



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

Marthe Larsen, MSc • Camilla F. Aglen, MA • Christoph I. Lee, MD, MS • Solveig R. Hoff, PhD • Håkon Lund-Hanssen, MD • Kristina Lång, PhD • Jan E. Nygård, PhD • Giske Ursin, PhD • Solveig Hofvind, PhD

From the Section for Breast Cancer Screening (M.L., C.F.A., S.H.) and Department of Register Informatics (J.E.N.), Cancer Registry of Norway (G.U.), P.O. Box 5313, 0304 Oslo, Norway; Department of Health and Care Sciences, Faculty of Health Sciences, The Arctic University of Norway, Tromsø, Norway (S.H.); Department of Radiology, University of Washington School of Medicine, Seattle, Wash (C.I.L.); Department of Health Systems and Population Health, University of Washington School of Public Health, Seattle, Wash (C.I.L.); Department of Radiology, Ålesund Hospital, Møre og Romsdal Hospital Trust, Ålesund, Norway (S.R.H.); Department of Circulation and Medical Imaging, Faculty of Medicine and Health Sciences, National University for Science and Technology, Trondheim, Norway (S.R.H.); Department of Radiology and Nuclear Medicine, St Olavs University Hospital, Trondheim, Norway (H.L.H.); Department of Translational Medicine, Lund University, Lund, Sweden (K.L.); and Unilabs Mammography Unit, Skåne University Hospital, Malmö, Sweden (K.L.). Received September 23, 2021; revision requested November 12; revision received January 12, 2022; accepted January 20. **Address correspondence** to S.H. (e-mail: sshh@kreftregisteret.no).

Supported by the Pink Ribbon campaign. C.I.L. supported in part by the National Cancer Institute (grant R37 CA240403).

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

Radiology 2022; 303:502–511 • <https://doi.org/10.1148/radiol.212381> • Content codes:  

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

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

Materials and Methods: In this retrospective study, 122 969 screening examinations from 47 877 women performed at four screening units in BreastScreen Norway from October 2009 to December 2018 were included. The data set included 752 screen-detected cancers (6.1 per 1000 examinations) and 205 interval cancers (1.7 per 1000 examinations). Each examination had an AI score between 1 and 10, where 1 indicated low risk of breast cancer and 10 indicated high risk. Threshold 1, threshold 2, and threshold 3 were used to assess the performance of the AI system as a binary decision tool (selected vs not selected). Threshold 1 was set at an AI score of 10, threshold 2 was set to yield a selection rate similar to the consensus rate (8.8%), and threshold 3 was set 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 screen-detected cancers (86.8%) and 92 of 205 interval cancers (44.9%) were given a score of 10 by the AI system (threshold 1). Using threshold 3, 80.1% of the screen-detected cancers (602 of 752) and 30.7% of the interval cancers (63 of 205) were selected. Screen-detected cancer with AI scores not selected using the thresholds had favorable histopathologic characteristics compared to those selected; opposite results were observed for interval cancer.

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

© RSNA, 2022

Worldwide, more than 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 past decades. These screening programs, along with improved treatment options, have resulted in a reduction of at least 30% 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, breast cancer is diagnosed in more than 25% of recalled women and about 0.6% of all screening examinations (5). Conversely, 99.4% of screening examinations are eventually determined to have a negative outcome.

Informed reviews of prior screening and diagnostic mammograms obtained by groups of radiologists have classified about 25% of 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 mammographic examinations (9–13). However, reported results are mainly from small studies with enriched data sets, and evidence gaps related to the use of AI in real screening settings remain (14). Retrospective studies on clinical data sets using consecutive examinations provide an opportunity to independently validate AI systems before evaluation in prospective studies. Furthermore, the histopathologic 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 the performance of a commercially available AI system with independent double

Abbreviations

AI = artificial intelligence, DCIS = ductal carcinoma in situ

Summary

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 that predicts the risk of cancer on a scale from 1 (lowest risk) to 10 (highest risk).
- A total of 86.8% of screen-detected cancers (653 of 752) and 44.9% of interval cancers (92 of 205) had the highest AI score of 10; 0.7% screen-detected cancers (five of 752) had the lowest AI score of 1.
- Interval cancers with high AI scores had favorable histopathologic tumor characteristics compared to those with low AI scores; the opposite was observed for screen-detected cancers.

reading as performed by radiologists in BreastScreen Norway. Furthermore, we explored the histopathologic characteristics of tumors with different AI scores.

