Artificial Intelligence Evaluation of 122 969 Mammography Examinations from a Population-based Screening Program

Article Type: Original Research

Summary Statement: The performance of the artificial intelligence system was promising for breast cancer detection in a large population-based mammography screening program.

Key Results:

- In this retrospective study of 122 969 examinations, mammograms were evaluated with an artificial intelligence (AI) system which predicts the risk of cancer on a scale from 1 (lowest risk) to 10 (highest risk).
- A total of 87.6% (653 of 752) of screen-detected and 44.9% (92 of 205) of interval cancers had the highest AI score of 10; 0.7% (five of 752) screen-detected cancers had the lowest AI score of 1.
- Interval cancers with high AI scores had favorable histopathological tumor characteristics compared to cancers with low AI scores; the opposite was observed for screen-detected cancers.

Abbreviations: AI = artificial intelligence, DCIS = ductal carcinoma in situ, IQR = interquartile range, T1 = Threshold 1, T2 = Threshold 2, T3 = Threshold 3

Abstract

Background: Artificial intelligence (AI) has shown promising results for cancer detection in mammographic screening. However, evidence related to the use of AI in real screening settings remain sparse.

Purpose: To compare performance of a commercially available AI system with routine independent double reading with consensus as performed in a population-based screening program. Further, we explored histopathological characteristics of tumors with different AI scores.

Materials and Methods: In this retrospective study, 122,969 screening examinations from 47,877 women performed at four facilities in BreastScreen Norway from October 2009 to December 2018 were included. The dataset included 752 (6.1/1000) screen-detected and 205 (1.7/1000) interval cancers. Each examination had an AI score between 1 and 10, where 1 indicated low risk of breast cancer and 10 high risk. Thresholds T1, T2, and T3 were used to assess the performance of the AI system as a binary decision tool (selected versus not selected). T1 was set at an AI score of 10, T2 was set to yield a selection rate similar to the consensus rate (8.8%) and T3 to yield a selection rate similar to an average individual radiologist (5.8%). Descriptive statistics were used to summarize screening outcomes.

Results: A total of 653 of 752 (86.8%) screen-detected and 92 of 205 (44.9%) interval cancers were given a score of 10 by the AI system (T1). Using T3, 80.1% (602 of 752) of the screen-detected and 30.7% (63 of 205) of the interval cancers were selected. Screen-detected cancer with AI scores not selected using the thresholds had favorable histopathological characteristic compared to those selected; opposite results were observed for interval cancer.

Conclusion: The proportion of screen-detected cancers not selected by the AI system at three evaluated thresholds was less than 20%. The overall performance of the AI system was promising according to cancer detection.

Introduction

World-wide, over half a million women die of breast cancer every year (1). To reduce this burden, mammographic screening has been implemented in many countries over the last decades. These screening programs, along with improved treatment options, has resulted in at least a 30% reduction in breast cancer mortality among participants (2).

Use of double reading is recommended and standard in most European screening programs (3, 4). Double reading interpretation is usually followed by consensus or arbitration, where the decision to recall the women for further assessment is made. In BreastScreen Norway, greater than 25% of recalled women and about 0.6% of all screening examinations result in breast cancer diagnosis (5). Conversely, 99.4% of screening examinations are eventually determined to have a negative outcome.

Informed reviews of prior screening and diagnostic mammograms performed by groups of radiologists have classified about 25% of the screen-detected and interval cancers as missed (6, 7). Also, it has been reported that 20% of screen-detected cancers were recommended for recall by one of two radiologists in independent double reading (8). More accurate and effective interpretive procedures may improve population-level outcomes of mammographic screening.

Artificial intelligence (AI) has shown promising results for cancer detection in mammography (9-13). However, reported results are mainly from small studies with enriched datasets, and evidence gaps related to the use of AI in real screening settings remain (14). Retrospective studies on clinical datasets using consecutive exams provide an opportunity to independently validate AI systems prior to evaluation in prospective studies. Furthermore, the histopathological characteristics of cancers identified by AI should be investigated to ensure detection of clinically significant breast cancers that would lead to a reduction in breast cancer mortality.

In this study, we compared performance of a commercially available AI system with independent double reading as performed by radiologists in BreastScreen Norway. Further, we explored histopathological characteristics of tumors with different AI scores.

Methods

The study was approved by the Regional Committee for Medical and Health Research Ethics (2018/2574). The data was disclosed with legal bases in the Cancer Registry of Norway Regulations of 21 December 2001 No. 47 (15). The requirement to obtain written consent was waived under the same regulations. Reporting cancer to the Cancer Registry is mandatory by law in Norway, and 99% of the breast cancers are histopatologically verified (16). Screening information from examinations

included in this study have been used in other studies from BreastScreen Norway, exemplified in the given references (8, 17-19). Data on AI scores was collected entirely for this study.

This study was based on retrospective data from four screening units in BreastScreen Norway, a population-based screening program (5). Digital mammograms performed between October 2009 and December 2018 with Siemens Mammomat Inspiration, Erlangen, Forchheim, Germany were included (Figure 1).

Study Setting

BreastScreen Norway offers all women aged 50–69 years biennial two-view mammographic screening (5). Two radiologists independently interpret the mammograms; these radiologists undergo dedicated training prior to entering the program and are recommended to go through continued training (4). Radiologist experience varied from first-year involvement to those with greater than 20 years of experience within the program. Screen-readings from 24 radiologists (including SRH and HLH) were included in the study. If available, prior mammograms are always used in interpretations. Each breast is assigned an interpretation score of 1–5 to indicate suspicion of malignancy: 1, negative for malignancy; 2, probably benign; 3, intermediate suspicion of malignancy; 4, probably malignant; 5, high suspicion of malignancy. If the interpretation score is 2 or higher by either radiologist, a consensus of at least two radiologists determines whether to recall the woman. The consensus rate (examinations discussed at consensus divided by the total number of examinations) is reported to be 7.4%, and recall rate 3.2% (5, 8).

