JAMA Oncology | Original Investigation

External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms

Mattie Salim, MD; Erik Wåhlin, MSc; Karin Dembrower, MD; Edward Azavedo, MD, PhD; Theodoros Foukakis, MD, PhD; Yue Liu, MSc; Kevin Smith, MSc, PhD; Martin Eklund, MSc, PhD; Fredrik Strand, MD, PhD

IMPORTANCE A computer algorithm that performs at or above the level of radiologists in mammography screening assessment could improve the effectiveness of breast cancer screening.

OBJECTIVE To perform an external evaluation of 3 commercially available artificial intelligence (AI) computer-aided detection algorithms as independent mammography readers and to assess the screening performance when combined with radiologists.

DESIGN, SETTING, AND PARTICIPANTS This retrospective case-control study was based on a double-reader population-based mammography screening cohort of women screened at an academic hospital in Stockholm, Sweden, from 2008 to 2015. The study included 8805 women aged 40 to 74 years who underwent mammography screening and who did not have implants or prior breast cancer. The study sample included 739 women who were diagnosed as having breast cancer (positive) and a random sample of 8066 healthy controls (negative for breast cancer).

MAIN OUTCOMES AND MEASURES Positive follow-up findings were determined by pathology-verified diagnosis at screening or within 12 months thereafter. Negative follow-up findings were determined by a 2-year cancer-free follow-up. Three AI computer-aided detection algorithms (AI-1, AI-2, and AI-3), sourced from different vendors, yielded a continuous score for the suspicion of cancer in each mammography examination. For a decision of normal or abnormal, the cut point was defined by the mean specificity of the first-reader radiologists (96.6%).

RESULTS The median age of study participants was 60 years (interquartile range, 50-66 years) for 739 women who received a diagnosis of breast cancer and 54 years (interquartile range, 47-63 years) for 8066 healthy controls. The cases positive for cancer comprised 618 (84%) screen detected and 121 (16%) clinically detected within 12 months of the screening examination. The area under the receiver operating curve for cancer detection was 0.956 (95% CI, 0.948-0.965) for AI-1, 0.922 (95% CI, 0.910-0.934) for AI-2, and 0.920 (95% CI, 0.909-0.931) for AI-3. At the specificity of the radiologists, the sensitivities were 81.9% for AI-1, 67.0% for AI-2, 67.4% for AI-3, 77.4% for first-reader radiologist, and 80.1% for second-reader radiologist. Combining AI-1 with first-reader radiologists achieved 88.6% sensitivity at 93.0% specificity (abnormal defined by either of the 2 making an abnormal assessment). No other examined combination of AI algorithms and radiologists surpassed this sensitivity level.

CONCLUSIONS AND RELEVANCE To our knowledge, this study is the first independent evaluation of several AI computer-aided detection algorithms for screening mammography. The results of this study indicated that a commercially available AI computer-aided detection algorithm can assess screening mammograms with a sufficient diagnostic performance to be further evaluated as an independent reader in prospective clinical trials. Combining the first readers with the best algorithm identified more cases positive for cancer than combining the first readers with second readers.

Invited Commentary page 1588

Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Fredrik Strand, MD, PhD, Department of Oncology-Pathology, Karolinska Institute, Karolinska vägen, A2:07, 171 64 Solna, Sweden (fredrik.strand@ki.se).

JAMA Oncol. 2020;6(10):1581-1588. doi:10.1001/jamaoncol.2020.3321
Published online August 27, 2020.

Population-wide mammography screening resulting in earlier detection of tumors has decreased breast cancer mortality by 20% to 40%.[1,2] Nevertheless, the workload for radiologists is high and the quality of assessment varies.[3,4] Having a computer algorithm that performs at, or above, the level of radiologists in mammography assessment would be valuable. An added benefit of artificial intelligence (AI) computer-aided detection (CAD) algorithms would be to reduce the broad variation in performance among human readers that has been shown in previous studies.[5,6] Computer-aided detection can take on 2 different roles: as a concurrent assistant directing the radiologist's attention to suspicious areas in the mammogram and as an independent reader making an overall assessment of the whole examination without radiologist intervention. Until recently, most commercial CAD systems operated as concurrent assistants and were based on a limited set of programmer-defined features used to identify suspicious areas in the mammogram.[7] This approach was never convincingly successful, with some early reports showing an increased sensitivity but later studies showing no clear improvement.[7,8] Furthermore, additional time was required from the radiologist to consider each CAD marking. During the last years, academic and commercial researchers have worked hard to leverage the capabilities of AI, or more specifically of deep neural networks, to enable CAD for independent mammography assessments.[9-12] The reported performance levels have in several cases been on par with radiologists. However, across these published studies there have been various issues: the source population was not a screening cohort,[9] the radiologists with which the AI CAD program was compared showed a poor performance,[10] and the AI CAD algorithms have often not been publicly available.[10-12] Most importantly, none of these studies involved third-party external validation with comparisons among competing AI CAD algorithms. In the present study, we compare the results of applying 3 different AI CAD algorithms as independent readers of a large set of mammographic examinations from a public mammography screening program for which the algorithm developers had no access to images and had no involvement in the evaluation process.

Key Points

Question Are there currently commercially available artificial intelligence (AI) algorithms that perform as well as or above the level of radiologists in mammography screening assessment?

