Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Performance of an Artificial Intelligence System for Breast Cancer Detection on Screening Mammograms from BreastScreen Norway

Marthe Larsen, MSc • Camilla F. Olstad, MS • Christoph I. Lee, MD, MS • Tone Hovda, MD • Solveig R. Hoff, PhD • Marit A. Martiniussen, MD • Karl Øyvind Mikalsen, PhD • Håkon Lund-Hanssen, MD • Helene S. Solli, MD • Marko Silberhorn • Åse Ø. Sulheim • Steinar Auensen, MSc • Jan F. Nygård, PhD • Solveig Hofvind, PhD

From the Section for Breast Cancer Screening (M.L., C.F.O., S.H.) and Department of Register Informatics (S.A., J.F.N.), Cancer Registry of Norway, Norwegian Institute of Public Health, PO 5313, Majorstuen, 0304 Oslo, Norway; 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, Vestre Viken Hospital Trust, Drammen, Norway (T.H.); Department of Radiology, Ålesund Hospital, Møre og Romsdal Hospital Trust, Ålesund, Norway (S.R.H.); Department of Circulation, Medical Imaging, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway (S.R.H.); Department of Radiology, Østfold Hospital Trust, Kalnes, Norway (M.A.M.); Institute of Clinical Medicine, University of Oslo, Oslo, Norway (M.A.M.); SPKI–The Norwegian Centre for Clinical Artificial Intelligence, University Hospital of North Norway, Tromsø, Norway (K.Ø.M.); Department of Clinical Medicine, Faculty of Health Sciences (K.Ø.M.), Department of Physics and Technology, Faculty of Science and Technology (J.F.N.), and Department of Health and Care Sciences, Faculty of Health Sciences (S.H.), UiT–The Arctic University of Norway, Tromsø, Norway; Department of Radiology and Nuclear Medicine, St Olavs University Hospital, Trondheim, Norway (H.L.H.); Department of Radiology, Hospital of Southern Norway, Kristiansand, Norway (H.S.S.); Department of Radiology, Innlandet Hospital Trust, Hamar, Norway (M.S.); and Department of Radiology, Innlandet Hospital Trust, Lillehammer, Norway (Å.Ø.S.). Received September 6, 2023; revision requested December 1; revision received February 18, 2024; accepted March 19. **Address correspondence to** S.H. (email: *sshh@kreftregisteret.no*).

Supported by the Norwegian Cancer Society and Pink Ribbon Campaign (grant 214931).

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

See also commentary by Bahl and Do in this issue.

Radiology: Artificial Intelligence 2024; 6(3):e230375 • https://doi.org/10.1148/ryai.230375 • Content codes: AI BR

Purpose: To explore the stand-alone breast cancer detection performance, at different risk score thresholds, of a commercially available artificial intelligence (AI) system.

Materials and Methods: This retrospective study included information from 661 695 digital mammographic examinations performed among 242 629 female individuals screened as a part of BreastScreen Norway, 2004–2018. The study sample included 3807 screen-detected cancers and 1110 interval breast cancers. A continuous examination-level risk score by the AI system was used to measure performance as the area under the receiver operating characteristic curve (AUC) with 95% CIs and cancer detection at different AI risk score thresholds.

Results: The AUC of the AI system was 0.93 (95% CI: 0.92, 0.93) for screen-detected cancers and interval breast cancers combined and 0.97 (95% CI: 0.97, 0.97) for screen-detected cancers. In a setting where 10% of the examinations with the highest AI risk scores were defined as positive and 90% with the lowest scores as negative, 92.0% (3502 of 3807) of the screen-detected cancers and 44.6% (495 of 1110) of the interval breast cancers were identified with AI. In this scenario, 68.5% (10 987 of 16 040) of false-positive screening results (negative recall assessment) were considered negative by AI. When 50% was used as the cutoff, 99.3% (3781 of 3807) of the screen-detected cancers and 85.2% (946 of 1110) of the interval breast cancers were identified as positive by AI, whereas 17.0% (2725 of 16 040) of the false-positive results were considered negative.

Conclusion: The AI system showed high performance in detecting breast cancers within 2 years of screening mammography and a potential for use to triage low-risk mammograms to reduce radiologist workload.

Supplemental material is available for this article.

Breast cancer is the most common cancer type among female individuals worldwide and is one of the leading causes of cancer-related deaths (1). Mammographic screening is well established in most European countries and all continents and has reduced breast cancer mortality (2,3). However, substantial radiologic resources are used to interpret screening mammograms, and more than 99% of the examinations are determined to be negative for breast cancer (4). Furthermore, interpreting mammograms is a subjective and perceptional task, and informed review studies have shown that 20%–30% of screen-detected and interval cancers are classified as false negative (missed) at previous screening (5,6).

