Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Improving Breast Cancer Detection Accuracy of Mammography with the Concurrent Use of an Artificial Intelligence Tool

Serena Pacilè, PhD • January Lopez, MD • Pauline Chone, MPhil • Thomas Bertinotti, MSc • Jean Marie Grouin, PhD • Pierre Fillard, PhD

From Therapixel SA, 39 Rue Claude Daunesse, 06560 Valbonne, France (S.P., P.C., T.B., P.F.); Radiology & Imaging Services, Hoag Memorial Hospital Presbyterian, Newport Beach, Calif (J.L.); and Department of Statistics, University of Rouen, Rouen, France (J.M.G.). Received November 20, 2019; revision requested February 16, 2020; revision received June 27; accepted July 7. **Address correspondence to** S.P.

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2020; 2(6):e190208 • **https://doi.org/10.1148/ryai.2020190208** • **Content codes:** AI BR

Purpose: To evaluate the benefits of an artificial intelligence (AI)–based tool for two-dimensional mammography in the breast cancer detection process.

Materials and Methods: In this multireader, multicase retrospective study, 14 radiologists assessed a dataset of 240 digital mammography images, acquired between 2013 and 2016, using a counterbalance design in which half of the dataset was read without AI and the other half with the help of AI during a first session and vice versa during a second session, which was separated from the first by a washout period. Area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and reading time were assessed as endpoints.

Results: The average AUC across readers was 0.769 (95% CI: 0.724, 0.814) without AI and 0.797 (95% CI: 0.754, 0.840) with AI. The average difference in AUC was 0.028 (95% CI: 0.002, 0.055, P = .035). Average sensitivity was increased by 0.033 when using AI support (P = .021). Reading time changed dependently to the AI-tool score. For low likelihood of malignancy (< 2.5%), the time was about the same in the first reading session and slightly decreased in the second reading session. For higher likelihood of malignancy, the reading time was on average increased with the use of AI.

Conclusion: This clinical investigation demonstrated that the concurrent use of this AI tool improved the diagnostic performance of radiologists in the detection of breast cancer without prolonging their workflow.

Supplemental material is available for this article.

Breast cancer screening programs are currently implemented in most developed countries and have been shown to increase earlier stage breast cancer detection leading to improved prognosis and reduced mortality (1). It is estimated that a small, node-negative tumor (less than 10 mm in size) can be successfully treated in about 90% of cases, whereas this value drops to about 55% in the case of local-regional nodal involvement and 18% in the case of distant metastases (2).

Mammography has been the frontline screening tool for breast cancer for decades with more than 200 million women being examined each year around the globe (3). However, limitations in sensitivity and specificity persist even in the face of the most recent technologic improvements. Up to 30% to 40% of breast cancers can be missed during screening and on average, only 10% of women recalled from screening for diagnostic workup are ultimately found to have cancer (4).

Traditional computer-aided detection systems were introduced in previous years with the intent to improve the performance of radiologists. In the United States, computer-aided detection systems are used in 83% of digital mammography examinations (5). However, literature on their efficacy is controversial. Some studies have shown improved cancer detection similar to that of double reading when using computer-aided detection systems (6–8). Others have demonstrated conflicting results, including a large study evaluating the performance of radiologists at 43 facilities over a 4-year period, which found that the use of computer-aided detection led to decreased accuracy in cancer detection as well as an increase in biopsy recommendations (9).

Given the growing interest in the use of artificial intelligence (AI) in medical imaging, several newer algorithms based on deep learning have been developed and applied to digital mammography. Preliminary investigations have demonstrated that the use of AI systems as concurrent readers for interpreting mammograms can improve efficiency of the radiologist in terms of time, sensitivity, and specificity (10–14).

Our work described a multireader, multicase clinical investigation carried out to test the hypothesis that the use of a new AI system can improve the performance of radiologists in breast cancer detection when reading digital screening mammography.

Abbreviations

AI = artificial intelligence, AUC = area under the ROC curve, BI-RADS = Breast Imaging Reporting and Data System, ICC = intraclass correlation coefficient, ROC = receiver operating characteristic

Summary

A multireader, multicase retrospective study demonstrated that the use of an artificial intelligence–based tool significantly improved the average area under the receiving operating characteristic curve across radiologists.

Key Points

- The investigation has shown that the use of artificial intelligence (AI) can decrease the false-negative rate of radiologists without affecting their specificity.
- The results of this study suggest also that incorporating AI-based machines into the process of evaluation of mammograms can improve the performance of radiologists.
- An improved diagnostic performance of radiologists in the mammographic detection of breast cancer is achievable without having an impact on their overall reading time.

Materials and Methods

Study Design

The retrospective study was conducted in accordance to the Health Insurance Portability and Accountability Act and approved by an institutional review board. The study was a multireader, multicase study with fully crossed design. Data were retrospectively collected spanning a 3-year period starting from January 1, 2016. Only examinations from women presenting for screening without clinical symptoms were included. Exclusion criteria included current and/or recent history of breast feeding, breast reduction or implant augmentation, and history of breast cancer. All examinations meeting the study criteria were classified into four categories: true positive, false negative, true negative, and false positive. True-positive cancer cases were verified by histopathologic evaluation, whereas false-positive cases were assessed either by a negative biopsy result (25%, 10 of 40) or a negative result at follow-up for at least 18 months (75%, 30 of 40). True-negative cases were verified by a negative result at follow-up of 18 months. False-negative cases were defined as screening negative cases with a positive result at follow-up within 18 months. Before being included in the final dataset, data underwent a quality check performed by an experienced breast radiologist not taking part in the reading sessions, with the aim of excluding examinations not meeting acquisition standards or presenting identifiable features (eg, nipple retraction, invasive cancer larger than approximately 2.5 cm, bilateral cancer, and others to minimize recall bias [15]), and confirming that, for false-negative examinations, malignant lesions were visible and identifiable in retrospect (Fig 1). The final selected dataset included 240 patient cases (average age, 59 years; range, 37–85 years) with 80 true-positive, 40 false-negative, 80 true-negative, and 40 false-positive cases. Demographic and histopathologic characteristics of the selected patients are summarized, respectively, in Tables 1 and 2.