Findings In this case-control study that included 8805 women, 1 of the 3 externally evaluated AI computer-aided detection algorithms was more accurate than first-reader radiologists in assessing screening mammograms. However, the highest number of cases positive for breast cancer was detected by combining this best algorithm with first-reader radiologists.

Meaning One commercially available AI algorithm performed independent reading of screening mammograms with sufficient diagnostic performance to act as an independent reader in prospective clinical studies.

Methods

The study sample was derived from the CSAW (Swedish Cohort of Screen-Age Women) data set,[13] which consists of all women 40 to 74 years of age in Stockholm county who were invited for screening examinations between 2008 and 2015. The screening interval in Stockholm county is 24 months. However, until 2012, the interval was only 18 months for women between 40 and 49 years of age. In the present study, all screening examinations were from one institution, the Karolinska University Hospital. This retrospective case-control study was approved by the Stockholm ethical review board, which waived the requirement for individual informed consent.

We included all women aged 40 to 74 years from the CSAW cohort who were diagnosed as having breast cancer between 2008 and 2015, who had a complete screening examination prior to diagnosis, who had no prior breast cancer, and who did not have implants. We excluded 419 examinations with a cancer diagnosis that had more than 12 months between the examination date and diagnosis owing to the lower likelihood of cancer being present at the time of screening. In a secondary analysis, we added 174 women who had received a cancer diagnosis between 12 and 23 months after screening. Random sampling of healthy women was carried out to enable efficient computer processing while maintaining representability. We included a random sample of 10 000 healthy women. Of those women, we excluded 995 who had less than 2-year cancer-free follow-up, 909 who had examinations after December 31, 2015, 26 who had implants, and 99 examinations with an unknown radiologist identification number (eFigure 1 in the Supplement). All images were acquired on full-field digital mammography Hologic equipment. Prospectively recorded screening assessments for each examination were extracted from the Regional Cancer Center Stockholm-Gotland screening register. The mammography screening system in Sweden requires a 2-view mammography of each breast. All examinations are assessed by double-reading, with a binary decision by each reader: normal or abnormal ("flagged" for discussion). There had been 25 different first-reader radiologists and 20 different second-reader radiologists. There is no defined designation of breast radiologists into first or second readers. However, the second reader is often more experienced than the first reader. In addition, when performing an assessment, the second reader can access the assessment already performed by the first reader. For any abnormal assessment, the examination proceeds to consensus discussion with another binary decision: normal or recall. Data on cancer diagnoses, including tumor characteristics and radiologic assessments, were obtained through linkage with the Regional Cancer Center Stockholm-Gotland breast cancer quality register and screening register using Swedish personal identity numbers. Positive follow-up findings were determined by pathology-confirmed diagnosis at screening or within 12 months thereafter.

All images were processed locally on our hardware by 3 different commercial AI CAD algorithms (AI-1, AI-2, and AI-3). The AI CAD algorithms have not been approved by the US Food and

Table 1. Area Under the Receiver Operating Characteristic Curve for the 3 Artificial Intelligence Algorithms

	AUC (95% CI)[a]		
Group (n = 8805)	Algorithm 1	Algorithm 2	Algorithm 3
Overall	0.956 (0.948-0.965)	0.92 (0.910-0.934)	0.92 (0.909-0.931)
By age, women, y			
Younger (<55)	0.925 (0.906-0.944)	0.882 (0.856-0.907)	0.889 (0.867-0.912)
Older (≥ 55)	0.974 (0.966-0.982)	0.943 (0.932-0.954)	0.938 (0.927-0.949)
By mammographic density[b]			
Dense area[c]			
Low	0.973 (0.964-0.981)	0.945 (0.932-0.959)	0.940 (0.926-0.954)
High	0.938 (0.923-0.954)	0.899 (0.879-0.918)	0.900 (0.882-0.917)
% Density[c]			
Low	0.976 (0.968-0.983)	0.954 (0.943-0.966)	0.950 (0.939-0.961)
High	0.933 (0.917-0.950)	0.886 (0.865-0.908)	0.886 (0.867-0.906)
By cancer detection mode			
Screen	0.984 (0.979-0.989)	0.959 (0.951-0.967)	0.952 (0.944-0.960)
Clinical	0.810 (0.767-0.852)	0.728 (0.677-0.779)	0.744 (0.696-0.792)

Abbreviation: AUC, area under the receiver operating characteristic curve.

[a] Test: Algorithm 1 has a higher AUC than the other 2 algorithms overall and for all subgroups ($P < .001$).

[b] Examination mean of all 4 views.

[c] Low represents below median; high, above median.

