Check for updates

Operating Point Optimization for Efficient Mammogram Triage Using Only Highly Elevated Probability scores

Mark Traill,[1,2,3] Jensen Jantz,[4] Blair Richards,[5] Shokoufeh Khalatbari,[5] Jeffrey Hoffmeister,[6] Julie Shisler,[7] Crystal Snyder[8]

Abstract

Artificial intelligence (AI) detection algorithms assign probability scores for malignancy on screening mammograms. Using 566 exams (53 cancers), we assessed detection, risk, and multi-score models to define efficient triage thresholds. ROC analysis showed very high scores achieved excellent specificity and positive predictive value with minimal false positives. A case-score threshold near 85 identified many cancers while limiting workflow burden. Initiating triage at these "super-specific" scores and lowering as logistics allow may expedite diagnosis and treatment.

Backgroud: Using AI breast cancer detection scores to triage screening mammogram patients has been described in several studies. This type of triage could expedite a cancer diagnosis and significantly shorten the time from screening mammogram to treatment. However, the optimal operating point, or score value, above which patients should be triaged, has not been well defined. Identifying practical operating points is important because the Positive Predictive Value (PPV), sensitivity, specificity, and false-negative (FN) rate at different operating points can vary greatly. **Materials and Methond:** Using patient datasets constructed from a clinical practice, the performance metrics of various AI scoring models were calculated at different operating point levels. **Results:** Mathematical analysis of the very highest cancer detection case scores revealed that they are highly specific for the presence of breast cancer with few false positives. Therefore, if only the few highest case scores are used for clinical triage, the expectation is that most patients triaged will have cancer. **Conclusion:** In clinical practice, this approach could be used to identify patients who would benefit from immediate diagnostic workup without overwhelming the clinic logistics with false-positive cases.

Clinical Breast Cancer, Vol. 26, No. 1, 73–85 © 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Keywords: Logistic regression models, Artificial intelligence, Image based risk assesment, Mammography scheduling, Delayed cancer treatment

Abbreviations: AI, Artificial intelligence; AUC, Area under the curve; CI, Confidence intervals; DBT, Digital breast tomosynthesis; FN, False Negative; FP, False Positive; NPV, Negative predicative value; OR, Odds ratio; PPV, Positive predictive value; ROC, Receiver operating characteristic curve; SD, Standard deviation; TN, True Negative; TP, True Positive.

[1]Department of Radiiology, Michigan State University College of Osteopathic Medicine, East Lansing, MI
[2]MammoBot PLLC, Niles, MI
[3]Department of Medical Imaging, Formerly at University of Michigan Health-West, Wyoming, MI
[4]Department of Clinical Research, University of Michigan, Health West, Wyoming, MI
[5]Michigan Institute of Clinical and Health Research, University of Michigan wide resource hub, Ann Arbor, MI
[6]iCAD, Inc. Manhattan Beach, CA
[7]JLS Consulting, Jupiter, FL
[8]Deephealth, Inc. Troy, OH

Submitted: Jul 8, 2025; Revised: Nov 11, 2025; Accepted: Nov 18, 2025; Epub: 21 November 2025

Address for correspondence: Mark Traill, MD, Department of Radiology, Michigan State University College of Osteopathic Medicine, 846 Service RdEast Lansing, East Lansing, 48824, MI.
E-mail contact: mtraill@mammobot.com

Introduction

AI algorithms have repeatedly been shown to be accurate at detecting breast cancer on a screening mammogram.[1-6] AI image-based algorithms can also estimate the risk of developing breast cancer in the short term of 1 to 5 years based on mammogram findings.[7-11] For both detection and risk assessment, a numerical probability score output is generated where the higher the score, the more likely the mammograms contains, or will soon contain, identifiable breast cancer respectively. However, the details of exactly how prevalent cancers are at the very high scores is often overlooked. For this study, the probability score for cancer detection is called the "case score," and the probability score for cancer risk is called the "risk score." Statistical analysis of score performance at the highest levels could provide insight into how these values might be applied for clinical triage.

1526-8209/$ - see front matter © 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.
https://doi.org/10.1016/j.clbc.2025.11.007

Recently, large data studies have demonstrated that algorithm predictive performance can be improved further by combining multiple scores, or by looking at how scores change over time.[12-15] These new techniques are referred to as multi-score fusion and time-series-based evolution, respectively. Newer commercial cancer detection algorithms are starting to incorporate a detection score change over time analysis, but models based on combining multiple scores are mostly research tools that are not available to practicing radiologists working in community clinics. Also, there has been little published on the change over time performance for a commercial risk assessment algorithm alone. While these advances improve AI precision, questions remain about how to translate such performance into practical, clinically deployable triage strategies.

Despite ongoing progress, there is little consensus regarding how to define or apply operating points. The score thresholds determine when a patient should be triaged for further evaluation. This analysis generated AUC, sensitivity, specificity, NPV, and PPV values that could be compared. This gives functional insight on score performance that would allow an optimal operating point to be chosen. This is a noninterpretive AI function, meaning a radiologist input is not necessary for the task as it is categorizing mammograms.

While using AI for triage has been discussed and validated in multiple prior studies,[16-22] these studies mostly used case score cut off values, ie, operating points, that were not well explained and difficult to understand. This resulted in large numbers of patients being triaged associated with marked logistic pressure on the existing workflow. In general, the triage approach focused on the presence or absence of abnormal mammogram findings with the expectation that some, but not the majority, of these findings were malignant. The focus of this approach is trying to achieve sensitivity. These were "bottom-up" models where the operating point started at a lower score than the department could handle and was adjusted up to make it more efficient and more specific. These inconsistencies highlight the need for a practical framework to determine AI score thresholds that maximize cancer detection efficiency while minimizing workflow burden.

With a theoretical triage model focusing on high specificity over sensitivity, there would be less impact on department logistics since fewer patients would enter the triage process. With this type of operating point, a larger percentage of patients sent for triage would be expected to currently have cancer, and not just questionable mammogram findings. More cancers would be missed at the higher operating point with less sensitivity, but there could also be a score range that would still expedite the care of many of the cancer patients without placing an unreasonable burden on department logistics or subjecting to many patients to the stress of a false positive triage experience. A high specificity triage model would be much easier to implement in a busy clinical practice and could be considered a "top-down" model where the case score operating point could be adjusted down from the highest level if department logistics allow it.

The purpose of this study was to evaluate AI-derived case and risk scores to identify operating points that achieve high specificity and positive predictive value for breast cancer detection while minimizing false positives and workflow disruption. We hypothesized that triaging only patients with the very highest AI scores would identify a small, high-risk subgroup likely to have cancer, enabling faster diagnosis and treatment without overwhelming clinical logistics.

Methods

This institutional review board (IRB) approved, Health Insurance Portability and Accountability Act (HIPAA) compliant retrospective study under waiver of informed consent used 2 different AI algorithms to obtain case and risk scores from mammogram studies on 566 women who underwent DBT mammography (Pristina; GE) at our institution, University of Michigan Health West. The collection period of study cases was from December 2020 to December 2022. Only DBT images were used for algorithm analysis. The patient cohort was intentionally cancer enriched, as this is a typical method used for algorithm testing. Scores from the detection and risk algorithms were obtained on each patient, and then the scores for the patients with biopsy proven cancer ($n = 53$) were separated and compared to non-cancer patients ($n = 513$). All cancer patients were biopsy proven no longer than a few weeks after the DBT exam. All non-cancer patients had a BI-RADS score of 1(negative) or 2 (benign) on the initial interpretation of the mammogram used for enrollment, and they remained cancer free for at least 1 year after enrollment. All of the initial 566 scores are called baseline scores for the study. All patients also had a second risk score, called "prior" risk score, obtained from the most recent DBT mammogram done before the mammogram used for the baseline scoring. The time interval between the baseline and most recent prior mammogram was recorded.