Fourteen reader participants read cases over two reading sessions separated by a washout period of 4 weeks. Cases were read twice using a counterbalance design in which half of the cases were read with AI and half without AI for the first reading session and vice versa for the second reading session. For each reading session, subgroups of cases to be read with and without AI contained the same distribution of true positive, false positive, true negative, and false negative, whereas case reading order was randomized separately for each reader. The design was based on similar studies used to test AI-based systems for breast imaging (11,16,17).

AI System

The AI system we used (MammoScreen V1; Therapixel, Nice, France) is designed to identify regions suspicious for breast cancer on two-dimensional digital mammograms and assess their likelihood of malignancy. The system takes as input the complete set of four views composing a mammogram (left and right craniocaudal and mediolateral oblique images) and outputs a set of image positions with a related suspicion score. The system uses two groups of deep convolutional neural networks combined together with an aggregation module. A detailed description of the system is given in Appendix E1 (supplement).

Reader Test

Fourteen radiologists were involved in the reader study, all of whom were certified by the American Board of Radiology and Mammography Quality Standards Act and Program and breast imaging fellowship trained. Readers' years in practice varied from 0 to 25 years (median, 8.5 years). Most readers (93%) devoted at least 50% of their practice to breast imaging, reading more than 3000 mammograms per year.

Readers evaluated the cases independently, with an individually randomized order (ie, each reader read the cases in a different order). They had no access to any information about the patient (eg, previous mammography and other imaging examinations) and were told that the dataset to be assessed was enriched with cancer cases, without specifying in which proportions. The AI system, as well as the readers, had access to and used both the craniocaudal and mediolateral oblique view. For each case, readers provided a forced Breast Imaging Reporting and Data System (BI-RADS) score of 1 to 5 and a level of suspicion of 0 to 100 at the time of interpretation. Starting from a level of suspicion of 40, cases were considered as assessed positive. Reading times per reader per case were recorded; the measure of the reading time per case went from the opening of a new case until the validation of the level of suspicion and forced BI-RADS attributed by the reader. Readers were informed of the time being recorded but were blinded to the measurement.

Statistical Analysis

The main goal was to demonstrate the superiority of the performance of the radiologists reading with the AI support with respect to radiologists reading unaided. The sample size to evaluate superiority was based on previous similar studies (11,18) and calculated using the Obuchowski-Rockette

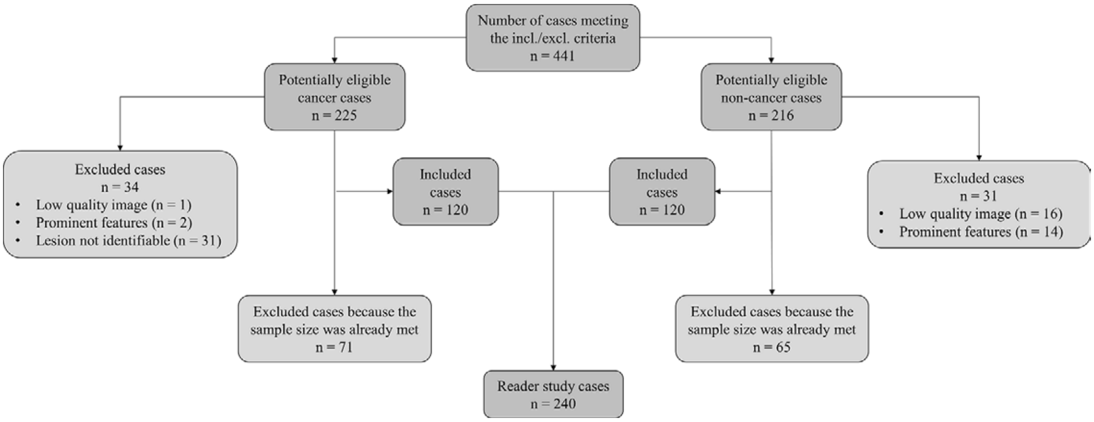
Number of cases meeting the incl./excl. criteria n = 441
Potentially eligible cancer cases n = 225
Potentially eligible non-cancer cases n = 216
Excluded cases n = 34
• Low quality image (n = 1)
• Prominent features (n = 2)
• Lesion not identifiable (n = 31)
Included cases n = 120
Included cases n = 120
Excluded cases n = 31
• Low quality image (n = 16)
• Prominent features (n = 14)
Excluded cases because the sample size was already met n = 71
Excluded cases because the sample size was already met n = 65
Reader study cases n = 240

Figure 1: Dataset selection flowchart.

Table 1: Demographic and Clinical Data of the Population and Digital Mammographic Examinations Selected for the Study

Characteristic	Value
Age (y)	
Mean	59
Median	60
Range	37–85
Interquartile range	50–68
Median breast thickness (mm)	59.88 (52–68)*
Mean glandular dose (mGy)	1.79 (1.27–2.18)*
Breast density†	
Category A	15 (36/240)
Category B	43.75 (105/240)
Category C	34.58 (83/240)
Category D	6.67 (16 /240)

* Interquartile range shown.
† Data are percentages with numbers in parentheses.

