Original Article

Comparison of Diagnostic Performances among Different Interpretation Schemes for Screening Mammography: A simulation study

Miribi Rho[1], Hye Sun Lee[2], Hee Jung Suh[3], Eun-Kyung Kim[1,4], Si Eun Lee[1,4], Jung Hyun Yoon[1]

[1]*Department of Radiology, Severance Hospital, Research Institute of Radiological Science, Center for Clinical Imaging Data Science, Yonsei University, College of Medicine*
[2]*Biostatistics Collaboration Unit, Yonsei University College of Medicine*
[3]*Department of Radiology, Severance Check-up Center*
[4]*Department of Radiology, Yongin Severance Hospital, Yonsei University, College of Medicine*

Purpose: To compare the diagnostic outcomes of different interpretation schemes simulated for interpreting screening mammography, adding AI-CAD vs. a second human reader to a single human reader, using a consecutive, screening study sample.

Materials and Methods: Between January 2018 and January 2019, 2,385 digital mammograms of 2,385 consecutive women (mean age: 50.0 ± 9.5 years) were included. As single reading is routine in our practice, interpretation reports were used as data for single reading. To simulate double reading, a second reader independently reviewed the screening mammograms with access to the interpretation reports. To simulate single reading interpretation with AI-CAD, one of the first readers re-evaluated the mammography images with positive AI-CAD results. Ground truth in terms of cancer/benign or absence of abnormality was confirmed according to histopathologic diagnosis or at least 1 year of follow-up.

Results: Among the 2385 mammograms, 6 (0.3%) were cancers, 32 (1.3%) were biopsy-confirmed benign, and 2347 (98.4%) were negative examinations. Reader 1+AI-CAD had significantly higher recall rates compared to reader 1, 2.6% (95% confidence interval [95% CI]: 2.0-3.3) vs. 2.4% (95% CI: 1.7-3.0) (p=0.008), respectively, that was lower than reader 1+2, 3.1% (95% CI: 2.4-3.8) (P=0.010). Specificity and accuracy were significantly higher in reader 1 compared to both reader 1+2 and reader 1+AI-CAD (all p<0.05, respectively). Reader 1+AI-CAD had significantly higher specificity (97.6% vs. 97.1%) and accuracy (97.5% vs. 97.0%) compared to reader 1+2 (p=0.010), respectively. High proportion of false-positive findings detected by AI-CAD were distortions, while calcifications were mostly the cause for false-positive findings detected by the readers.

Conclusion: Adding readers, either AI-CAD or human second readers, results in higher recalls with significantly lower specificity and accuracy compared to a single human reader. When comparing the effect of adding AI-CAD vs. human second reader, AI-CAD had significantly lower recall and higher specificity and accuracy compared to the scheme of two human readers.

Index words: Mammography; Breast neoplasms; Artificial intelligence; Computer-Assisted detection/diagnosis

Correspondence to: Jung Hyun Yoon, MD, PhD
Department of Radiology, Severance Hospital, Research Institute of Radiological Science, Yonsei University, College of Medicine 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea
Tel: 82-2- 2228-7400, Fax: 82-2-2227-8337
e-mail:lvjenny@yuhs.ac

Introduction

Breast cancer is the most common cause of death for women worldwide (1). In order to detect breast cancers at a treatable stage, breast cancer screening using mammography has been implemented in many countries, resulting in approximately 20% reduction in breast cancer-related mortality (2). Although the use of mammography in breast cancer screening has proven effective, several drawbacks of mammography have also been recognized: 1) false-positive recalls where the reader recalls abnormalities that are eventually proven as benign and 2) false-negative interpretation where the cancer is either overlooked by the reader or interpreted erroneously. These limitations are the focus of our investigation to improve the efficiency of screening mammography by reducing the economic and emotional burdens of screening for patients while at the same time detecting cancers before they become untreatable.