Image Data and AI System

The Cancer Registry identified the screening examinations to be included in this study, and the examination accession numbers were given to the Picture Archiving and Communication System vendor to extract the mammograms. Image data were pseudonymized before being processed with the AI system. Outputs from the AI system were merged with pseudonymized screening information using random study identification numbers.

We used Transpara version 1.7.0, a commercially available AI system for automated mammography interpretation, developed by ScreenPoint Medical, Nijmegen, the Netherlands. The AI system uses convolutional neural networks to analyze mammograms and is trained on mammograms from different screening programs and mammograms from several vendors (20). The AI system provides one score for each view of each breast. We used the highest score of all views to assign an overall exam-level score (AI score). The AI score ranges from 1 to 10 and is based on a “raw score” with the accuracy of four or five decimal points. AI scores are raw scores rounded up to the nearest integer

(Figure 2). The system aims to distribute the examinations equally across the AI scores, with about 10% of examinations assigned each score.

AI Decision Thresholds

We explored the performance of the AI system as a binary decision tool with three different thresholds for selecting examinations to be suspicious or not suspicious (Figure 2). The thresholds were defined prospectively. With threshold 1 (T1), a raw score above 9.00 (an AI score of 10) was defined as "selected" by the AI system and examinations with a score lower than 10 as "not selected". We allowed a higher selection rate than the consensus rate of 8.8% in the study sample since we know that cancers are missed at screening. Threshold 2 (T2) represented a selection rate equal to the consensus rate (raw score > 9.13) and was used to explore the performance of AI when the system selected a similar number of examinations as suspicious as the two radiologists. Threshold 3 (T3) corresponded to a selection rate equal to the observed average individual rate of positive interpretations by the radiologists of 5.8% in the study sample (raw score > 9.43). The lower proportion of selected examinations was explored with an aim of reducing false positive screening results.

Examination Variables

The women's first attendance in BreastScreen Norway was referred to as the prevalent examination, while returning attendance was considered subsequent. An examination was defined as negative if the mammograms had a negative assessment by both radiologists, concluded negative after consensus or after a recall with negative outcome. We defined recalls as screening examinations resulting in further assessments due to abnormal mammographic findings. Screen-detected cancer was defined as breast cancer diagnosed after a recall and within 6 months after the screening examination, and interval cancer as breast cancer diagnosed within 24 months after a negative screening or 6-24 months after a recall with a negative outcome (18). Mammograms from prior screening examination were processed with the AI system for interval cancers. Both ductal carcinoma in situ (DCIS) and invasive carcinoma were considered breast cancer.

Screening data included radiologist interpretation, consensus outcome, procedures performed at recall, and final outcomes including histopathological tumor characteristics. Characteristics of invasive cancers included histological type, tumor diameter, Nottingham grade 1-3, lymph node involvement and immune histochemical subtype. Subtype was classified into five groups (21). Histopathological characteristics of DCIS included tumor diameter and van Nuys grade 1-3 (22).

Statistical Analysis

Categorical variables were presented as frequencies and percentages, and continuous variables were presented as means and standard deviations (SD) or medians and interquartile ranges (IQR) according to the distribution. Results on tumor characteristics were stratified by examinations selected and not selected by the AI-system based on T1, T2 and T3. Stata version 17.0 for Windows (StataCorp, TX, USA) was used to analyze the data.

Results

Patient Overview

A total of 122 969 examinations from 47 877 women were included in the final study sample (Figure 1). Examinations performed in Ålesund and Molde during the period from 2011 to 2018, were interpreted by five radiologists at Ålesund Hospital, and examinations performed in Namsos and Levanger during the period from 2009-2018 were interpreted by 19 radiologists at St. Olavs Hospital, Trondheim University Hospital. The sample included women with implants, which is reported to be about 1.3% of women in the program (23).

Mean age at screening was 60 (SD=6) years and 14.1% (17 350 of 122 969) of the examinations were performed among prevalent attendees. Prevalent and subsequent examinations followed the same distribution of AI scores (Table 1).

AI Scores for Screen-detected and Interval Cancers

Our study sample included 752 screen-detected and 205 interval cancers (Table 2). A total of 77.9% (745 of 957) of the cancers had the highest AI score of 10, including 86.8% (653 of 752) of the screen-detected and 44.9% (92 of 205) of the interval cancers. For illustration, see Figure 3. Among all examinations with an AI score of 10, 5.3% (653 of 12 383) were screen-detected and 0.74% (92 of 12 383) were interval cancers.

Five screen-detected cancers had the lowest AI score of 1: three were invasive and two DCIS. Median tumor diameter was 9 mm (IQR: 9-18) for invasive cancers, with one grade 3 tumor and none with positive lymph node involvement. Figure 4 shows a screen-detected cancers with an AI score of 1. Among the 12 screen-detected cancers with an AI score of 4 or 5, 10 were invasive and two DCIS. Median tumor diameter was 8 mm (IQR: 6-11) for invasive cancers, with one grade 3 tumor and none with positive lymph node involvement.

The consensus rate was 8.8% (10 787 of 122 969) and the recall rate 3.2% (3896 of 122 969) in the study sample (Table 3). Of examinations discussed at consensus, 26.0% (2805 of 10 787) had an AI score of 10, and of the recalled cases, 36.9% (1438 of 3896) had an AI score of 10. Among the screen-

detected cancers with an AI score of 10, 80.9% (528 of 653) had a positive interpretation by both radiologists, while 19.1% (125 of 653) had a positive interpretation by only one radiologist. In comparison, for the 99 screen-detected cancers with an AI score of less than 10, 48.9% (45 of 99) had a positive interpretation by only one radiologist. The five screen-detected cancers with an AI score of 1 had a positive interpretation by only one of the two radiologists. Of interval cancers, 10.2% (21 of 205) were recalled with a negative outcome.

