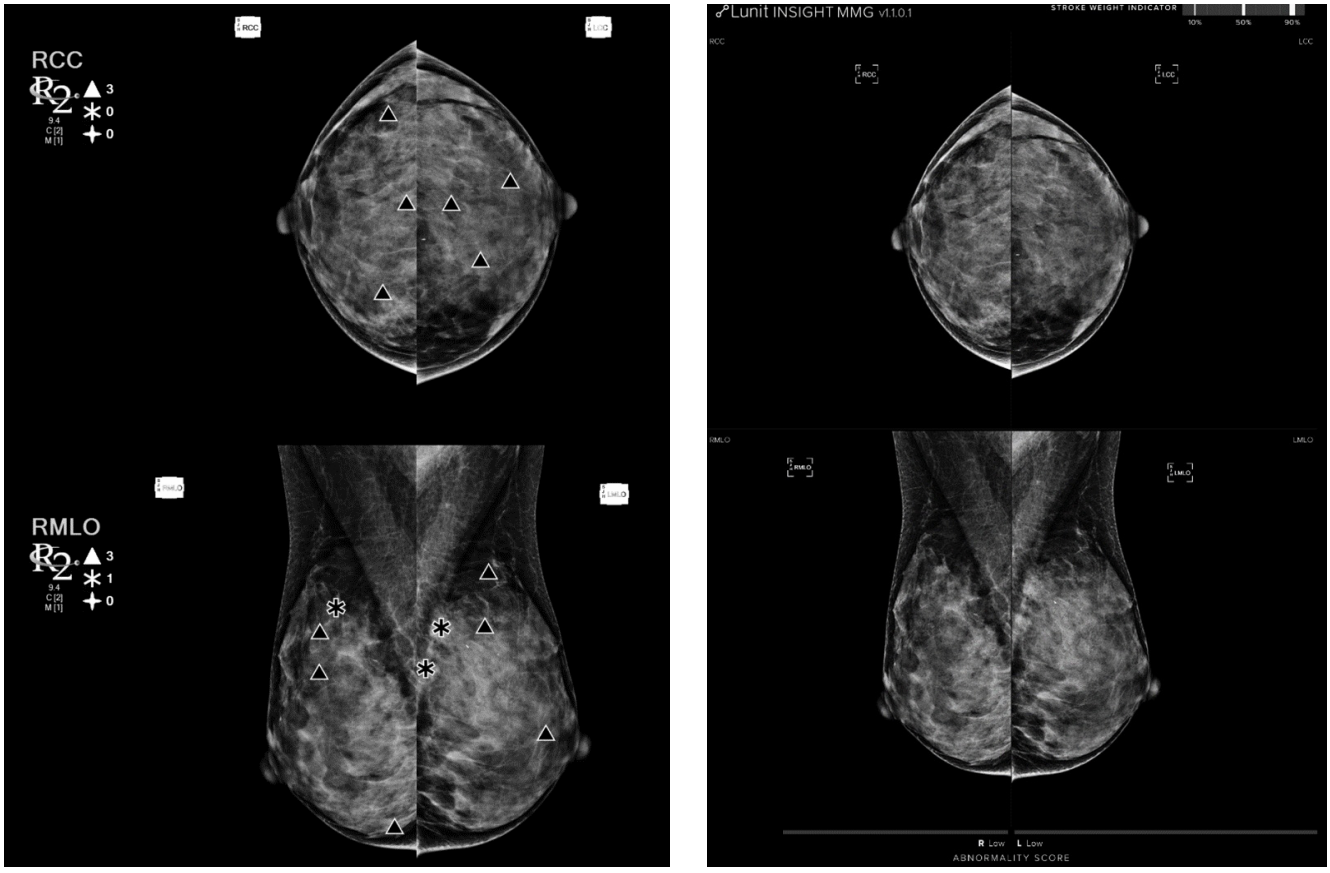
A total of 19 cases of breast cancer were diagnosed at initial visit and during 2 years of follow-up (Table 4). Among 11 immediately diagnosed cancers, AI-CAD recognized 9 with abnormality scores ranging from 27 to 99 o with correct localization. Two cancers missed by AI-CAD were diagnosed as BI-RADS 2 on initial interpretation and diagnosed through concurrent screening US examinations. Conventional CAD detected 10 of the 11 cancers, with one more cancer being detected than AI-CAD (Fig. 3, 4).