Table 2: Characteristic of the 120 Malignant Cancers Included in the Selected Dataset

Characteristic	No. of Cases
Histologic type	
Invasive ductal carcinoma	75
Ductal carcinoma in situ	27
Invasive lobular carcinoma	6
Other	12
Lesion type	
Mass	64
Calcification	30
Asymmetry	13
Architectural distortion	13

approach (19). Endpoints were the area under the receiver operating characteristic (ROC) curve (AUC), specificity, sensitivity, and reading time. The difference in AUC, specificity, and sensitivity for each reader under each condition was estimated with the trapezoidal method and analyzed using two-sided 95% CI and *P* value. *P* values less than .05 were considered indicative of statistical significance. The difference in mean AUC between the two reading conditions was estimated with the Obuchowski-Rockette model, assuming a 0.03 difference in mean AUCs, with an 80% power at the nominal two-sided level of significance (20). The covariances of the errors were estimated using the jackknife method.

For the analysis of the reading time, a generalized linear model with Poisson distribution, including random case, reader, condition, session factors, and their interactions, was used. Outliers (defined as values extending beyond 10 minutes) were initially removed and considered as not representative of the real clinical practice (eg, due to interruptions, breaks, or technical problems). The two reading sessions made it possible to take into account the learning effect between the first time that readers used the AI system and the second reading session in which the readers were more familiar with the system and the entire workflow.

A secondary analysis was conducted to better understand the obtained results. This analysis was based on four subgroups of examination according to lesion type (soft tissue or calcifications), breast density (lower density [BI-RADS categories a and b] or higher density [BI-RADS categories c and d]), radiologists' years of experience (less than 10 years or more than 10 years), and reading time as a function of three MammoScreen score categories (MammoScreen score ≤ 4, MammoScreen score = 5 or 6, and MammoScreen score ≥ 7). The software used for conducting the statistical analysis was Obuchowski-Rockette and Dorfman-Berbaum-Metz software (version 2.5; Medical Image Perception Laboratory–University of Iowa, Iowa City, Iowa; available from *https://perception.lab.uiowa.edu/software-0*).

Finally, the intraclass correlation coefficient (ICC) was computed to assess the agreement between the 14 radiologists in rating the 240 included cases with and without the use of the AI system (21,22). ICC estimates and their 95% CIs were calculated using ICC R package (23) based on a single-rater, absolute-agreement, two-way random-effects model.

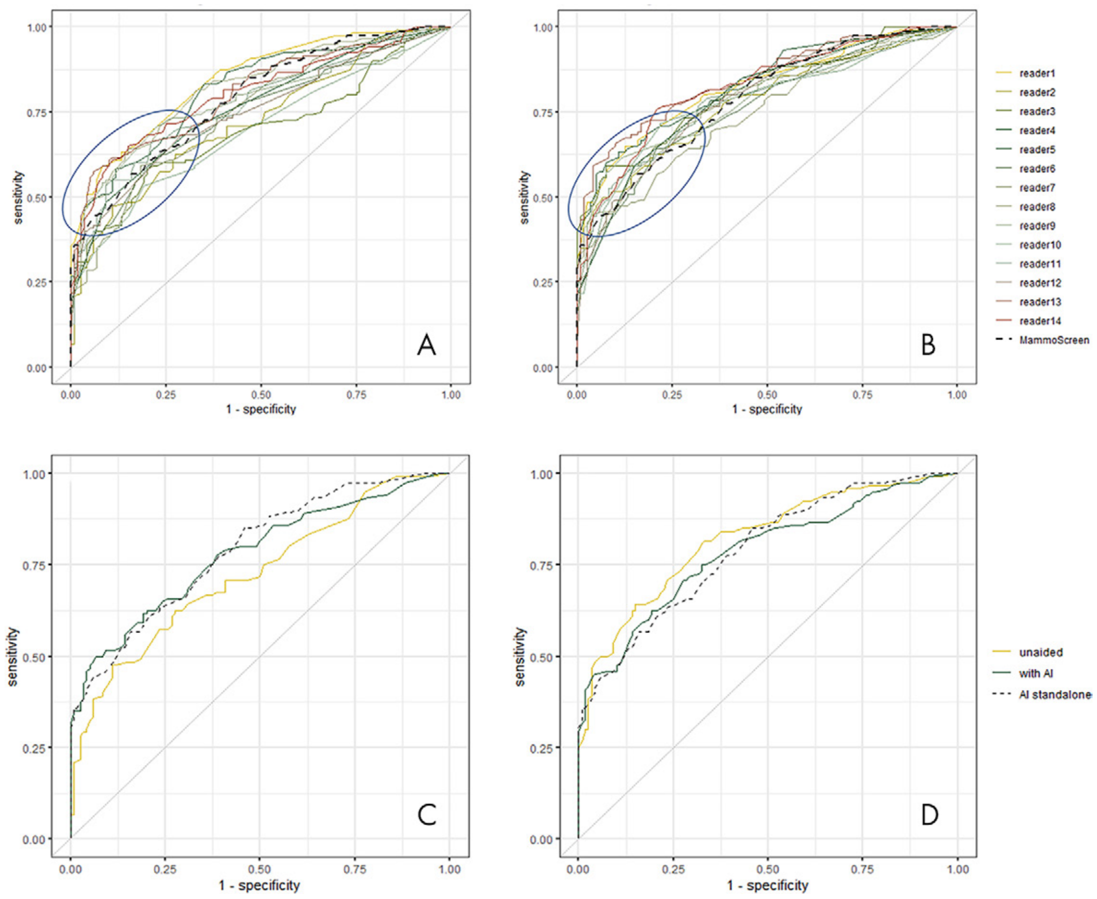

Figure 2: *A,* Receiver operating characteristic (ROC) curves of all readers in unaided reading condition and, *B,* reading with the help of artificial intelligence (AI). The blue circle underlines the area where ROC curves are most improved. Examples of ROC curve of, *C,* reader 2 and, *D,* reader 11 with and without using the help of AI.