Use of T1 Threshold

The T1 threshold corresponds to selecting examinations with AI score 10. T1 selected 86.8% (653 of 752) of the screen-detected cancers and 82.2% (537 of 653) of these were invasive (Table 4). The percentage of invasive interval cancers selected was 93.5% (86 of 92). The median tumor diameter of the invasive screen-detected cancers selected by the AI system was 13 mm (IQR: 9-19) versus 10 mm (IQR: 7-17) for cancers not selected. The percentage of histological grade 3 cancers was 24.6% (131 of 532) for those selected and 20.3% (16 of 79) for those not selected. Lymph node involvement was observed for 22.9% (120 of 524) for those selected and 17.7% (14 of 79) for those not selected. Based on histological grade, lymph node involvement, and subtype, interval cancers selected by AI had favorable tumor characteristics compared to interval cancers not selected by AI.

Use of the T2 Threshold

The T2 threshold mirrors the consensus rate in the study sample, i.e., positive interpretation by one or both radiologists. Using T2, 85.1% (640 of 752) of the screen-detected and 41.5% (85 of 205) of the interval cancers were selected by the AI system (Table 5). Among the 112 screen-detected cancers not selected, 42.9% (48 of 112) had a positive interpretation by one of the two radiologists. The percentage of cancers with histological grade 3 was 24.5% (128 of 523) among the invasive screen-detected cancers selected by AI versus 21.6% (19 of 90) among those not selected by AI. Lymph node involvement was observed for 23.3% (120 of 515) of the selected and 15.9% (14 of 88) of the non-selected cases.

Use of the T3 Threshold

The T3 threshold mirrors the average individual radiologist rate of positive interpretation. Using T3, 80.1% (602 of 752) of the screen-detected and 30.7% (63 of 205) of interval cancers were selected by the AI system (Table 6). Among the 150 screen-detected cancers not selected by the AI system, 43.3% (65 of 150) had a positive interpretation by one of the two radiologists. The median tumor diameter of the invasive screen-detected cancers was 13 mm (IQR: 9-20) for cancers selected by the AI system and 9 mm (IQR: 7-15) for the cancers not selected. The percentage of histological grade 3

cancers was 25.3% (124 of 491) for those selected and 19.2% (23 of 120) for the non-selected cancers, while lymph node involvement was observed for 24.3% (117 of 482) and 14.0% (17 of 121), respectively.

Including screen-detected and interval cancers as true positives for T1, T2 and T3, AI selected 77.9% (745 of 957), 75.8% (725 of 957) and 69.5% (665 of 957) of the cancers. The rate of selected cases without cancer ("false positives") were 94.0% (11 638 of 12 383), 93.3% (10 064 of 10 789) and 90.7% (6471 of 7316), respectively.

Discussion

The purpose of this study was to evaluate an artificial intelligence (AI) system for breast cancer detection on mammography. The performance of the AI system was compared to radiologists in an independent double reading setting with consensus. A total of 77.9% of all breast cancers (86.8% of screen-detected and 44.9% of interval cancers) had the highest AI score of 10. With a threshold that mirrors the average individual radiologist rate of positive interpretation (called threshold 3; T3), 80.1% of screen-detected and 30.7% of the interval cancers were selected by the AI system.

To our knowledge, this is the largest AI evaluation study to date, including more than 120 000 examinations, 752 screen-detected, and 205 interval cancers from a real screening setting. There are several publications describing the performance of the AI-system in other, smaller screening cohorts (11, 13, 26, 27). Use of this same system in a population from Malmö, Sweden, found that none of the 68 screen-detected cancers had an AI score below 3 (11). Similar results were obtained in a study from Spain (26). None of the 76 screen-detected cancers had an AI score below 3. In our larger sample, five of 752 screen-detected cancers had a score below 3 (five had AI score 1 and none had AI score 2). Differences in cancer detection across these studies may be related to our use of an updated version of the AI system or differences in characteristics of the screening populations and interpreting radiologists (11, 27).

The high percentage of true negative examinations classified with a low AI score may indicate that the AI system could safely select examinations not to be interpreted by radiologists. In such an approach, the interpretive volume would be substantially reduced while a small proportion of cancers not selected by the AI system would remain undetected. If AI is used as one of the two readers in a double reading setting, the radiologist may still identify the small number of missed

cancers. Further, 23% of screen-detected cancers in the study had a positive assessment by only one radiologist and it may thus be acceptable that some cancers have a low AI score.

Similar to the challenge in defining the ideal combination of two radiologists in double reading, more research is needed to find the optimal combination of radiologists and AI systems. For instance, when using AI as a standalone system to identify true negative cases that can forego radiologist interpretation altogether, an accurate low score on mammograms without missed cancers is critical. Using an AI score of 10 as a threshold in a standalone setting could lead to 10% of examinations requiring radiologist interpretation before eventual consensus or 10% of the examinations discussed at consensus. In the latter scenario, the consensus rate would be higher than usual in BLINDED and likely result in a higher recall rate. If radiologists are using an AI system in a screening setting, it is expected that their assessment and the recall rates will depend on AI scores. The optimal timing of and format of being presented with AI scores is unknown and need further investigation to find the optimal settings. The effect of being presented with a high AI score may lead to overreliance on the AI system without a radiologist maintaining their own vigilance or lead to reduced attention to other suspicious areas (automation bias) (28).

Our results indicate favorable histopathologic characteristics for screen-detected cancers with low versus high AI scores. Studies have shown less than 10% of screen-detected cancers are clinically insignificant, indicating a low risk of breast cancer death (29). An AI system which is able to differentiate between clinically significant and non-significant cancers could be beneficial for individual women and the screening program. Currently there is limited data on the progression of small, low proliferation cancers, but such information could help women and clinicians make informed choices on the intensity and extent of treatment.

