The Case of Missed Cancers: Applying AI as a Radiologist's Safety Net

Michal Chorev[1](✉), Yoel Shoshan[1], Adam Spiro[1], Shaked Naor[1], Alon Hazan[1], Vesna Barros[1], Iuliana Weinstein[2], Esma Herzel[3], Varda Shalev[3], Michal Guindy[2], and Michal Rosen-Zvi[1]

[1] Department of Healthcare Informatics, IBM Research, IBM R&D Labs, University of Haifa Campus, Mount Carmel, 3498825 Haifa, Israel
{michalc,yoels,rosen}@il.ibm.com

[2] Department of Imaging, Assuta Medical Centers, Tel Aviv, Israel
michalgu@assuta.co.il

[3] MaccabiTech, MKM, Maccabi Healthcare Services, Tel Aviv, Israel

Abstract. We investigate the potential contribution of an AI system as a safety net application for radiologists in breast cancer screening. As a safety net, the AI alerts on cases suspected to be malignant which the radiologist did not recommend for a recall. We analyzed held-out data of 2,638 exams enriched with 90 missed cancers. In screening mammography settings, we show that a system alerting on 11 out of every 1,000 cases, could detect up to 10.7% of the radiologists' missed cancers. Thus, significantly increasing radiologist's sensitivity to 80.3%, while only slightly decreasing their specificity to 95.3%. Importantly, the safety net demonstrated a significant contribution to their performance even when radiologists utilized both mammography and ultrasound images. In those settings, it would have alerted 8.5 times per 1,000 cases, and detected 11.7% of the radiologists' missed cancers. In an analysis of the missed cancers by an expert, we found that most of the cancers detected by the AI were visible post-hoc. Finally, we performed a reader study with five radiologists over 120 exams, 10 of which were originally missed cancers. The AI safety net was able to assist 3 out of the 5 radiologists in detecting missed cancers without raising any false alerts.

Keywords: Computer-aided diagnosis · Deep learning · Breast imaging

1 Introduction

1.1 Radiologists Performance in Screening Digital Mammography

Breast cancer (BC) is the most commonly diagnosed cancer among women worldwide, and the second leading cause of cancer-related deaths. As treatment options improve, early detection may have a larger impact on morbidity and mortality. Presently, digital mammography (DM) is the most common method of screening being used globally. Women undergo a DM exam every 1–3 years depending on their familial history and national policy. These exams are then interpreted by radiologists based on the Breast

A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12266, pp. 220–229, 2020.
https://doi.org/10.1007/978-3-030-59725-2_22

Imaging Report and Data System (BIRADS). According to the breast cancer surveillance consortium (BCSC) benchmark for radiologists in DM screening [1], the average radiologist's sensitivity and specificity are 87% and 89%, respectively. While 97.1% of radiologists are within the acceptable range of sensitivity ≥75%, only 63.0% met the acceptable range for specificity of 88%–95%. Indeed, analyzing mammograms is a challenging task. Previous works have shown the agreement between radiologists to be slight to moderate at best [2–4]. A second reading of mammograms by an additional radiologist has been proven to increase sensitivity and specificity [5, 6]. However, lack of trained radiologists, budget, and time limitations often make it inapplicable to the standard screening procedure [7]. AI systems may help close the gap of readily available second readers, but their real-world efficacy is still a matter of debate [8–10].

1.2 AI Performance in Screening Digital Mammography

Since June 2019, six different papers [11–16] reported results of AI models trained on retrospective screening mammography data for the detection of BC within 12 months. Most studies were based on large-scale data (ranging from an order of 15K to 150K women) and reported impressive results, often at the range of the radiologists' performance or even surpassing it. The radiologists' performance on the held-out data in the above papers (when reported) is presented in Fig. 1 by the full circles. The reported numbers vary between sensitivity of 77% (readers in Italy and the Netherlands) to 98.5% (UK consensus reader) and specificity between 67% (Italian readers) and 97.4% (Israeli readers), which is consistent with the performance of radiologists derived by BCSC [1]. The performance of AI models is illustrated by empty circles.