To enhance the outcomes of screening mammography, various screening protocols have been applied in different countries according to the availability of medical resources, such as increasing screening frequencies, using supplementary imaging modalities, or adding readers. For instance, the European Guidelines recommend 'double reading', i.e., screening mammograms being interpreted independently by two readers to reduce recalls and improve diagnostic sensitivity (3), while many other countries including the United States accommodate 'single reading' with or without the aid of computers (4). Both interpretation strategies have their pros and cons. Double reading requires a considerable amount of medical resources, while single reading leaves patients at risk for increased recalls or missed cancers. Computer-assisted detection/diagnosis (CAD) algorithms have emerged as a possible solution for providing interpretive assistance or even as a substitute reader, with these roles strengthening with continuous developments in artificial intelligence (AI) technology. Studies have shown that AI-CAD for mammography improves interpretive performance when used by the radiologists (5, 6). In this aspect, little has been evaluated on the strengths and weaknesses among different interpretation schemes, and especially when using AI-CAD.

In this background, we compared the diagnostic outcomes of different interpretation schemes simulated for interpreting screening mammography, adding AI-CAD vs. a second human reader to a single human reader, using a consecutive, screening study sample.

Materials & Methods

This retrospective study was approved by the institutional review board (IRB) of Severance Hospital, Seoul, South Korea, with a waiver for informed consent.

Study sample

Between January 2018 and January 2019, 2,635 consecutive women underwent mammograms for screening purposes at a single screening facility. We excluded women who had a personal history of breast cancer and surgery (n=41), augmented mammoplasty (n=34), clip insertion after biopsy (n=4), interstitial or autologous fat injection (n=3), and pacemaker implantation (n=1). Also, we excluded women who were younger than 35 years old (n=115), were not followed (n=40), and did not have bilateral mammograms available (n=12). Finally, 2,385 bilateral, four-view digital mammograms of 2,385 women who underwent breast cancer screening were included in this study. The mean age of the 2,385 women was 50.0 ± 9.5 years (range, 35-85 years). The mean follow-up interval after the mammography examination was 13.4 ± 3.1 months (range, 0-27.2 months).

Mammography acquisition and interpretation

Screening mammography was obtained with the bilateral routine mediolateral oblique (MLO) and craniocaudal (CC) views using a dedicated digital mammography unit (Lorad Selenia, Hologic Inc., Danbury, CT, USA). Two radiologists (1 with fellowship training in breast imaging and 1 general radiologist) with 9 and 3 years of experience in breast imaging, respectively individually interpreted the screening mammograms using the American College of Radiology Breast Imaging Reporting And Data System (ACR BI-RADS) final assessments (7). Single reading is currently routine protocol for mammography interpretation in Korea, and medical records were retrospectively reviewed for the interpretation results which were used as the analytic data for the first reader, ′reader 1′ (Fig. 1). For breast parenchymal density, a four-grade system was used; grade A: almost entirely fat, grade B: scattered areas of fibroglandular density, grade C: heterogeneously dense, and grade D: extremely dense breast (7).

AI-based Clinical Decision Support Software for Mammography

AI-CAD dedicated to breast cancer detection on digital mammography (Lunit INSIGHT for Mammography, version 1.1.0.0, Lunit Inc., Seoul, Korea) was used for image processing. Deep convolutional neural networks (CNNs) were used to develop this software, and it was trained and validated with over 170,000 mammography examinations and tested with a separate external mammography dataset (7, 8). The AI-CAD software provides per-breast malignancy scores with four-view region-of-interests, ′AI-CAD marks′, for suspicious lesions on each input mammogram. Along with the AI-CAD marks, the software provides a continuous abnormality score ranging between 0 - 100% (100% meaning high likelihood