Interval cancers are known to be less prognostically favorable compared to screen-detected cancers (7, 18) and it is essential to keep the rate as low as possible to reduce breast cancer mortality. We observed that the invasive interval cancers selected using T1, T2, and T3 by the AI system had more favorable tumor characteristics compared to those not selected. This may indicate that interval cancers with low AI scores are true interval cancers and not visible on the screening mammograms. Similar results were observed in a retrospective study on a large cohort of interval cancers using the same AI system (24).

Strengths of our study are the large study population from a real screening setting and capture of all cancers through registry linkage. Limitations are related to the retrospective approach; however, this limitation is ameliorated by a complete follow-up of all screened women. Additional limitations include evaluation of mammograms from a single manufacturer, the regional homogeneous

population, an AI system not considering prior mammograms, the limited number of radiologists, and not including laterality, mammographic features or density.

In conclusion, the proportion of screen-detected cancers not selected by the AI system at the three evaluated thresholds was less than 20%, and several of these would probably be detected at an early stage also in the next screening round. However, there are also tumor characteristics of examinations not selected indicative of clinically significant cancers. Prospective studies are needed to better understand the prognostic characteristics of AI selected and AI non-selected cases. Further research is also needed to understand how the relatively large number of negative examinations with a high AI score can influence the recall rate and rate of false-positive results. Future studies should also examine mammographic features identified by AI, evaluate multiple AI algorithms in a comparative manner, examine AI in more diverse screening populations, and include cost-effectiveness analyses of using AI in screening.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018;68(6):394-424.
2. Lauby-Secretan B, Scoccianti C, Loomis D, et al. Breast-cancer screening--viewpoint of the IARC Working Group. N Engl J Med. 2015;372(24):2353-8.
3. European Commision Initiative on Breast Cancer. Cited September 2021: https://healthcare-quality.jrc.ec.europa.eu/european-breast-cancer-guidelines/screening-ages-and-frequencies/women-50-69.
4. Hofvind S, Bennett RL, Brisson J, et al. Audit feedback on reading performance of screening mammograms: An international comparison. J Med Screen. 2016;23(3):150-9.
5. Hofvind S, Tsuruda KM, Mangerud G, et al. The BLINDED Breast Cancer Screening Program, 1996-2016: Celebrating 20 years of organised mammographic screening. Cancer in BLINDED 2016 - Cancer incidence, mortality, survival and prevalence in BLINDED Cancer Registry of BLINDED. ISBN 978-82-473-0055-8; 2017.
6. Hoff SR, Abrahamsen AL, Samset JH, Vigeland E, Klepp O, Hofvind S. Breast cancer: missed interval and screening-detected cancer at full-field digital mammography and screen-film mammography-- results from a retrospective review. Radiology. 2012;264(2):378-86.
7. Hovda T, Hoff SR, Larsen M, Romundstad L, Sahlberg KK, Hofvind S. True and Missed Interval Cancer in Organized Mammographic Screening: A Retrospective Review Study of Diagnostic and Prior Screening Mammograms. Acad Radiol 2021 doi:101016/jacra202103022.
8. Hofvind S, Geller BM, Rosenberg RD, Skaane P. Screening-detected breast cancers: discordant independent double reading in a population-based screening program. Radiology. 2009;253(3):652-60.
9. Akselrod-Ballin A, Chorev M, Shoshan Y, et al. Predicting Breast Cancer by Applying Deep Learning to Linked Health Records and Mammograms. Radiology. 2019;292(2):331-42.
10. 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-e74.
11. Lång K, Dustler M, Dahlblom V, Åkesson A, Andersson I, Zackrisson S. Identifying normal mammograms in a large screening population using artificial intelligence. European Radiology. 2021;31(3):1687-92.
12. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. JNCI: Journal of the National Cancer Institute. 2019;111(9):916-22.
13. Salim M, Wåhlin E, Dembrower K, et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. JAMA Oncol. 2020;6(10):1581-8.
14. Freeman K, Geppert J, Stinton C, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ. 2021;374:n1872.
15. Lovdata, Kreftregisterforskriften. Cited Sept 2021: https://lovdata.no/dokument/SF/forskrift/2001-12-21-1477.
16. Larsen IK, Småstuen M, Johannesen TB, et al. Data quality at the Cancer Registry of BLINDED: an overview of comparability, completeness, validity and timeliness. Eur J Cancer. 2009;45(7):1218-31.
17. Hofvind S, Skaane P, Elmore JG, Sebuødegård S, Hoff SR, Lee CI. Mammographic performance in a population-based screening program: before, during, and after the transition from screen-film to full-field digital mammography. Radiology. 2014;272(1):52-62.
18. Hofvind S, Sagstad S, Sebuødegård S, Chen Y, Roman M, Lee CI. Interval Breast Cancer Rates and Histopathologic Tumor Characteristics after False-Positive Findings at Mammography in a Population-based Screening Program. Radiology. 2018;287(1):58-67.

19. Hoff SR, Myklebust TÅ, Lee CI, Hofvind S. Influence of Mammography Volume on Radiologists' Performance: Results from BreastScreen Norway. Radiology. 2019;292(2):289-296.
20. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. Medical Image Analysis. 2017;35:303-12.
21. Goldhirsch A, Winer EP, Coates AS, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol. 2013;24(9):2206-23.
22. Silverstein MJ, Poller DN, Waisman JR, et al. Prognostic classification of breast ductal carcinoma-in-situ. Lancet. 1995;345(8958):1154-7.
23. Sondén ECB, Sebuødegård S, Korvald C, et al. Cosmetic breast implants and breast cancer. Tidsskr Nor Laegeforen 2020 doi:104045/tidsskr190266.140(3).
24. Lång K, Hofvind S, Rodriguez-Ruiz A, Andersson I. Can artificial intelligence reduce the interval cancer rate in mammography screening? Eur Radiol. 2021;31:5940-7.
25. Sasaki M, Tozaki M, Rodríguez-Ruiz A, et al. Artificial intelligence for breast cancer detection in mammography: experience of use of the ScreenPoint Medical Transpara system in 310 Japanese women. Breast Cancer. 2020;27(4):642-51.
26. Raya-Povedano JL, Romero-Martín S, Elías-Cabot E, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. AI-based Strategies to Reduce Workload in Breast Cancer Screening with Mammography and Tomosynthesis: A Retrospective Evaluation. Radiology. 2021;300(1):57-65.
27. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. European Radiology. 2019;29(9):4825-32.
28. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012;19(1):121-7.
29. Bulliard JL, Beau AB, Njor S, et al. Breast cancer screening and overdiagnosis. Int J Cancer. 2021;149:846-53.