The ability to achieve radiologists' level in specificity and sensitivity indicates that the technology could have assisted radiologists. Unfortunately, it is almost impossible to compare the AI models. Not only that each paper is leveraging different datasets from different geographies (USA, South Korea, Israel, and multiple countries in western Europe), but also different evaluation methods and performance measures were used for reporting the results. Most importantly, some key elements that could demonstrate the usefulness of these models as part of a screening routine were often missing from the report. Namely, their expected performance in real-life settings, including what composition of cases was considered real-life settings, and the existence of false negative (FN) cases by radiologists (definition, count, and performance). Moreover, two of the six papers conducted reader studies [13, 14] that only shed light on part of the picture. Kim et al. only tested the different readers on cancers that were detectable by the original radiologist by DM or ultrasound (US). McKinney et al. did not explicitly report the number of missed cancer cases in their reader study.

In this work, we focused on a safety net application, where the aim of the technology was to alert on FN cases within 12 months from the index exam, while maintaining a low number of false alerts. Reducing AI's false alarms is key, as computer-aided diagnosis systems have been shown in the past to generate a large number of false positive findings, slowing the radiologist's work without contributing to their performance [8]. False positive recalls may induce extra costs, unnecessary anxiety and additional procedures for a healthy population [17, 18]. For that purpose, our system worked in a high specificity operation point, in which it was expected to identify normal cases with high confidence.

However, high specificity operation point is not necessarily optimal for the detection of all cancer cases, let alone those that were missed by the first reader.

Here, we analyzed the overall contribution of the AI system in a safety net application to the radiologist's performance; first, on a held-out set based on the original radiologists' performance, and second, in a multi-reader study to examine its contribution to individual readers. Both the held-out data and the reader study were enriched with radiologists' FN cases.

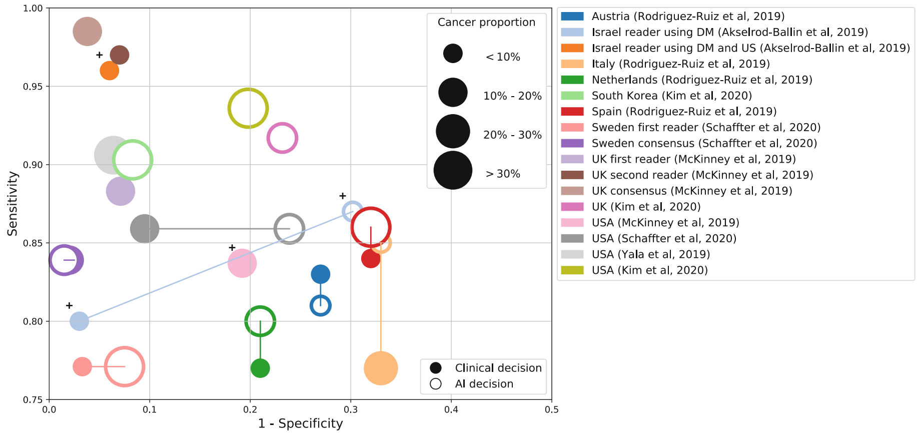

Fig. 1. Performance of radiologists and AI systems within a 12-month follow-up period as reported in recent publications. Straight lines correspond to studies who reported performance of both AI and radiologists on the same cohort. +Indicates adjusted results for screening scenario. Results with specificity below 50% and sensitivity below 75% are not shown.

2 Methods

This work was approved by the research ethics review board of Assuta Medical Centers (AMC), who also waived the need to obtain a written informed consent. Data were collected, managed and anonymized by Maccabi Health Services (MHS).