Scoring was done using a commercially available breast cancer detection algorithm, iCAD ProFound AI Detection v2, iCAD, Inc., Nashua, NH, which was trained to identify existing cancer on the mammogram based on morphologic findings and asymmetries. Only v2 detection software was used, and there were no hardware upgrades during the data accumulation period. The overall chance of malignancy, or probability score, on the mammogram from this detection algorithm is called a case score. The probability score is also sometime referred to as a detection score. Mammograms were also scored using an image-based risk assessment algorithm, Profound AI Risk v2, iCAD, Inc., Nashua, NH. The algorithm assesses a woman's 1 to 2-year short-term, absolute risk of developing breast cancer based on age, breast density, and mammographic breast features such as asymmetric findings, masses, calcifications, and other morphologic changes. The risk assessment algorithm was only available for retrospective research and is currently undergoing FDA review. These baseline scores were collected after the mammograms had been read by 1 of 2 radiologists; 1 radiologist was fellowship trained in mammography.

The radiologists used the ProFound detection algorithm at the time of initial image interpretation, before the scores were retrospectively collected. There was no awareness of the study by the radiologist at the time of initial interpretation. At the time of the study and submission, the product was developed by iCAD, Inc. During peer review, iCAD, Inc. was acquired by DeepHealth, which now owns the products.

AI algorithm scores and clinical information was collected from the hospital PACS by a single clinical research nurse who was also responsible for data organization. Patients had the timelines for all

mammograms recorded. Patient age and breast density were noted. Patients with a history of any prior breast surgery were excluded from the study because postoperative changes can mimic the appearance of malignancy and cause false positive very high scores. Patients with a history of needle biopsy were not excluded as they typically do not develop morphologic changes identical to malignancy that would cause a false positive high score. There were no male patients in the study.

Cohort A was the non-cancer group. Starting in December 2020, consecutive mammograms were enrolled from women who had a screening DBT mammogram with no cancer detected, BI-RADS 1 or 2, by the interpreting radiologist at the time of the mammogram interpretation. All non-cancer women also had at least a 1-year cancer free follow up period that was confirmed by review of our institution's cancer registry.

Cohort B consisted of nonconsecutive women who had a DBT mammogram associated with a new breast cancer after a no-cancer, BI-RADS 1 or 2, mammogram 0.7 to 2.5 years prior to the new cancer diagnosis. For 12/53 of these patients, the DBT mammogram used was part of a diagnostic workup, either symptomatic or nonsymptomatic. One cancer could only be seen on MRI, and 1 could only be seen by ultrasound. All cancers were confirmed with biopsy. Final pathology reports were reviewed and the characteristics of women diagnosed with breast cancer were summarized for estimated percentages of breast density, histology, invasive grade, stage and tumor subtypes. Cancer cohort patients were collected between December 2020 and December 2022. This was a longer period of collection than the non-cancer cases as the cases were less frequently encountered, and some were not added until late in the collection period.

Combining the 2 Cohorts resulted in a cancer enriched test population, with the non-cancer to cancer ratio was approximately 10:1. Cancer enriched populations are traditionally and commonly used to test algorithm performance.

The collected AI scores for each patient included baseline detection score, baseline risk score, most recent prior risk score, and change in risk score from most recent prior to the baseline exam. Logistic regression models were used to generate odds ratios (ORs) and 95% confidence intervals (CI) for accessing how well the scores were able to separate the cancer and non-cancer mammograms. ROC curves were also constructed to obtain AUC performance numbers. In addition, sensitivity, specificity, positive predictive value, and negative predictive value were obtained on each model at different operating points.

ROC curve analysis assessed predictive discrimination for 3 single score models: baseline case score, prior risk score, and risk score change. ROC curve analysis also assessed predictive discrimination for 2 models that had multiple scores: baseline case score + risk score change, and baseline case score + risk score change + prior risk score.

Given the novel nature of the risk score change metric, additional comparisons and observations were made. For this, Sankey diagrams displaying risk score changes from prior to baseline mammogram within cohorts of cancers and non-cancer were constructed. Sankey diagram groups were created by quartiles of the range in 1-year absolute risk at prior and baseline mammograms.

Results

The study Cohort A was composed of 513 women without breast cancer (mean age = 55.7 years; SD = 8.2 years), and Cohort B was composed of 53 women with breast cancer (mean age, 57.0 years SD = 8.9 years). All patients had a baseline mammogram, dated between January 2020 and December 2022, that was compared with their most recent prior mammogram. The time between the 2 comparison mammograms was slightly longer in the cancer group with a mean of 1.4 years (SD = 0.4) compared to 1.1years (SD = 0.1) for the non-cancer group, for a difference of 0.3 years (SD = 0.2). The longest period between mammograms was 2.5 years for the cancer group and 1.7 years for the non-cancer group, see Table 1.

The prior risk score mean for non-cancers was 0.30 (SD = 0.33), and the baseline mean was 0.31 (SD = 0.35), a difference of 0.01. The prior risk score mean on cancer cases was 0.47 (SD = 0.40) and the mean on the baseline exam at the time of new cancer diagnosis was 0.98 (SD = 0.45), for a difference of 0.51.

The women with breast cancer detected were analyzed for breast density and cancer characteristics. 62% (33/53) of the women had dense breast tissue. Eighty five percentage (45/53) of the cancer patients had an invasive malignancy with the most common cell type (77%; 35/45) being ductal carcinoma. Only 15% (8/53) of the cancer cohort were diagnosed with ductal carcinoma in situ (DCIS), see Table 2.

Table 3 contains the comparison of the 5 model ROC curves. The lowest AUC of 0.662 was the model based only on the prior risk score. The best single score model, with the third highest AUC of 0.891, was based on baseline case score alone. The highest AUCs overall were the multicore models of baseline case score + risk score change and baseline case score + risk score change + prior risk score models with AUCs of 0.901 and 0.903 respectively.

The comparison of the logistic regression ROC graphs for the single score models are shown in, see Figure 1. These results show that there is a very high specificity in the models for identifying which baseline studies have cancer and which do not at the high score probability thresholds. Table 4 breaks down the specific predictive probability values for base line case score ROC results. Based on the estimated probability threshold of 0.53 from the baseline case score ROC results, the positive predicted value (PPV) was 71.4, the specificity was 98.4%, and the negative predictive value was 93.9%. The table shows that increasing the score threshold will extract progressively fewer total patients, but at a higher specificity. So, the higher the predicted probability threshold, the fewer patients extracted, the higher the PPV, and the more specific the results. As the score threshold increases to the 71.4 PPV level, 28 patients would be extracted which is just below 5% of the total. Of these patients, 20/28 have cancer. At the 0.53 probability estimate threshold, 33 cancer patients were below the score threshold as false negatives. Going to the next highest probability level in Table 4, 23 patients would be extracted with 19 cancers. Additional graphic representation of these results for the baseline case score predictive probability is shown in Figure 2, which includes 95% confidence limits.