Drug Administration for use as independent readers. The vendors asked to remain anonymous, with the possibility for each vendor to later decide to waive anonymity. Each algorithm was described by the vendor according to the structure devised by the authors of this study (eAppendix in the Supplement). All 3 AI CAD algorithms processed the images, and no other data, and yielded a prediction score for each breast ranging between 0 and 1 for the suspicion of cancer, where 1 denotes the highest suspicion level. All analyses in this study were carried out on the examination level, which is equivalent to the patient level, based on the maximum score of the left or the right breast, whichever scored highest. The area under the receiver operating curve (AUC) was calculated for each of the 3 AI systems overall and by subgroups of age, mammographic density, and detection mode. To enable a comparison with the recorded binary decisions of radiologists, the output of each algorithm was dichotomized at a cut point corresponding to a specificity as close as possible to that of the first-reader radiologists (ie, 96.6%). Because our study sample was enriched with positive cases, we applied stratified bootstrapping (1000 samples) with a 14:1 ratio of healthy women to women receiving a diagnosis of cancer to mimic the ratio in the source screening cohort (approximately 0.5% screen-detected cancer among all screened women). We determined performance levels for all AI CAD algorithms and for all radiologist assessments (first reader, second reader, and consensus) for the following diagnostic metrics: sensitivity (number of true positives divided by all true), specificity (number of true negatives divided by all negative), accuracy (number of true positives plus true negatives divided by all), abnormal interpretation rate (number positive divided by all, multiplied by 1000), cancer detection rate (number of true positives divided by all, multiplied by 1000), false-negative rate (number of false negatives divided by all, multiplied by 1000) and positive predictive value (number of true positives divided by all positive, multiplied by 1000). We also investigated whether an association existed between the number of abnormal interpretations and the number of cases positive for cancer detected for all 3 AI CAD algorithms alone and also combined with the assessment of the first or second reader or both readers (the joint assessment was considered abnormal if at least 1 of the AI CAD algorithms or readers made an abnormal assessment). We examined the sensitivity and specificity when combining all 3 AI CAD algorithms (the joint assessment was considered abnormal if at least 1 AI CAD algorithm made an abnormal assessment), as well as when combining all algorithms and radiologists (the joint assessment was considered abnormal if at least 2 of the AI CAD algorithms or radiologists made an abnormal assessment).

The computer software Stata, version 15.1 (StataCorp), was used for all statistical analyses. All statistical tests were 2-sided. The level for statistical significance was set at $\alpha = .05$. We tested for differences in the AUC using the DeLong method. The AUC CIs were estimated by the sandwich variance estimator.

Results

The final study sample, as described in eTable 1 in the Supplement, consisted of 8805 women and screening examinations, of whom 739 women received a diagnosis of breast cancer (positive) and a random sample of 8066 women were healthy controls (negative). The median age at screening was 54.5 years (interquartile range, 47.4-63.5 years), and the median age at diagnosis was 59. 8 years (interquartile range, 49.8-65.8 years). The median age for healthy controls was 54 years (interquartile range, 47-63 years). The median age for cases was 60 years (interquartile range, 50-66 years). The positive cases consisted of 618 (84%) screen-detected cancer cases and 121 (16%) clinically detected cancer cases within 12 months of the screening examination. Of those, 640 cases of cancer had an invasive component and 85 were in situ only.

Table 1 reports the AUC for cancer detection for each AI algorithm overall and by subgroups. Overall, the AUC was 0.956 (95% CI, 0.948-0.965) for AI algorithm 1 (AI-1), 0.922 (95% CI, 0.910-0.934) for AI-2, and 0.920 (95% CI, 0.909-0.931) for AI-3.

Table 2. Screening Performance Benchmarks for Artificial Intelligence Algorithms and for Radiologists in 739 Women Who Received a Diagnosis of Breast Cancer and 112 924 Healthy Women

	Benchmark point estimate (95% CI)[a]					
	Algorithm[b]			Reader		
Benchmark	1	2	3	First	Second	Consensus
Specificity, %	96.6 (96.5-96.7)	96.6 (96.5-96.7)	96.7 (96.6-96.8)	96.6 (96.5-96.7)	97.2 (97.1-97.3)	98.5 (98.4-98.6)
Sensitivity, %	81.9 (78.9-84.6)	67.0 (63.5-70.4)	67.4 (63.9-70.8)	77.4 (74.2-80.4)	80.1 (77.0-82.9)	85.0 (82.2-87.5)
Accuracy, %	96.5 (96.4-96.6)	96.4 (96.3-96.5)	96.5 (96.4-96.6)	96.5 (96.4-96.6)	97.1 (97.0-97.1)	98.4 (98.3-98.5)
PPV, %	13.6 (12.5-14.7)	11.4 (10.5-12.4)	11.8 (10.8-12.8)	13.0 (12.0-14.0)	15.9 (14.7-17.1)	27.2 (25.4-29.1)
AIR	39.1 (38.0-40.2)	38.1 (37.0-39.2)	37.3 (36.2-38.4)	38.8 (37.7-39.9)	32.8 (31.8-33.9)	20.3 (19.5-21.1)
CDR	5.32 (4.91-5.76)	4.36 (3.98-4.76)	4.38 (4.00-4.78)	5.03 (4.63-5.46)	5.21 (4.80-5.64)	5.53 (5.10-5.97)
FNR	0.181 (0.154-0.211)	0.330 (0.296-0.364)	0.330 (0.296-0.364)	0.226 (0.196-0.256)	0.177 (0.150-0.205)	0.150 (0.124-0.176)

Abbreviations: AIR, abnormal interpretation rate (per 1000 examinations); CDR, cancer detection rate (per 1000 examinations); FNR, false-negative rate (per cancer diagnosed within 12 months); PPV, positive predictive value.

[a] Benchmark estimates based on stratified bootstrapping to attain a proportion of women who received a diagnosis of breast cancer to healthy women similar to the source screening cohort (approximately 0.5%).