Artificial intelligence (AI) has potential to reduce the interpretation volume in mammographic screening by replacing one of the two radiologists in a double-reading setting with AI, or by triaging the examinations into different risk groups that do not require any human readers, one reader, or two readers. In a study of commercially available AI systems from 2021, 12 systems for mammography were reported (7), and more than 20 had been cleared by the U.S. Food and Drug Administration (FDA) by August 2022 (8). This underscores the need for further validation of available algorithms to gather knowledge and explore generalizability of the reported performance. In addition, evidence from retrospective studies is crucial for planning

Abbreviations

AI = artificial intelligence, AUC = area under the ROC curve, BI-RADS = Breast Imaging and Reporting Data System, DCIS = ductal carcinoma in situ, FDA = U.S. Food and Drug Administration, HER2 = human epidermal growth factor receptor, ROC = receiver operating characteristic

Summary

A commercially available artificial intelligence system showed high performance in detecting breast cancers within 2 years of screening mammography and may help triage low-risk mammograms to reduce radiologist workload.

Key Points

- A commercially available artificial intelligence (AI) system for breast cancer detection on screening mammograms had an area under the receiver operating characteristic curve of 0.93 (95% CI: 0.92, 0.93) when screen-detected and interval breast cancers were included.
- After the 661 695 screening examinations were divided into two equal parts using a threshold of 3.1 for the AI risk score, 99.3% (3781 of 3807) of the screening cancers detected with independent double reading and consensus and 85.2% (946 of 1110) of the interval cancers were included in the half with the highest AI risk score.

Keywords

Mammography, Breast, Screening, Convolutional Neural Network (CNN), Deep Learning Algorithms

prospective studies, cost-effectiveness analyses, and implementation of AI in mammographic screening.

In addition to AI, more personalized screening schemes based on the patient's individual risk for the disease are likely to be implemented to improve sensitivity and specificity, reduce patient harm, and make breast cancer screening programs more cost-effective. Screening with MRI has been suggested for female individuals with extremely dense breasts at mammography because of the increased risk of the disease and low sensitivity of mammography (9). However, offering MRI to all female individuals with dense breasts is not feasible in most screening programs with current resources. AI might have the potential to increase sensitivity of mammography for those with extremely dense breasts (10,11). Exploring AI performance across different mammographic densities is thus important to ensure that the sensitivity is at an acceptable level across all densities. The objective measure of mammographic density that AI and other non–AI-based quantitative algorithms can provide is critical if a personalized approach is to be based on density.

Implementation of personalized screening and AI are possible improvements in breast cancer screening programs. We wanted to contribute to the knowledge needed for safe implementation of these efforts. Using image and screening data collected from BreastScreen Norway, we explored the performance of a commercially available AI system for identification of breast cancer. To make estimations more relevant for clinical use, we evaluated sensitivity of different AI risk score thresholds for defining positive and negative cases. We also investigated distribution and cancer detection by mammographic density measured by the AI system.

Materials and Methods

This retrospective registry study included Digital Imaging and Communications in Medicine image data of screening mammograms and information about the screening examinations from BreastScreen Norway. The study was approved by the Regional Committees for Medical and Health Research Ethics (#2018/2574) and had a legal basis in accordance with Articles 6(1)(e) and 9(2)(j) of the General Data Protection Regulation. The data were disclosed with legal basis in the Cancer Registry Regulations section 3–1 and the Personal Health Data Filing System Act section 19a to 19h (12,13). The target group in the screening program is informed that data related to participation are used for quality assurance and research and that they may decline this (opt out).

The Cancer Registry has a research agreement with Lunit, allowing free access to the AI software. None of the authors were employees of or consultants for Lunit. The authors had full control of the data and the information submitted for publication.