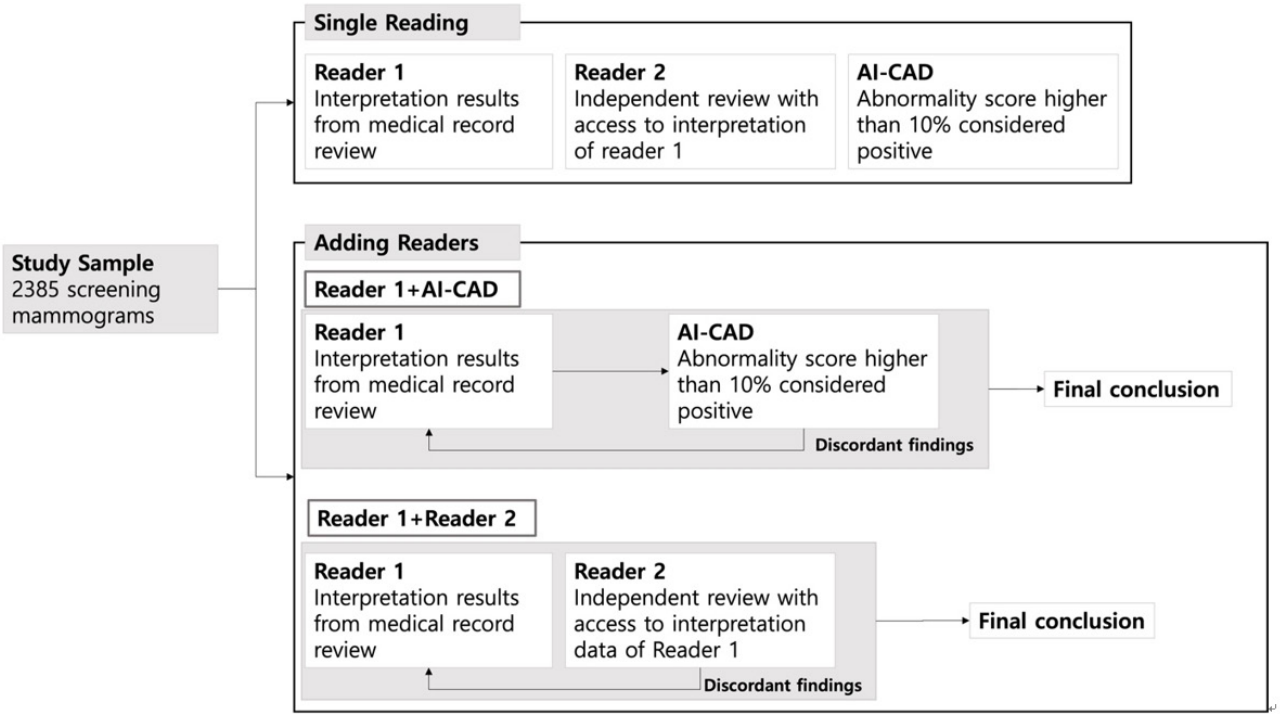

Fig. 1. Diagram showing the workflows of single reading and schemes adding AI-CAD or a second human reader in this study.

for cancer being present in the breast) for each breast that represents the suspicion level of the abnormality detected in that breast on imaging.

Simulation of double reading and AI-CAD integration in mammography interpretation

To simulate double reading, a second reader, 'reader 2' (J.H.Y., 14 years of experience in breast imaging) independently reviewed the mammograms of the study sample with access to the interpretation reports of reader 1. Since this study was of retrospective design using mammography images already interpreted in the past, a discussion for consensus was not possible. Instead, one of the interpreting radiologists (H.J.S., 9 years of experience in breast imaging) re-evaluated the mammography images of cases with discordant interpretations between readers 1 and 2 and called the final conclusion that were analyzed as data from 'reader 1+2'. To simulate interpretation with AI-CAD, this same radiologist re-evaluated the mammography images with positive AI-CAD results. After reviewing AI-CAD marks and abnormality scores, reader 1 chose to either change or maintain the initial interpretation, and the final conclusions were analyzed as data from 'reader 1+AI-CAD' (Fig. 1).

Data and Statistical Analysis

Ground truth in terms of cancer, benign diagnosis or absence of abnormality was confirmed according to histopathologic diagnosis via biopsy/surgery or at least 12 months of follow-up. As the ACR BI-RADS final assessment categories were used for mammography interpretation, BI-RADS categories 1 and 2 were classified as 'negative', and BI-RADS categories 0, 3, 4 and 5 were classified as 'positive'. For statistical analysis, breast parenchymal density was dichotomized as 'fatty', including grades A and B, and 'dense', including grades C or D (7). Abnormality scores calculated by AI-CAD were dichotomized as 'positive' and 'negative' using a threshold value of 10% (8). The following diagnostic metrics were calculated for the individual readers, AI-CAD, and double reading settings: recall rates (number of positive assessments divided by all examinations), sensitivity, specificity, and accuracy. Generalized estimated equation (GEE) methods were used to compare performance metrics between individual readers, AI-CAD, and combined reading settings.

All analyses were conducted using SAS statistical software (version 9.2, SAS Inc., Cary, NC, USA). $p<0.05$ was considered to indicate statistical significance.