Table 3: AUC Values for Each Radiologist and Average AUC for Reading Conditions

Reader	Years of Experience	$\mathrm{AUC}_{\mathrm{R}}$	$\mathrm{AUC}_{\mathrm{R+A}}$	Δ
1	8	0.845	0.824	−0.021
2	13	0.724	0.779	0.055
3	12	0.697	0.812	0.115
4	5	0.762	0.796	0.034
5	25	0.816	0.831	0.015
6	23	0.744	0.767	0.023
7	5	0.751	0.745	−0.006
8	3	0.782	0.797	0.015
9	6	0.694	0.783	0.089
10	0	0.768	0.781	0.013
11	21	0.820	0.785	−0.035
12	7	0.747	0.791	0.044
13	10	0.809	0.847	0.038
14	9	0.803	0.825	0.022
Average*	…	0.769 (0.724, 0.814)	0.797 (0.754, 0.840)	0.028 (0.002, 0.055)

Note.—Area under the receiver operating characteristic curve (AUC) for reading conditions with (R + A) and without (R) the AI system.

* Values in parentheses in the last line of the table (the one related to the average values) are the 95% CIs. The *P* value between the observed average values was .035.

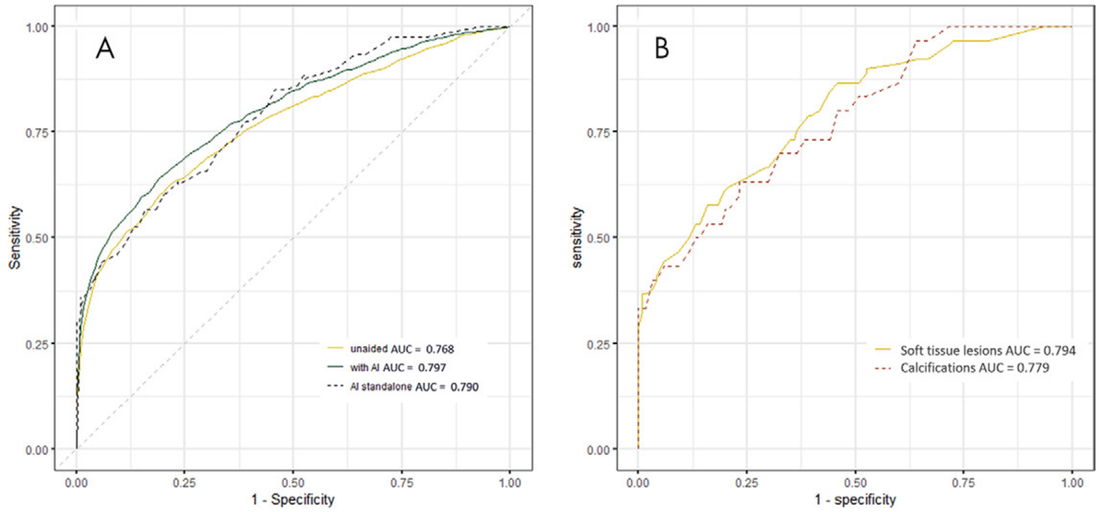

Figure 3: *A,* Average receiver operating characteristic (ROC) curves of all readers when unassisted (yellow) and assisted with artificial intelligence (AI) (dark green) and ROC curve of the AI system as stand-alone (dashed black). ROC curves are averaged using linear interpolation between sampled points of each curve (the area under the ROC curve [AUC] of the average ROC curve is similar to the average area under the curve of all readers [difference of 1 × 10^{-3}]). *B,* ROC curve of the AI as a stand-alone system for soft-tissue lesions (yellow) and calcifications (red).

Table 4: AUC Values for Subgroups Secondary Analysis

Subgroup	Average AUC_R	Average AUC_{R+A}	Δ	*P* Value
Lesion type				
Soft-tissue lesions	0.775 (0.726, 0.824)	0.802 (0.756, 0.849)	0.027 (−0.003, 0.058)	.073
Calcifications	0.749 (0.669, 0.830)	0.782 (0.699, 0.865)	0.033 (−0.006, 0.072)	.099
Breast density				
Low breast density	0.782 (0.729, 0.836)	0.813 (0.761, 0.866)	0.031 (< 1 × 10^{-4}, 0.062)	.050
High breast density	0.750 (0.683, 0.817)	0.776 (0.705, 0.846)	0.026 (−0.009, 0.061)	.144
Experience level				
Least experienced	0.776 (0.725, 0.826)	0.793 (0.748, 0.839)	0.018 (−0.017, 0.052)	.281
Most experienced	0.760 (0.703, 0.817)	0.803 (0.755, 0.851)	0.043 (0.002, 0.084)	.041

Note.—Area under the receiver operating characteristic curve (AUC) for reading conditions with (R + A) and without (R) the AI system. Values in parentheses are 95% CIs.

Results

Use of AI Increased the AUC

ROC curves and their related AUCs were computed using the level of suspicion estimated by the reader during the reading sessions (Fig 2, *A, B*). The AUC of each reader aided by the AI system, each reader unaided, and the difference between them, together with average values and *P* values are reported in Table 3.

Among the 14 readers, 11 (79%) had an increase in AUC using the AI system. Figure 2, *C* and *D,* show the change of the ROC curve for reader 2 and reader 11, respectively. In the first case, there was an increase of AUC going from 0.724 to 0.779, while in the second case the AUC decreased from 0.820 to 0.785. The average AUC across readers reading unaided was 0.769, while the average AUC across readers when using the AI system was 0.797. The average difference in AUC was 0.028 (95% CI: 0.002, 0.055 and *P* = .035) (Fig 3). The increase in AUC was also seen in all the considered subgroups; results are shown in Table 4.