Risk scores change over time was demonstrated to be associated with new malignancy. The Sankey diagram comparison shown in

Table 1 Patient Characteristics of the Study Dataset and Cohort

Characteristic	Overall ($n = 566$)	Yes ($n = 53$)	No ($n = 513$)
Estimated age (Yrs.) at prior mammogram			
Mean (SD)	53.8 (8.3)	54.9 (8.9)	53.7 (8.2)
Median (IQR)	54 (47-61)	55 (49-63)	54 (47-60)
Range	36-69	36-68	39-69
Years from prior to baseline mammogram			
Mean (SD)	1.16 (0.18)	1.38 (0.41)	1.14 (0.12)
Median (IQR)	1.13 (1.03-1.24)	1.28 (1.11-1.49)	1.12 (1.03-1.23)
Range	0.65-2.46	0.65-2.46	0.89-1.69
Prior risk score (%)			
Mean (SD)	0.31 (0.34)	0.47 (0.40)	0.30 (0.33)
Median (IQR)	0.17 (0.07-0.41)	0.35 (0.15-0.71)	0.15 (0.07-0.39)
Range	0.00-1.49	0.01-1.42	0.00-1.49
Baseline risk score (%)			
Mean (SD)	0.37 (0.41)	0.98 (0.45)	0.31 (0.35)
Median (IQR)	0.17 (0.07-0.51)	1.21 (0.67-1.28)	0.14 (0.07-0.40)
Range	0.01-1.89	0.06-1.74	0.01-1.89
Risk score change (baseline-prior risk)			
Mean (SD)	0.06 (0.35)	0.51 (0.47)	0.01 (0.30)
Median (IQR)	0.01 (−0.09 to 0.17)	0.46 (0.05-0.90)	0.00 (−0.10 to 0.11)
Range	−1.21 to 1.59	−0.17 to 1.59	−1.21 to 1.12

Figure 1 Single score model ROC curves including novel risk score change approach.

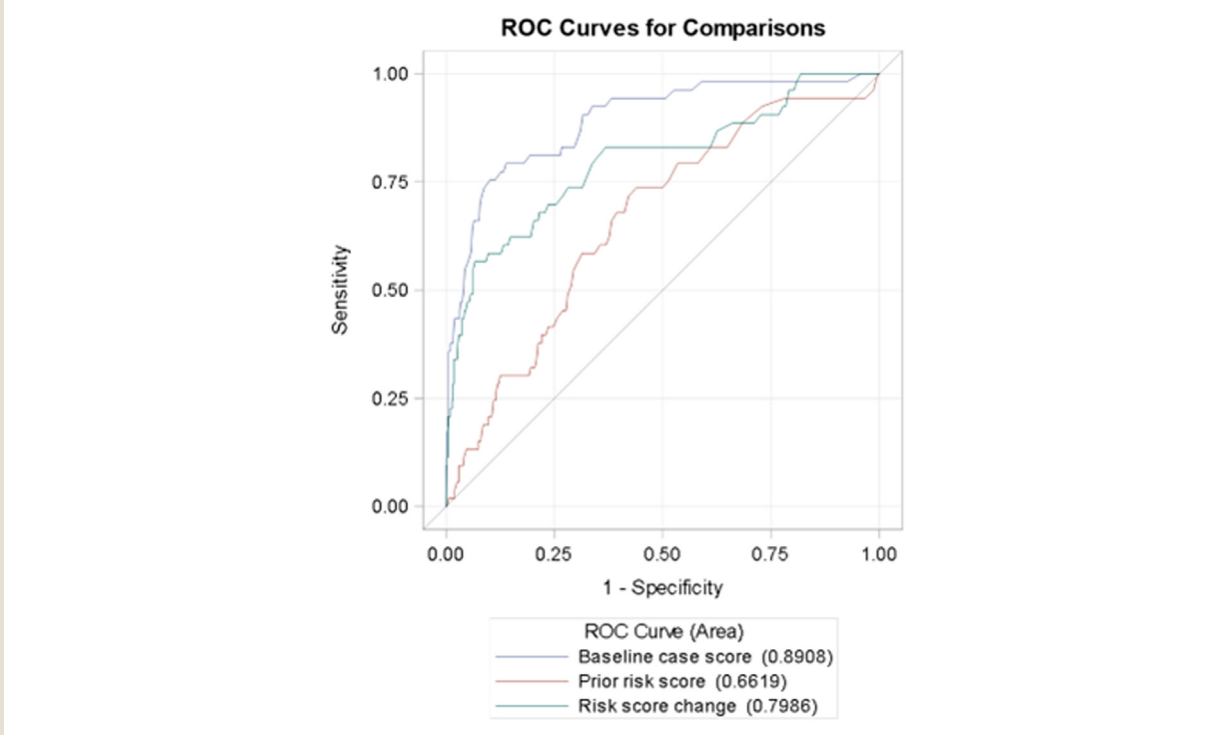


Table 2 Cancer Cohort Characteristics

No. of Women	53
Age at Baseline Mammogram	**57.0 ± 8.9**
Density	
Fatty	8 (15%)
Scattered fibro glandular	12 (23%)
Heterogeneously dense	21 (40%)
Extremely dense	12 (23%)
Histology	
Invasive	45 (85%)
Ductal	35
Lobular	8
Other/mixed	2
DCIS	8 (15%)
Invasive grade	
1	21 (47%)
2	13 (29%)
3	11 (24%)
stage	
0	8 (15%)
1	35 (66%)
2	5 (9%)
3	5 (9%)
4	0 (0%)
ER positive	40 (75%)
ER negative	7 (13%)
ER mixed	6 (12%)
HER2 negative	39 (87%)
HER2 positive	6 (13%)

Abbreviations: DCIS = ductal carcinoma in situ; ER = estrogen receptor; HER2 = human epidermal growth factor receptor 2.

Figure 3 demonstrates that many of the cancers showed an increase in risk score quartiles compared to the non-cancer group. In the non-cancer cohort, 73% of the patients are in the bottom quartile of risk scores on the prior mammogram with 74% still in the bottom quartile of risk scores on the baseline mammogram. For the cancer cohort, 55% are in the bottom quartile on the prior mammogram, with this number shrinking to only 15% in the bottom quartile on the baseline study. For the top quartile of risk scores, there is no percentage difference between the prior and baseline mammograms for the non-cancer group, while the cancer cohort has a jump from 13% to 57% in the top quartile. In the non-cancer group, only 2.8% of patients (13/457) move from the lowest 2 quartiles (< 0.75%) to the highest quartile (≥ 1.125%). This compares to 50.0% (20/40) of patients in the cancer cohort, making this jump in scores.

Table 3 Comparison of Model ROC Curves

Model	AUC (95% Confidence Limits)
Case score (baseline)	0.891 (0.841, 0.941)
Risk score (prior)	0.662 (0.587, 0.737)
Risk score change	0.799 (0.726, 0.872)
Case score (baseline) + risk score change	0.901 (0.854, 0.948)
Baseline case + risk score change + prior risk score	0.903 (0.856, 0.950)

Abbreviation: AUC = area under the curve.

Specific predictive probability values for the risk score change model are presented in Table 5. At the higher probability levels, the numbers are very similar to the case score model. Likewise, the predictive probability curve for risk score change is very similar to the case score model, see Figure 4.

Figure 5 shows the ROC curves for the 2 multi-score models. The overall AUC values were only slightly improved from case score alone. But the curves do show better performance, a higher and flatter curve, in the 0.5 to 0.75 specificity range. Predictive probability values were similar to the single variable model based on risk score change. Since the improvement over the case score model was small, and combining multiple scores more complicated, the single variable model based on case score was used for the discussion of how the upper limit high scores could be used in a theoretical triage model focused on specificity.

