대한유방검진의학회지 2020;17:70-76

원 저

유방촬영의 위양성 판정에 관한 전통적 진단보조프로그램과 인공지능 기반 진단보조프로그램의 비교

이시은 · 김명현 · 김은경

연세대학교 의과대학 용인세브란스병원 영상의학과

목적: 같은 유방촬영술 영상에 적용된 전통적 진단보조프로그램과 인공지능 기반 진단보조프로그램의 결과를 분석하여 위양성 판정을 비교한다.

대상 및 방법: 2020년 5월부터 7월까지 본원에서 유방촬영술을 시행받은 환자 중 BI-RADS 범주 1 또는 2 였던 환자에서 1 개월 전후에 시행한 유방초음파상 BI-RADS 범주 1 또는 2 였던 환자 256명을 대상으로 하였다. 각 유방촬영술 검사에 대해 두 종류의 진단보조프로그램이 적용되었고 4개의 영상 (RMLO, LLMO, RCC, LCC)에 대해 양성으로 표시된 부위를 비교하였다.

결과: 환자당 위양성 마커의 개수는 전통적 진단보조프로그램에서 2.23개, 인공지능 기반 진단보조프로그램에서 0.14개로 현저한 차이를 보였다 (P <0.001). 전체 환자 중 마커없이 정상으로 나타난 환자는 전통적 진단보조프로그램에서 42명 (16%), 인공지능 기반 진단보조프로그램에서 239명 (93%) 이었다.

결론: 인공지능 기반 진단보조프로그램은 현저히 낮은 위양성 판정을 보여 기존 프로그램에 비해 판독의 효율성을 높이는데에 기여할 수 있을 것으로 사료된다.

Index words: Computer-assisted diagnosis; Artificial Intelligence; Digital Mammography

서 론

유방촬영술에 대한 전통적 진단보조프로그램 (Conventional computer-aided detection, 이하 C-CAD)은 1990년대 미국 식품의약국 (U.S. Food and Drug Administration) 및 유럽 CE (Conformité Européene) 허가를 취득한 이래 전세계적으로 널리 이용되고 있다 (1). 2000년대 초반 진행된 여러 연구에서는 C-CAD를 이용함으로써 유방암의 발견율을 높여 유방촬영술의 진단 능력을 향상시킬 수 있다고 보고하였다 (2-4). 그러나 이어진 대규모 연구에서 유방암 발견율에서 유의미한 차이가 없었거나 오히려 위양성 진단을 증가시켜 불필요한 추가 검사를 유도했다는 결과가 보고되기도 하여, 실제 임상환경에서의 역할에 대해서는 논란이 되어왔다 (5-7). C-CAD가 전문가에 의해 제시된 수제 특징 (hand-crafted feature)를 추출하여 이를 학습하는 방식이라면, 최근 개발된 인공지능 기반의 진단보조프로그램 (Artificial intelligence based computer aided

통신저자: Eun-Kyung Kim, M.D., Ph.D.
Department of Radiology, Yongin Severance Hospital Yonsei University College of Medicine 363, Dongbaekjukjeon-daero, Giheung-gu, Yongin-si, Gyeonggi-do 16995, Korea
Tel. (031) 5189-8321, Fax. (02) 2227-8337
E-mail: ekkim@yuhs.ac

detection/diagnosis, 이하 AI-CAD)은 디지털 유방촬영술이 정착한 이후의 방대한 데이터를 이용하여 표준참조값 (Standard reference)을 기준으로 자체적으로 학습한다 (8). 최근의 여러 연구에서는 AI-CAD가 C-CAD의 취약점이었던 위양성진단을 늘리지 않으면서도 유방암의 진단 정확도를 높일 수 있다는 결과를 발표하였다 (9-12). C-CAD의 주된 단점으로 여겨지는 높은 위양성율은 판독시간을 지연시킬 뿐만 아니라 판독자에게 상당한 피로감을 줄 수 있다 (13). 실제 판독자의 업무를 경감시키고 보조프로그램으로써 효율적으로 활용되기 위해서는 유방암 발견율을 높이는 것만큼이나 위양성 소견을 줄이는 것이 중요하다고 여겨진다 (14).