Materials and Methods

The study was approved by the Regional Committee for Medical and Health Research Ethics (2018/13294). The data were disclosed with legal bases in the Cancer Registry of Norway Regulations of December 21, 2001, number 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 histopathologically 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 were 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 obtained between October 2009 and December 2018 with MAMMOMAT Inspiration (Siemens Healthcare) were included (Fig 1).

Study Setting

BreastScreen Norway offers all women aged 50–69 years biennial two-view mammographic screenings (5). Two radiologists independently interpret the mammograms; these radiologists undergo dedicated training before entering the program and are recommended to go through continued training (4). Radiologist experience varied from 1st-year involvement to those with greater than 20 years of experience within the program. Screen readings from 24 radiologists (including S.R.H. and H.L.H.) were included in the study. If available, then prior mammograms are always used in interpretations. Each breast is assigned an interpretation score of 1–5 to indicate suspicion of malignancy, as follows: 1, negative for malignancy; 2, probably benign; 3, intermediate suspicion of malignancy; 4, probably malignant; and 5,

high suspicion of malignancy. If the interpretation score is 2 or higher by either radiologist, then 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 is 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 mammogram interpretation developed by ScreenPoint Medical. The AI system uses convolutional neural networks to analyze mammograms and is trained on mammograms from different screening programs and 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 examination-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 (Fig 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 (Fig 2). The thresholds were defined prospectively. With threshold 1, 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 were defined as “not selected.” We allowed a higher selection rate than the consensus rate of 8.8% in the study sample because we know that cancers are missed at screening examination. Threshold 2 represented a selection rate equal to the consensus rate (raw score >9.13) and was used to explore the performance of AI when the number of examinations selected by the system as suspicious was similar to the number of examinations selected by the two radiologists. Threshold 3 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 a subsequent examination. An examination was defined as negative if the mammograms had a negative assessment by both radiologists, had a negative

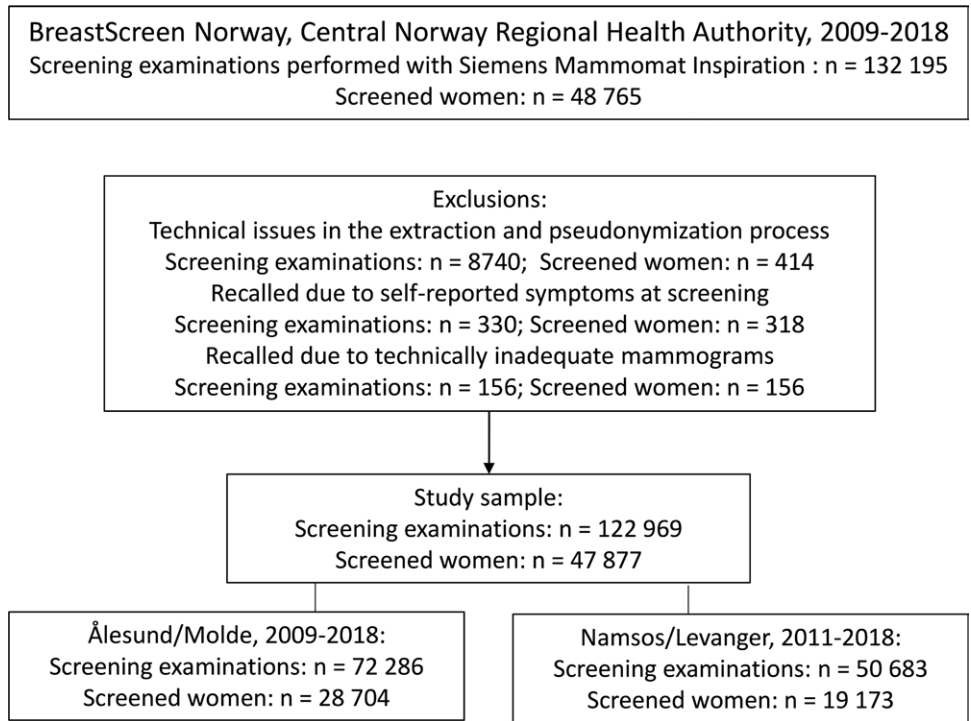


Figure 1: Flowchart of the study sample.