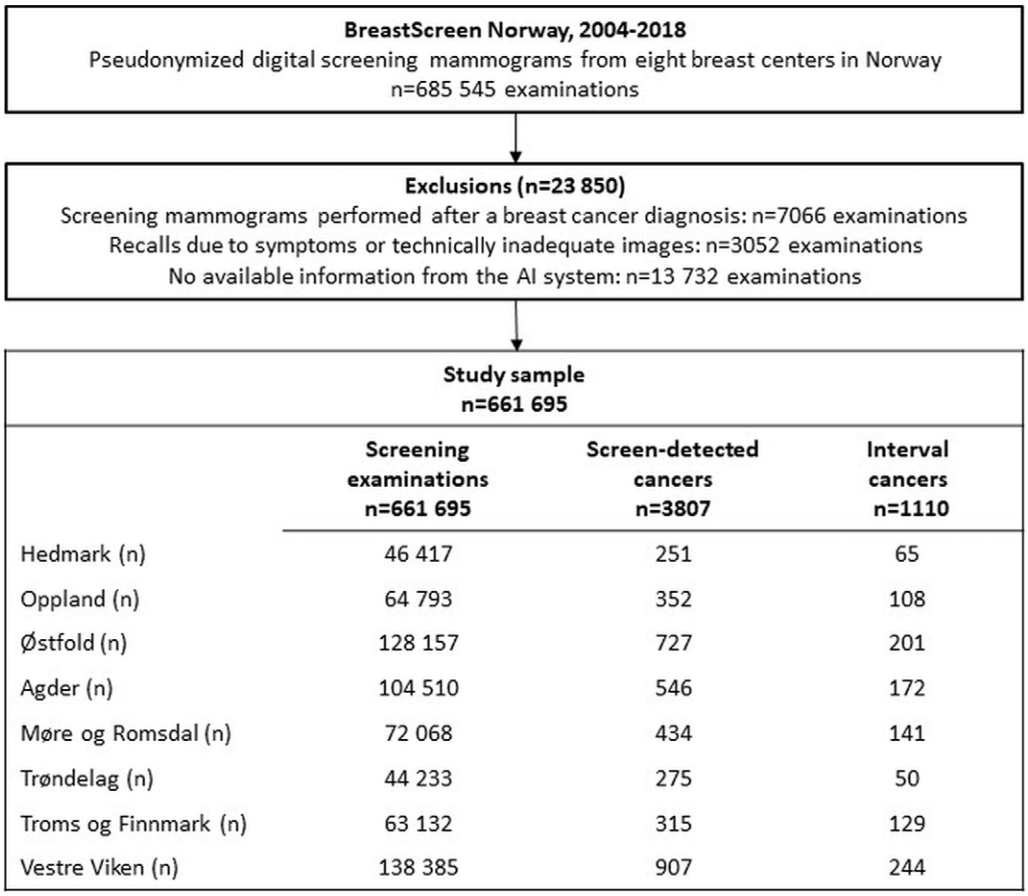
Study Sample

BreastScreen Norway invites all female individuals age 50–69 years to two-view biennial mammography screening. The program is administered by the Cancer Registry of Norway (4). The reporting of cancer information to the Cancer Registry of Norway is mandatory by law, and completeness has been reported to be 98.8% (14). The screening mammograms are independently interpreted by two breast radiologists, and each radiologist assigns each breast a score from 1 to 5. A score of 1 indicates normal findings; 2, probably benign; 3, intermediate suspicion; 4, probably malignant; and 5, high suspicion of malignancy. If either or both radiologists give a score of 2 or higher, the examination is discussed in a consensus meeting to decide whether to recall. From 2017 to 2021, the screening attendance rate was 76%, recall rate was 3.3%, screen-detected cancer rate was 6.2 per 1000 screening examinations, and interval cancer rate was 1.8 per 1000 screening examinations (4).

Our study included data from 685 545 digital screening examinations performed with MAMMOMAT Inspiration (Siemens Healthcare) from 2004 to 2018 at eight of BreastScreen Norway's 17 breast centers (Fig 1). The AI system retrospectively analyzed all examinations. Interpretation of results and final cancer diagnosis and characteristics were available before AI analysis, meaning that radiologists had no AI results available. The patients were followed up for interval cancers for 2 years after the screening examination, and the AI score on the latest screening mammogram was used in the analyses.

AI System

All examinations were retrospectively processed with INSIGHT MMG, version 1.1.7.2 (Lunit). The AI system provided a continuous risk score for each view of each breast, from 0 to 100, with higher scores indicating higher risk of suspicious findings (15,16). The highest risk score for each examination in our sample was used as an overall examination level risk score (AI score). The algorithm is Conformité



	Screening examinations n=661 695	Screen-detected cancers n=3807	Interval cancers n=1110
Hedmark (n)	46 417	251	65
Oppland (n)	64 793	352	108
Østfold (n)	128 157	727	201
Agder (n)	104 510	546	172
Møre og Romsdal (n)	72 068	434	141
Trøndelag (n)	44 233	275	50
Troms og Finnmark (n)	63 132	315	129
Vestre Viken (n)	138 385	907	244

Figure 1: Flowchart of the study sample. AI = artificial intelligence.

Européenne marked (CE-marked) and is FDA cleared (8). Furthermore, the AI system classified mammographic breast density using a 10-point scale, where 1 indicated fatty breast tissue and 10 indicated extremely dense breast tissue (17).