이번 연구에서는 정상 혹은 전형적 양성 소견을 가진 환자를 대상으로 두 개의 상용화된 C-CAD와 AI-CAD를 함께 적용시켜보았을 때 나타나는 위양성 소견을 비교해 보고자 하였다.

대상 및 방법

2020년 5월에서 2020년 7월까지 본원에서 양측 유방촬영술을 시행한 환자 중 ACR BI-RADS 범주 1, 2의 결과를 보인 881명 중, 유방촬영술 시행 전후 1개월 내에 초음파 검사를 시행하였고 역시 범주 1, 2의 결과를 보였던 258명의 영상을 수집하였다. 이 중 1명은 내외사위 영상만 촬영한 27세 환자로 제외하였고 또 다른 1명은 유방암에 대한 부분 절제술을 받은 환자로 제외하여 총 256명이 포함되었다. 유방촬영술 검사는 한 종류의 기기 (Senographe Pristina Mammography System, GE Healthcare, Chicago, IL, USA)로 촬영되었고, 2-25년의 유방촬영검사 판독 경험이 있는 영상의학과 의사 5명이 BI-RADS에서 제시하는 기준에 따라 판독하였다.

해당 기간 동안 전향적으로 C-CAD (SecondLook, version 7.2H, iCAD, Nashua, NH, USA) 및 AI-CAD (Lunit insight MMG, version 1.1.1.0, Lunit, Seoul, Korea)가 유방촬영술에 적용되었고 이 결과를 후향적으로 분석하였다. 전자의 프로그램은 유방촬영술에 포함된 네 장의 사진 (RMLO, LLMO, RCC, LCC)에 종괴와 석회 소견을 구분하여 각각 원과 사각형으로 표기하며, 후자의 프로그램은 소견을 구별하지 않고 각 위치의 위험도에 따라 녹색에서 붉은색까지의 영역 (heatmap)으로 표기한다. 이 위험도는 10-100까지의 숫자로 함께 제시되며, 10 미만은 영역으로 표시되지 않는다. 본 연구에서 제시한 한 환자의 악성예측도는 네 장의 사진에서 표시된 영역의 위험도 중 가장 높은 값을 이용하였다.

본 연구에 포함된 전체 256명은 유방촬영 및 함께 시행한 초음파 검사상 범주 1, 2로 진단된 환자로 모두 음성으로 판단하였고, 각 환자에서 보인 진단보조프로그램의 마커는 위양성 소견으로 정의하였다. 각 사진당 관찰된 위양성 소견의 개수를 확인하여 대응표본 T 검정을 통해 비교하였다.