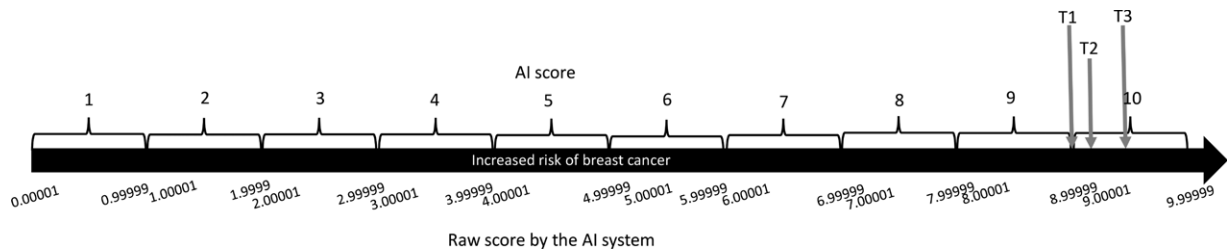


Figure 2: Diagram shows the artificial intelligence (AI) scoring system (raw score and AI score) with the three different thresholds (threshold 1 [T1], threshold 2 [T2], threshold 3 [T3]) defined for this study. T1 corresponds to AI score 10, T2 corresponds to a raw score of 9.13 and results in selecting 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.

assessment after consensus, or had a recall with a 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 was defined as breast cancer diagnosed within 24 months after a negative screening examination or 6–24 months after a recall with a negative outcome (5,18). Mammograms from prior screening examinations 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 histopathologic tumor characteristics.

Characteristics of invasive cancers included histologic type, tumor diameter, Nottingham grade 1–3, lymph node involvement, and immunohistochemical subtype. Subtype was classified into five groups (21). Histopathologic characteristics of DCIS included tumor diameter and Van Nuys grade 1–3 (22).

Statistical Analysis

Categorical variables are presented as frequencies and percentages, and continuous variables are presented as means and SDs or medians and IQRs, according to the distribution. Results on tumor characteristics were stratified by examinations selected and not selected by the AI system based on thresholds 1, 2, and 3. Stata for Windows (version 17.0, StataCorp) 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 (Fig 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 to 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).

Table 1: Characteristics of Included Examinations Stratified according to AI Score

AI Score	Prevalent Screening Examinations	Subsequent Screening Examinations
1	1908 (11.0)	13 285 (12.6)
2	911 (5.3)	5303 (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)

Note.—Data in parentheses are percentages. Percentages were calculated among the total number of prevalent and subsequent screening examinations. Artificial intelligence (AI) score is defined as the overall examination-level score from the AI system, and a score of 1 indicates low probability of breast cancer and 10 indicates high probability.

Mean age at screening was 60 years (SD = 6 years), and 14.1% of the examinations (17 350 of 122 969) 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% of the cancers (745 of 957) had the highest AI score of 10, including 86.8% of the screen-detected cancers (653 of 752) and 44.9% of the interval cancers (92 of 205). For illustration, see Figure 3. Among all examinations with an AI score of 10, 5.3% were screen-detected cancers (653 of 12 383) and 0.74% were interval cancers (92 of 12 383).

Five screen-detected cancers had the lowest AI score of 1: three were invasive and two were DCIS. Median tumor diameter was 9 mm (IQR, 9–18 mm) for invasive cancers, with one grade 3 tumor and none with positive lymph node involvement. Figure 4 shows a screen-detected cancer 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 were DCIS. Median tumor diameter was 8 mm (IQR, 6–11 mm) 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 examinations), and the recall rate was 3.2% (3896 of 122 969 examinations) 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, 45% (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.

Table 2: Screening Examinations and Results Stratified according to AI Score

AI Score	All Screening Examinations	Examinations with 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.—Data in parentheses are percentages. All other data are numbers of examinations or numbers of cancer cases. Percentages were calculated from the number of screening examinations and cancers. Negative screening results included a negative screening result and recall for further assessments with a negative outcome. Artificial intelligence (AI) score is defined as the overall examination-level score from the AI system, and a score of 1 indicates low probability of breast cancer and 10 indicates high probability.