2.1 AI as a Safety Net Application

The AI system being used in this work was trained on the index exam DM images and detailed clinical history of the women. Its architecture consisted of an ensemble of three deep learning (DL) models on the image level, these models were used to produce a malignancy score and as feature extractors. Another machine learning model utilized the DL models' output and incorporated the clinical information (including imaging history, gynecological and familial history, medications, diagnoses, and lab results) into a final malignancy score on the breast-level. For a study-level decision, the maximum between the prediction for each side was taken. For more details see [11].

The model's output is a continuous score in the range between 0 and 1. For the system to provide a final binary decision, thresholds were set at specific operation points using a validation set from a previous study [11] without overlap with this work's datasets. One operation point was 87% sensitivity, consistent with the average reader sensitivity according to BCSC benchmark [1]. Here, we focused mostly on a second operation point of 99% specificity in a safety net application. In such an application, the AI system analyzes the cases independently of the radiologist and is activated after the radiologist's BI-RADS score assignment. When the AI system deems a case malignant, only then it checks its original BI-RADS category. If the case was assigned BI-RADS 1–2, it raises an alarm for a recall. A 99% specificity operation point was selected to reduce the number of false alarms.

2.2 Retrospective Held-Out Set

The dataset was collected from five AMC imaging facilities and from MHS database. The cohort was composed of women who underwent at least one DM examination between 2013 and 2017 consecutively and had at least one year of clinical history. Most women in AMC undergo a screening mammography biannually, with the exception of women with familial history who are offered an annual exam. Mammograms are typically read by a single fellowship-trained reader and interpreted using the BI-RADS scale. We excluded exams of women with a history of breast cancer; exams post breast operations (e.g., lumpectomy or mammoplasty); and exams of a single breast. Exams were considered positive if there was an indication of a biopsy positive for cancer or they appeared in the cancer registry within 12 months. Exams were considered negative if there was an indication of a negative biopsy within 12 months or they had a BI-RADS 1–3 index exam and a completely clean follow-up for at least two years (i.e. all follow-up exams in that period were BI-RADS 1–2, without biopsy recommendations or procedures). For each woman, the first exam meeting the inclusion/exclusion criterion was selected as the index examination. All FN in the retrospective test set were reviewed post-hoc by a breast radiology specialist with more than 20 years of experience. The retrospective test set analyzed in this work was never used to train or tune the AI system.

2.3 Reader Study

In the reader study, five AMC board-certified radiologists with breast mammography fellowship have interpreted 120 exams. Two of the readers had 20 years of experience, two had 10 years of experience, and one had more than a year of experience reading breast mammograms. Another >20 years experienced reader conducted a post-hoc review of the FN cases in the study. Readers used their regular system and screens. Readers and AI model were exposed to the index exam images and the entire set of clinical data. They had no access to previous exams or to other modalities taken in the index examination date. However, we excluded cases with high BI-RADS due only to high US BI-RADS (i.e., the retrospective DM required a recall). Additionally, when an exam had a negative biopsy indicating a healthy tissue, we made sure that there isn't a follow-up positive biopsy exam or a record in the cancer registry. Here the definition of FN was loosened in comparison to the retrospective test set, to introduce cancers that were diagnosed within a two years window. We made no distinction whether a finding was visible in DM, US, or neither. Cases were assigned in a random order to each reader, and each has covered the entire set of cases.

2.4 Bootstrapping and Statistical Analysis

Ideally, AI system's performance and contribution to a human reader should be assessed in conditions as close to real-world prevalence as possible. However, in most studies, this is not the case. Roughly 98% of mammograms in a screening population are normal. The manner in which AI models are trained often results in datasets enriched in abnormal cases. Here too, the retrospective test set is not reflective of real-world prevalence of breast cancer as reported by AMC or the BCSC benchmark [1]. The data is enriched with biopsy cases (880/2,638, 33%) and especially FN (90/2,638, 3%).