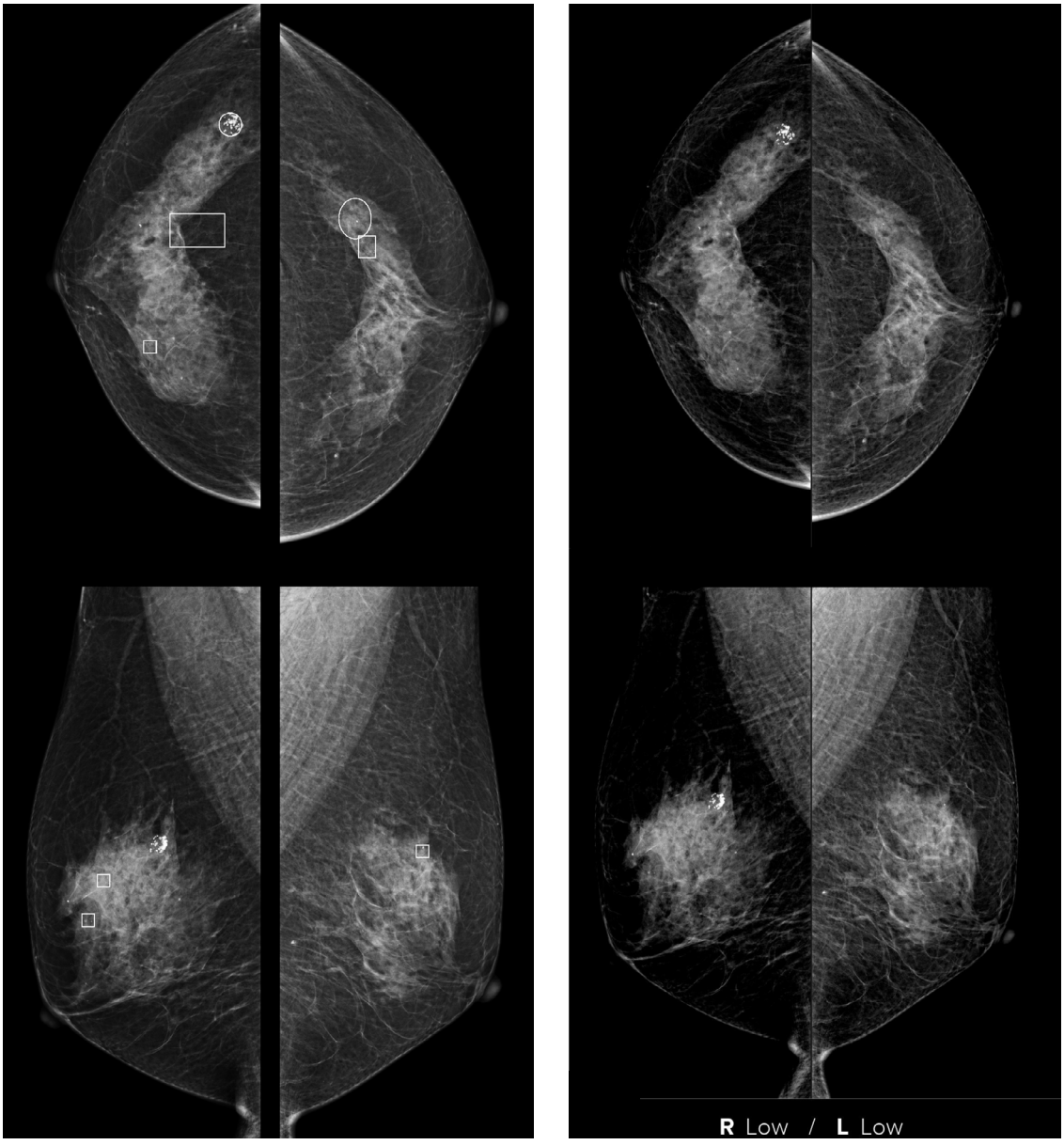
결 과

총 256명의 환자가 포함되었으며 평균 54세 여성이었다 (범위 30-77, SD 9). 판독문에 근거한 유방치밀도는 A 3명 (1.2%), B 44명 (17.2%), C 156명 (60.9%), D 53명 (20.7%)으로 분포하였다. 유방 촬영술 판독상 BI-RADS 1이 201명 (78.5%), BI-RADS 2가 55명 (21.5%) 이었고 유방 촬영 시행 전후 1개월에 시행한 초음파 검사 판독상 BI-RADS 1이 111명 (43.4%), BI-RADS 2가 145명 (56.6%) 이었다.