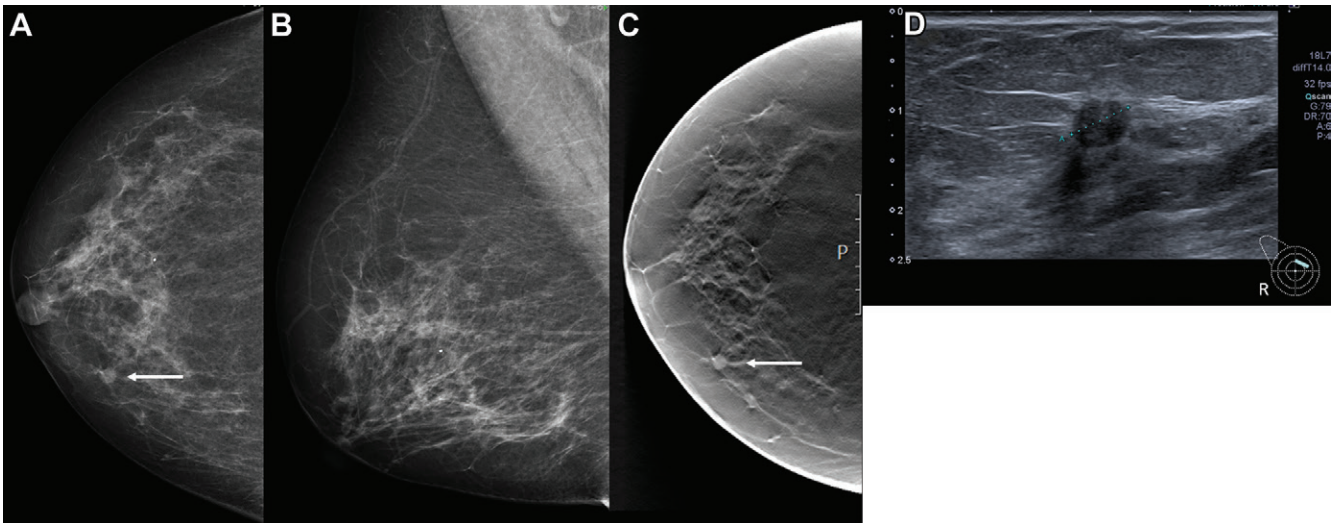


Figure 3: Images in a 68-year-old woman with a screen-detected ductal carcinoma in situ with an artificial intelligence (AI) score of 10 on the screening mammograms. **(A)** Mammogram of right breast from craniocaudal view. **(B)** Mammogram of right breast from mediolateral oblique view. **(C)** Craniocaudal digital breast tomosynthesis image of right breast. **(D)** US image of right breast. AI score is defined as the overall examination-level score from the AI system, and a score of 1 is indicative of low probability of breast cancer and 10 of high probability. The arrows in **A** and **C** indicate the malignancy, and the dotted line in **D** indicates the tumor diameter.