Results

Among 2,385 mammograms, 6 (0.3%) were cancer, 32 (1.3%) were biopsy-confirmed benign lesions, and 2,347 (98.4%) were negative examinations. The mean follow-up interval of the negative/benign examinations was 13.5 ± 3.1 months (range, 12.1 to 27.2 months). Of the 6 cancers, 3 were ductal carcinoma in situ (DCIS) and 3 were invasive cancers (2 invasive ductal carcinoma and 1 invasive lobular carcinoma). Of the 2,385 mammograms, 2,329 (97.7%) were initially assessed as negative and the remaining 56 (2.3%) were assessed as positive by the interpreting radiologists (Table 1). The AI-CAD abnormality score was <10% in 2,355 (94.5%) and ≥10% in 130 (5.5%). Supplementary screening ultrasonongraphy (US) examinations are commonly performed in Korea, and 1,396 (58.5%) of the mammography examinations had corresponding US examinations performed.

Table 2 summarizes the clinicopathologic features of the 6 cancer examinations included in this study. Two of the 6 cancer examinations were recalled by the interpreting radiologist due to the presence of mass and distortions, of which the abnormality score of AI-CAD was 89.11% and

Table 1. Patient Characteristics and Imaging Features Among the 2,385 Screening Mammograms Included in this Study

	Negative (n=2,347)	Benign (n=32)	Cancer (n=6)	Total (n=2,385)
Age				
<50 years	1,107 (47.2)	19 (59.4)	2 (33.3)	1,128 (47.3)
≥50 years	1,240 (52.8)	13 (40.6)	4 (66.7)	1,257 (52.7)
Parenchymal density*				
Fatty breast	315 (13.4)	3 (9.4)	1 (16.7)	319 (13.4)
Dense breast	2,032 (86.6)	29 (90.6)	5 (83.3)	2,066 (86.6)
Initial BI-RADS†				
Negative	2,308 (98.3)	20 (62.5)	1 (16.7)	2,329 (97.7)
Positive	39 (1.7)	12 (37.5)	5 (83.3)	56 (2.3)
AI-CAD Abnormality score				
<10%	2,223 (94.7)	28 (87.5)	4 (66.7)	2,255 (94.5)
≥10%	124 (5.3)	4 (12.5)	2 (33.3)	130 (5.5)
US Examinations				
No US	985 (42.0)	4 (12.5)	0 (0.0)	989 (41.5)
US at screening	1,362 (58.0)	28 (87.5)	6 (100.0)	1,396 (58.5)

Percentages are in parentheses, BI-RADS: Breast Imaging Reporting And Data System, US: ultrasonography
*: fatty breast including grades A and B, dense breast including grades C and D
†: negative including BI-RADS 1 and 2, positive including BI-RADS 0, 3, 4, and 5

Table 2. Summary of the Clinicopathologic Features of the 6 Cancer Examinations with Mammography

No.	Age	Breast parenchymal density*	BI-RADS assessment for reader 1	Reason for recall	AI Abnormality score (%)	Cancer diagnosis interval after screening mammography (months)	Pathologic diagnosis	Cancer size (mm)	Axilla lymph node metastasis
1	49	Grade C	0	Calcifications	0.27	1.2	DCIS	31	-
2	46	Grade C	1	No recall : US detected	0.30	0.7	DCIS	15	-
3	53	Grade D	0	Calcifications	2.13	1.2	DCIS	25	-
4	51	Grade C	0	Asymmetry	5.61	0.5	ILC	15	No
5	64	Grade A	0	Mass	89.11	7.0	IDC	13	No
6	55	Grade C	0	Distortion	98.70	1.2	IDC	21	No

BI-RADS: Breast Imaging Reporting And Data System, AI: artificial intelligence, US; ultrasonography, DCIS: Ductal carcinoma in situ, ILC: Invasive lobular carcinoma, IDC: Invasive ductal carcinoma
* According to the ACR BI-RADS

98.70%, respectively. Three cancer examinations were recalled by the interpreting radiologists due to the presence of asymmetry (n=1) and calcifications (n=2), of which the abnormality scores were <10%. One cancer case was not recalled by the two readers or AI-CAD (abnormality score, 0.30%). The patient in this case had undergone supplementary US on which a 7 mm-sized ductal carcinoma in situ (DCIS) was detected.