C-CAD의 경우 영상 당 0.5-0.62개의 위양성 마커를 표시하였고, 석회화 (0.17-0.22) 보다는 종괴 (0.32-0.44)로 표시된 경우가 많았다. 총 4개의 영상을 포함하는 한 환자의 검사당 평균 2.23개로 나타났다. AI-CAD의 경우 석회화와 종괴를 구별하여 표기하지 않지만, 영상당 0.03-0.04개의 위양성 마커가 표기되었고 한 환자당 평균 0.14개로 나타났다. 개별 영상 및 환자당 표기된 위양성 마커는 AI-CAD에서 현저히 낮게 나타났다 (모두 P <0.001, Table 1, Fig. 1). 두 CAD에서 보인 환자당

Table 1. False-Positive Marks per Image and Patient between the Conventional CAD and AI-CAD

		RCC	LCC	RMLO	LMLO	Total (per patient)
Conventional CAD	Mass	0.33	0.32	0.44	0.37	2.23
	Microcalcification	0.17	0.20	0.18	0.22	
	Total	0.50	0.52	0.62	0.59	
AI-CAD		0.03	0.04	0.04	0.03	0.14
P-value		<0.001	<0.001	<0.001	<0.001	<0.001



a b

Fig. 1. A 62-years-old patient with false-positive marks on the conventional CAD. (**a**) The conventional CAD shows 1 mass (circle) and 4 microcalcifications (square) on the right breast, and 1 mass (circle) and 2 microcalcifications (square) on the left breast. (**b**) The AI-CAD shows no mark.

위양성 마커수에 대한 대응표본 상관계수는 0.107 (P = 0.087)로 의미있는 상관관계를 보이지 않았다.

전체 256개 검사 중 마커가 전혀 표시되지 않은 영상은 C-CAD의 경우 42명으로 16%에 불과한 반면 AI-CAD의 경우 239명, 전체의 93%에서 정상에 준하는 결과를 보였다. 위양성 마커가 환자당 3개 이상 보인 경우는 C-CAD에서 96명으로 38%였던 반면에 AI-CAD의 경우 4명으로 2%에 불과했다 (Fig. 2).

AI-CAD에서 위양성 마커를 보인 17명에서는 악성예측도가 평균 42.1점 (범위 13-94)으로 나타났다. 17명 중 13명은 평균 5.3년 (범위 3.1-8.1)간 정상 혹은 변화 없는 양성 소견으로 검진을 받아왔던 환자였다. 이전 검사가 없었던 나머지 4명 중 최고 94점을 보인 환자는 양측에서 대칭적인 음영을 보였으며 오랜 기간 당뇨로 투약 중이었던 환자로, 초음파 및 임상적인 판단으로 당뇨병성 유방병증으로 의심되었다 (Fig. 3). AI-CAD에서 위양성 마커를 보