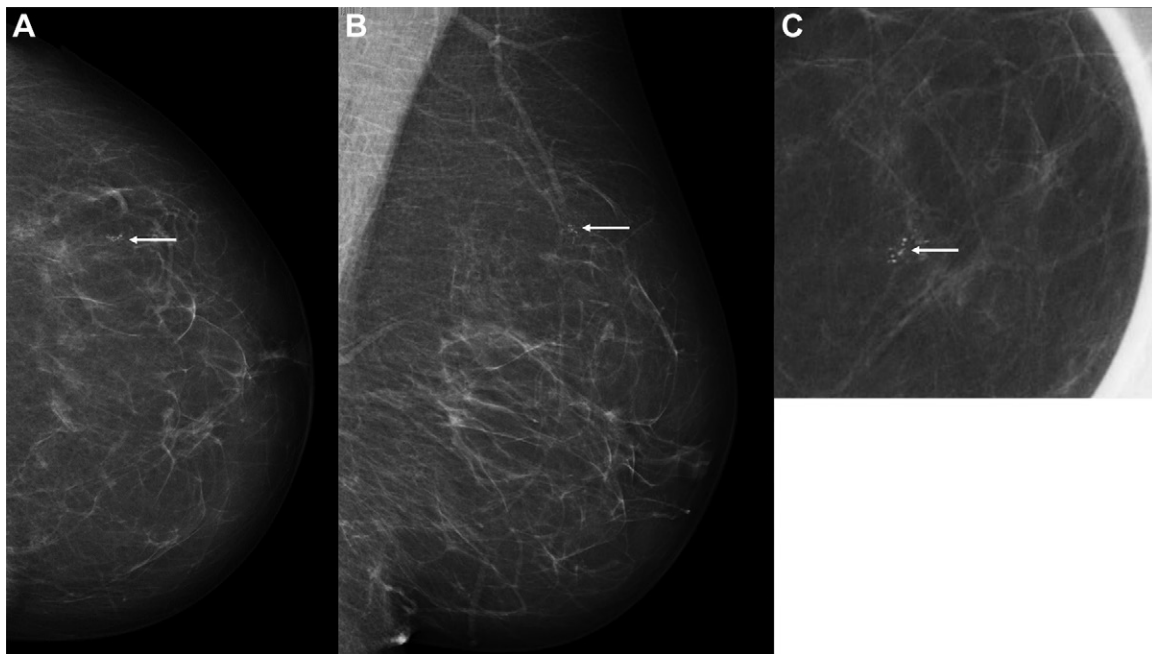


Figure 4: Images in a 60-year-old woman with an invasive screen-detected cancer with an artificial intelligence (AI) score of 1 on the screening mammograms. **(A)** Mammogram of left breast from craniocaudal view. **(B)** Mammogram of left breast from mediolateral oblique view. **(C)** Craniocaudal cone view mammogram with magnification. AI score is defined as the overall examination-level score from the AI system, and a score of 1 is indicative of low probability of breast cancer and 10 of high probability. The arrows indicate the malignancy.

Use of Threshold 1

Threshold 1 corresponds to selecting examinations with AI score 10. Threshold 1 selected 86.8% of the screen-detected cancers (653 of 752), and 82.2% of these were invasive (537 of 653) (Table 4). The percentage of invasive interval cancers selected was 93% (86 of 92). The median tumor diameter of the invasive screen-detected cancers selected by the AI system was 13 mm (IQR, 9–19 mm) versus 10 mm (IQR, 7–17 mm) for cancers

not selected. The percentage of histologic grade 3 cancers was 24.6% for those selected (131 of 532) and 20% for those not selected (16 of 79). Lymph node involvement was observed in 22.9% for those selected (120 of 524) and 18% for those not selected (14 of 79). On the basis of histologic grade, lymph node involvement, and subtype, interval cancers selected by AI had favorable tumor characteristics compared with interval cancers not selected by AI.

Table 3: Screening Outcome Stratified according to AI Score

AI Score	Examinations Discussed at Consensus after Positive Assessment by One or Both Radiologists	Examinations Recalled after Consensus	Screen-detected Cancers		Interval Cancers		
			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	10787 (100)	3896 (100)	170	582	21	16	5

Note.—Data in parentheses are percentages. All other data are numbers of examinations or numbers of cancer cases. Percentages were calculated from the total number of cases recalled after consensus and recalled cancers. Cancers were stratified by a positive assessment (interpretation score of 2 or higher) by one or both radiologists. Artificial intelligence (AI) score is defined as the overall examination-level score from the AI system, and a score of 1 indicates low probability of breast cancer and 10 indicates high probability.

Use of Threshold 2

Threshold 2 mirrors the consensus rate in the study sample, that is, positive interpretation by one or both radiologists. With the use of threshold 2, 85.1% of the screen-detected cancers (640 of 752) and 41.5% of the interval cancers (85 of 205) 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 histologic grade 3 was 24.5% among the invasive screen-detected cancers selected by AI (128 of 523) versus 22% among those not selected by AI (19 of 88). Lymph node involvement was observed for 23.3% of the selected cases (120 of 515) and 16% of the nonselected cases (14 of 88).

Use of Threshold 3

Threshold 3 mirrors the average individual radiologist rate of positive interpretation. Using threshold 3, 80.1% of the screen-detected cancers (602 of 752) and 30.7% of interval cancers (63 of 205) 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 mm) for cancers selected by the AI system and 9 mm (IQR, 7–15 mm) for the cancers not selected. The percentage of histologic grade 3 cancers was 25.3% for those selected (124 of 491) and 19.2% for the nonselected cancers (23 of 120), 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 with true-positive results for threshold 1, threshold 2, and threshold 3, AI selected 77.9% (745 of 957), 75.8% (725 of 957), and 69.5% (665 of 957) for thresholds 1, 2, and 3, respectively. The rates of selected cases without cancer (false-positive findings) were 94.0%

(11 638 of 12 383), 93.3% (10 064 of 10 789), and 90.7% (6471 of 7136) for thresholds 1, 2, and 3, respectively (the numbers for the two latter percentages are not given in tables or figures).