Discussion

Our analysis identified operating points at the highest case score values that categorize mammograms as containing cancer with a specificity in the high 90's. At the highest operating points, most

Table 4 Baseline Case Score

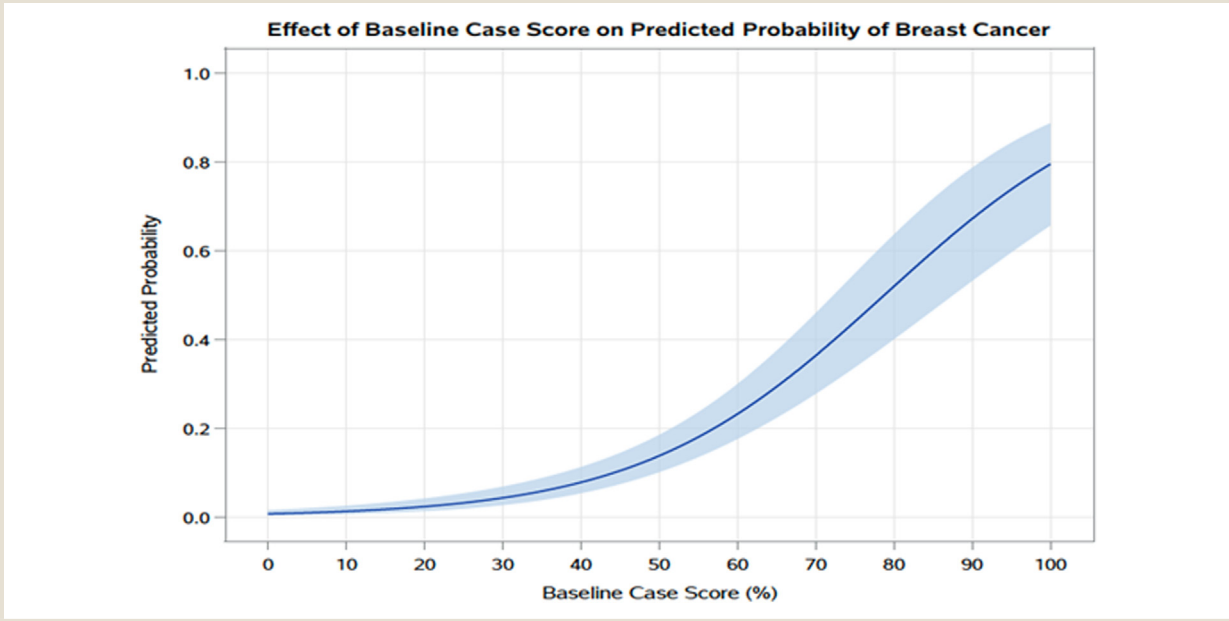
Probability Estimate Level	Correct		Incorrect		Percentages				
	TP	TN	FP	FN	Correct	Sensitivity	Specificity	Pos Pred	Neg Pred
0.160	40	456	57	13	87.6	75.5	88.9	41.2	97.2
0.180	38	468	45	15	89.4	71.7	91.2	45.8	96.9
0.220	35	475	38	18	90.1	66.0	92.6	47.9	96.3
0.250	32	483	30	21	91.0	60.4	94.2	51.6	95.8
0.310	27	492	21	26	91.7	50.9	95.9	56.3	95.0
0.430	23	499	14	30	92.2	43.4	97.3	62.2	94.3
0.530	20	505	8	33	92.8	37.7	98.4	71.4	93.9
0.600	19	509	4	34	93.3	35.8	99.2	82.6	93.7

Table 5 Logistic Regression for Risk Score Change

Classification Table									
Prob Level	Correct		Incorrect		Percentages				
	TP	TN	FP	FN	Correct	Sensitivity	Specificity	Pos Pred	Neg Pred
0.060	39	365	148	14	71.4	73.6	71.2	20.9	96.3
0.120	31	448	65	22	84.6	58.5	87.3	32.3	95.3
0.160	30	472	41	23	88.7	56.6	92.0	42.3	95.4
0.250	24	490	23	29	90.8	45.3	95.5	51.1	94.4
0.440	18	501	12	35	91.7	34.0	97.7	60.0	93.5
0.610	8	511	2	45	91.7	15.1	99.6	80.0	91.9

Tie probability levels back to risk score change value.

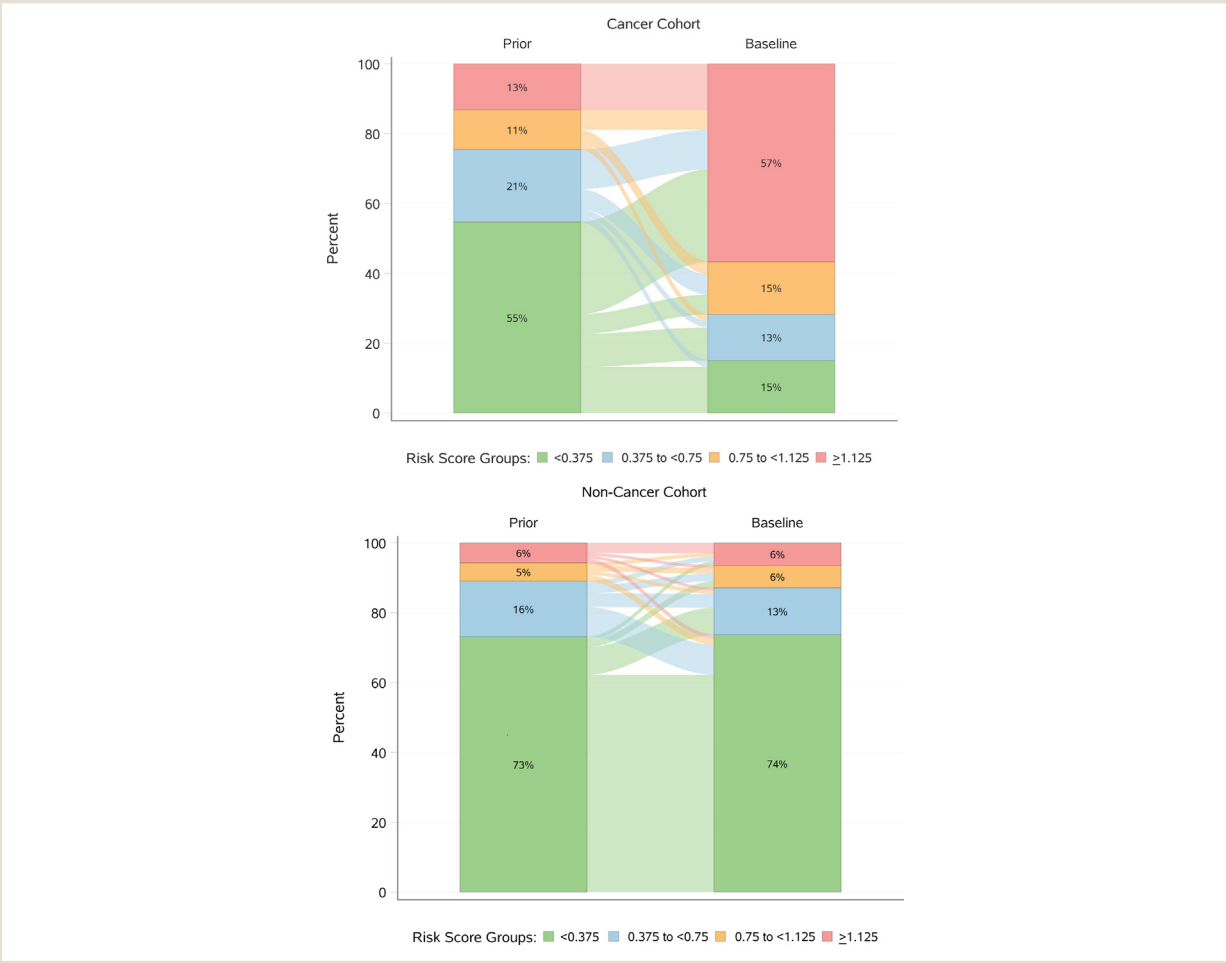
Figure 2 Predictive probability curve for baseline case score illustrates strong performance for predicting cancer or no cancer on screening mammogram at the highest score levels.



of the mammograms selected would contain cancer with the true positives outnumbering the false positives by 2:1 or higher. In a clinical practice, triaging patients with a score of 99 or 100 would expedite cancer care in most of that group without overwhelming the department with additional triage patients. After implementing and validating this approach over time, if the clinic wishes to capture more cancer cases and increase triage sensitivity, the operating point could be lowered to 97 or 98 and re-evaluated. Multiple prior studies[16-23] have looked at using a percentage of the highest scores to triage screening mammograms. This triage would hopefully put the most suspicious mammograms at the head of the reading list to decrease the turnaround time to biopsy and possible treatment. In the studies, the operating point involved extracting patients who had scores in the top 10% to 30%. It was typically not clear exactly what score value was used as the cut off, or if this percentage meant the top 10% of scores or the top 10% of patients based on their scores.