[b] The operating point of each algorithm was set at a specificity as close as possible to that of the first reader (96.6%).

The differences between AI-1 and each of the other 2 AI CAD algorithms (AI-2 and AI-3) were statistically significant ($P < .001$), whereas there was no significant difference between AI-2 and AI-3 ($P = .68$). Within all analyzed subgroups, AI-1 had a significantly higher AUC than AI-2 and AI-3 ($P < .001$), whereas there was no significant difference between AI-2 and AI-3 for any subgroup. Specifically, the AUC for clinically detected cancer after negative radiologist assessment was 0.810 (95% CI, 0.767-0.852) for AI-1, 0.728 (95% CI, 0.677-0.779) for AI-2, and 0.744 (95% CI, 0.696-0.792) for AI-3. In addition, we observed that the AUC for younger vs older, and for higher vs lower breast density, were significantly lower for all algorithms. For AI-1, the AUC was 0.974 for women 55 years or older and 0.925 for women younger than 55 years; 0.933 for mammograms with high percent density and 0.976 for mammograms with low percent density. In a secondary analysis, after extending the study population with the 174 women who received a diagnosis of cancer between 12 and 23 months after screening, the AUC was 0.916 (95% CI, 0.905-0.928) for AI-1, 0.859 (95% CI, 0.843-0.874) for AI-2, and 0.877 (95% CI, 0.964-0.890) for AI-3. A box plot of the raw estimated scores of each algorithm is shown in eFigure 4 in the Supplement.

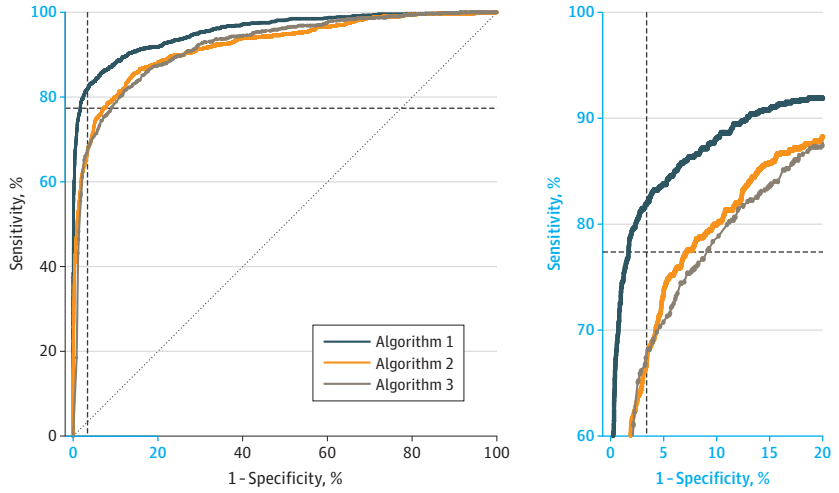
The results of the comparisons with radiologists' assessments are presented in **Table 2**. The total simulated screening population consisted of 113 663 examinations (of which 739 were diagnosed as positive for breast cancer). The sensitivity was 81.9% (95% CI, 78.9%-84.6%) for AI-1, 67.0% (95% CI, 63.5%-70.4%) for AI-2, and 67.4% (95% CI, 63.9%-70.8%) for AI-3, 77.4% (95% CI, 74.2%-80.4%) for the first reader, 80.1% (95% CI, 77.0%-82.9%) for the second reader, and 85.0% (95% CI, 82.2%-87.5%) for the consensus discussion. There was a significant sensitivity difference between AI-1 and the other 2 AI CAD algorithms ($P < .001$) as well as between AI-1 and the first reader ($P = .03$). However, the analysis did not show a difference between AI-1 and the second reader ($P = .40$) or the consensus discussion ($P = .11$). Specificity for the AI CAD algorithms was preselected to match the specificity of the first reader and should therefore not be compared. The specificity for the second reader was 97.2% (95% CI, 97.1%-97.3%), and for the consensus discussion, it was 98.5% (95% CI, 98.4%-98.6%). Potential cut points for the continuous score of each algorithm to achieve various sensitivity levels are presented in eTable 2 in the Supplement. When choosing an operating point corresponding to the Breast Cancer Surveillance Consortium benchmark of 88.9% specificity, the sensitivity was 88.6% for AI-1, 80.0% for AI-2, and 80.2% for AI-3 (eTable 3 in the Supplement). Examples of mammograms of cancer identified by AI CAD but missed by radiologists, and vice versa, are shown in eFigure 2 and eFigure 3 in the Supplement.

The results of the combined assessment across all 3 algorithms showed a sensitivity of 86.7% (95% CI, 84.2%-89.2%) and a specificity of 92.5% (95% CI, 92.3%-92.7%). Compared with the best algorithm, that is, AI-1, the combined system had a marginally higher sensitivity ($P = .01$) but a much lower specificity ($P < .001$). As a comparison, AI-1 alone achieved 86.3% sensitivity at 92.5% specificity, and 79.3% sensitivity at 98.0% specificity (**Figure**).