Discussion

The purpose of this study was to evaluate an artificial intelligence (AI) system for breast cancer detection on mammograms. The performance of the AI system was compared with that of 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 (threshold 3), 80.1% of screen-detected and 30.7% of 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 (11,13,24,25). 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 (24)—none of the 76 screen-detected cancers had an AI score below 3. In our larger sample, five of the 752 screen-detected cancers had a score below 4 (five had AI score 1 and none had AI score 2 or 3). 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 (11,25).

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

Table 4: Histopathologic Characteristics of Screen-detected and Interval Cancers Stratified by Using Threshold 1

Characteristic	Screen-detected Cancers		Interval Cancers	
	Selected with Threshold 1	Not Selected with Threshold 1	Selected with Threshold 1	Not Selected with Threshold 1
No. of cancers (DCIS and invasive cancer)	653 (86.8)	99 (13.2)	92 (44.9)	113 (55.1)
No. of DCIS lesions	116 (17.8)	17 (17)	6 (6.5)	8 (7.1)
Characteristics of DCIS				
Tumor diameter (mm)*	20 (10–30)	11 (10–15)	19 (12–25)	11 (7–17)
Data not available	18	4	0	1
Van Nuys grade				
1	14 (13.2)	5 (36)	0 (0)	2 (40)
2	11 (10.4)	3 (21)	0 (0)	0 (0)
3	81 (76.4)	6 (43)	6 (100)	3 (60)
Data not available	10	3	0	3
No. of invasive cancers	537 (82.2)	82 (83)	86 (94)	105 (92.9)
Characteristics of invasive cancers				
Tumor diameter (mm)*	13 (9–19)	10 (7–17)	17 (11–28)	16 (11–25)
Data not available	8	1	2	3
Nottingham grade				
1	175 (32.9)	33 (42)	19 (22)	14 (13.5)
2	226 (42.5)	30 (38)	37 (44)	37 (35.6)
3	131 (24.6)	16 (20)	29 (34)	53 (51.0)
Data not available	5	3	1	1
Lymph node involvement	120 (22.9)	14 (18)	26 (32)	38 (37.6)
Data not available	13	3	5	4
Immunohistochemical subtype				
Luminal A-like	313 (60.1)	48 (62)	45 (53)	44 (42.7)
Luminal B-like (HER2 negative)	85 (16.3)	11 (14)	12 (14)	26 (25.2)
Luminal B-like (HER2 positive)	77 (14.8)	11 (14)	16 (19)	13 (12.6)
HER2 positive (nonluminal)	14 (2.7)	3 (3.8)	3 (3.5)	7 (6.8)
Triple negative	32 (6.1)	5 (6.4)	9 (11)	13 (12.6)
Data not available	16	4	1	2

Note.—Except where indicated, data are numbers of cancers, with percentages in parentheses. Threshold 1 corresponds to cancers given an artificial intelligence (AI) score of 10. AI score is defined as the overall examination-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 DCIS lesions were calculated from the total number of cancers. DCIS = ductal carcinoma in situ, HER2 = human epidermal growth factor receptor 2.

* Data are medians, with IQR in parentheses.

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, then the radiologist may still identify the small number of missed cancers. Furthermore, 23% of screen-detected cancers in the study had a positive assessment by only one radiologist, and, thus, it may 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 result in 10% of the examinations requiring radiologist interpretation or 10% of the examinations directly selected for consensus. In the latter scenario, the consensus rate would be higher than usual in BreastScreen Norway and likely result in a higher recall rate. If radiologists are using

an AI system in a screening setting, then 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 are 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) (26).

Our results indicate favorable histopathologic characteristics for screen-detected cancers with low versus high AI scores. Studies have shown that less than 10% of screen-detected cancers are clinically insignificant, indicating a low risk of breast cancer death (27). An AI system that is able to differentiate between clinically significant and nonsignificant cancers could be beneficial for individual women and the screening program. Currently, there are limited data on the progression of small low-proliferation cancers, but such information could help women and clinicians to make informed choices on the intensity and extent of treatment.