For this purpose, we utilized the entire set of clinical data in MHS (69,149 cases) to estimate real-world prevalence in the population. Performance of the original radiologist was estimated once according to DM alone (TP: 1,444/69,149, 2.08%; FN: 405/69,149, 0.59%; TN: 64,832/69,149, 93.76%; FP: 2,468/69,149, 3.57%), and once according to DM and US (TP: 1,692/69,149, 2.45%; FN: 157/69,149, 0.23%; TN: 61,667/69,149, 89.17%; FP: 5,633/69,149, 8.15%). Using these proportions, we bootstrapped with replacements a sample set of 1,000 cases in 1,000 iterations and calculated the reader's and AI's performance on the retrospective dataset. We report mean and 95% confidence interval (CI) for each measure.

Fisher exact test and Wilcoxon signed-rank test were used to evaluate significant differences in performance. For multiple hypotheses we used Benjamini-Hochberg adjustment. Inter-reader agreement was estimated using Cohen's Kappa statistic. P-values less than 0.05 were considered statistically significant.

3 Results and Discussion

3.1 Safety Net Application on Held-Out Retrospective Data

A held-out set of 2,638 individual exams was collected (age 55 [47–63], BMI 26 [23–30], median and interquartile range). Each exam in the dataset included the four standard mammography images as well as detailed clinical history of the women. The dataset consisted of: 1,688/2,638 (64%) BI-RADS 1–2 cases with a clean follow-up, 70/2,638 (3%) BI-RADS 3 cases with a clean follow-up, 501/2,638 (19%) negative-biopsy cases, and 379/2,638 (14%) positive-biopsy cases. The dataset was intentionally enriched in FN cases, with 24% (90/379) of cancer cases originally missed by the radiologist. A screening US exam was performed in 75% (1,967/2,638) of the cases, and BI-RADS were reported separately for DM and US. Performance of the original radiologist was estimated twice; based on DM alone and based on both modalities, when US was available. The AI system's performance was estimated at an operation point of 99% specificity as a safety net application (see Sect. 2.1). We used a bootstrapping analysis, mean and CI, to estimate performance on real-word prevalence (see Sect. 2.4).

Based on DM alone, the radiologists obtained a sensitivity of 77.9% [60.9%–92.6%] (or 20.8 cancers detected per 1,000 cases) and specificity of 96.4% [95.1%–97.5%]. With the addition of US, their sensitivity increased to 91.3% [79.2%–100.0%] (or 24.3 cancers detected per 1,000 cases) and their specificity decreased to 91.7% [89.8%–93.3%]. The AI system obtained a sensitivity of 47.0% [27.3%–66.7%] at a specificity of 98.7% [98.0%–99.4%]. We then simulated a safety net application of the AI system. In this scenario, the AI system analyzed the cases independently after the radiologist work is done. For cases it deemed likely to be missed cancer, it raised an alert (see Sect. 2.1).

For radiologists reading only DM, the AI system would have raised 10.68 [4.00–20.00] alerts for every 1,000 cases. From the alerts, 0.63 would be valid. In other words, for an average of 5.87 [1.00–13.00] radiologists' FN per 1,000 cases, the AI could have identified 10.73% [0.00%–23.07%]. The remaining 10.05/1000 of the alerts would be false alarms. Hence, the safety net application is able to increase the reader's sensitivity significantly (80.3% vs. 77.9%, p-value $= 1.0 \times 10^{-78}$), with a slight but significant decrease in specificity (95.3% vs. 96.4%, p-value $= 3.3 \times 10^{-165}$). When the reader's interpretation was based on both DM and US, the AI's contribution was significant still. It would have raised an alarm on 8.51 [3.00–16.00] out of every 1,000 cases, of which 0.27 [0.00–2.00] would have been valid. For an average of 2.30 [0.00–7.00] radiologists' FN, the AI could have identified 11.73% [0.00%–28.57%] for every 1,000 cases. The remaining 8.24 alarms would have been false. The AI safety net was still able to increase sensitivity (92.3% vs. 91.3% p-value $= 5.6 \times 10^{-39}$) with a slight decrease in specificity (90.8% vs 91.7%, p-value $= 3.3 \times 10^{-165}$).