Table 1. Characteristics of Included Examinations Stratified by AI Scores

AI score	Prevalent screening examinations	Subsequent screening examinations
1	1908 (11.0%)	13 285 (12.6%)
2	911 (5.3%)	5393 (5.0%)
3	1835 (10.6%)	11 959 (11.3%)
4	1855 (10.7%)	10 948 (10.4%)
5	1759 (10.1%)	10 709 (10.1%)
6	1634 (9.4%)	9907 (9.4%)
7	1604 (9.2%)	10 253 (9.7%)
8	1870 (10.8%)	10 953 (10.4%)
9	1993 (11.5%)	11 900 (11.3%)
10	1981 (11.4%)	10 402 (9.9%)
Total	17 350 (100%)	105 619 (100.0)

Note.— Percentages were calculated among the total number of prevalent and subsequent screening examinations.AI score is defined as the overall exam-level score from the AI system, and a score of 1 indicates low probability of breast cancer and 10 indicates high probability.

Table 2. Screening Examinations and Results Stratified by AI Scores

AI score	All screening examinations	Negative screening results	Screen-detected cancers	Interval cancers	Screen-detected and interval cancers
1	15 193 (12.4%)	15 179 (12.4%)	5 (0.7%)	9 (4.4%)	14 (1.5%)
2	6214 (5.1%)	6213 (5.1%)	0 (0%)	1 (0.5%)	1 (0.1%)
3	13 794 (11.2%)	13 785 (11.3%)	0 (0%)	9 (4.4%)	9 (0.9%)
4	12 803 (10.4%)	12 786 (10.5%)	8 (1.1%)	9 (4.4%)	17 (1.8%)
5	12 468 (10.1%)	12 453 (10.2%)	4 (0.5%)	11 (5.4%)	15 (1.6%)
6	11 541 (9.4%)	11 523 (9.4%)	9 (1.2%)	9 (4.4%)	18 (1.9%)
7	11 857 (9.6%)	11 836 (9.7%)	7 (0.9%)	14 (6.8%)	21 (2.2%)
8	12 823 (10.4%)	12 788 (10.5%)	21 (2.8%)	14 (6.8%)	35 (3.7%)
9	13 893 (11.3%)	13 811 (11.3%)	45 (6.0%)	37 (18.1%)	82 (8.6%)
10	12 383 (10.1%)	11 638 (9.5%)	653 (86.8%)	92 (44.9%)	745 (77.9%)
Total	122 969 (100%)	122 012 (100%)	752 (100%)	205 (100%)	957 (100%)

Note.— Percentages were calculated from the number of screening examinations and cancers. Negative screening results included negative screening result and recall for further assessments with a negative outcome. AI score is defined as the overall exam-level score from the AI system, and a score of 1 indicates low probability of breast cancer and 10 indicates high probability.

Table 3. Screening Outcome Stratified by AI Score

AI score	Examinations discussed at consensus after positive assessment by one or both radiologists	All cases recalled after consensus	Screen-detected cancer		Interval cancer		
			Positive assessment by one radiologist	Positive assessment by both radiologists	Recalled, negative outcome	Positive assessment by one radiologist	Positive assessment by both radiologists
1	363 (3.4%)	57 (1.5%)	5	0	0	0	0
2	265 (2.5%)	68 (1.8%)	0	0	0	0	0
3	603 (5.6%)	146 (3.8%)	0	0	0	0	0
4	764 (7.1%)	201 (5.2%)	4	4	0	0	0
5	840 (7.8%)	223 (5.7%)	2	2	1	1	0
6	957 (8.9%)	296 (7.6%)	4	5	0	0	0
7	1103 (10.2%)	341 (8.8%)	3	4	0	0	0
8	1320 (12.2%)	465 (11.9%)	8	13	1	1	0
9	1767 (16.4%)	661 (17.0%)	19	26	3	3	0
10	2805 (26.0%)	1438 (36.9%)	125	528	16	11	5
Total	10 787 (100%)	3896 (100%)	170	582	21	16	5

Note. — Percentages were calculated of total number of consensus and recalled. Cancers were stratified by a positive assessment (interpretation score of 2 or higher) by one or both radiologists. AI score is defined as the overall exam-level score from the AI system, and a score of 1 indicates low probability of breast cancer and 10 indicates high probability.