Table 5: Histopathologic Characteristics of Screen-detected and Interval Cancers Stratified by Using Threshold 2

Characteristic	Screen-detected Cancers		Interval Cancers	
	Selected with Threshold 2	Not Selected with Threshold 2	Selected with Threshold 2	Not Selected with Threshold 2
No. of cancers (DCIS and invasive cancer)	640 (85.1)	112 (14.9)	85 (41.5)	120 (58.5)
No. of DCIS lesions	112 (17.5)	21 (18.8)	6 (7.1)	8 (6.7)
Characteristics of DCIS				
Tumor diameter (mm)*	20 (10–30)	10 (7–15)	19 (12–25)	11 (7–17)
Data not available	16	6	0	1
Van Nuys grade				
1	11 (10.7)	8 (47)	0 (0)	2 (40)
2	11 (10.7)	3 (18)	0 (0)	0 (0)
3	81 (78.6)	6 (35)	6 (100)	3 (60)
Data not available	9	4	0	3
Characteristics of invasive cancers				
Tumor diameter (mm)*	13 (9–19)	10 (7–17)	17 (11–26)	16 (11–25)
Data not available	8	1	2	3
Nottingham grade				
1	172 (32.9)	36 (40.9)	18 (23)	15 (13.5)
2	223 (42.6)	33 (37.5)	35 (45)	39 (35.1)
3	128 (24.5)	19 (21.6)	22 (32)	57 (51.4)
Data not available	5	3	1	1
Lymph node involvement				
Lymph node involvement	120 (23.3)	14 (15.9)	24 (32)	40 (37.0)
Data not available	13	3	5	4
Immunohistochemical subtype				
Luminal A-like	307 (60.0)	54 (62.1)	42 (54)	47 (42.7)
Luminal B-like (HER2 negative)	83 (16.2)	13 (14.9)	12 (15)	26 (23.6)
Luminal B-like (HER2 positive)	77 (15.0)	11 (12.6)	13 (17)	16 (14.6)
HER2 positive (nonluminal)	14 (2.7)	3 (3.5)	3 (3.9)	7 (6.4)
Triple negative	31 (6.1)	6 (6.9)	8 (10)	14 (12.7)
Data not available	16	4	1	2

Note.—Except where indicated, data are numbers of cancers, with percentages in parentheses. Threshold 2 corresponds 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 the highest scores by the artificial intelligence (AI) system were selected. The percentage of invasive cancers and DCIS lesions are calculated from the total number of cancers. DCIS = ductal carcinoma in situ, HER2 = human epidermal growth factor receptor 2.

* Data are medians, with IQR in parentheses.

Interval cancers are known to be less prognostically favorable compared with 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 threshold 1, threshold 2, and threshold 3 by the AI system had more favorable tumor characteristics compared with 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 (28).

The strengths of our study are the large study population from a real screening setting and the capture of all cancers through registry linkage. The 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 artificial intelligence (AI) system at the three evaluated thresholds was less than 20%, and several of these would probably also be detected at an early stage 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-nonselected 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.

Table 6: Histopathologic Characteristics of Screen-detected and Interval Cancers Stratified by Using Threshold 3

Characteristic	Screen-detected Cancers		Interval Cancers	
	Selected with Threshold 3	Not Selected with Threshold 3	Selected with Threshold 3	Not Selected with Threshold 3
No. of cancers (DCIS and invasive cancer)	602 (80.1)	150 (19.9)	63 (30.7)	142 (59.3)
No. of DCIS lesions	107 (17.8)	26 (17.3)	5 (7.9)	9 (6.3)
Characteristics of DCIS				
Tumor diameter (mm)*	20 (10–30)	10 (7–15)	20 (18–25)	12 (8–16)
Data not available	15	7	0	1
Van Nuys grade				
1	9 (9.1)	10 (48)	0 (0)	2 (33)
2	11 (11.1)	3 (14)	0 (0)	...
3	79 (79.8)	8 (38)	5 (100)	4 (67)
Data not available	8	5	0	3
No. of invasive cancers	495 (82.3)	124 (82.7)	58 (92)	133 (93.7)
Characteristics of invasive cancers				
Tumor diameter (mm)*	13 (9–20)	9 (7–15)	17 (11–26)	16 (12–25)
Data not available	7	2	1	4
Nottingham grade				
1	157 (32.0)	51 (42.5)	17 (30)	16 (12.1)
2	210 (42.8)	46 (38.3)	21 (37)	53 (40.2)
3	124 (25.3)	23 (19.2)	19 (33)	63 (47.7)
Data not available	4	4	1	1
Lymph node involvement	117 (24.3)	17 (14.0)	18 (33)	47 (36.7)
Data not available	13	3	4	5
Immunohistochemical subtype				
Luminal A-like	283 (59.1)	78 (65.0)	30 (53)	59 (45.0)
Luminal B-like (HER2 negative)	82 (17.1)	14 (11.7)	9 (16)	29 (22.1)
Luminal B-like (HER2 positive)	75 (17.1)	13 (10.8)	8 (14)	21 (16.0)
HER2 positive (nonluminal)	13 (2.8)	4 (3.3)	2 (3.5)	8 (6.1)
Triple negative	26 (5.4)	11 (9.2)	8 (14)	14 (10.7)
Data not available	16	4	1	2