From the 90 cases that were missed by the radiologist, the AI safety net was able to identify 11. We asked an expert breast radiologist to review all FN cases and determine whether the malignant lesion was visible in the index exam's DM or not. The expert reader analyzed first the index exam, and only then utilized any other US or follow-up examinations images as well as pathology reports to localize the malignant finding. The AI was able to identify a larger proportion of visible cancers than not (Fisher exact test, p-value $= 1.76 \times 10^{-2}$; see Table 1).

Table 1. False negative cases' post-hoc visibility vs. identification by AI safety net

	Visible	Not visible	Total
AI identified	8	3	11
AI missed	27	52	79
Total	35	55	90

3.2 Reader Study

We then continued to examine the AI safety net application in a separate reader study with five certified breast radiologists from AMC. The study consisted of 120 cases (age 53 [47–65], BMI 26 [23–29], median and interquartile range), including 36/120 (30%) normal cases (original BI-RADS 1–2 with a clean two years follow-up), 13/120 (11%)

original BI-RADS 3 cases without biopsy and a clean two years follow-up, 35/120 (29%) with a negative biopsy with one year, and 36/120 (30%) with a positive biopsy. The cancer cases further included 10/36 (28%) FN. The original BI-RADS assigned to the FN was either 1 or 2.

Readers were asked to assign each case a BI-RADS 1–5 score. Their answers were compared to the ground truth and individual sensitivity/specificity measures were calculated. We compared the readers performance to the AI system in two operation points: 1) 87% sensitivity, and 2) 99% specificity, for the safety net application (Fig. 2a, see Sect. 2.3).

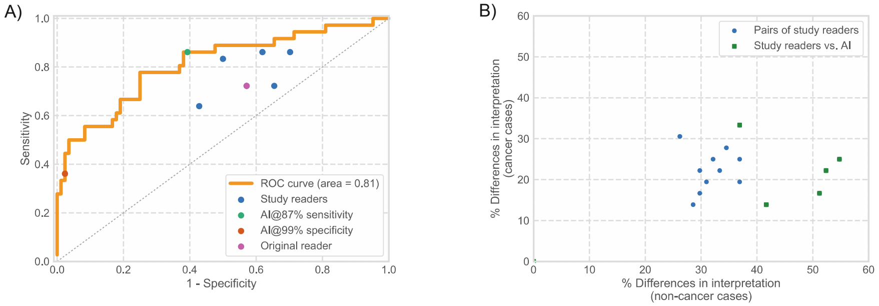

Fig. 2. Performance of the readers and AI system in the reader study. A) ROC curve of the AI system (AUC = 0.81). Sensitivity/specificity of each study reader (in blue), the original reader (purple) and the AI at 87% sensitivity and 99% specificity (green and red, respectively). B) Differences in interpretation (recall/no-recall) of cancer and non-cancer cases between pairs of readers (blue dots) and readers and AI (green squares). (Color figure online)

At an operation point of 87% sensitivity, the AI system has exceeded the average radiologist performance (sensitivity 86.1% vs. 78.3%; specificity 60.7% vs. 41.9%). Interestingly, the readers' average sensitivity matched the retrospective average sensitivity estimated for AMC radiologists on DM (see Sect. 2.4). However, the retrospective specificity was much higher than the one obtained in the reader study (96% vs. 42%). Indeed, in regular settings, the radiologist could compare the index exam to previous exams of the same woman or to the US, if either existed. Moreover, the reader study is enriched with negative-biopsy cases, which are more contestable. The average level of agreement between radiologists based on BI-RADS score was only fair (Cohen's Kappa of 0.34 [0.28–0.42]), with a slight increase when it was based on recall/no-recall bins instead (0.37 [0.30–0.47]). In general, the radiologists tended to agree more between themselves on cancer cases than on non-cancer cases (Fig. 2b). The average agreement with the AI system was even lower (0.19 [0.02–0.35], Table 2). Similarly, most of the disagreement between readers and the AI system was rooted in the non-cancer cases (which the AI more often classified correctly) rather than the cancer-cases (Fig. 2b).