Comparison of performance metrics according to different interpretation settings

Table 3 summarizes the performance metrics

Table 3. Comparison of Interpretive Performances of Screening Mammography in Different Interpretation Strategies

(%)	Single Reading			Addition of Readers				
	Reader 1	Reader 2	AI-CAD	Reader 1+2	P*	Reader 1+AI-CAD	P†	P‡
TP	5	5	2	5	-	5	-	-
TN	2328	2326	2251	2309	-	2321	-	-
FP	51	53	128	70	-	58	-	-
FN	1	1	4	1	-	1	-	-
Recall rates	2.348 (1.740-2.956)	2.432 (1.814-3.050)	5.451 (4.540-6.362)	3.145 (2.444-3.845)	<0.001	2.642 (1.998-3.285)	0.008	0.010
Sensitivity	83.3 (53.5-100.0)	83.3 (53.5-100.0)	33.3 (0.0-71.1)	83.3 (53.5-100.0)	>0.999	83.3 (53.5-100.0)	>0.999	>0.999
Specificity	97.9 (97.3-98.4)	97.8 (97.2-98.4)	94.6 (93.7-95.5)	97.1 (96.4-97.7)	<0.001	97.6 (96.9-98.2)	0.008	0.010
Accuracy	97.8 (97.2-98.4)	97.7 (97.1-98.3)	94.5 (93.6-95.4)	97.0 (96.3-97.7)	<0.001	97.5 (96.9-98.2)	0.008	0.010

95% confidence intervals are in parentheses

AI-CAD: artificial intelligence-based computer assisted diagnosis/detection, TP: true positive, TN: true negative, FP: false positive, FN: false negative

*: comparison between Reader 1 vs. Reader 1+2

†: comparison between Reader 1 vs. Reader 1+AI-CAD

‡: comparison between Reader 1+2 vs. Reader 1+AI-CAD

Table 4. Distribution of False Positive Findings Detected by Readers and AI-CAD in Different Interpretation Schemes

	Reader 1	AI-CAD	Reader 1+2	Reader 1+ AI-CAD
Negative	0 (0.0)	33 (25.8)	0 (0.0)	0 (0.0)
Soft tissue lesions*	36 (70.6)	60 (46.8)	49 (70.0)	39 (67.2)
Distortion	3 (5.9)	18 (14.1)	3 (4.3)	5 (8.7)
Calcifications	10 (19.6)	12 (9.4)	14 (20.0)	12 (20.7)
Combined	2 (3.9)	5 (3.9)	4 (5.7)	2 (3.4)
Total	51	128	70	58

Percentages are in parentheses

*: including mass and asymmetry

for screening mammograms according to different interpretation settings. When reader 1 was compared to AI-CAD, AI-CAD alone had significantly higher recall rates compared to reader 1, 5.5% (95% confidence intervals (95% CIs) 4.5-6.4) vs. 2.4% (95% CI: 1.7-3.0), respectively (p<0.001). Reader 1+2 had significantly higher recall rates than reader 1 alone, 3.1% (95% CI: 2.4-3.8) vs. 2.4% (95% CI: 1.7-3.0) (p<0.001), respectively. Reader 1+AI-CAD had significantly higher recall rates compared to reader 1, 2.6% (95% CI: 2.0-3.3) vs. 2.4% (95% CI: 1.7-3.0) (p=0.008), respectively. Reader 1+AI-CAD had significantly lower recall rates compared to reader 1+2 (p=0.010).

Specificity (97.9% vs. 97.1%) and accuracy (97.8% vs. 97.0%) were significantly higher in reader 1 compared to both reader 1+2 and reader 1+AI-CAD (all p<0.05, respectively). Similarly, reader 1+AI-CAD showed significantly higher specificity (97.6% vs. 97.1%) and accuracy (97.5% vs. 97.0%) compared to reader 1+2 (p=0.010), respectively.

Summary of false-positive findings detected in different interpretation settings

Table 4 summarizes the distribution of false-

positive findings detected by readers or AI-CAD in different interpretation settings. Soft tissue lesions such as mass and asymmetry were the most common causes for false-positive findings. High proportion of false-positive findings detected by AI-CAD were distortions, while calcifications were mostly the cause for false-positive findings detected by the readers (Fig. 2).