Among 8 patients who were diagnosed with delayed breast cancers, five were diagnosed with breast cancer before their regular screening (Table 5). One patient had a lesion first assessed as BI-RADS 3 due to microcalcifications that was then diagnosed as ductal carcinoma *in situ* at the 6-month follow-up. Both AI-CAD and conventional CAD marked

Table 4. Results of AI-CAD and conventional CAD (CAD) for the 11 cancers that were immediately diagnosed by radiologists

Case/Age	MG BI-RADS	AI-CAD (score)	CAD
1/45	Category 0	Yes (40)	Yes
2/58	Category 0	Yes (27)	Yes
3/47	Category 0	Yes (97)	Yes
4/61	Category 0	Yes (99)	Yes
5/44	Category 0	Yes (83)	Yes
6/48	Category 0	Yes (98)	Yes
7/52	Category 0	Yes (96)	Yes
8/57	Category 0	Yes (91)	Yes
9/44	Category 0	Yes (97)	Yes
10/74*	Category 2	No (1)	No
11/47*	Category 2	No (0)	Yes

*Detected by concurrent screening ultrasound
MG = mammography, BI-RADS = Breast Imaging-Reporting and Data System, AI-CAD = artificial-intelligence computer-assisted diagnosis, C-CAD = conventional-computer assisted diagnosis



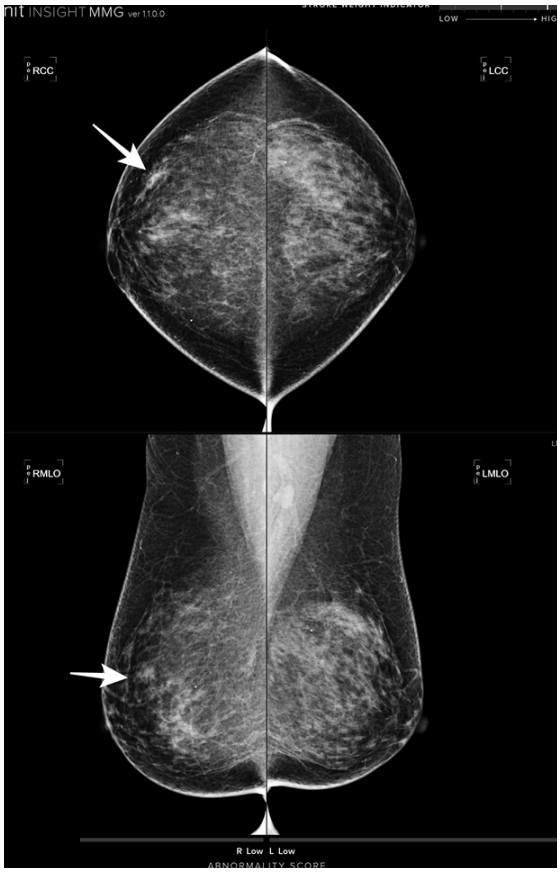
a **b**

Fig. 2. A 40-year-old woman underwent screening mammography. A radiologist diagnosed the lesion as BI-RADS 2 due to benign microcalcifications on mammography, and no lesion newly developed on follow-up screening mammography and ultrasound after 2 years. (**a**) Conventional CAD drew a total of 15 markers on mammography. (**b**) AI-CAD did not indicate anything specific on this mammography.