Table 4. Histopathological Characteristics of Screen-detected and Interval Cancers Stratified by use of the T1 Threshold

	Screen-detected cancers		Interval cancers	
	Selected with T1	Not selected with T1	Selected with T1	Not selected with T1
Total	653 (86.8%)	99 (13.2%)	92 (44.9%)	113 (55.1%)
Characteristics of in situ cancers	116 (17.8%)	17 (17.2%)	6 (6.5%)	8 (7.1%)
Tumor diameter, mm	20 (10-30)	11 (10-15)	19 (12-25)	11 (7-17)
Data not available, n	18	4	0	1
Van Nuys grade for DCIS				
Grade 1	14 (13.2%)	5 (35.7%)	0 (0%)	2 (40%)
Grade 2	11 (10.4%)	3 (21.4%)	0 (0%)	0 (0%)
Grade 3	81 (76.4%)	6 (42.9%)	6 (100%)	3 (60%)
Data not available	10	3	0	3
Characteristics of invasive cancers	537 (82.2%)	82 (82.8%)	86 (93.5%)	105 (92.9%)
Tumor diameter, mm	13 (9-19)	10 (7-17)	17 (11-28)	16 (11-25)
Data not available, n	8	1	2	3
Nottingham grade				
Grade 1	175 (32.9%)	33 (41.8%)	19 (22.4%)	14 (13.5%)
Grade 2	226 (42.5%)	30 (38.0%)	37 (43.5%)	37 (35.6%)
Grade 3	131 (24.6%)	16 (20.3%)	29 (34.1%)	53 (51.0%)
Data not available	5	3	1	1
Lymph node involvement	120 (22.9%)	14 (17.7%)	26 (32.1%)	38 (37.6%)
Data not available	13	3	5	4
Immune histochemical subtype				
Luminal A-like	313 (60.1%)	48 (61.5%)	45 (52.9%)	44 (42.7%)
Luminal B-like (HER2 negative)	85 (16.3%)	11 (14.1%)	12 (14.1%)	26 (25.2%)
Luminal B-like (HER2 positive)	77 (14.8%)	11 (14.1%)	16 (18.8%)	13 (12.6%)
HER2 positive (non-luminal)	14 (2.7%)	3 (3.8%)	3 (3.5%)	7 (6.8%)
Triple negative	32 (6.1%)	5 (6.4%)	9 (10.6%)	13 (12.6%)
Data not available	16	4	1	2

Note.— The T1 threshold corresponded to cancers given an AI score of 10. AI score is defined as the overall exam-level score from the AI system, and a score of 1 indicates low probability of breast cancer and 10 indicates high probability. The percentage of invasive cancers and Ductal Carcinoma in situ (DCIS) were calculated from the total number of cancers. Continuous variables shown as median (interquartile range).

Table 5. Histopathological Characteristics of Screen-detected and Interval Cancers Stratified by use of the T2 Threshold

	Screen-detected cancers		Interval cancers	
	Selected with T2	Not selected with T2	Selected with T2	Not selected with T2
Total	640 (85.1%)	112 (14.9%)	85 (41.5%)	120 (58.5%)
Characteristics of in situ cancers	112 (17.5%)	21 (18.8%)	6 (7.1%)	8 (6.7%)
Tumor diameter, mm	20 (10-30)	10 (7-15)	19 (12-25)	11 (7-17)
Data not available, n	16	6	0	1
Van Nuys grade for DCIS				
Grade 1	11 (10.7%)	8 (47.1%)	0 (0%)	2 (40%)
Grade 2	11 (10.7%)	3 (17.7%)	0 (0%)	0 (0%)
Grade 3	81 (78.6%)	6 (35.3%)	6 (100%)	3 (60%)
Data not available	9	4	0	3
Characteristics of invasive cancers	528 (82.5%)	91 (81.3%)	79 (92.9%)	112 (93.3%)
Tumor diameter, mm	13 (9-19)	10 (7-17)	17 (11-26)	16 (11-25)
Data not available, n	8	1	2	3
Nottingham grade				
Grade 1	172 (32.9%)	36 (40.9%)	18 (23.1%)	15 (13.5%)
Grade 2	223 (42.6%)	33 (37.5%)	35 (44.9%)	39 (35.1%)
Grade 3	128 (24.5%)	19 (21.6%)	22 (32.1%)	57 (51.4%)
Data not available	5	1	1	1
Lymph node involvement	120 (23.3%)	14 (15.9%)	24 (32.4%)	40 (37.0%)
Data not available	13	3	5	4
Immune histochemical subtype				
Luminal A-like	307 (60.0%)	54 (62.1%)	42 (53.9%)	47 (42.7%)
Luminal B-like (HER2 negative)	83 (16.2%)	13 (14.9%)	12 (15.4%)	26 (23.6%)
Luminal B-like (HER2 positive)	77 (15.0%)	11 (12.6%)	13 (16.7%)	16 (14.6%)
HER2 positive (non-luminal)	14 (2.7%)	3 (3.5%)	3 (3.9%)	7 (6.4%)
Triple negative	31 (6.1%)	6 (6.9%)	8 (10.3%)	14 (12.7%)
Data not available	16	4	1	2

Note.— The T2 threshold corresponded to the consensus rate (score of 2 or higher by either or both radiologists) of 8.8% in the study sample, meaning that 8.8% of the examinations with highest score by the AI system was selected. The percentage of invasive cancers and in situ cancers (DCIS) are calculated from the total number of cancers. Continuous variables shown as median (interquartile range). DCIS = ductal carcinoma in situ

Table 6. Histopathological Characteristics of screen-detected and Interval Cancers Stratified by use of the T3 Threshold