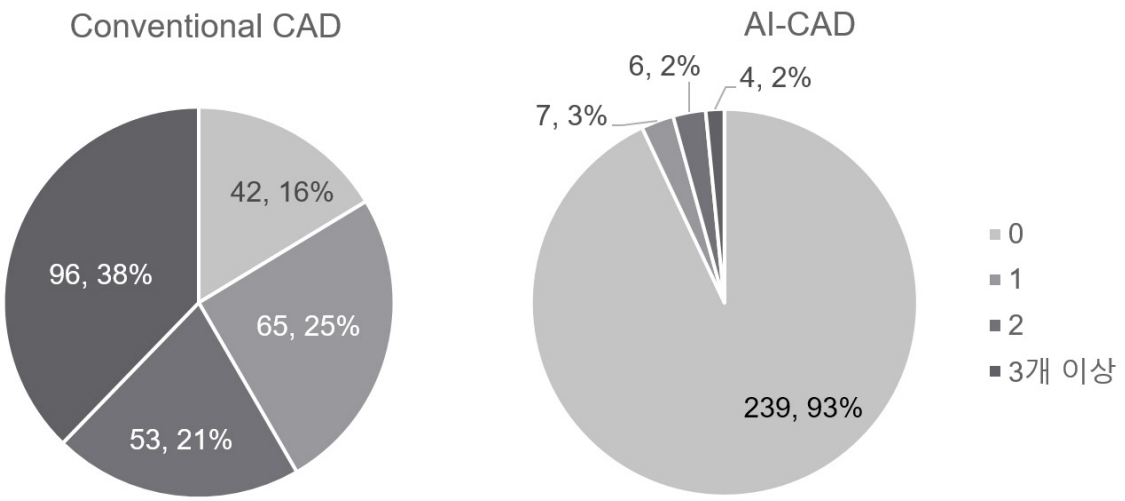
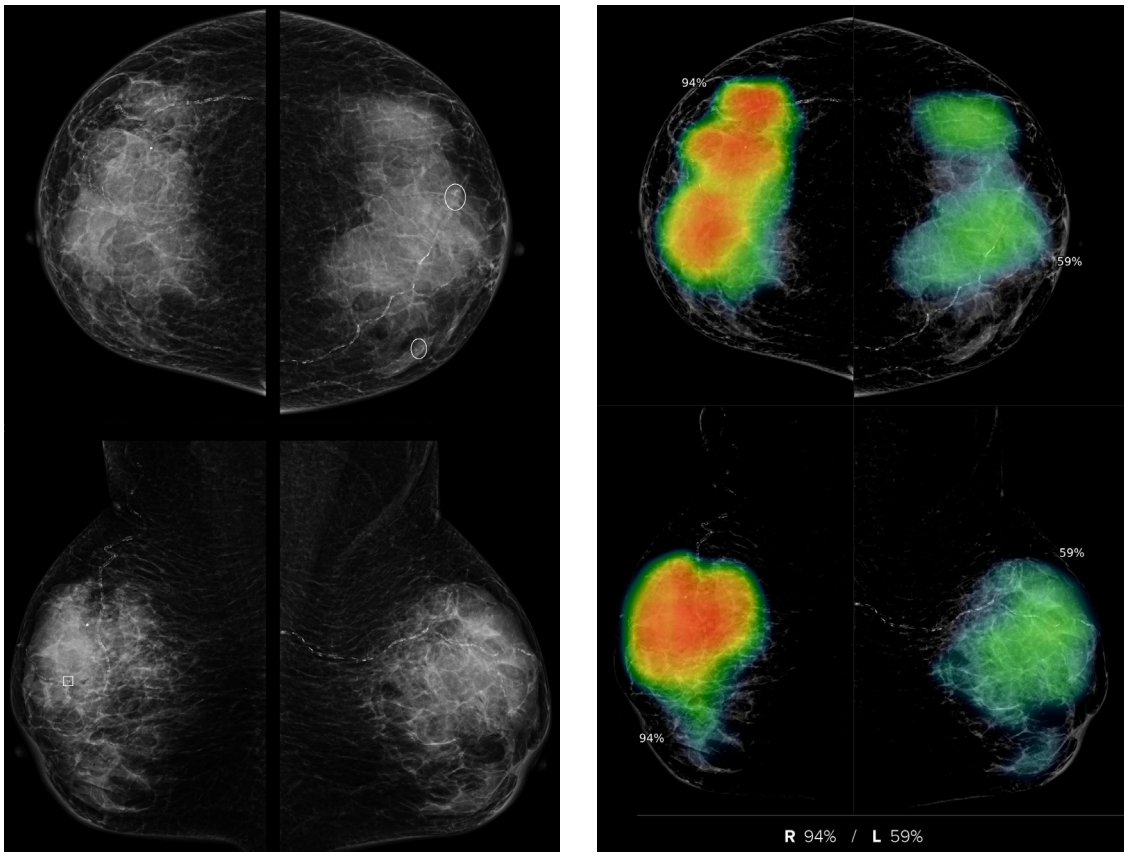


Fig. 2. Two graphs show the number and percentage of patients according to the number of false-positive marks.



a **b**

Fig. 3. A 71-years-old patient with false-positive marks on both CADs. (**a**) The conventional CAD shows 1 microcalcification (square) on the right MLO view and two masses (circle) on the left CC view. (**b**) The AI-CAD shows broad heat maps (total 4 markers) on the right and left breast with 94% and 59% malignancy risks, respectively.