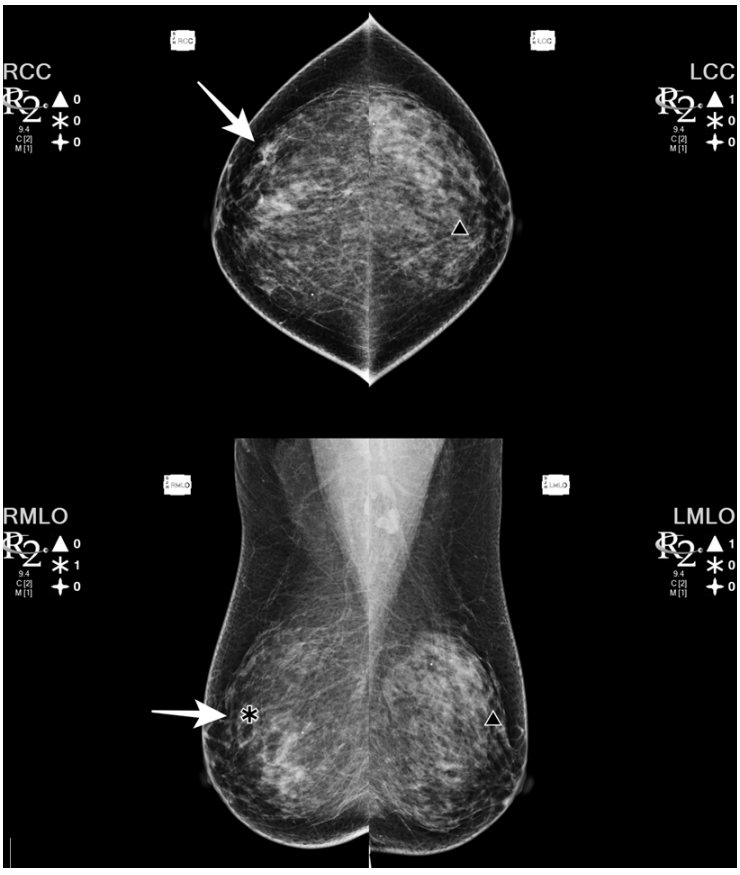
Table 5. Population demographics and results of AI-CAD and conventional CAD (CAD) for the 8 cancers that were delayed diagnosed after initial mammography

Cases/Age	Initial MG BI-RADS	Initial US BI-RADS	AI-CAD (score)	CAD	Diagnosis interval (month)	Detection modality
Interval cancers						
1/43	Category 3	Category 3	Yes (31)	Yes	5	MG and US
2/51	Category 2	Not done	Yes (72)	Yes	9	MG and US
3/37	Category 2	Not done	Yes (47)	No	15	MG and US
4/50	Category 2	Not done	No (3)	No	8	US
5/46	Category 2	Not done	No (1)	No	11	Microdochectomy
Cancers detected at next screening round						
1/48	Category 0	Category 3	No (2)	No	16	MG
2/51	Category 1	Category 2	No (0)	No	23	MG and US
3/45	Category 1	Category 2	No (0)	No	18	US

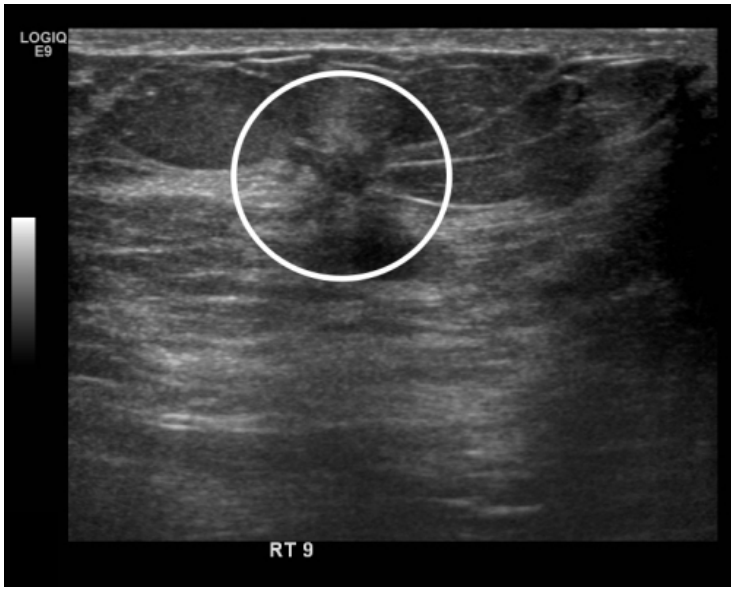
MG = mammography, US = ultrasound, BI-RADS = Breast Imaging-Reporting and Data System, AI-CAD = artificial-intelligence computer-assisted diagnosis, C-CAD = conventional-computer assisted diagnosis