Table 3 gives the simulated scenarios in which the binary decisions by the 3 AI CAD algorithms and the readers were combined. Of 739 total cancer cases, there were 655 screening examinations assessed as abnormal for the first reader combined with AI-1 (88.6% sensitivity), 620 combined with AI-2 (83.9% sensitivity), 623 combined with AI-3 (84.3% sensitivity), and 640 combined with the second reader (86.6% sensitivity). Of 113 663 total examinations in the simulated screening cohort, there were 7851 examinations assessed as abnormal for the first reader combined with AI-1 (93.0% specificity), 7998 combined with AI-2 (92.9% specificity), 7847 combined with AI-3 (92.9% specificity), and 5484 combined with the second reader (95.1% specificity). For the first reader, the relative increase in cancer detection was 15% when adding AI-1 and 12% when adding the second reader; the relative increase in abnormal interpretations was 78% when adding AI-1, and 24% when adding the second reader. To examine these results separated into in situ cancer, invasive cancer, and stage II or higher breast cancer, please see eTable 4 in the Supplement.

When combining all 3 algorithms and 2 reader radiologists (at least 2 had to make a positive assessment), the estimated sensitivity was 87.4% (95% CI, 85.0%-89.8%), and the estimated specificity was 95.9% (95% CI, 95.7%-96.0%).

Figure. Receiver Operating Characteristic Curves for the 3 Artificial Intelligence Computer-Aided Detection Algorithms



As a comparison, for first-reader radiologists, the vertical dashed line represents the mean 1 − specificity (1 − 0.966 = 0.034); and the horizontal dashed line, the mean sensitivity (0.774). The right panel is a magnification of the vertical line and horizontal line intersection.

Table 3. Number of Abnormal Interpretations and Cases Positive for Cancer Detected by Algorithms and Readers Alone and by Algorithms Combined With the Assessment of the First, Second, or Both Readers

	No. (% increase vs alone)				
	Algorithm			Reader	
Assessment	1	2	3	First	Second
Abnormal interpretation[a]					
Alone	4441	4331	4236	4408	3728
With first reader	7851 (77)	7998 (85)	7847 (85)	NA	5484 (47)
With second reader	7188 (62)	7260 (68)	7139 (69)	5484 (24)	NA
With both readers	8745 (97)	8885 (105)	8762 (107)	NA	NA
Cancer detected[b]					
Alone	605	495	498	572	592
With first reader	655 (8)	620 (25)	623 (25)	NA	640 (8)
With second reader	664 (10)	638 (29)	643 (29)	640 (12)	NA
With both readers	667 (10)	653 (32)	656 (32)	NA	NA

Abbreviation: NA, not applicable.

[a] Based on a total of 113 663 screenings. Observations of healthy women have been duplicated to attain a similar proportion as in the source screening cohort (0.5% with a diagnosis of cancer).

[b] Actual screen-detected cancer (n = 618); actual clinically detected cancer (n = 121).

The sensitivity of the combined algorithms and radiologists was higher than AI-1 (P = .003) and higher than second readers (P < .001). The specificity was somewhat lower than AI-1 (P < .001), the second readers (P < .001), and the consensus decision (P < .001).

Discussion

The present study observed 3 main findings. First, a difference was found in the AUC among the 3 AI CAD algorithms, from 0.920 to 0.956. Second, the best computer algorithm reached, and in some comparisons surpassed, the performance level of radiologists in assessing screening mammograms, obtaining 81.9% sensitivity when operating at 96.6% specificity in a simulated study population of 113 663 screening examinations based on an original sample of 8805 women from a population-based screening cohort. Third, combining the first reader with the best algorithm identified more cancer cases than combining the first and second readers.

The proportion of clinically detected interval cancer in our study was 16%, which is lower than the 28% reported in a prior European study.[14] The lower proportion may be explained by our exclusion criteria for cancer diagnosed later than 12 months after screening, which was chosen to increase the likelihood that cancer was present in the breast at the time of examination.

The best-performing algorithm, AI-1, had an overall AUC of 0.956 for the detection of cancer at screening or within 12 months thereafter. The 2 other AI CAD algorithms had an overall AUC of 0.922 and 0.920. Prior studies have reported AUC values between 0.840 and 0.959.[9-11,15-18] The subgroup analysis of the AUC in our study showed a decreased performance for younger vs older women and for higher vs lower breast density on mammography. This is in line with prior studies showing that there is an increase of interval cancer cases, that is, decreased mammographic sensitivity, for younger women and for women with higher mammographic density.[19-22] In our specific analysis of interval cancer detected within 12 months after a negative screening examination, AI-1 achieved an AUC

of 0.810, suggesting that there is potential for the AI algorithms to promote earlier cancer detection and that there are suspicious findings present in many of those mammograms. The AI-1 algorithm was superior to the other 2 algorithms across all subgroups. The differences between AI-2 and AI-3 were minimal across all subgroups. The cause of the stronger performance of AI-1 was not the subject of this study. However, our reading of the algorithm descriptions submitted by the vendors (eAppendix in the Supplement) revealed the following differences, which might be part of the explanation; AI-1 was trained on more data than the other 2 were, had pixel-level annotations for training, and had a higher capacity backbone (ResNet34); in addition, the data augmentation included adjustment of contrast and brightness. The largest training population for the superior performing AI-1 consisted of images from GE equipment and images of South Korean women. Although we do not have ethnic descriptors of our study population, the vast majority of women in Stockholm are White, and all images in our study were acquired on Hologic equipment. Against this background, the superior performance of AI-1 is an interesting example of robustness. In training AI algorithms for mammographic cancer detection, matching ethnic and equipment distributions between the training population and the clinical test population may not be of highest importance.