In a 2025 study, Friedewald et al.[23] reported their experience with this type of triage using AI scores. In this 1000 patient prospective study, they triaged patients for same day diagnostic study based on probability score value. They found a decrease in turnaround time to completion of diagnostic studies and to completion of biopsy of 25% and 30% respectively for the triaged patient cohort that had suspicious high scores and were extracted for expedited reading. Specifically, the time until completion of biopsy decreased from 55.9 days to 39.2 days. However, the score operating point they used for extracting a suspicious case for immediate review resulted in a total patient extraction rate of 28%. The paper acknowledged that such a high extraction rate was a definite burden on the department's schedule logistics that required staffing modification. They

Figure 3 Sankey diagram of the non-cancer and cancer cohort risk score quartiles. Comparison shows how risk scores jumped from the lowest quartile to the highest much more frequently in the patients who developed cancer. Temporal risk score elevation as an indicator of interval change in the breast morphology and possible development of malignancy.



also reported that some radiologists found it more difficult to call some cases normal just because they were categorized in the triage group. In conclusion, this study did support triage based on AI scores, and the authors felt there were the additional benefits of less patient stress while waiting for results. Also mentioned was the possibility of better patient compliance since the patient did not have to schedule a second return trip to the department for diagnostic studies. They did not, however, mention the additional patient anxiety that may be generated by triaging a patient only to tell them nothing is wrong.

After further reviewing their results, the researchers did an ad hoc simulation using a more specific extraction rate of 10%. This was done to define a triage model that would be less disruptive to department logistics. Using this lower extraction rate, they found that they were still able to extract 2 out of 3 experimental cancers with less disruption to the schedule. This would be consistent with the data for our analysis that shows that by increasing the score operating point, you will still triage a significant number of cancers, but with fewer false negatives. Our data would support decreasing the number of patients triaged to an even more practical and efficient level.

In all models analyzed for this study, the ROC curves show that at the highest score values, specificity becomes very high and PPV reaches its highest levels. Using only the highest score values would yield the highest specificity for triage. The multi-score fusion models had the best performance, but by only a small margin. The risk score change model also performed well but was below the other model's performance. In the typical clinical setting, the cancer detection case score would be the most readily available and simplest option for use in this type of selective triage. The triage focus would be on only extracting patients with a very high probability of cancer, accepting that not all cancers would be caught. Specificity would be favored over sensitivity.

Reviewing the classification for baseline case score (see Table 4) shows how specificity, sensitivity, PPV, and patient extraction number change as at the highest score levels. At a probability level of 0.53, 28 patients would be extracted with 20 cancer and 8

Figure 4 Risk score change probability curve with results similar to baseline case curve.

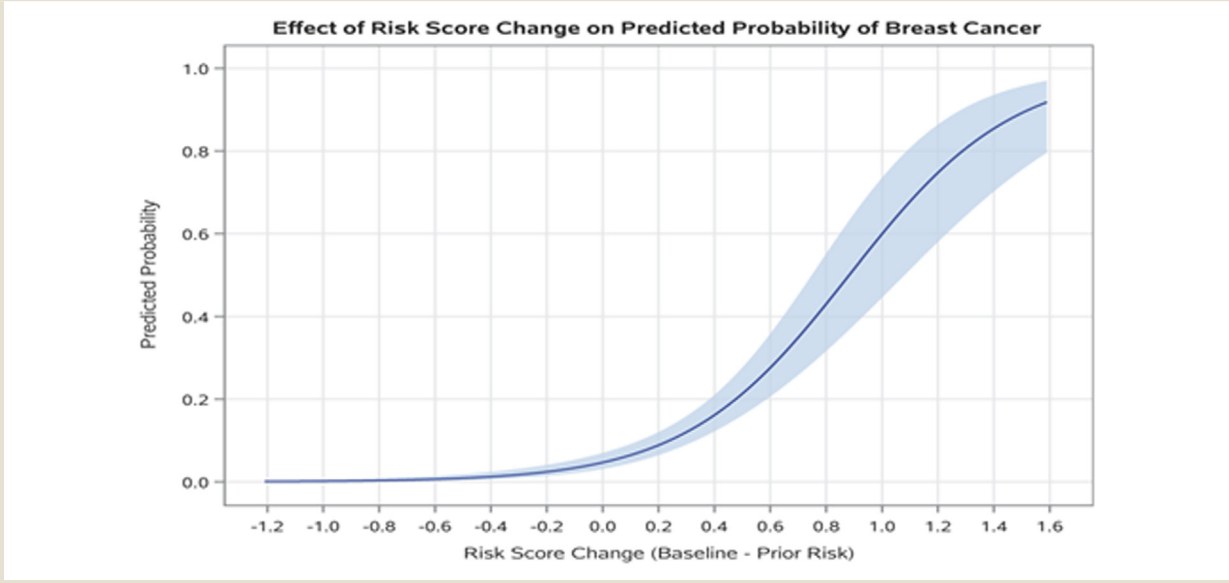
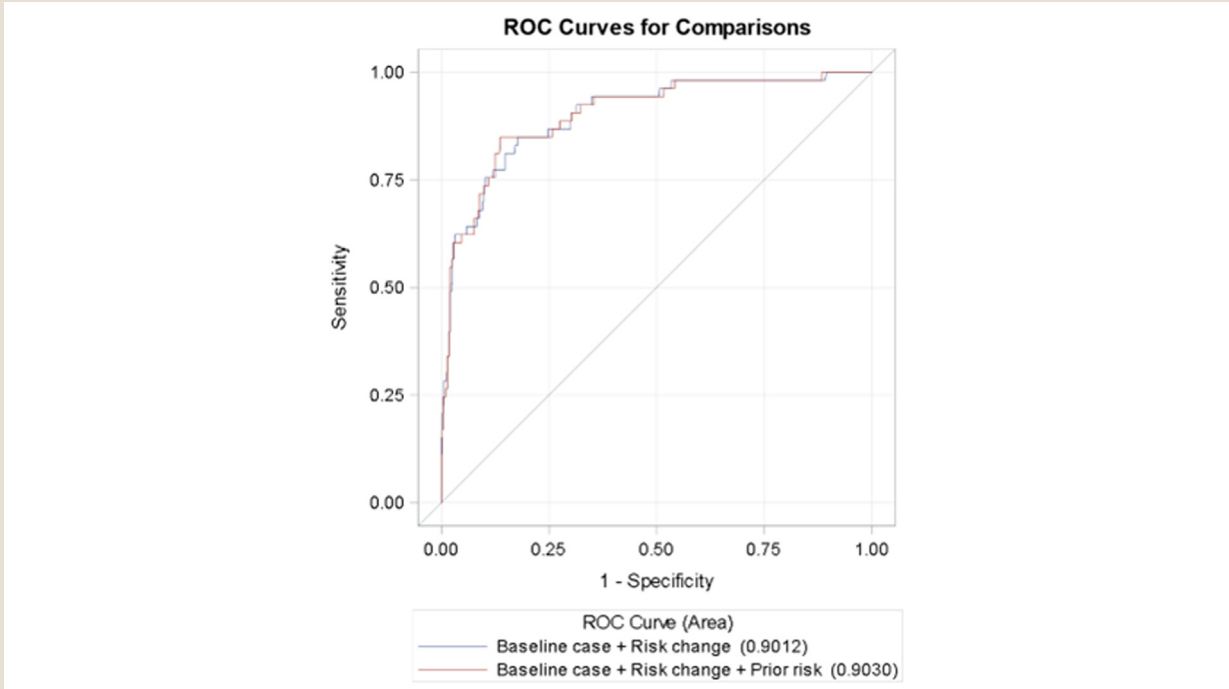
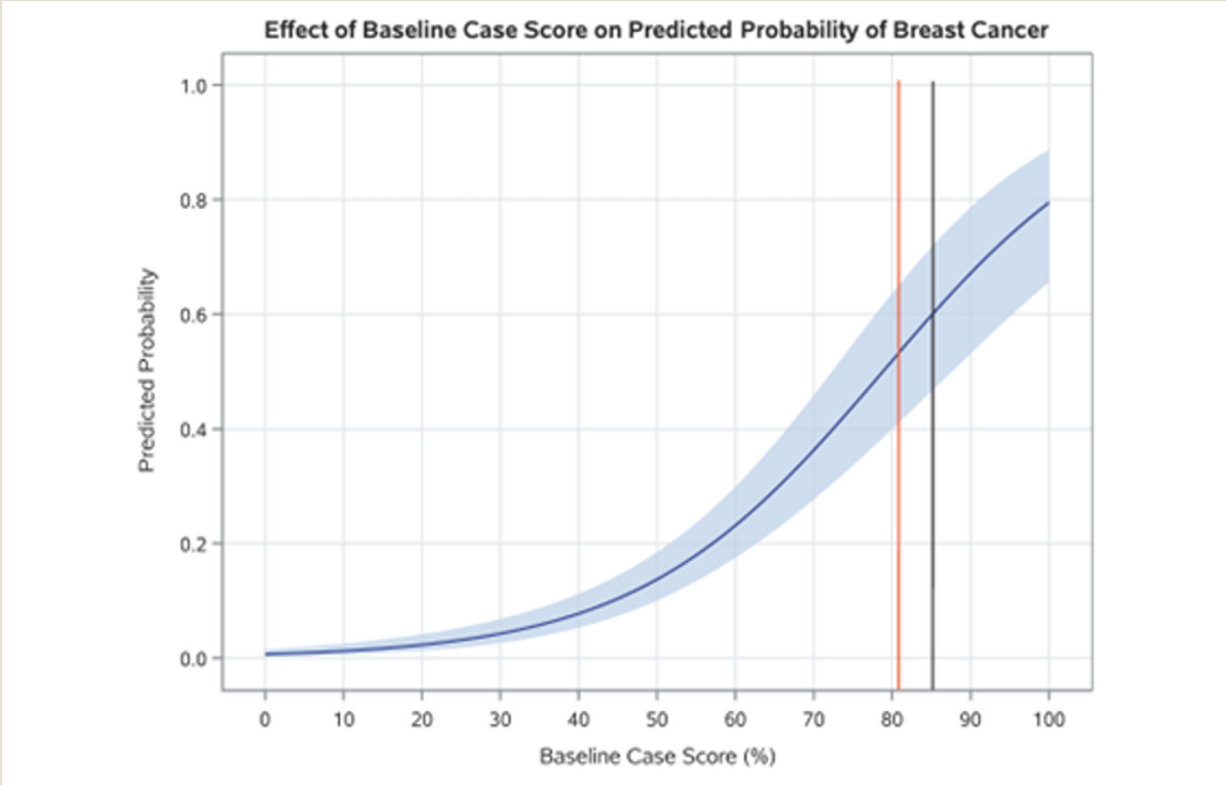