Use of AI Increased Sensitivity and Specificity

Average sensitivity was shown to be statistically significantly increased by 0.033 when using AI support (*P* = .021); average specificity showed a lower level of improvement (*P* = .634). Results are reported in Table 5. The use of the AI tool resulted in a trend toward lowering the false-negative rate for 11 of 14 readers with an average improvement of 18% (range, 2%–50%). Similarly, the false-positive rate was decreased by the use of AI for eight radiologists by an average of 25% (range, 9%–42%). Figure 4 reports the cancer detection rate and the false-positive rate for each reader along with the percentage of improvement brought by the use of the AI system. Figure 5 shows an example in which nine of the 14 radiologists detected an invasive ductal carcinoma when reading the case using the AI tool while in the unaided reading condition only three radiologists detected the cancer.

Table 5: Average Sensitivity and Specificity across Readers

Parameter	R	R+A	Δ	*P* Value
Sensitivity	0.658 (0.574, 0.743)	0.691 (0.600, 0.782)	0.033 (0.017, 0.072)	.021
Specificity	0.725 (0.656, 0.794)	0.735 (0.656, 0.815)	0.010 (−0.030, 0.038)	.634

Note.—Sensitivity and specificity for reading conditions with (R + A) and without (R) the AI system. Number in parentheses are 95% CI values.

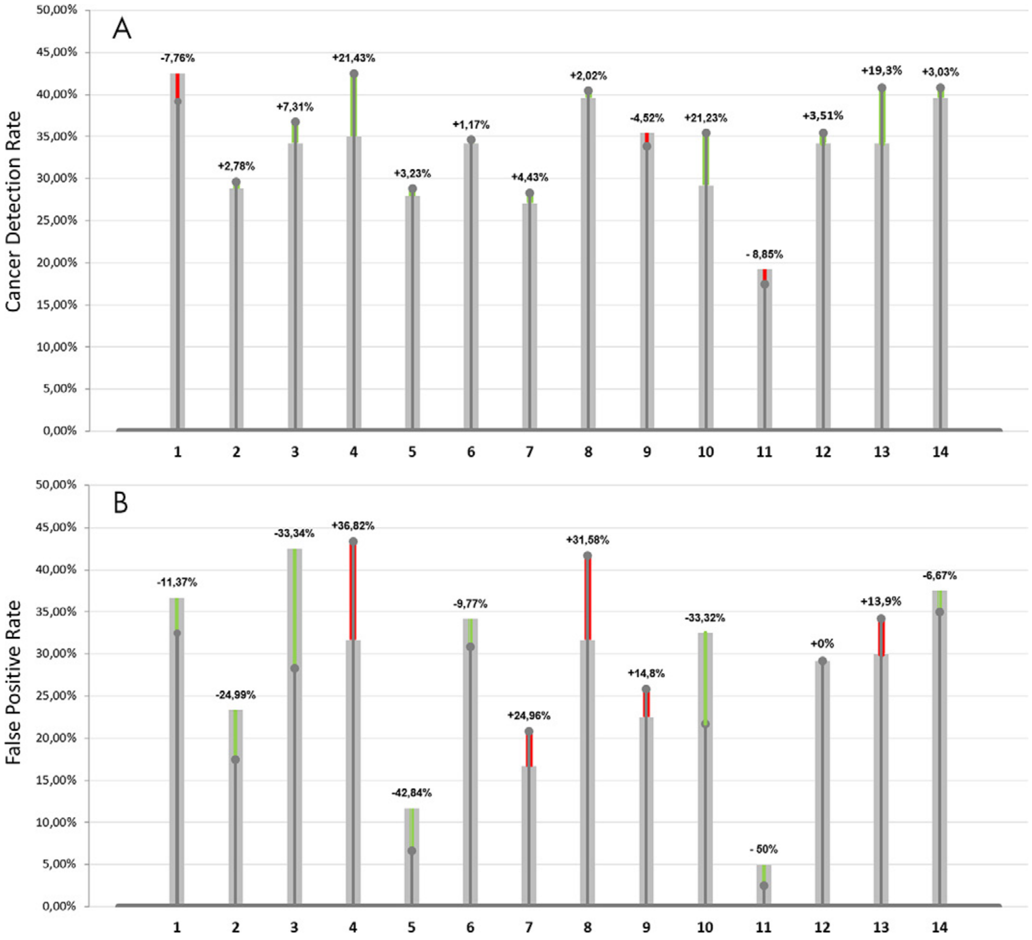

Figure 4: *A*, Cancer detection rate and percentage improvement brought by the use of the artificial intelligence (AI) system and, *B*, false-positive rate and percentage decrease as a result of the use of AI. Green bars indicate the percentage improvement brought by the help of AI, thus an increase in cancer detection rate and a decrease in the false-positive rate. Similarly, red bars indicate a deterioration of performances, thus a, *A*, decrease in cancer detection rate, *B*, and an increase in false-positive rate.

Reading Time Changes with Use of AI Were Dependent on Likelihood of Malignancy

Of all reading times, 51 of 6720 (0.8%) were defined as outliers and were excluded. On the first reading session, the average reading time per case was 62.79 seconds for the unaided readings (95% CI: 60.77, 64.80) and 71.93 seconds for the readings with the AI support (95% CI: 69.52, 74.33) (Table E3, Appendix E2 [supplement]). The difference was statistically significant ($P < .001$); the reading time increased for 11 radiologists and decreased for three radiologists. The analysis of the reading time as a function of MammoScreen score categories is reported in Figure 6, *A*.

For the second reading session, the average reading time per case was 57.22 seconds for the unaided readings (95% CI: 55.10, 59.33) and 62.16 seconds for the readings with AI (95% CI: 60.04, 64.29) (Table E3, Appendix E2 [supplement]). The difference was statistically significant ($P < .001$); the mean reading time increased for eight radiologists and decreased for six radiologists. For this second session, the analysis of the reading time as a function of MammoScreen score categories showed a learning effect, with a decrease in the reading time for MammoScreen score lower than 4, and an increase in reading time of less than 10 seconds for scores higher than 4 (Fig 6, *B*).