a



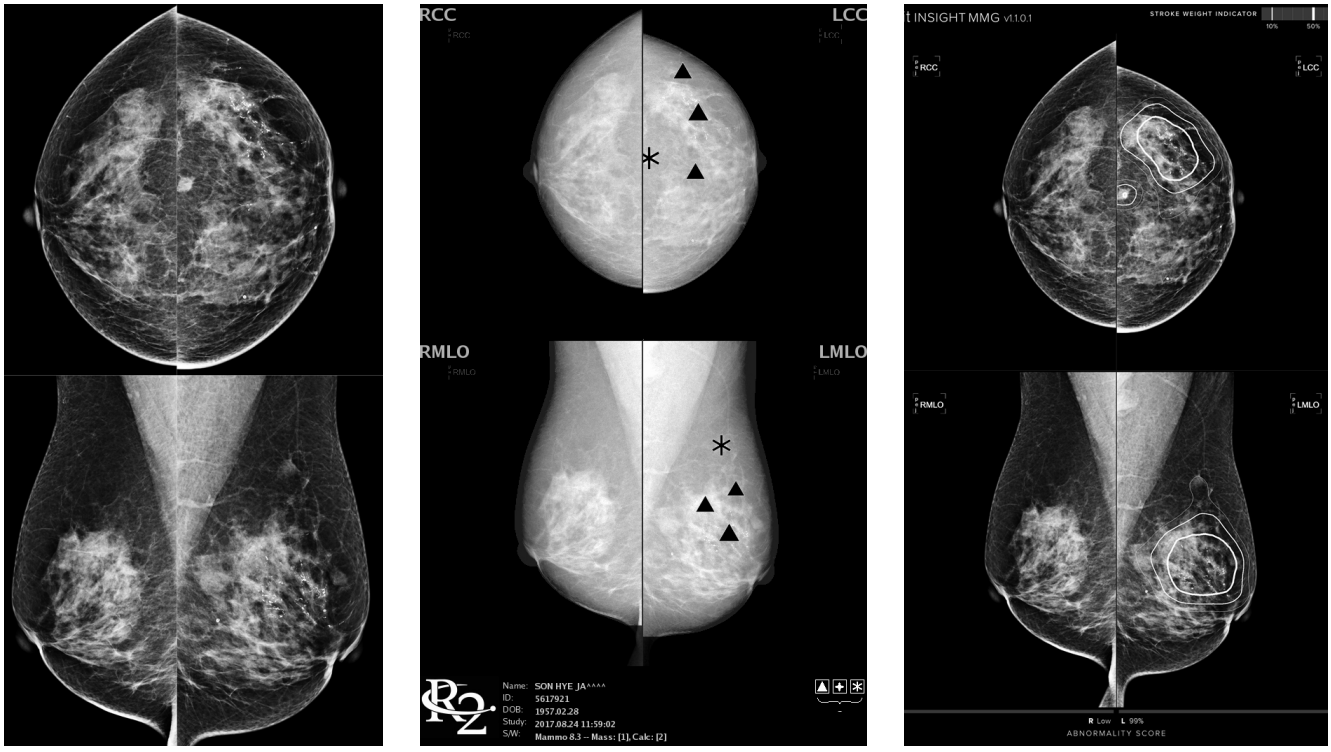
b



c

d

Fig. 3. A 47-year-old patient diagnosed with invasive ductal carcinoma in the right breast. Her initial mammography interpretation was BI-RADS 2, but was diagnosed by concurrent screening ultrasound. (**a**) AI-CAD did not recognize the cancer site (arrow). (**b**) Conventional CAD correctly marked the cancer at the RMLO view. (**c-d**) The known malignancy on the right breast at 9 o'clock on US and MRI.