Figure 5 ROC cures for multi-score fusion models.



false positives with a PPV of 71.4. In a clinical triage application, limiting the triage to a small percentage of the highest case scores would extract a smaller number of patients, but you would expect a higher percentage of those patients to have malignancy with a lower percentage of false positives. Assuming a similar Baseline Case Score distribution in a hypothetical sample with the typical clinical incidence of 6 cancer cases per 1000 screened, the probability of having a false positive would remain the same (8/513 = 0.01559). So extrapolating this to a sample with a cancer rate of 6 per 1000 would be 0.01559×994 non-cancer patients = 15.50. So,

Figure 6 The red vertical line shows a specific case score value of 81 for 0.53 probability estimate level operating point.



the estimated number of false positives would be 15.50 at the 0.53 probalility level . And the number of cancer cases identified would be estimated by $0.377 \times 6 = 2.26$. So about 2.26 cases would be correctly identified, with 15.5 false positives. The specific operating point for this probability level would be a case score of 81 (see Figure 6).

However, if you move to the 0.60 probability level threshold, there would be very similar numbers of cancer patients estimated to be identified ($0.358 \times 6 = 2.15$), but considerable improvement in the number of false positives estimated ($4/513 = 0.0078$; $0.0078 \times 994 = 7.75$). So about 2 cases would still be correctly identified, but with about half the number of false positives.[8] In practical terms, if a clinic did 100 mammograms per day for 10 days, 1% of the patients[10] would be triaged with 33% of the expected cancer patients identified. The specific operating point for this probability level would be 85 (see Figure 6).

This result expands on the similar finding previously reported by Fiedewald et al.[23] in their ad hoc analysis. Our numbers indicate that even if you only triage 1% of the total patients, you will still find a significant number of cancers. These results are due to the very high specificity in the highest score values.

The exact case score to use as the operating point in a real-life, top-down, clinical triage application would not be expected to exactly match our study results as different software versions, patient population demographic differences, and hardware configurations can all affect high score specificity. However, the best operating point could easily be found by starting the triage program with a score near the highest probability level, monitoring how the triage is progressing, and then adjusting to a slightly lower score as indicated, see Figure. 7. Starting at the top of the score range would still be expected to triage only a small percentage of patients with most of them having cancer. Post-op cases can easily be removed from the triage process by the technologist performing the mammogram procedure as they have access to the surgical history and are able to clinically examine the patient for the presence of scars. Also, many of the post-op patients will likely have a cancer history.

Importantly, integrating such a triage model into clinical workflow could do more than just expedite diagnostic imaging. It could also expedite biopsy for patients flagged with high suspicion of having cancer. This is especially important for women with fast-growing malignancies, where a lengthy delay in treatment could impact outcomes.

The process of "fast-tracking" patients with a suspicious clinical finding for immediate diagnostic studies, specifically a palpable mass, is already the standard of care in the United States.[24] Recently published research on what percentage of palpable masses turn out to be malignant is not extensive, but available studies report a frequency of approximately 10%.[25,26] A 2017 study evaluated 935 palpable areas in 861women found that 39.3% (367/935) had no abnormal mammogram or ultrasound findings. Only 8.2%

Figure 7 Workflow for clinical validation of AI at a new site or clinical site. The Blue vertical line shows the 0.60 probability estimate level corresponds to a specific case score of 85. Moving the operating point/case score for triage from a case score of 81 to 85 reduced the false negative triage cases from 8 to 4 in a theoretical population of 6 cancers per 1000. The number of cancers triaged remained stable at 2 for both operating points.



(77/935) of these were diagnosed as cancer.[27] Our study suggests that releasing a patient to go home after a screening mammogram with a very high case score may lead to a greater delay in diagnosis and worse outcomes than releasing a patient with a newly identified palpable mass at the time of screening.