Variables of Interest

A negative screening result included examinations with an interpretation score of 1 by both readers and those selected for consensus but determined to be negative (ie, no recall for further assessment). A false-positive screening result was defined as a recall assessment with a negative outcome (no cancer). Screen-detected cancer was defined as breast cancer (ductal carcinoma in situ [DCIS] or invasive breast cancer) diagnosed after recall assessment. Interval cancer was defined as breast cancers detected within 24 months after a negative screening or 6–24 months after a false-positive screening result.

All cancer cases were based on surgical pathology (ie, histologically verified). Histopathologic tumor characteristics included histologic type (DCIS or invasive), tumor diameter in millimeters, histologic grade 1–3, and lymph node involvement for invasive cases. Immunohistochemical subtypes were based on estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2 (HER2) status, given as luminal A–like, luminal B–like HER2-negative, luminal B–like HER2-positive, HER2-positive, and triple-negative for invasive cancers (18). Information about mammographic features for invasive cancer cases was reported by the breast radiologists at assessment. A modified Breast Imaging and Reporting Data System (BI-RADS) was used, classifying the mammographic features of the tumors as mass, spiculated mass, architectural distortion, asymmetric density, density with calcifications, and calcifications alone (19–21).

Statistical Analysis

Categorical variables are presented as frequencies and percentages, and continuous variables are presented as means with SDs or medians with IQRs. Overall *P* values for the association between the different tumor characteristics variables and classification variable (high- or low-risk group) were calculated with a nonparametric test for tumor diameter and χ^2 test for categorical tumor characteristics variables. Distribution and cancer detection by mammographic density are presented as numbers and rates. All tests were two-sided and considered significant if the *P* value was less than .05. Performance was assessed by the area under the receiver operating characteristic (ROC) curve (AUC) for the AI system and calculated with 95% CIs for screen-detected cancers and screen-detected plus interval cancers separately. CIs were computed as asymptotic normal CI and considered according to a method described elsewhere (22). For

Table 1: Characteristics of Study Sample and Area under the Receiver Operating Characteristic Curve for AI System

Breast Center	Age at Screening (y)	Age at Diagnosis (y)	No. of Baseline Screening Examinations	AUC (95% CI)	
				Screen-detected and Interval Cancers	Screen-detected Cancers
Hedmark	59.4 ± 5.7	60.3 ± 6.1	7040/46 417 (15.2)	0.93 (0.91, 0.95)	0.98 (0.97, 0.99)
Oppland	59.5 ± 5.9	60.5 ± 6.0	9521/64 793 (14.7)	0.95 (0.93, 0.96)	0.98 (0.98, 0.99)
Østfold	59.2 ± 5.8	60.0 ± 5.9	18 833/128 157 (14.7)	0.93 (0.92, 0.94)	0.97 (0.97, 0.98)
Agder	59.4 ± 5.7	60.6 ± 6.0	14 651/104 510 (14.0)	0.93 (0.92, 0.94)	0.98 (0.97, 0.98)
Møre og Romsdal	60.1 ± 5.7	61.0 ± 5.9	10 337/72 068 (14.3)	0.91 (0.89, 0.92)	0.96 (0.95, 0.97)
Trøndelag	59.3 ± 5.7	60.7 ± 6.0	5583/44 233 (12.6)	0.94 (0.92, 0.96)	0.97 (0.96, 0.98)
Troms og Finnmark	59.7 ± 5.6	60.8 ± 5.6	8840/63 132 (14.0)	0.93 (0.91, 0.94)	0.97 (0.95, 0.98)
Vestre Viken	59.3 ± 5.8	60.2 ± 6.0	19 758/138 385 (14.3)	0.93 (0.92, 0.94)	0.97 (0.96, 0.98)
Total	59.5 ± 5.8	60.4 ± 5.9	94 563/661 695 (14.3)	0.93 (0.92, 0.93)	0.97 (0.97, 0.97)

Note.—Unless otherwise indicated, data are means ± SDs or numbers with percentages in parentheses. AI = artificial intelligence, AUC = area under the receiver operating characteristic curve.

the independent double reading, the AUC was calculated as 0.5 × (sensitivity + specificity) based on the individual radiologist's detection of screen-detected cancers and screen-detected plus interval cancers, considering all interval cancers as false negatives.