a **b** **c**

Fig. 4. A 61-year-old patient diagnosed with invasive ductal carcinoma in her left breast, which presented as suspicious mass with microcalcifications in her left upper center to outer breast (**a**). Both conventional CAD (**b**) and AI-CAD (**c**) indicated a similar extent of suspicious findings.

the delayed diagnosed site on initial mammography, and the abnormality score of AI-CAD was 31. Four patients revisited the hospital with new symptoms, and AI-CAD drew markers for two of them on their initial mammography at the site of the developed cancer. One of them also had markings on the same site drawn by conventional CAD, and this patient was regarded as a case of missed breast cancer on retrospective review.

The remaining 3 patients were diagnosed with breast cancer on the next screening mammography or ultrasound examination performed within 2 years of the initial screening examination. In the retrospective review done by radiologists, all of the cancers were occult on initial mammography. None of them were marked by AI-CAD, but conventional CAD had false-positive marks for 2 of the 3 patients.

Discussion

The most unfortunate drawback of conventional CAD for screening mammography is the exhausting number of false-positive markings which curtails its ability to improve cancer detection. The number of false-positive results decreased with AI-CAD compared to conventional CAD was beyond expectations in our study, with sensitivity maintained.

Conventional CAD gave marks for more than half of the women without cancers, while AI-CAD surprisingly presented positive results on 7% of them, which was a percentage even less than the benchmark of recall rates for screening mammography in BI-RADS. This result suggests that AI-CAD will be a time saver for radiologists by reducing misleading results (12). Conventional CAD is trained by hand-crafted features which

are suggested by professional radiologists, and can subsequently be biased by the way a human professional interprets images. However, AI-CAD self-trains with correct answers and has shown superior diagnostic performance in a previous large-scale study (18). Also, a prior study found findings similar to ours, by reporting a 69% reduction in false-positive markings when another AI-CAD was compared to another conventional CAD (17).

To evaluate the role of CADs in interval cancer detection or cancer prediction, we included patients who were diagnosed with breast cancer 24 months after undergoing screening mammography. This inevitably decreased the sensitivity of both CADs, but we found that three of the five interval cancers could be detected on screening mammography. Conventional CAD also detected two interval cancers, which had been ignored at initial interpretation. The two CADs showed a similar level of sensitivity, and detected more cancers than radiologists.

In addition, we compared two different versions of conventional CADs. Unfortunately, the updated version of conventional CAD did not show improvement. One of the advantages of AI-CAD is that additional training can make it much smarter than conventional CAD with more qualified big data being collected and used to train the AI-CAD system. Thus, we anticipate better performances for later versions of AI-CAD as future updates to the system are expected to continuously improve its overall performance.

There were some limitations to our study. First, this was a retrospective study performed in a single screening center. Second, since AI-CAD can present a single area with multifocal abnormalities, simply counting the number of markings is disadvantageous for conventional CAD. However, the fact that a much higher portion of women without cancers had negative AI-CAD results is still meaningful. Third, we used one commercial AI-CAD and two conventional CAD programs, and our results cannot be directly generalized to other platforms without further comparative studies. Lastly, we included normal/benign and cancer groups from different time periods due to the very small number of cancers, and the small quantity of data being collected could have affected the assessment of diagnostic performances, especially for sensitivity.

In conclusion, AI-CAD showed better diagnostic performance than conventional CAD, especially for lowering false-positive results with higher specificity.

References

1. Tabar L, Vitak B, Chen TH, et al. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. Radiology 2011;260:658-663
2. Paci E, Broeders M, Hofvind S, Puliti D, Duffy SW, Group EW. European breast cancer service screening outcomes: a first balance sheet of the benefits and harms. Cancer Epidemiology and Prevention Biomarkers 2014;23:1159-1163
3. 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
4. Sechopoulos I, Teuwen J, Mann R. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. Semin Cancer Biol 2020
5. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. Radiology 2001;220:781-786
6. Birdwell RL, Bandodkar P, Ikeda DM. Computer-aided detection with screening mammography in a university hospital setting. Radiology 2005;236:451-457
7. Morton MJ, Whaley DH, Brandt KR, Amrami KK. Screening mammograms: interpretation with computer-aided detection—prospective evaluation. Radiology 2006;239:375-383
8. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. New England Journal of Medicine