There is no MQSA rule for turnaround time limits. A study presented at the American Society of Breast Surgeons Annual Meeting in April 2024 evaluated the timeliness of care at National Accreditation Program for Breast Centers (NAPBC) facilities.[28] The study analyzed data from 373 breast centers. For 2021, they found the median time from screening to diagnostic study to be 12 days, from diagnostic mammogram to biopsy to be 9 days, and from biopsy to first surgery to be 42 days, see Table 6. They also noted that all these times were longer than patients expected. Taken together, the median time from screening mammogram to first surgery was 63 days, with half of the centers studied taking longer. 19% of the delay to first treatment would be the 12 days waiting for a diagnostic mammogram after an abnormal screening study. It is also noted that the times reported are the median, so in an understaffed department, an individual patient could experience much longer delays. The lead author of the PROMPT (Patient-Reported Observations on Timeliness of Breast Services) survey study, Daniella Thompson MD from the University of Chicago, commented "Study data suggests that there are significant opportunities to improve timeliness of care throughout a process that is often rife with patient anxiety."[28-29]

The American Society of Breast Surgeons (ASBrS) does use 60 days until first treatment as a quality metric.[30] This standards statement also notes that patient satisfaction is improved with timely care while outcomes may be worse if care is delayed. A literature review study done by Belcher et al.[31] in 2018 also concluded that delays in getting to the first treatment can impact outcomes. Multiple studies have also described increased patient anxiety associated with mammography and especially waiting for a diagnosis.[32-34]

Triage can make a difference for patients with breast cancer, especially if it is aggressive and rapidly growing. The shortage of mammographers[35,36] imposes limits on the ability of a department to accommodate the extra diagnostic studies that would result from using an operating point that is less specific with too many false positives to be practically implemented.

Including other scores in the triage model could potentially improve the specificity of the process. One additional metric that

Table 6 Breast Cancer Care Timeliness Metrics From Patient and Center Perspectives From PROMPT Study

		Screening MGM to Diagnostic MGM	Diagnostic MGM to Biopsy	Biopsy to First Surgery	Biopsy to Neoadjuvant Therapy
Patient perspective		5 d	7 d	21 d	Not collected
Center perspective		7 d	7 d	28 d	21 d
Reported timelines metric					
	2019	11 d	8 d	39 d	33 d
	2020	11 d	8 d	40 d	32 d
	2021	12 d	9 d	42 d	34 d

could be helpful is the interval change in risk score. At the minimum, any score changes of 1.2 or higher should trigger extra concern for cancer on the mammogram. It appears that while detection and risk algorithms are trained differently, using them together can be complimentary.

Limitations of this study include a retrospective design and a relatively small patient cohort. However, our results did have ROC curves very similar to those found in our prior references, and our observations on specificity agreed with findings from the 2023 RSNA Screening Mammography Breast Cancer Detection AI Challenge that tested over 1500 different algorithms against a very large patient data base and consistently found very high specificity at a high operating point where the sensitivity would be low.[37] Clinical patient variability, different reader abilities, and different gantry types, reflecting actual U.S. breast center environments could produce different study results. There was also variability in the time between the baseline exam and the most recent prior.

In our unique test population, logistic regression analysis demonstrated that the very high case score number were highly specific. Working within the 95% confidence levels, our analysis identified 85 as a potential operating point for efficient screening mammogram triage as long as the hardware, software, and radiologists remained unchanged. In practice, every clinical setting is unique, and the operating point will always need to be clinically validated at each clinical site. Fortunately, the highest scores can always be used as the starting operating point for triage as they are "super" scores when it comes to specificity. Then, after a period of monitoring how many cancers and non-cancers are being triaged, the operating point can then be adjusted lower to triage more patients if desired. This clinical validation approach would not require extensive statistical analysis. Very high specificity at the highest score levels is typical for any AI breast cancer detection algorithm.

Conclusion

The clinical footprint of AI assisted breast cancer detection, and the associated research, continues to expand across the globe. However, the outstanding predictive performance of the few, very highest case scores remains under-utilized in the context of screening mammogram triage at a busy breast clinic. Anyone currently using a detection algorithm should be able to efficiently triage at least some of the cancer cases by simply looking at the very highest scores and using a top-down triage approach. This approach has the potential to significantly decrease the time to treatment for some cancer patients without causing unmanageable disruptions to the department logistics.

Clinical Practice Points

- AI-generated cancer detection scores can be used to triage screening mammograms for expedited diagnostic workup.
- Patients with the highest AI scores (≥85) have a significantly increased likelihood of breast cancer.
- Implementing an AI score-based threshold enables same-day diagnostic evaluation with minimal impact on clinical operations.
- This strategy may reduce time to diagnosis and treatment for a subset of patients, potentially improving clinical outcomes.
- Results support integrating AI into breast imaging workflows to improve efficiency and patient care through risk-based triage.

Glossary

AI (Artificial Intelligence): Computer algorithms designed to perform tasks that typically require human intelligence, such as image recognition and decision-making.

AUC (Area Under the Curve): A performance metric derived from the receiver operating characteristic (ROC) curve, representing the ability of a model to distinguish between classes (eg, cancer vs. no cancer).

Baseline Mammogram: The screening mammogram used as a reference point in this study for assessing changes in score or findings.

Case Score / Detection Score: The likelihood, as determined by an AI detection algorithm, that a mammogram shows signs of current malignancy.

Confidence Interval (CI): A range of values derived from sample data that is likely to contain the true population parameter.

Digital Breast Tomosynthesis (DBT): A 3D imaging technique used in breast cancer screening that takes multiple X-ray images of the breast from different angles.

False Negative (FN): A test result that incorrectly indicates no presence of disease when cancer is actually present.

False Positive (FP): A test result that incorrectly indicates the presence of disease when no cancer is present.

Logistic Regression: A statistical model used to estimate the probability of a binary outcome, such as presence or absence of cancer.

Negative Predictive Value (NPV): The probability that patients with a negative test result truly do not have the disease.

Odds Ratio (OR): A statistic that quantifies the strength of association between an exposure (eg, high AI score) and an outcome (eg, breast cancer).

Positive Predictive Value (PPV): The probability that patients with a positive test result actually have the disease.

Receiver Operating Characteristic (ROC) Curve: A graphical representation of a model's diagnostic ability across different thresholds, plotting sensitivity against 1-specificity.

Risk Score: A numerical value generated by an AI-based risk assessment algorithm estimating the short-term probability of developing breast cancer.

Sankey Diagram: A type of flow diagram used to visualize changes between categories or groups over time or between variables.

Sensitivity: The proportion of actual positives (eg, patients with cancer) that are correctly identified by the model.

Specificity: The proportion of actual negatives (eg, patients without cancer) that are correctly identified by the model.

Funding

Supported by funding from the National Center for Advancing Translational Sciences (NCATS Grant Number: UM1TR004404) for the Michigan Institute for Clinical and Health Research (MICHR)"

Disclosure

Dr. Mark Traill is a paid advisor and speaker for DeepHealth (formerly iCAD, Inc.) and has received honoraria for these activities. Dr. Traill has no financial interest in DeepHealth Inc., or the parent company Radnet Inc.. Julie Shisler served as a consultant for iCAD, Inc. Jeff Hoffmeister reports a prior relationship with iCAD, Inc. that included employment and ownership of equity or stock. At the time this work was being written, Crystal Snyder served as a consultant to iCAD, Inc. She is now a full-time employee of DeepHealth. (formerly iCAD, Inc.) and holds equity in the company. Authors Blair Richards, Jensen Jantz, and Shokoufeh Khalatbari declare no financial interests.

CRediT authorship contribution statement

Mark Traill: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jensen Jantz:** Writing – original draft, Resources, Project administration, Data curation. **Blair Richards:** Writing – original draft, Formal analysis. **Shokoufeh Khalatbari:** Writing – original draft. **Jeffrey Hoffmeister:** Writing – original draft. **Julie Shisler:** Writing – original draft. **Crystal Snyder:** Writing – review & editing, Writing – original draft.