Interrater Reliability

A moderate interrater reliability was found in both reading conditions. For the unaided reading condition, ICC was equal

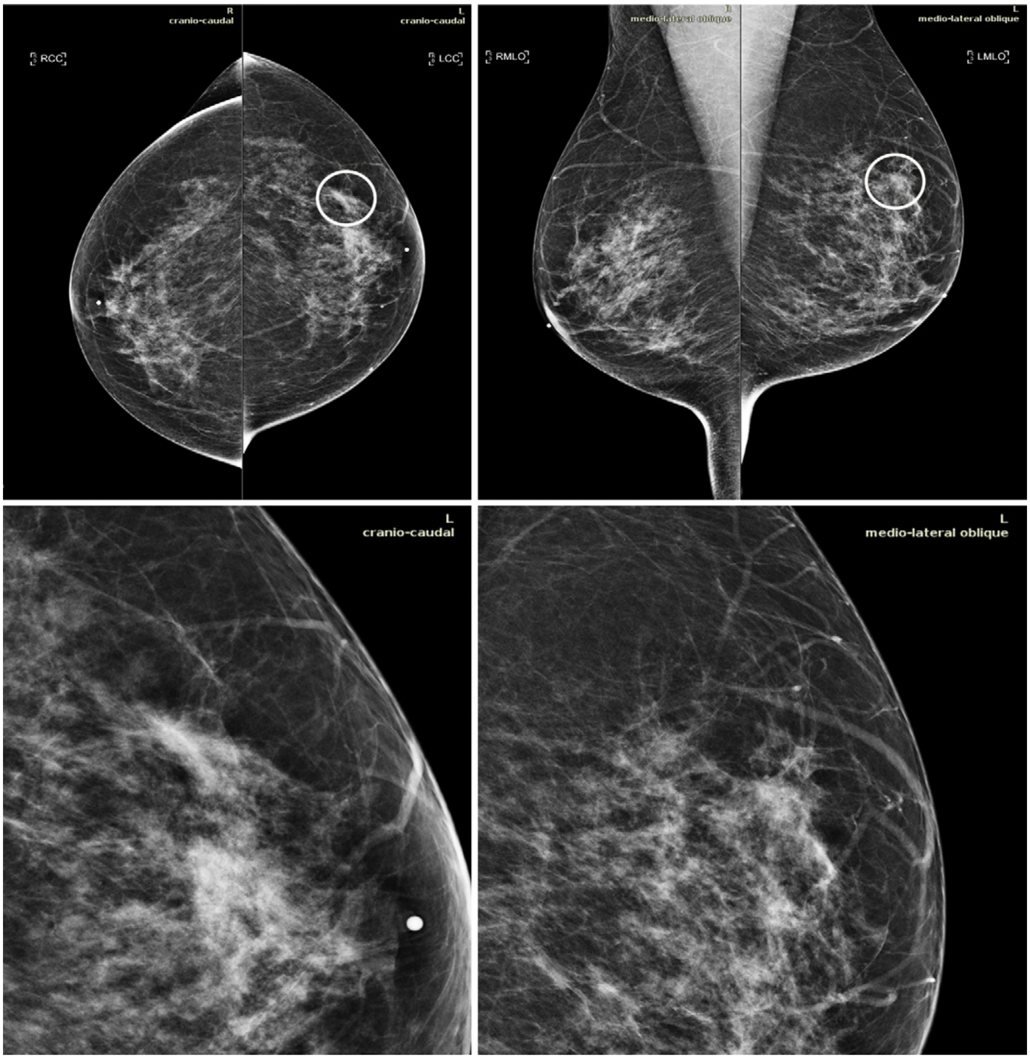

Figure 5: Mammograms in a 51-year-old woman with invasive ductal carcinoma. The upper panels show the craniocaudal and the mediolateral oblique views. The lower panels show a close-up of the left breast area containing the lesion. The case is one of the false-negative cases included in the dataset. Accordingly, the initial screening assessment was a BI-RADS 2, meaning visible findings were judged as benign. After 1 year, the patient presented for another screening examination. This time, a focal asymmetry with associated distortion within the left breast was noticed; the patient was recalled and diagnosed with a 1.5-cm mass in the upper outer quadrant of the left breast on the craniocaudal view (circle).

to 0.586 (95% CI: 0.528, 0.642), while for the reading condition using the AI system the ICC value was 0.679 (95% CI: 0.62, 0.732).

Discussion

This clinical investigation demonstrated that the performance of radiologists in reading two-dimensional breast cancer screening mammograms can be improved with the concurrent use of an AI-based tool. The improvement was seen on a cancer-enriched dataset of 240 digital mammography examinations including different types of abnormalities.

Looking at the overall trend of all ROC curves with and without the aid of AI, it was observed that all curves exhibited less dispersion (variability) when AI was used, which highlights the influence of the system on radiologist decision. Interreader reliability appeared to increase in aided reading conditions, meaning that AI would provide a more standardized, expert-independent result.

Reading time increased in both reading sessions when using AI. For low MammoScreen scores (1 to 4), the time was about the same in the first session and slightly decreased in the second session; for higher MammoScreen scores, the reading time increased, on average, with the use of AI. However, because of the presence of measurement errors, we were not able to quantify these time differences in a reliable way.

The decrease in reading times was observed for cases that received a MammoScreen score of less than 4. The AI-based tool has the potential to increase overall efficiency of radiologists on these cases, allowing them to focus their attention on the most suspicious examinations, while reassuring them on less suspicious examinations, which are far more numerous.