	Screen-detected cancers		Interval cancers	
	Selected with T3	Not selected with T3	Selected with T3	Not selected with T3
Total	602 (80.1%)	150 (19.9%)	63 (30.7%)	142 (59.3%)
Characteristics of in situ cancers	107 (17.8%)	26 (17.3%)	5 (7.9%)	9 (6.3%)
Tumor diameter, median	20 (10-30)	10 (7-15)	20 (18-25)	12 (8-16)
Data not available, n	15	7	0	1
Van Nuys grade for DCIS				
Grade 1	9 (9.1%)	10 (47.6%)	0 (0%)	2 (33%)
Grade 2	11 (11.1%)	3 (14.3%)	0 (0%)	-
Grade 3	79 (79.8%)	8 (38.1%)	5 (100%)	4 (67%)
Data not available	8	5	0	3
Characteristics of invasive cancers	495 (82.3%)	124 (82.7%)	58 (92.1%)	133 (93.7%)
Tumor diameter, median	13 (9-20)	9 (7-15)	17 (11-26)	16 (12-25)
Data not available, n	7	2	1	4
Nottingham grade				
Grade 1	157 (32.0%)	51 (42.5%)	17 (29.8%)	16 (12.1%)
Grade 2	210 (42.8%)	46 (38.3%)	21 (36.8%)	53 (40.2%)
Grade 3	124 (25.3%)	23 (19.2%)	19 (33.3%)	63 (47.7%)
Data not available	4	4	1	1
Lymph node involvement	117 (24.3%)	17 (14.0%)	18 (33.3%)	47 (36.7%)
Data not available	13	3	4	5
Immune histochemical subtype				
Luminal A-like	283 (59.1%)	78 (65.0%)	30 (52.6%)	59 (45.0%)
Luminal B-like (HER2 negative)	82 (17.1%)	14 (11.7%)	9 (15.8%)	29 (22.1%)
Luminal B-like (HER2 positive)	75 (17.1%)	13 (10.8%)	8 (14.0%)	21 (16.0%)
HER2 positive (non-luminal)	13 (2.8%)	4 (3.3%)	2 (3.5%)	8 (6.1%)
Triple negative	26 (5.4%)	11 (9.2%)	8 (14.0%)	14 (10.7%)
Data not available	16	4	1	2

Note.— The T3 threshold corresponded to the average individual rate of positive scores (score of 2 or higher) of 5.8% by study sample radiologists, meaning that 5.8% of the examination with highest score by the AI system was selected. The percentage of invasive cancers and in situ cancers (DCIS) are calculated from the total number of cancers. Continuous variables shown as median (interquartile range). DCIS = ductal carcinoma in situ

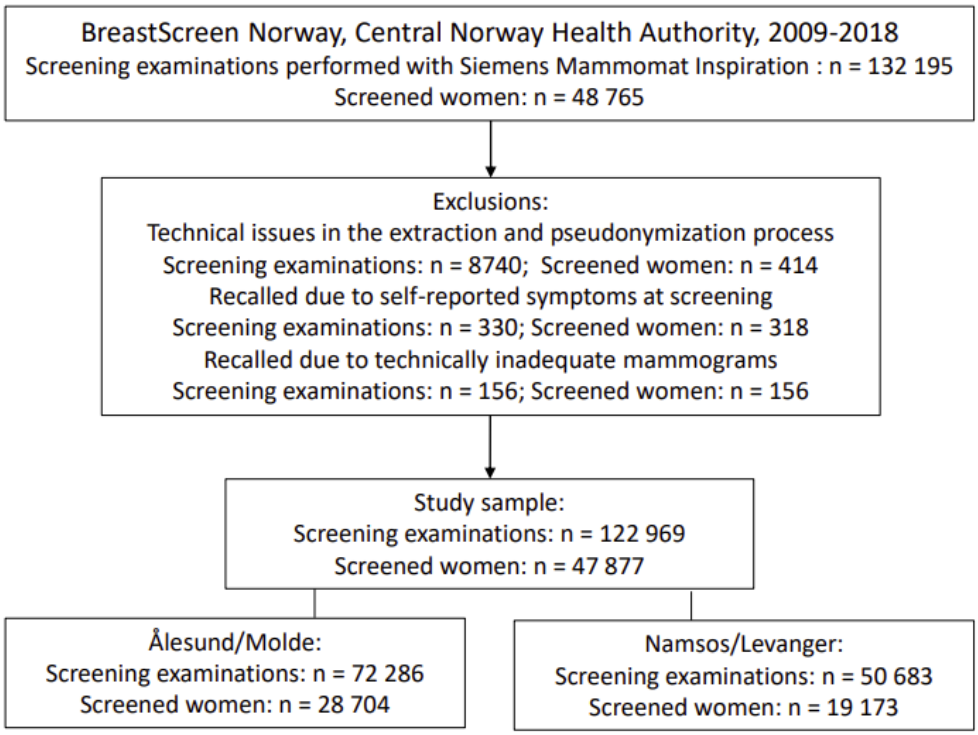

Figure 1. Flowchart of the study sample

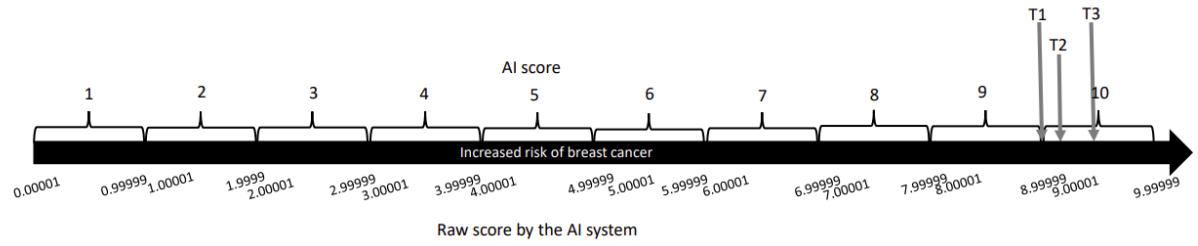

Figure 2. The artificial intelligence scoring system (raw score and AI score) with the three different thresholds (T1, T2, T3) defined for this study. T1 corresponds to AI score 10, T2 corresponds to a raw score of 9.13 and results in selecting 8.8% of the examinations with the highest score by the AI system, and T3 corresponds to a raw score of 9.43 and results in selecting 5.8% of the examinations with the highest score by the AI system.