Acknowledgments

Clinical Research Department, University of Michigan Health West, Wyoming MI provided patient data collection and verification.

Michigan Institute of Clinical and Health Research, Ann Arbor MI provided biostatistical analysis

DeepHealth, Inc., Cambridge, MA (formerly iCAD, Inc., Nashua, NH) provided breast cancer detection and image based risk assesment algorithms

References

1. Sechopoulos I, Teuwen J, Mann R. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: state of the art. *Sem Cancer Biol*. 2021;72:214–225.
2. Abdelrahman L, Ghamdi M, Collado-Mesa F, Adbel-Mottaleb M. Convolutional neural networks for breast cancer detection in mammography: a survey. *Comput Biol Med*. 2021;131:104248.
3. McKinney S, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577:89–94.
4. Conant E, Toledano A, Periaswamy S, et al. Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. *Radiol Artif Intell*. 2019;1(4):e18009.
5. VanWinkel S, Rodriguez-Ruiz A, Appelman L, et al. Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study. *Eur Radiol*. 2021;31:8682–8691.
6. Salim M, Wahlin E, Dembrower K, et al. External evaluation of 3 commercial artificial intelligence algorithms for intendent assessment of screening mammograms. *JAMA Oncol*. 2020;6(10):1581–1588.
7. Arasu VA, Habel LA, Achacoso NS, et al. Comparison of mammography AI algorithms with a clinical risk model for 5-year breast cancer risk prediction: an observational study. *Radiology*. 2023;307(5):e222733.
8. Eriksson M, Czene K, Strand F, et al. Identification of women at high risk of breast cancer who need supplemental screening. *Radiology*. 2020;297:327–333.
9. Yala A, Hughes KS. Rethinking risk modeling with machine learning. *Ann Surg Oncol*. 2023;30(12):6950–6952.
10. Eriksson M, Destounis S, Czene K, et al. A risk model for digital breast tomosynthesis to predict breast cancer and guide clinical care. *Sci Transl Med*. 2023;14(644):eabn3971.
11. Gastounioti A, Eriksson M, Cohen EA, et al. External validation of a mammography derived AI based risk model in a US breast cancer screening cohort of white and black women. *Cancers*. 2022;14:4803.
12. Park J, Witowski J, Xu Y, et al. A multi-modal AI system for screening mammography: integrating 2D and 3D imaging to improve breast cancer detection in a prospective clinical study. arXiv preprint arXiv:2504.05636 (2025).
13. Zuckerman SP, Periaswamy S, Shisler JL, et al. Evaluating the impact of changes in artificial intelligence–derived case scores over time on digital breast tomosynthesis screening outcomes. *Radiol Artif Intell*. 2025;7:2.
14. Lauritzen AD, von Euler-Chelpin MC, Lynge E, et al. Robust cross-vendor mammographic texture models using augmentation-based domain adaptation for long-term breast cancer risk. *J Med Imaging*. 2023;10(5):054003.
15. Lauritzen AD, von Euler-Chelpin MC, Lynge E, et al. Assessing breast cancer risk by combining AI for lesion detection and mammographic texture. *Martin Lillholm Radiology*. 2023;308(2):230227.
16. Houssami N, Marinovich ML. AI for mammography: making double screen-reading history. *Lancet Digital Health*. 2025;7(3):e168–e169.
17. Watanabe AT, Vu H, Chim CY, Litt AW, Retson T, Mayo RC. Potential impact of an artificial intelligence–based mammography triage algorithm on performance and workload in a population-based screening sample. *J Breast Imaging*. 2025;7(1):45–53. doi:10.1093/jbi/wbae056.
18. Dembrower K, Wåhlin E, Liu Y, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. *Lancet Digit Health*. 2020;2(9):e468–e474. doi:10.1016/S2589-7500(20)30185-0.
19. Elhakim MT, Stougard SW, Graumann O, et al. AI-integrated screening to replace double reading of mammograms: a population-wide accuracy and feasibility study. *Radiol Artif Intell*. 2024;6(6). doi:10.1148/ryai.230529.
20. Lauritzen AD, Lillholm M, Lynge E, Nielson M, Karssemeijer N, Vejborg I. Early indicators of the impact of using AI in mammography screening for breast cancer. *Radiology*. 2024;311. doi:10.1148/radiol.232479.
21. Morant R, Gräwingholt A, Subelack J, et al. The possible benefit of artificial intelligence in an organized population-related screening program: initial results and perspective. *Radiologie (Heidelberg)*. 2024;64(10):773–778. doi:10.1007/s00117-024-01345-6.
22. Lang K, Josefsson V, Larsson AM, et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol*. 2023;24(8):936–944. doi:10.1016/S1470-2045(23)00298-X.
23. Friedewald SM, Sieniek M, Jansen S, et al. Triaging mammography with artificial intelligence: an implementation study. *Breast Cancer Res Treat*. 2025;211. doi:10.1007/s10559-025-07616-7.
24. American College of Radiology Appropriateness Criteria/Palpable Breast Masses. Available at https://gravitas.acr.org/ACPortal/GetDataForOneScenario?senarioId=6802. Accessed December 2024.
25. Salzman B, Collins E, Hersh L. Common breast problems. *Am Fam Physician*. 2019;99(8):505–514.

26. Stachs A, Stubert J, Reimer T, Hartmann S. Benign breast disease in women. *Dtsch Arztebl Int*. 2019;116(33-34):565–574. doi:10.3238/arztebl.2019.0565.
27. Brown AL, Phillips J, Slanetz PJ, et al. Clinical value of mammography in the evaluation of palpable breast lumps in women 30 years old and older. *Am J of Roet*. 2017;209(4). doi:10.2214/AJR.16.17088.209.
28. Thompson DM, Fefferman ML, Nicholson KM, et al. Time from screening to treatment at accredited breast centers in the United States. *JCO Oncol Pract*. 2024;21:OP2400516. doi:10.1200/OP-24-00516.
29. Thompson DM, Fefferman ML, Nicholson KM, et al. Patient-reported observations on timeliness of breast services (PROMPT): results from a national survey. Presented at: American Society of Breast Surgeons Annual Meeting; April 10-14, 2024; Orlando, FL.
30. The American Society of Breast Surgeons Quality Measure: timeliness of care for Breast cancer. Available at https://www.breastsurgeons.org/docs/statements/asbrs-qm-timeliness-of-care-for-breast-cancer.pdf. May 2022
31. Bleicher RJ. Timing and delays in breast cancer evaluation and treatment. *Ann Surg Oncol*. 2018;25(10):2829–2838. doi:10.1245/s10434-018-6615-2.
32. Jaafar Sidek MA, Amajid K, Loh YS, et al. The prevalent factors of anxiety in women undergoing mammography. *Front Psychiatry*. 2023;14:1085115. doi:10.3389/fpsyt.2023.1085115.
33. Rimer BK, Bluman LG. The psychosocial consequences of mammography. *JNCI Monographs*. 1997;22:131–138. doi:10.1093/jncimon/1997.22.131.
34. Keyzer-Dekker CM, van Esch L, de Vries J, et al. An abnormal screening mammogram causes more anxiety than a palpable lump in benign breast disease. *Breast Cancer Res Treat*. 2012;134(1):253–258. doi:10.1007/s10549-012-2025-5.
35. Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. *BrMed J*. 2017;359:359:j4683.
36. Wing P, Langelier MH. Workforce shortages in breast imaging: impact on mammography utilization. *AJR Am J Roentgenol*. 2009;192(2):370–378. doi:10.2214/AJR.08.1665.
37. Chen Y, Partridge GJW, Vazirabad M, et al. Performance of algorithms submitted in the 2023 RSNA screening mammography breast cancer detection AI challenge. *Radiology*. 2025;316(2):e241447. doi:10.1148/radiol.241447.