인 17명 중 1명은 C-CAD상 마커가 없었고 16명에서는 1-7개의 위양성 마커를 보였다.

고 찰

의료영상 분야에서 유방촬영에 대한 진단보조프로그램은 가장 널리, 오랜 기간 사용되고 있는 소프트웨어 중 하나이다 (15). 가장 중요한 역할은 유방암의 발견율을 높이는 것이지만, 이로 인해 위양성 판정이 불가피하게 증가된다는 우려가 있었다 (5-7). 본 연구에서는 정상 혹은 전형적 양성 (benign) 소견으로 진단 받은 환자에서 기존 C-CAD와 최근 개발된 AI-CAD의 위양성 표시가 얼마나 나타나는지를 비교해보고자 하였다. C-CAD는 한 환자당 평균 2.23개의 위양성 표시를 나타냈고, AI-CAD는 평균 0.14개로, 94%의 위양성 표시가 감소되는 효과가 있었다. 또한 표시가 없는 환자가 C-CAD에서는 42명 (16%)에 불과했으나, AI-CAD의 경우 239명 (93%)가 정상에 준하는 결과를 보였다.

C-CAD는 영상의학과 의사가 판단하는 영상 소견과 연관된 여러 수제 특징 (hand-crafted feature)을 추출하여 판단을 하는 방식으로 사람이 생각하는 방향으로의 편견 (bias)을 유발할 수 있는 반면, AI-CAD는 정답지에 대해 프로그램 자체적으로 학습을 하는 방식으로 후자의 방식이 진단 성능에 있어 우월하다는 대규모 연구 결과가 발표된 바 있다 (16). 2019년 발표된 논문에서는 본 연구와 다른 종류의 C-CAD 및 AI-CAD를 비교하였으며 (ImageChecker CAD, version 10.0, Hologic, Sunnyvale, USA vs. cmAssist, prototype AI-CAD, CureMetrix, La Jolla, USA), AI-CAD에서 69%의 위양성 마크가 감소하였다 (17). 과도한 위양성 마크는 판독시간을 지연시키고, 판독자의 피로도를 증가시킨다. C-CAD에 대한 이전의 연구에서는 CAD의 결과를 추가로 확인하여 판독하는데 평균 23초의 시간이 더 소모된다고 분석하였다 (13). 이번 연구 결과상 C-CAD에서 1개 이상의 위양성 표시를 보인 환자가 214명 (84%) 임을 감안하면 이를 확인하는 데에 총 1.4시간이 추가로 소모되었다고 볼 수 있다.

AI-CAD는 병변의 위치를 보여줌과 동시에 각각의 악성가능성을 함께 제시하는데, 이 점수는 대규모 연구에서 유방암에 대한 높은 진단 성능을 보여주었다 (9-12). 그러나 본 연구에서 AI-CAD는 7% (17/256)의 환자에서 위양성 소견을 보였으며, 평균 42.1점(범위 13-94)의 다소 높은 악성가능성을 제시하였다. 이 환자 모두에서 초음파 검사상 정상 소견이거나 전형적 양성 소견을 보였다. 이들 중 76% (13/17)의 환자는 평균 5년 간 꾸준히 검진을 받았던 환자들로 이전 검사와 비교 판독이 가능했고, 다른 1명은 오랜 기간 당뇨병을 앓고있어 임상적으로 당뇨병성 유방병증을 의심할 수 있었던 환자였다. CAD의 활용에 더하여 이전의 검사 결과와 함께 얻을 수 있는 임상적 정보는 위양성진단을 줄이는 데에 중요한 역할을 할 수 있을 것으로 생각된다.