Discussion

In this simulation study, we investigated and compared the diagnostic outcomes of different interpretation schemes simulated for interpreting screening mammography, single reading and double reading by using AI-CAD and a second human reader. Our study results showed that adding readers, either AI-CAD or human second readers, results in higher recalls with significantly lower specificity and accuracy compared to a single human reader. When comparing the effect of adding AI-CAD vs. human second reader, AI-CAD had significantly lower recall and higher specificity and accuracy compared to the scheme of two human readers.

Double reading was applied to overcome the pitfalls of mammography, theoretically with expectations that false-negative interpretation, 'missed cancers', and recalls for additional investigation would decrease (7). Conflictingly, some previous studies have claimed that double reading shows lower recall rates (9, 10) while others have reported higher recalls compared to single reading (11). Our results are consistent with those prior studies showing increased recall rates for double reading compared to single reading. Since this was a simulation study using retrospective interpretation data, consensus discussion between the two readers were not possible that may have been the cause for the higher recall rates and lower specificity/accuracy compared to interpretation results of a single human reader.

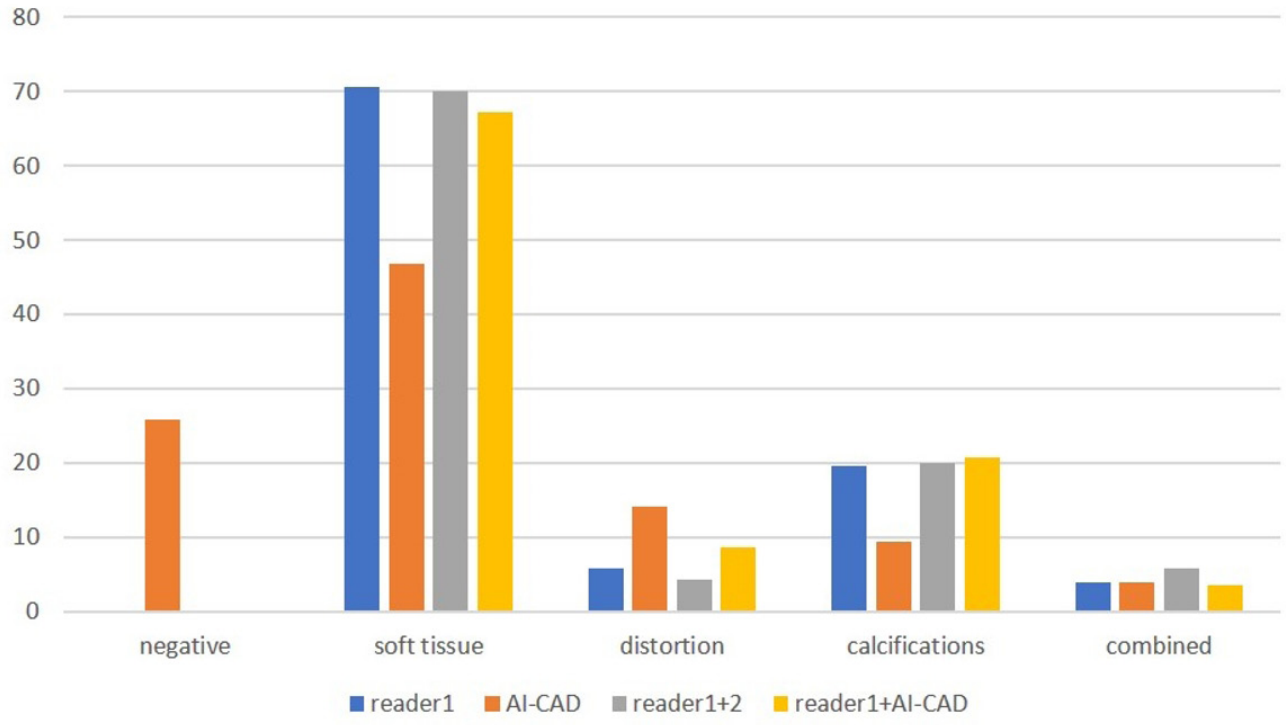

Fig. 2. Box plots of false-positive findings detected by readers or AI-CAD according to different interpretation schemes.