Note.—Except where indicated, data are numbers of cancers, with percentages in parentheses. Threshold 3 corresponds 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 examinations with the highest scores by the artificial intelligence (AI) system were selected. The percentage of invasive cancers and DCIS lesions are calculated from the total number of cancers. DCIS = ductal carcinoma in situ, HER2 = human epidermal growth factor receptor 2.

* Data are medians, with IQR in parentheses.

Author contributions: Guarantors of integrity of entire study, **M.L., C.F.A., S.H.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **M.L., C.F.A., S.R.H., K.L., S.H.**; statistical analysis, **M.L., S.H.**; and manuscript editing, all authors

Disclosures of conflicts of interest: **M.L.** No relevant relationships. **C.F.A.** No relevant relationships. **C.I.L.** Participant on data safety monitoring board for GRAIL; deputy editor of the *Journal of the American College of Radiology*. **S.R.H.** No relevant relationships. **H.L.H.** No relevant relationships. **K.L.** Participant on Siemens Healthineers Advisory Board. **J.F.N.** No relevant relationships. **G.U.** No relevant relationships. **S.H.** No relevant relationships.

References

- 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 Cancer J Clin* 2018;68(6):394–424 [Published correction appears in *CA Cancer J Clin* 2020;70(4):313].
- 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):2353–2358.
- European Commission Initiative on Breast Cancer. Screening for women aged 50–69. <https://healthcare-quality.jrc.ec.europa.eu/european-breast-cancer-guidelines/screening-ages-and-frequencies/women-50-69>. Accessed September 2021.
- 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):150–159.
- Hofvind S, Tsuruda KM, Mangerud G, et al. The Norwegian Breast Cancer Screening Program, 1996–2016: Celebrating 20 years of organised mammographic screening. *Cancer in Norway 2016 - Cancer incidence, mortality, survival and prevalence in Norway*. Oslo, Norway: Cancer Registry of Norway, 2017.
- 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):378–386.
- 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. 10.1016/j.acra.2021.03.022. Published online April 26, 2021.
- 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):652–660.

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):331–342.
10. Dembrower K, Wählin 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.
11. Lång K, Dustler M, Dahlblom V, Åkesson A, Andersson I, Zackrisson S. Identifying normal mammograms in a large screening population using artificial intelligence. *Eur Radiol* 2021;31(3):1687–1692.
12. Rodríguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *J Natl Cancer Inst* 2019;111(9):916–922.
13. Salim M, Wählin E, Dembrower K, et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. *JAMA Oncol* 2020;6(10):1581–1588.
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. Krefregisterforskriften. <https://lovdata.no/dokument/SF/forskrift/2001-12-21-1477>. Accessed September 2021.
16. Larsen IK, Småstuen M, Johannesen TB, et al. Data quality at the Cancer Registry of Norway: an overview of comparability, completeness, validity and timeliness. *Eur J Cancer* 2009;45(7):1218–1231.
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. *Med Image Anal* 2017;35:303–312.
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):2206–2223.
22. Silverstein MJ, Poller DN, Waisman JR, et al. Prognostic classification of breast ductal carcinoma-in-situ. *Lancet* 1995;345(8958):1154–1157.
23. Sondén ECB, Sebuødegård S, Korvald C, et al. Cosmetic breast implants and breast cancer [in Norwegian]. *Tidsskr Nor Laegeforen* 2020;140(3).
24. 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.
25. Rodríguez-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. *Eur Radiol* 2019;29(9):4825–4832.
26. 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):121–127.
27. Bulliard JL, Beau AB, Njor S, et al. Breast cancer screening and overdiagnosis. *Int J Cancer* 2021;149(4):846–853.
28. Lång K, Hofvind S, Rodríguez-Ruiz A, Andersson I. Can artificial intelligence reduce the interval cancer rate in mammography screening? *Eur Radiol* 2021;31(8):5940–5947.