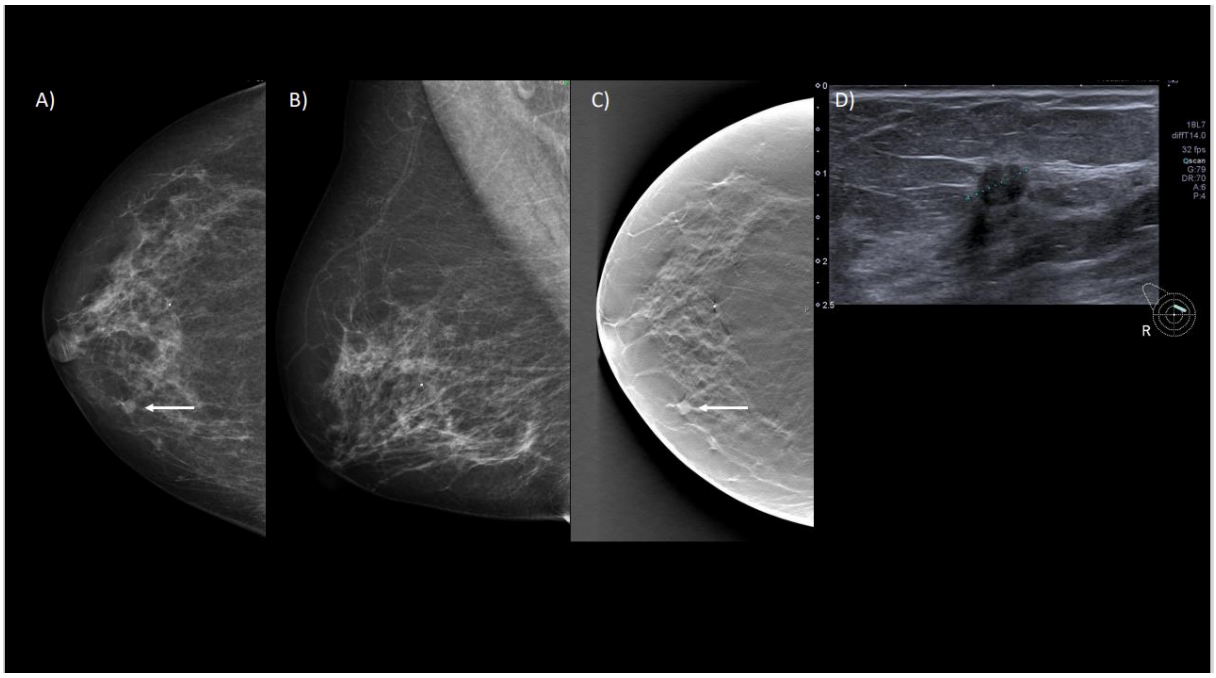

Figure 3. Sixty-eight-year-old woman with a screen-detected ductal carcinoma in situ with an AI score of 10 on the screening mammograms. A) Mammogram from craniocaudal view of right breast B) Mammogram from medio-lateral oblique view of right breast C) Craniocaudal tomosynthesis of right breast D) Ultrasound of right breast. AI score is defined as the overall exam-level score from the AI system, and a score of 1 is indicative of low probability of breast cancer and 10 high probability. The arrows indicate the malignancy.

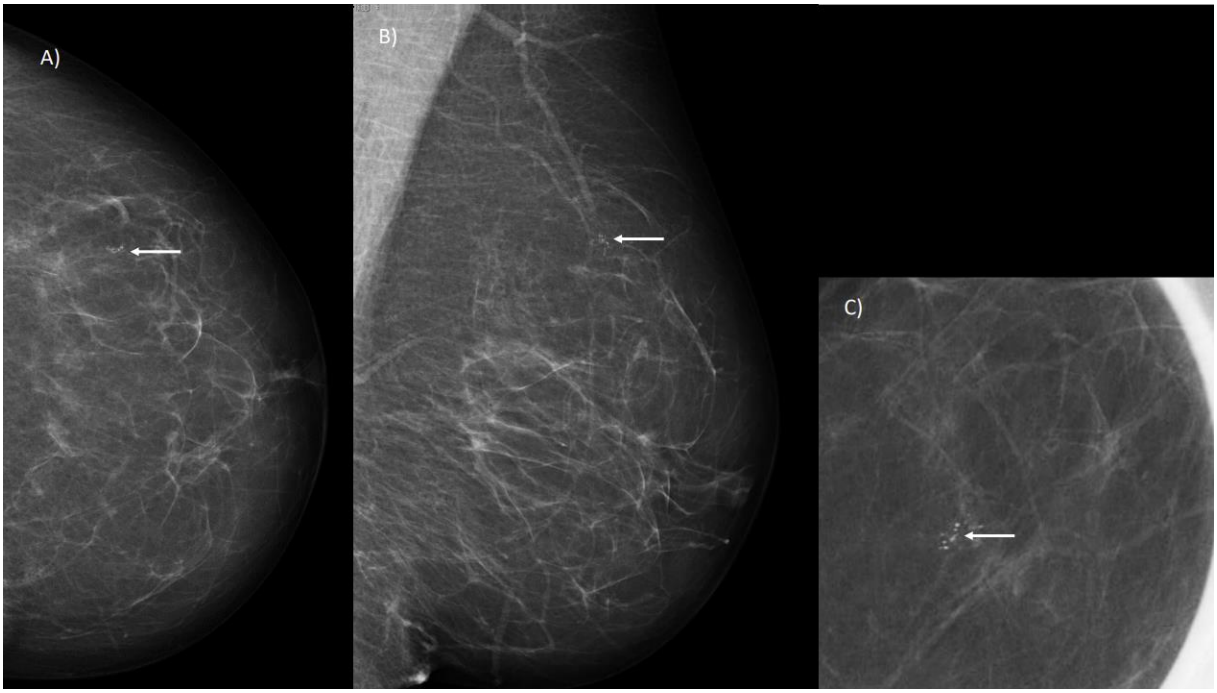

Figure 4: Sixty-year-old woman with an invasive screen-detected cancer with an AI score of 1 on the screening mammograms. A) Mammogram from craniocaudal view of left breast B) Mammogram form medio-lateral oblique view of left breast C) Cone with magnification. AI score is defined as the

overall exam-level score from the AI system, and a score of 1 is indicative of low probability of breast cancer and 10 high probability. The arrows indicate the malignancy.