At an operation point of 99% specificity (sensitivity of 36.1% at specificity of 97.6%), the AI system was still able to detect four missed cancers (two cases for reader #3, one case for reader #2, and one case for reader #5). Importantly, the safety net application in the reader study did not add additional false alarms to any of the readers (Table 2).

In a post-hoc visibility analysis of then 10 FN cases (see Sect. 2.3), an independent expert has determined that 6 out of 10 FN were not visible at the index exam. Moreover, there was no association between the FN cases each reader has suspected to be malignant and their post-hoc visibility (p-value > 0.05, Fisher exact test). As such, the suspected lesions identified by the readers in those cases were most likely benign, and if biopsied, would have returned negative.

Table 2. Individual readers performance in the reader study.

	Reader 1	Reader 2	Reader 3	Reader 4	Reader 5
Specificity	50.0%	38.1%	57.1%	29.8%	34.5%
Sensitivity	83.3%	86.1%	63.9%	86.1%	72.2%
TP (with AI)	30 (30)	31 (32)	23 (25)	31 (31)	26 (27)
FP (with AI)	42 (42)	52 (52)	36 (36)	59 (59)	55 (55)
Mean κ with other readers	0.41 [0.25–0.57]	0.39 [0.22–0.56]	0.39 [0.24–0.54]	0.34 [0.17–0.50]	0.35 [0.18–0.52]
κ with AI	0.32 [0.16–0.49]	0.16 [–0.01–0.33]	0.28 [0.11–0.46]	0.10 [–0.06–0.26]	0.06 [–0.11–0.23]

Note—TP = true positives out of the 36 cancer cases. In parentheses – TP identified with AI safety net. Similarly, FP = false positives, in parentheses with AI safety net. κ refers to Cohen's Kappa agreement statistic, first with other readers (mean and CI), and then with the AI at sensitivity 87% (κ and CI).

4 Conclusions

In this work, we evaluated an AI system as a radiologist's safety net application. The safety net has contributed significantly to reader's sensitivity, especially when the analysis was based on DM alone, but also when combined with US. In a reader study, we demonstrated that even in a challenging dataset enriched with biopsy and FN cases, a safety net application could have benefited the readers. When the AI operated at the average sensitivity level of radiologists according to the BCSC, it had a low agreement with the readers, and as such, was in a better position to give useful insights, especially when there were no prior images or US available, such as in the case of women undergoing their first exam.

This analysis was not without limitations. Data originated from five different facilities of a single provider in one country, using a single mammography vendor (Hologic). In some cases, US was performed prior to the mammogram, and the radiologist may have been aware of the US report before analyzing the DM. Even so, according to the data, the existence of an US did not guarantee that the DM's BI-RADS was equal to or higher than the US's. In the reader study, the readers have operated in their regular environment, but did not have access to prior images or US, both essential tools in their daily work known to have impact on their performance. The AI system did not use those either.

To account for the lack of US, cases with original high BI-RADS due only to US were excluded from the study. Readers and AI had access to the same clinical data.

The AI safety net application was designed to interfere as little as possible with the radiologist routine; analyzing cases independently and raising a minimal amount of alarms as a second reader. Even under those restrictions, it demonstrated useful abilities. This is only one possible application of AI systems, but one we believe to be practical for immediate use.

Funding. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 813533.