When comparing between the schemes of adding AI-CAD or a second human reader, reader 1+AI-CAD had a significantly lower recall rate with higher specificity and accuracy compared to reader 1+2. This result represents the promising aspects of AI-CAD as a replacement for the second reader in double reading, which is different from the rather disappointing results observed with the use of 'conventional' CAD. After CAD was introduced for mammography interpretation, studies investigated conventional CAD as a substitute for the second reader (12-15). Adding conventional CAD to single reading resulted in increased cancer detection rates with equivocal performance metrics, but with increased recall rates compared to double reading (12, 13). One systematic review showed that conventional CAD did not significantly affect cancer detection, while the overall recall rates increased in single reading settings that added CAD to its reading strategy (15). Our results reflect the different strengths of human readers and AI-CAD for lesion detection, i.e., the second human reader may have similar detection sensitivity with the first reader while AI-CAD analyzes images beyond the human eye, enabling the detection of abnormalities with different characteristics.

As with the differences in detection ability between human readers and AI-CAD, differences between conventional CAD and AI-CAD may be one reason for contrasting outcomes when used for mammographic interpretation. In a study evaluating the abnormality features of conventional CAD-detected cancers, high accuracy was seen for microcalcifications (100%), masses (87%), and asymmetry (80%) (14). In contrast to conventional CAD, AI-CAD has been reported as advantageous when detecting soft tissue abnormalities such as masses and asymmetry because deep learning algorithms show better contrast between abnormal findings and normal parenchymal tissue (8). These features are represented in our results as AI-CAD showed lower rates of false-positive results presenting as soft tissue lesions (Fig. 2). In addition, 2 of the 3 proven cancers that were overlooked by AI-CAD (abnormality score was 0.27% and 2.13%, respectively) but detected by the radiologists presented as calcifications that were surgically-proven as DCIS. AI-CAD also had higher false-positive results presenting as distortions, in which most were dismissed by the human readers (Table 4). Mammographic features that are marked or dismissed by AI-CAD needs characterization according to their final diagnosis, and we anticipate future investigation in this topic.

There are several limitations to this study. First, we retrospectively simulated different interpretation schemes using a small study sample, and different results may be seen where active consensus or arbitration is possible during actual clinical practice. Second, the follow-up period for non-cancer cases was relatively short (mean, 13.5 months). Third, approximately 58.5% of our study population had supplementary US performed that may have affected the initial interpretation results, which was an issue not considered. Last, the relatively low recall rates from readers and low cancer rate (0.3%, 6 of 2385) may have affected the statistical power of the comparison.

In conclusion, adding readers, either AI-CAD or human second readers, results in higher recalls with significantly lower specificity and accuracy compared to single reading of a human reader. When comparing the effect of adding AI-CAD vs. human second reader, AI-CAD had significantly lower recall and higher specificity and accuracy compared to the scheme of two human readers.

Acknowledgement

This study was supported by a faculty research grant of Yonsei University College of Medicine for 2020 (2020-32-0041).

References

1. International Agency for Research on Cancer IARC handbooks of cancer prevention. Vol. 15. Breast cancer screening: IARC Press, 2015.
2. 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.
3. Taylor-Phillips S, Stinton C. Double reading in breast cancer screening: considerations for policy-making. Br J Radiol 2020;93:20190610.
4. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med 2015;175:1828-1837.
5. 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.
6. Rodriguez-Ruiz A, Krupinski E, Mordang JJ, Schilling K, Heywang-Kobrunner SH, Sechopoulos I, et al. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. Radiology 2019;290:305-314.
7. American College of Radiology. Breast imaging reporting and data system. 5th ed. Reston, VA: American College of Radiology, 2013.
8. Kim H-E, Kim HH, Han BK, Kim K-H, Nam H, Lee EH, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digital Health 2020;2:e138-e148.
9. Euler-Chelpin MV, Lillholm M, Napolitano G, Vejborg I, Nielsen M, Lynge E. Screening mammography: benefit of double reading by breast density. Breast Cancer Res Treat 2018;171:767-776.
10. Taylor-Phillips S, Jenkinson D, Stinton C, Wallis MG, Dunn J, Clarke A. Double Reading in Breast Cancer Screening: Cohort Evaluation in the CO-OPS Trial. Radiology 2018;287:749-757.
11. Coolen AMP, Voogd AC, Strobbe LJ, Louwman MWJ, Tjan-Heijnen VCG, Duijm LEM. Impact of the second reader on screening outcome at blinded double reading of digital screening mammograms. Br J Cancer 2018;119:503-507.
12. Gilbert FJ, Astley SM, Gillan MG, Agbaje OF, Wallis MG, James J, et al. Single reading with computer-aided detection for screening mammography. N Engl J Med 2008;359:1675-1684.
13. Gilbert FJ, Astley SM, McGee MA, Gillan MG, Boggis CR, Griffiths PM, et al. Single reading with computer-aided detection and double reading of screening mammograms in the United Kingdom National Breast Screening Program. Radiology 2006;241:47-53.
14. James JJ, Gilbert FJ, Wallis MG, Gillan MG, Astley SM, Boggis CR, et al. Mammographic features of breast cancers at single reading with computer-aided detection and at double reading in a large multicenter prospective trial of computer-aided detection: CADET II. Radiology 2010;256:379-386.
15. Taylor P, Potts HW. Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. Eur J Cancer 2008;44:798-807.