When comparing binary decisions by AI CAD algorithms, operating at the specificity level of the first reader, with the actual recorded assessment by radiologists, we concluded that AI-1 showed superior performance to the other 2 AI CAD algorithms and to the first readers. The specificity was 96.6% by design, and the resulting sensitivity of AI-1 was 81.9%; when using an operating point corresponding to 88.9% specificity, the resulting sensitivity was 88.6%. This can be compared with the Breast Cancer Surveillance Consortium benchmarks of 86.9% sensitivity at 88.9% specificity.[5] In this retrospective analysis, it is apparent that AI-1 fulfills the specificity and sensitivity criteria. It is well known that the specificity of European breast radiologists is generally higher and the sensitivity lower than US colleagues.[23] It is important to bear in mind that in contrast to the original readers, the AI CAD algorithms did not benefit from information from prior mammograms nor of patient reports of breast symptoms.

The foregoing discussion focuses on comparing various algorithms and radiologists when applied separately. However, based on double-reading screening programs, we know that combining assessments can improve performance. Taking the first-reader assessments as the starting point, we found that 2 human readers showed more agreement regarding abnormal interpretations and for false positives than an algorithm and a human reader did. When adding an AI CAD algorithm to the first reader, more true-positive cases would likely be found, but a much larger proportion of false-positive examinations would have to be handled in the ensuing consensus discussion. Changing the perspective by taking the AI CAD assessments as the starting point, cancer detection was estimated to increase by merely 8% when combined with the first reader, whereas the abnormal assessments (true positives plus false positives) increased by 77%. Even though there is an absolute diagnostic gain when adding the first reader to the AI CAD algorithm, a cost-benefit analysis is required, in any given setting, to determine the economic implications of adding a human reader at all.

When assessing the combination of 3 algorithms based on voting systems, we found that combining all 3 algorithms did not achieve a markedly higher performance than using the best algorithm alone. Likewise, compared with the diagnostic performance of combining the best algorithm with the first reader, we found no clear advantage of having a voting system involving all algorithms and all readers. Given that commercial AI systems will likely come with a price that is a notable proportion of the cost of the corresponding radiologist time, we view the most realistic implementation to be 1 radiologist and 1 algorithm. However, as shown by the 77% increase in abnormal findings, even though this implementation will obviate the need for 1 radiologist in screening assessment, it would increase the workload for the 2 radiologists involved in the consensus discussions. Before clinical practice can change, it is critically important to conduct prospective clinical studies as well as a thorough examination of ethical, legal, and societal aspects of replacing a medical professional by computer software.

As a development of regular, 2-dimensional (2-D) mammography, on which the present study was based, digital breast tomosynthesis, or 3-D mammography, has become increasingly accepted as an alternative screening modality. There are reports that the interval cancer rate decreases with the use of tomosynthesis.[24] Since tomosynthesis appears to make some signs of cancer more conspicuous and easier to identify for radiologists, it will be an interesting topic for future studies to examine whether AI algorithms applied to 2-D mammography reduces the clinical utility of 3-D mammography, or whether AI algorithms trained on 3-D mammography can further improve the diagnostic performance.

Strengths and Limitations

The major strength of our study is that we performed an independent evaluation of several AI CAD algorithms, none of which had ever been exposed to images from our institution. Additional strengths included the comparative aspect between AI CAD algorithms and radiologists, and that our large set of examinations was chosen in a representative manner from a population-based screening cohort. The study limitations are that our results applied to a version of each algorithm that has already been replaced by a more recent algorithm, that the examinations were from a Swedish setting, and that we did not analyze the performance for women with implants or prior breast cancer. For computational reasons, we used a cancer-enriched cohort. We therefore used inverse probability weighted bootstrapping to simulate a study population with a cancer prevalence matching a screening cohort to address issues raised for studies of cancer-enriched study populations.[25] A weakness of our study is that the AI CAD algorithms did not consider prior mammograms, hormonal medication, or breast symptoms—which puts AI CAD algorithms at a disadvantage compared with radiologists.

Conclusions

In conclusion, our results suggested that the best computer algorithm evaluated in this study assessed screening mammograms with a diagnostic performance on par with or exceeding that of radiologists in a retrospective cohort of women undergoing regular screening. This achievement is considerable, bearing in mind that radiologists, but not AI algorithms, had proprietary access to certain information. We believe that the time has come to evaluate AI CAD algorithms as independent readers in prospective clinical studies in mammography screening programs.

ARTICLE INFORMATION

Accepted for Publication: June 2, 2020.