References

1. Lehman, C.D., et al.: National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. Radiology **283**, 49–58 (2016)
2. Antonio, A.L.M., Crespi, C.M.: Predictors of interobserver agreement in breast imaging using the Breast Imaging Reporting and Data System. Breast Cancer Res. Treat. **120**, 539–546 (2010). https://doi.org/10.1007/s10549-010-0770-x
3. Nishikawa, R.M., Comstock, C.E., Linver, M.N., Newstead, G.M., Sandhir, V., Schmidt, R.A.: Agreement between radiologists' interpretations of screening mammograms. In: Tingberg, A., Lång, K., Timberg, Pontus (eds.) IWDM 2016. LNCS, vol. 9699, pp. 3–10. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41546-8_1
4. Katalinic, A., Bartel, C., Raspe, H., Schreer, I.: Beyond mammography screening: quality assurance in breast cancer diagnosis (The QuaMaDi Project). Br. J. Cancer **96**, 157 (2007)
5. Karssemeijer, N.: Effect of independent double and multiple reading of screening mammograms by breast density. 1136 words (2014). https://doi.org/10.1594/ecr2014/c-0358
6. Taylor-Phillips, S., Jenkinson, D., Stinton, C., Wallis, M.G., Dunn, J., Clarke, A.: Double reading in breast cancer screening: cohort evaluation in the CO-OPS trial. Radiology **287**, 749–757 (2018). https://doi.org/10.1148/radiol.2018171010
7. Leivo, T., et al.: Incremental cost-effectiveness of double-reading mammograms. Breast Cancer Res. Treat. **54**, 261–267 (1999). https://doi.org/10.1023/A:1006136107092
8. Lehman, C.D., Wellman, R.D., Buist, D.S.M., Kerlikowske, K., Tosteson, A.N.A., Miglioretti, D.L.: Breast cancer surveillance consortium: diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern. Med. **175**, 1828–1837 (2015). https://doi.org/10.1001/jamainternmed.2015.5231
9. Cole, E.B., Zhang, Z., Marques, H.S., Edward Hendrick, R., Yaffe, M.J., Pisano, E.D.: Impact of computer-aided detection systems on radiologist accuracy with digital mammography. Am. J. Roentgenol. **203**, 909–916 (2014). https://doi.org/10.2214/AJR.12.10187
10. Gao, Y., Geras, K.J., Lewin, A.A., Moy, L.: New frontiers: an update on computer-aided diagnosis for breast imaging in the age of artificial intelligence. AJR Am. J. Roentgenol. **212**, 300–307 (2019). https://doi.org/10.2214/AJR.18.20392
11. Akselrod-Ballin, A., et al.: Predicting breast cancer by applying deep learning to linked health records and mammograms. Radiology **292**, 331–342 (2019). https://doi.org/10.1148/radiol.2019182622
12. Yala, A., Lehman, C., Schuster, T., Portnoi, T., Barzilay, R.: A deep learning mammography-based model for improved breast cancer risk prediction. Radiology **292**, 60–66 (2019). https://doi.org/10.1148/radiol.2019182716

13. McKinney, S.M., et al.: International evaluation of an AI system for breast cancer screening. Nature **577**, 89–94 (2020). https://doi.org/10.1038/s41586-019-1799-6
14. Kim, H.-E., et al.: Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit. Health **2**, e138–e148 (2020). https://doi.org/10.1016/S2589-7500(20)30003-0
15. Schaffter, T., et al.: Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw. Open **3**, e200265 (2020). https://doi.org/10.1001/jamanetworkopen.2020.0265
16. Rodriguez-Ruiz, A., et al.: Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J. Natl Cancer Inst. **111**, 916–922 (2019). https://doi.org/10.1093/jnci/djy222
17. Siu, A.L.: Screening for breast cancer: U.S. preventive services task force recommendation statement. Ann. Intern. Med. **164**, 279 (2016). https://doi.org/10.7326/M15-2886. On behalf of the U.S. Preventive Services Task Force
18. Alcusky, M., Philpotts, L., Bonafede, M., Clarke, J., Skoufalos, A.: The patient burden of screening mammography recall. J. Womens Health **23**, S-11 (2014). https://doi.org/10.1089/jwh.2014.1511