이 연구의 몇 가지 제한점은 다음과 같다. 첫째, 2020년 유방촬영 검사를 받은 환자를 대상으로 하여 경과관찰을 위한 기간이 확보되지 않았다. 때문에 유방촬영 및 1개월 전후에 시행한 초음파 검사상 정상 혹은 전형적 양성소견을 보였던 환자만을 대상으로 하였고, 위양성 판정에 초점을 맞추고자 악성으로 진단된 환자 역시 제외하였다. 둘째, 사용한 두 개의 상용 CAD는 C-CAD와 AI-CAD 전체를 대표하지는 못하며, 현재 상용화된 프로그램들 사이에서도 진단 성능에 차이가 있다는 결과가 발표된 바 있다. 그러나 본 연구는 유사한 주제로 발표된 이전 논문과 같은 경향성을 보여주고 있다. 셋째, 유방촬영 및 초음파 판독 결과는 후향적으로 재분석하지 않고 검사 시행 당시의 판정 결과를 기준으로 하였다. 그러나 BI-RADS 1 혹은 2인 검사만 분석에 포함되었으므로 검사자간 발생할 수 있는 다양성은 최소화되었을 것으로 보인다.

결론적으로 유방촬영 영상에 적용된 AI-CAD는 C-CAD에 비해 현저히 낮은 위양성 표시를 보임으로써 판독의 효율성을 높일 수 있을 것으로 기대한다.

참 고 문 헌

1. Castellino RA. Computer aided detection (CAD): an overview. Cancer imaging 2005;5:17-19
2. 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
3. Birdwell RL, Bandodkar P, Ikeda DM. Computer-aided detection with screening mammography in a university hospital setting. Radiology 2005;236:451-457
4. Morton MJ, Whaley DH, Brandt KR, Amrami KK. Screening mammograms: interpretation with computer-aided detection--prospective evaluation. Radiology 2006;239:375-383
5. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med 2007;356:1399-1409

6. 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. J Natl Cancer Inst 2004;96:185-190
7. Hall FM. Breast imaging and computer-aided detection. N Engl J Med 2007;356:1464-1466
8. Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw 2015;61:85-117
9. Rodríguez-Ruiz A, Krupinski E, Mordang J-J, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology 2019;290:305-314
10. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst 2019;111:916-922
11. 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. Lancet Digit Health 2020;2:e138-e148
12. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577:89-94
13. Tchou PM, Haygood TM, Atkinson EN, et al. Interpretation Time of Computer-aided Detection at Screening Mammography. Radiology 2010;257:40-46
14. Mayo RC, Leung J. Artificial intelligence and deep learning - Radiology's next frontier? Clin Imaging 2018;49:87-88
15. Rao VM, Levin DC, Parker L, Cavanaugh B, Frangos AJ, Sunshine JH. How widely is computer-aided detection used in screening and diagnostic mammography? J Am Coll Radiol 2010;7:802-805
16. Kooi T, Litjens G, Van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 2017;35:303-312
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. J Digit Imaging 2019;32:618-624

Comparison of conventional CAD and AI-CAD applied to digital mammography in respect of false-positive marks

Si Eun Lee, MD, Myeong Hyun Kim, MD, Eun-Kyung Kim, MD, PhD

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

Purpose: To retrospectively compare false-positive marks of mammography applied by the conventional computer-assisted detection/diagnosis (CAD) program and artificial intelligence based CAD (AI-CAD) program through head-to-head setting.

Materials and Methods: Between May 2020 and July 2020, 256 patients who were reported as BI-RADS 1 and 2 on mammography and ultrasound which was performed 1 month before or after mammography. We compared false-positive marks per image and patient between the conventional CAD and AI-CAD.

Results: The number of false-positive marks was markedly decreased in the AI-CAD compared with the conventional CAD (0.14, 2.23, respectively, P <0.001). The number of patients without any mark was only 42 (16%) in the conventional CAD while it was 239 (93%) in the AI-CAD.

Conclusion: The AI-CAD showed far fewer false-positive marks than the conventional CAD and it may reduce the reading time and fatigue level of radiologists.

Index words: Computer-assisted diagnosis; Artificial Intelligence; Digital Mammography

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