Furthermore, the learning curve observed between the first and the second session, together with the fact that the maximum increment of time did not exceed 15 seconds, suggested that the introduction of this tool into screening programs may not

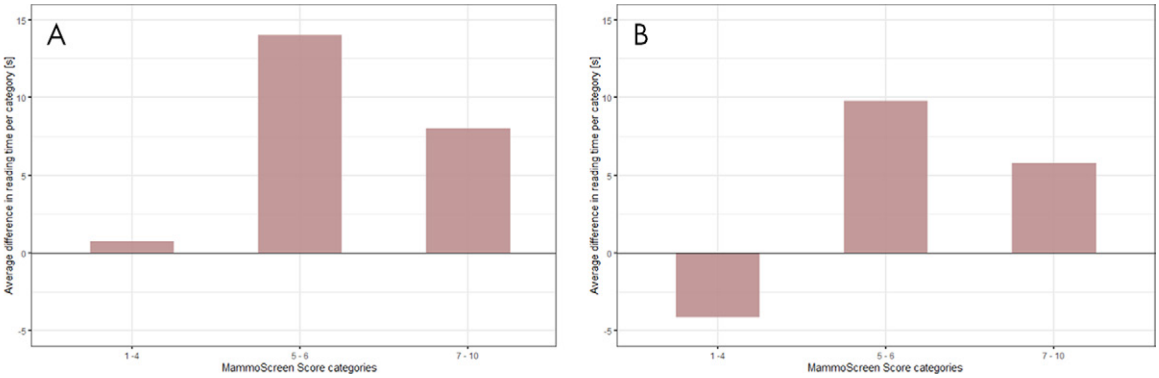

Figure 6: Difference in reading time per MammoScreen for the, *A*, first reading session and the, *B*, second reading session. The upper part of the plot indicates an increment in reading time when using MammoScreen. The lower part of the plot indicates a decreased reading time when using MammoScreen. The maximum increase in reading time for the first reading session does not exceed 15 seconds. A session effect is noticeable with a gain in reading time for low MammoScreen score (1 to 4), while the increase does not exceed 10 seconds for the higher scores.

prolong the workflow of the radiologists and possibly even lead to a shorter average reading time. It should be noted that in real conditions, additional factors may have an impact on reading time (ie, stress, tiredness, etc), and that those factors were obviously not considered in the present analysis.

Because the use of AI systems to help read screening mammograms is in the early stage of investigation, there is sparse literature on their clinical performance (16,24–27). Important results have been achieved by Rodríguez-Ruiz et al (11) who demonstrated an improvement in the average AUC of two percentage points with the use of the AI system with respect to the unassisted reading. A significant difference of this study, with respect to the work of Rodríguez-Ruiz et al, was the inclusion of false-negative cases within the dataset assessed by the readers, which explains the differences in absolute values of AUC (with and without AI). The choice of including false-negative cases was driven by the fact that we believe the AI tool can aid in the detection of very early signs of cancer that may be overlooked in a regular screening setting. Despite the different distribution of examination categories (true positive, true negative, false positive) and the inclusion of false negative in the dataset, the present study confirms the observed trend that AI algorithms are able to improve radiologists' success rate in breast cancer detection, supporting the conclusion that radiologists and AI achieve better performance together than each of them individually. Another important difference with this study was the reading setup. Rodríguez-Ruiz et al tested two configurations (ie, half the readers used the AI system integrated in the reading workstation, and the other half used the AI system on a separate screen from the workstation). In our study, the results (ie, suspicious region, level of suspicion related to each suspicious region, level of suspicion per breast, and overall assessment of the case) were displayed on a separate screen and presented at the same time as the mammography. In addition, contrary to the AI system tested by Rodríguez-Ruiz et al, the system used in the present study does not allow for interaction between the radiologist and the system, thus resulting in much shorter average reading times. Other recent remarkable results have been published by Kim et al (26) and by McKinney et al (27). Both describe reader studies carried out on similar datasets and with similar designs. However, apart from the use-case tested (AI tested as a second reader by Kim et al, and AI tested as stand-alone system by McKinney et al), the major difference that emerged with the present study was that mammograms collected in the very same centers used for algorithm training were used during validation, while independent, geographically different (United States vs Europe) centers were used in the present study. Choosing data from centers independent of those used during algorithm training is especially important when dealing with neural networks that generally contain several millions of parameters. Indeed, algorithms tested and trained with data from the same center have the capacity to learn center-specific biases, often indistinguishable for humans, and performances of such models tend to be overestimated when evaluated on data originating from the training centers. This type of validation is often referred to as external geographic validation and shall be preferred to other types of validation when evaluating generalizability of AI models (28).

As with the study of Rodríguez-Ruiz et al, the main limitations of this investigation were due to the used dataset that was not representative of the normal screening practice. First, it was enriched with cancer cases and because readers were aware of that fact, this could have caused a laboratory effect, inducing a high rate of false-positive assessments (29,30). Second, all subcategories (eg, high or low density, lesion type) were not homogeneously distributed. In addition, because readers had no access to prior mammographic examinations of the same patient, additional imaging examinations, or any other kind of information, the assessment was more challenging than a typical screening mammography reading workflow. However, this scenario is representative of a baseline screening examination (eg, a patient who does not have prior studies), which accounts for 12% of all screening mammography per year (31). It has been demonstrated that the callback rate of baseline mammograms is higher than the recall rate of nonbaseline patients (31–33); thus, having a relevant benefit on this subgroup of patients would have an important impact on the global recall rate and false-positive reduction. Finally, the overall conclusion of this clinical investigation was that the concurrent use of this AI tool improved the diagnostic performance of radiologists in the mammographic detection of breast cancer. In addition, the use of AI was shown to reduce false negatives without affecting the specificity.