Published Online: August 27, 2020. doi:10.1001/jamaoncol.2020.3321

Author Affiliations: Department of Oncology-Pathology, Karolinska Institute, Stockholm, Sweden (Salim, Azavedo, Foukakis, Strand); Department of Radiology, Karolinska University Hospital, Stockholm, Sweden (Salim, Foukakis); Department of Medical Radiation Physics and Nuclear Medicine, Karolinska University Hospital, Stockholm, Sweden (Wåhlin); Department of Physiology and Pharmacology, Karolinska Institute, Stockholm, Sweden (Dembrower); Department of Radiology, Capio Sankt Görans Hospital, Stockholm, Sweden (Dembrower); Department of Molecular Medicine and Surgery, Karolinska Institute, Stockholm, Sweden (Azavedo); Division of Computational Science and Technology, KTH Royal Institute of Technology, Science for Life Laboratory, Solna, Sweden (Liu); KTH Royal Institute of Technology, Science for Life Laboratory, Solna, Sweden (Smith); Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden (Eklund); Breast Radiology, Karolinska University Hospital, Stockholm, Sweden (Strand).

Author Contributions: Drs Salim and Strand had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Salim, Azavedo, Smith, Eklund, Strand.
Acquisition, analysis, or interpretation of data: Salim, Wåhlin, Dembrower, Foukakis, Liu, Eklund, Strand.
Drafting of the manuscript: Salim, Eklund, Strand.
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Salim, Smith, Eklund, Strand.
Obtained funding: Strand.
Administrative, technical, or material support: Wåhlin, Dembrower, Liu, Strand.
Supervision: Azavedo, Foukakis, Smith, Eklund, Strand.

Conflict of Interest Disclosures: Dr Foukakis reported receiving grants from Pfizer outside the submitted work. Dr Eklund reported receiving grants from Swedish Research Council and from the Swedish Cancer Society during the conduct of the study. Dr Strand reported receiving grants from Stockholm City Council during the conduct of the study; and receiving personal fees from Collective Minds Radiology outside the submitted work. No other disclosures were reported.

Funding/Support: This study was funded by the Stockholm County Council Dnr 20170802 award.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Contributions: The 3 commercially available artificial intelligence algorithms assessed in this study were provided by the 3 companies that developed them; no exchange of financial funds or other benefits were involved.

Additional Information: This work originated in the Department of Oncology-Pathology, Karolinska Institute, Karolinska.

REFERENCES

1. 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(11):2205-2240. doi:10.1038/bjc.2013.177

2. Seely JM, Alhassan T. Screening for breast cancer in 2018—what should we be doing today? *Curr Oncol*. 2018;25(suppl 1):S115-S124. doi:10.3747/co.25.3770

3. Giess CS, Wang A, Ip IK, Lacson R, Pourjabbar S, Khorasani R. Patient, radiologist, and examination characteristics affecting screening mammography recall rates in a large academic practice. *J Am Coll Radiol*. 2019;16(4, pt A):411-418. doi:10.1016/j.jacr.2018.06.016

4. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. *J Natl Cancer Inst*. 2004;96(24):1840-1850. doi:10.1093/jnci/djh333

5. Lehman CD, Arao RF, Sprague BL, et al. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. *Radiology*. 2017;283(1):49-58. doi:10.1148/radiol.2016161174

6. Elmore JG, Jackson SL, Abraham L, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology*. 2009;253(3):641-651. doi:10.1148/radiol.2533082308

7. Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson AN, Miglioretti DL; Breast Cancer Surveillance Consortium. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med*. 2015;175(11):1828-1837. doi:10.1001/jamainternmed.2015.5231

8. Ciatto S, Del Turco MR, Risso G, et al. Comparison of standard reading and computer aided detection (CAD) on a national proficiency test of screening mammography. *Eur J Radiol*. 2003;45(2):135-138. doi:10.1016/S0720-048X(02)00011-6

9. 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(9):916-922. doi:10.1093/jnci/djy222

10. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94. doi:10.1038/s41586-019-1799-6

11. Akselrod-Ballin A, Chorev M, Shoshan Y, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology*. 2019;292(2):331-342. doi:10.1148/radiol.2019182622

12. Wu K, Wu E, Wu Y, et al. Validation of a deep learning mammography model in a population with low screening rates. Preprint. arXiv: 1911.00364v1. Posted online November 1, 2019.

13. Dembrower K, Lindholm P, Strand F. A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks—the Cohort of Screen-Aged Women (CSAW). *J Digit Imaging*. 2020;33(2):408-413. doi:10.1007/s10278-019-00278-0

14. Törnberg S, Kemetli L, Ascunce N, et al. A pooled analysis of interval cancer rates in six European countries. *Eur J Cancer Prev*. 2010;19(2):87-93. doi:10.1097/CEJ.0b013e32833548ed

15. 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 Digital Health*. Published online February 6, 2020. doi:10.1016/S2589-7500(20)30003-0

16. Wu N, Phang J, Park J, et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans Med Imaging*. 2020;39(4):1184-1194. doi:10.1109/TMI.2019.2945514

17. 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(2):305-314. doi:10.1148/radiol.2018181371

18. Schaffter T, Buist DSM, Lee CI, et al; and the DM DREAM Consortium. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Netw Open*. 2020;3(3):e200265. doi:10.1001/jamanetworkopen.2020.0265

19. Buist DSM, Porter PL, Lehman C, Taplin SH, White E. Factors contributing to mammography failure in women aged 40-49 years. *J Natl Cancer Inst*. 2004;96(19):1432-1440. doi:10.1093/jnci/djh269

20. Boyd NF, Guo H, Martin LJ, et al. Mammographic density and the risk and detection of breast cancer. *N Engl J Med*. 2007;356(3):227-236. doi:10.1056/NEJMoa062790