대한유방검진의학회지 2023;20:52-61

검진 유방촬영술 판독체계별 진단 성능 비교: 시뮬레이션 연구

노미리비[1] · 이혜선[2] · 서희정[3] · 김은경[1,4] · 이시은[1,4] · 윤정현[1]

[1]연세대학교 의과대학 세브란스병원 영상의학과, 방사선의과학연구소
[2]연세대학교 의과대학 의생명시스템정보학교실 통계지원실
[3]세브란스헬스체크업센터 영상의학과
[4]연세대학교 의과대학 용인세브란스병원 영상의학과

배경: 본 연구는 검진 유방촬영술 판독체계별로 진단 성능을 알아보는데 목적이 있다. 영상의학과 의사 단독 판독과 비교하여 인공지능 진단 보조프로그램을 독립된 판독의로 활용한 이중 판독과 두명의 영상의학과 전문의가 시행한 이중 판독을 시뮬레이션 하여 판독 체계의 차이에 따른 진단 성적의 차이를 비교하고자 하였다.

방법: 2018년 1월부터 2019년 1월까지 건강검진을 위해 유방촬영술을 시행한 여성 2,385명 (평균 연령: 50.0±9.5세)에서 시행한 유방촬영술에 대한 판독문을 후향적으로 분석하여 단독 판독 결과로 사용하였다. 이중 판독은 다음과 같이 1) 단독 판독을 시행한 영상의학과 전문의 중 1인이 기존 판독 보고와 인공지능 진단 보조프로그램 분석 결과를 참고하여 재판독한 결과와 2) 단독 판독에 참여하지 않은 독립적인 두번째 영상의학과 전문의가 기존 판독 결과를 참고한 이중 판독으로 설정하였다. 암/양성 또는 이상 소견은 조직학적 진단 또는 최소 1년간의 추적 관찰을 통해 확인하였다.

결과: 2,385개의 유방촬영술 중 6건(0.3%)은 암으로 진단되었으며, 32건(1.3%)은 조직검사 생검을 통해 확인된 양성, 2,347건(98.4%)은 음성 판정이었다. 인공지능 진단 보조프로그램을 활용한 이중 판독은 기존 단독 판독과 비교하여 유의하게 높은 재검률을 보였으나 (2.6% vs. 2.4%, p=0.008) 두명의 전문의가 판독한 설정과 비교하여 유의하게 낮은 재검률을 보였다(2.6% vs. 3.1%, p=0.010). 특이도와 정확도는 단독 판독이 이중 판독의 두 가지 설정과 비교해 유의하게 높았다. 인공지능 진단 보조프로그램을 활용한 이중 판독의 경우 전문의 두명이 판독한 경우와 비교하여 특이도 (97.6% vs. 97.1%)와 정확도 (97.5% vs. 97.0%)가 유의하게 높았다 (p=0.010). 인공지능 진단 보조프로그램으로 검출된 위양성 소견은 높은 비율로 왜곡인 반면 이중판독으로 인한 위양성소견은 대부분 석회화 소견이었다.

결론: 단독 판독과 비교하여 인공지능 진단 보조프로그램을 활용하거나 두명의 전문의가 판독한 이중 판독이 높은 재검률과 낮은 특이도, 정확도를 보였다. 이중 판독 설정 중 인공지능 진단보조프로그램을 활용한 경우가 두명의 전문의 판독 보다 높은 특이도와 정확도를 보였다.

Index words: Mammography; Breast neoplasms; Artificial intelligence; Computer-Assisted detection/diagnosis

Corresponding author: Jung Hyun Yoon, M.D., Ph.D.