2007;356:1399-1409
9. Gur D, Sumkin JH, Rockette HE, et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. Journal of the National Cancer Institute 2004;96:185-190
10. Hall FM. Breast imaging and computer-aided detection. Mass Medical Soc, 2007
11. Lehman CD, Wellman RD, Buist DS, et al. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med 2015;175:1828-1837
12. Tchou PM, Haygood TM, Atkinson EN, et al. Interpretation Time of Computer-aided Detection at Screening Mammography. Radiology 2010;257:40-46
13. Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks 2015;61:85-117
14. Kim H-E, Kim HH, Han B-K, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. The Lancet Digital Health 2020;2:e138-e148
15. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577:89-94
16. Schaffter T, Buist DSM, Lee CI, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. JAMA Netw Open 2020;3:e200265
17. Mayo RC, Kent D, Sen LC, Kapoor M, Leung JW, Watanabe AT. Reduction of false-positive markings on mammograms: a retrospective comparison study using an artificial intelligence-based CAD. Journal of Digital Imaging 2019;32:618-624
18. Kooi T, Litjens G, Van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis 2017;35:303-312
19. D'Orsi CJ SE, Mendelson EB, Morris EA. ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. Reston, VA: American College of Radiology, 2013

대한유방검진의학회지 2023;20:19-29

검진 유방촬영술에 적용한 전통적 진단보조프로그램과 인공지능 진단보조프로그램의 일대일 비교

이시은[1] · 윤정현[2] · 홍한표[1] · 손낙훈[3] · 김은경[1]

[1]연세대학교 의과대학 용인세브란스병원 영상의학과
[2]연세대학교 의과대학 세브란스병원 영상의학과
[3]계명대학교 통계학과

배경: 인공지능 진단보조프로그램은 유방촬영술에서 유방암 진단의 정확도를 향상시키는 것으로 보고되고 있다. 우리는 동일 환자 군에 전통적 진단보조프로그램을 함께 적용하여 두 프로그램 간의 진단 성능을 비교해보고자 하였다.

방법: 2017년 한해동안 건강검진을 위해 유방촬영술을 시행한 997명 환자를 후향적으로 분석하였다. 이 중 978명은 2년 이후 유방촬영술에서 정상 혹은 양성 소견을 보였고, 19명은 2년 내에 유방암을 진단받았다. 전통적 진단보조프로그램은 유방촬영과 동시에 PACS에 결과가 분석되었으며, 인공지능 진단보조프로그램은 후향적으로 적용하여 결과값을 얻었다. 우리는 두 프로그램의 진단 성적을 비교하고, 즉시 진단된 암과 2년 내 지연 진단된 암에 대해 증례별 분석하였다.

결과: 인공지능 진단보조프로그램은 유의하게 높은 특이도 (92.7% vs. 48.7%, P <0.001), 양성예측도 (14.5%, 2.3%, P <0.001), 정확도 (92.2% vs. 48.9%, P <0.001)를 보였다. 978명의 정상 혹은 양성 질환 환자들에서 전통적 진단보조프로그램은 51.3%, 502명의 여성에서 표지를 나타낸 반면, 인공지능 진단보조프로그램은 7.2%, 71명의 환자에서 표지를 나타냈다. 인공지능 진단보조프로그램은 8개의 지연진단 암 중 3개를 미리 검출할 수 있었고, 전통적 진단보조프로그램은 2개를 미리 검출했다.

결론: 인공지능 진단보조프로그램은 위양성 결과를 줄여 특이도를 높임으로써 전통적 진단 보조프로그램에 비해 더 좋은 진단 성능을 보였다.

Index words: Digital Mammography; Breast cancer; Diagnosis, Computer-Assisted; Artificial Intelligence

Corresponding author: Eun-Kyung Kim, M.D., Ph.D.