21. Klemi PJ, Toikkanen S, Räsänen O, Parvinen I, Joensuu H. Mammography screening interval and the frequency of interval cancers in a population-based screening. *Br J Cancer*. 1997;75(5):762-766. doi:10.1038/bjc.1997.135

22. Tabár L, Faberberg G, Day NE, Holmberg L. What is the optimum interval between mammographic screening examinations? an

analysis based on the latest results of the Swedish two-county breast cancer screening trial. *Br J Cancer*. 1987;55(5):547-551. doi:10.1038/bjc.1987.112

23. Domingo L, Hofvind S, Hubbard RA, et al. Cross-national comparison of screening mammography accuracy measures in U.S., Norway, and Spain. *Eur Radiol*. 2016;26(8):2520-2528. doi:10.1007/s00330-015-4074-8

24. McDonald ES, Oustimov A, Weinstein SP, Synnestvedt MB, Schnall M, Conant EF. Effectiveness of digital breast tomosynthesis compared with digital mammography: outcomes analysis from 3 years of breast cancer screening. *JAMA Oncol*. 2016;2(6):737-743. doi:10.1001/jamaoncol.2015.5536

25. Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. *Expert Rev Med Devices*. 2019;16(5):351-362. doi:10.1080/17434440.2019.1610387

Invited Commentary

Artificial Intelligence to Support Independent Assessment of Screening Mammograms—The Time Has Come

Constance Dobbins Lehman, MD, PhD

Screening mammography is our best method currently available to detect breast cancer early, when it can be cured. However, global access to high-quality, affordable screening mammography is constrained by the limited supply of radiologists subspecialized in breast imaging to interpret each individual examination. The need for interpretation of each mammogram by a subspecialist not only increases costs and limits access to screening but also adds the element of human error to even the most advanced screening programs. Owing to well-documented human error and variation, there is no "diagnostic accuracy" of screening mammography but rather a wide range of performance outcomes based on the individual radiologist interpreting the mammogram. In a study of more than 1.6 million modern, all-digital screening mammograms, investigators of the Breast Cancer Surveillance Consortium found a wide range of interpretive performance across radiologists, with more than 40% of certified, specialized radiologists failing to meet recommended recall rates.[1] Recognition of these challenges supported early efforts to develop deep learning models to assist humans in mammographic interpretation.[2-4] However, the outcomes have been mixed, with wide variation in quantity and quality of data used for model development, variable methods to train, test, and internally and externally validate models developed, and inconsistent use of peer-reviewed publications to share discoveries.

Related article page 1581

In this issue of *JAMA Oncology*, Salim and colleagues[5] use a large, curated screening mammography database to carefully measure the performance of 3 commercial artificial intelligence (AI) algorithms. The test set used for this external validation included 739 mammograms associated with breast cancer and 8066 randomly sampled mammograms that were negative for breast cancer (for an enriched prevalence of 84 cancers per 1000 screening mammograms). Human performance estimates were derived from prior studies in a double-reader setting (in which the first reading, second reading, and consensus readings were recorded). The AI performance was estimated in a variety of settings. First, the AI algorithm score for suspicion of cancer on the mammogram (a continuous scale from 0 to 1.0) was translated to a binary "positive" or "negative" examination result by using a threshold set for a specificity equal to the mean of the first human reader (96.6%). How the human reader would respond to the input of the AI algorithm continuous score of suspicion of cancer was assumed by counting all cancers on mammograms with AI scores above the set threshold as "detected." The areas under the receiver operating characteristic curves were estimated for each of the algorithms, as was the full spectrum of mammography performance metrics (including sensitivity, specificity, accuracy, positive predictive value, abnormal interpretation rate, cancer detection rate, and false-negative rate). Finally, the AI model performance with or without first and second readers' inputs was estimated in diverse simulations, which included enriching the testing set by duplicating the true-negatives from the original 8066 mammograms negative for breast cancer up to 112 914 negative mammograms (to reduce the original rate of 84 cancers per 1000 to 6.5 cancers per 1000).

Using these methods, the authors found that 1 of the 3 models achieved a sensitivity of 81.9% with the specificity set at 96.6%. The authors also simulated an operating point corresponding to 88.9% specificity, which produced a sensitivity of 88.6%. Those results compare favorably with the US Breast Cancer Surveillance Consortium benchmarks of 86.9% sensitivity at 88.9% specificity. As a final analysis, the authors simulated a double-reading scenario in which the AI algorithm served as a first reader, followed by a human second reader. The assumption was made that any exam with an AI risk estimate above the threshold would result in a cancer diagnosis by the final human reader. When added to the first reader assessment, cancer detection increased by 8% but at a cost of increasing examination results considered abnormal by 77%.

The authors are to be commended for using a modern, all-digital screening mammography database to compare performance of 3 commercial AI algorithms. They provide comprehensive performance metrics, enabling the reader to compare the balance of false-negatives and false-positives as the algorithm thresholds are adjusted. They provide interesting insights that challenge existing assumptions in the field. For example, their results suggest that the volume of cases may be more important than the diversity of vendors or patient populations in the databases used to develop the algorithm. The highest performing algorithm was developed from the largest database of screening mammograms (72 000 cancer images and 680 000 normal images in the top performing model compared with 6000 cancer images and 106 000 normal im-