Author contributions: Guarantor of integrity of entire study, J.L.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, S.P., P.F.; clinical studies, S.P., J.L., P.C., T.B., P.F.; statistical analysis, S.P., J.M.G., P.F.; and manuscript editing, S.P., J.L., P.F.

Disclosures of Conflicts of Interest: **S.P.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: employed by Therapixel. Other relationships: disclosed no relevant relationships. **J.L.** disclosed no relevant relationships. **P.C.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: employed by Therapixel. Other relationships: disclosed no relevant relationships. **T.B.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: employed by Therapixel. Other relationships: disclosed no relevant relationships. **J.MG.** Activities related to the present article: consulting fee from Therapixel. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **P.F.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: employed by Therapixel and owns stock in Therapixel. Other relationships: disclosed no relevant relationships.

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.
2. Sant M, Allemani C, Berrino F, et al. Breast carcinoma survival in Europe and the United States. Cancer 2004;100(4):715–722.
3. Breast Cancer Screening Market | Global Analysis by Population, Screening Tests [Mammography, MRI, Ultrasound], Countries & Forecast (2018-2024). https://www.renub.com/breast-cancer-screening-mammography-mri-and-ultrasound-market-and-forecast-worldwide-63-p.php. Accessed November 18, 2019.
4. Rawashdeh MA, Lee WB, Bourne RM, et al. Markers of good performance in mammography depend on number of annual readings. Radiology 2013;269(1):61–67.
5. Lehman CD, Wellman RD, Buist DSM, et al. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med 2015;175(11):1828–1837.
6. Gilbert FJ, Astley SM, Gillan MGC, et al. Single reading with computer-aided detection for screening mammography. N Engl J Med 2008;359(16):1675–1684.
7. James JJ, Gilbert FJ, Wallis MG, 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(2):379–386.
8. Gromet M. Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. AJR Am J Roentgenol 2008;190(4):854–859.
9. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med 2007;356(14):1399–1409.
10. 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.
11. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. Radiology 2019;290(2):305–314.
12. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Sci Rep 2019;9(1):12495.
13. Wu N, Phang J, Park J, et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. arXiv:190308297 [cs, stat]. [preprint]. http://arxiv.org/abs/1903.08297. Posted March 19, 2019. Accessed October 3, 2019.
14. Sahran S, Qasem A, Omar K, et al. Machine Learning Methods for Breast Cancer Diagnostic. Breast Cancer and Surgery.. Published November 5, 2018. Accessed October 3, 2019.
15. Pannucci CJ, Wilkins EG. Identifying and avoiding bias in research. Plast Reconstr Surg 2010;126(2):619–625.
16. Conant EF, Toledano AY, Periaswamy S, et al. Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis. Radiol Artif Intell 2019;1(4):e180096.
17. Benedikt R, Toledano AY, Boatsman J, et al. Concurrent CAD with Digital Breast Tomosynthesis Improves Reading Time and Maintains Performance for Dedicated Breast Radiologists and General Radiologists. ECR 2017 PosterNG. https://posterng.netkey.at/esr/viewing/index.php?module=viewing_poster&task=&pi=138664. Published 2017. Accessed October 29, 2019.
18. Hupse R, Samulski M, Lobbes MB, et al. Computer-aided detection of masses at mammography: interactive decision support versus prompts. Radiology 2013;266(1):123–129.
19. Hillis SL, Obuchowski NA, Berbaum KS. Power estimation for multireader ROC methods an updated and unified approach. Acad Radiol 2011;18(2):129–142.
20. Hillis SL, Schartz KM. Multireader sample size program for diagnostic studies: demonstration and methodology. J Med Imaging (Bellingham) 2018;5(4):045503.
21. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;15(2):155–163 [Published correction appears in J Chiropr Med 2017;16(4):346.]
22. Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation – A discussion and demonstration of basic features. PLoS ONE 2019;14(7):e0219854.
23. GitHub - matthewwolak/ICC: R package to assist in the estimation of the Intraclass Correlation Coefficient (ICC). https://github.com/matthewwolak/ICC. Accessed February 27, 2020.
24. Rodriguez-Ruiz A, Mordang JJ, Karssemeijer N, Sechopoulos I, Mann RM. Can radiologists improve their breast cancer detection in mammography when using a deep learning based computer system as decision support? In: Krupinski EA, ed. 14th International Workshop on Breast Imaging (IWBI 2018). Vol 10718. Bellingham, Wash: International Society for Optics and Photonics, 2018; 1071803.
25. Watanabe AT, Lim V, Vu HX, et al. Improved Cancer Detection Using Artificial Intelligence: a Retrospective Evaluation of Missed Cancers on Mammography. J Digit Imaging 2019;32(4):625–637.
26. Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health 2020;2(3):E138–E148.
27. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89–94.
28. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology 2018;286(3):800–809.
29. Evans KK, Birdwell RL, Wolfe JM. If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening. PLoS One 2013;8(5):e64366.
30. Gur D, Bandos AI, Cohen CS, et al. The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. Radiology 2008;249(1):47–53.
31. McDonald ES, McCarthy AM, Akhtar AL, Synnestvedt MB, Schnall M, Conant EF. Baseline Screening Mammography: Performance of Full-Field Digital Mammography Versus Digital Breast Tomosynthesis. AJR Am J Roentgenol 2015;205(5):1143–1148.
32. Gur D, Sumkin J, Zuley M, Klym A, Brown E, Lederman D. The Baseline Mammogram: Are We Doing Enough to Reduce Recall Rates? In: Radiological Society of North America 2011 Scientific Assembly and Annual Meeting, Chicago, IL, 2011. Oak Brook, Ill: Radiological Society of North America, 2011.
33. Sumkin JH, Ganott MA, Chough DM, et al. Recall Rate Reduction with Tomosynthesis During Baseline Screening Examinations: An Assessment From a Prospective Trial. Acad Radiol 2015;22(12):1477–1482.