We explored AI performance for different AI risk score thresholds. The top 1% corresponded to an AI risk score above 93.0; a 5% threshold, to a score above 62.4; a 10% threshold, to a score above 39.3; a 30% threshold, to a score above 9.8; and a 50% threshold, to a score above 3.1. Histopathologic tumor characteristics and mammographic features were presented for screen-detected and interval cancers stratified by the highest 5% and lowest 95% AI scores to illustrate different characteristics for cancers with high versus low scores. A selection rate of 5% corresponded well to the mean rate of positive interpretations for each reader (reader 1 and 2) in the study sample (5.2%). In the analysis of mammographic density, ROC curves were presented for density categories 2, 6, and 10 to visualize differences across densities. AUC and 95% CI were calculated for each of these three categories.

Stata for Windows, version 17.0 (Stata), was used to analyze the data.

Results

Study Sample

Of 685 545 initial screening examinations, we excluded 7066 examinations performed after a breast cancer diagnosis, 3052 examinations resulting in a recall due to technical reasons or self-reported symptoms, and 13 732 examinations without AI results available. The final study sample included data from 661 695 screening examinations from 262 489 female individuals, including 3807 screen-detected and 1110 interval breast cancers (Fig 1). The mean age of individuals in the final study sample was 59.5 years ± 5.8 (SD) (Table 1).

Overall Performance

When we included screen-detected and interval cancers as positive cancer cases, the overall AUC for the AI system was 0.93 (95% CI: 0.92, 0.93), ranging from 0.91 (95% CI: 0.89, 0.921) to 0.95 (95% CI: 0.93, 0.96) for the different breast centers (Table 1). For screen-detected cancers, the AUC was 0.97 (95% CI: 0.97, 0.97), ranging from 0.96 (95% CI: 0.95, 0.97) to 0.98 (95% CI: 0.98, 0.99) between the breast centers. The AUC for the independent double reading in the regular screening setting was 0.88 (95% CI: 0.87, 0.88).

Triage Settings

In a triage scenario where 5% of the examinations with the highest AI scores were defined as positive and the remaining 95% were negative, 86.1% (3276 of 3807) of the screen-detected and 30.0% (333 of 1110) of the interval cancer cases were classified as positive (Fig 2, Table 2). When the highest 10% of examinations were defined as positive, 92.0% (3502 of 3807) of the screen-detected and 44.6% (495 of 1110) of the interval cancers were identified by the AI system. In the scenario of defining 50% of examinations as positive, 99.3% (3781 of 3807) of the screen-detected cancers and 85.2% (946 of 1110) of the interval cancers were identified by AI (Fig 3). A total of 26 (0.7%) of the 3807 screen-detected cancers were classified as negative in this scenario (Table 2), including 42.3% (11 of 26) of the DCIS cases. Among the 15 invasive cases considered negative, one was lymph node positive and one was histologic grade 3.

With a 90% threshold for negative cancer cases, 68.5% (10 987 of 16 040) of the false positives were included. This means that the rate of false-positive screening results could be reduced by 68.5% if AI alone were used to select 90% of the cases defined as negative and no radiologists were involved in the reading (Table 2). If 50% were considered negative, the percentage of false-positive cases among these would be 17.0% (2725 of 16 040). Furthermore, 26.1% (8973 of 34 379) of the cases

discussed and dismissed at consensus could be reduced in the 50% scenario.

Histopathologic Tumor Characteristics

When the 5% of screening examinations with the highest AI scores were defined as positive, 84.3% (2762 of 3276) of the screen-detected cancers classified as positive by the AI system were invasive and 84.4% (448 of 531) of those classified as negative were invasive. For the invasive cancers classified as positive, median tumor diameter was 13 mm (IQR, 9–19 mm), 21.1% (578 of 2746) were histologic grade 3, 21.9% (595 of 2720) were lymph node positive, and 46.6% (1065 of 2285) were luminal A–like (Table 3). For invasive cancers classified as negative, median tumor diameter was 9 mm (IQR, 7–13 mm), 14.9% (66 of 442) were histologic grade 3, 10.3% (46 of 445) were lymph node positive, and 52.6% (193 of 367) were luminal A–like.

Among the 5% of examinations with the highest AI score, 94.3% (314 of 333) of the interval cancer cases classified as positive by the AI system were invasive, and 93.1% (723 of 777) of those classified as negative were invasive (Table 3). For the invasive cancers classified as positive, median tumor diameter was 20 mm (IQR, 13–30 mm), 31.8% (99 of 311) were histologic grade 3, 40.6% (123 of 303) were lymph node positive, and 29.5% (80 of 271) were luminal A–like. For those classified as negative, median tumor diameter was 17 mm (IQR, 12–25 mm), 37.7% (266 of 705) were histologic grade 3, 32.1% (221 of 689) were lymph node positive, and 30.1% (191 of 634) were luminal A–like. Calcifications were more common for positive cases than negative for screen-detected and interval cancers (Table S1).

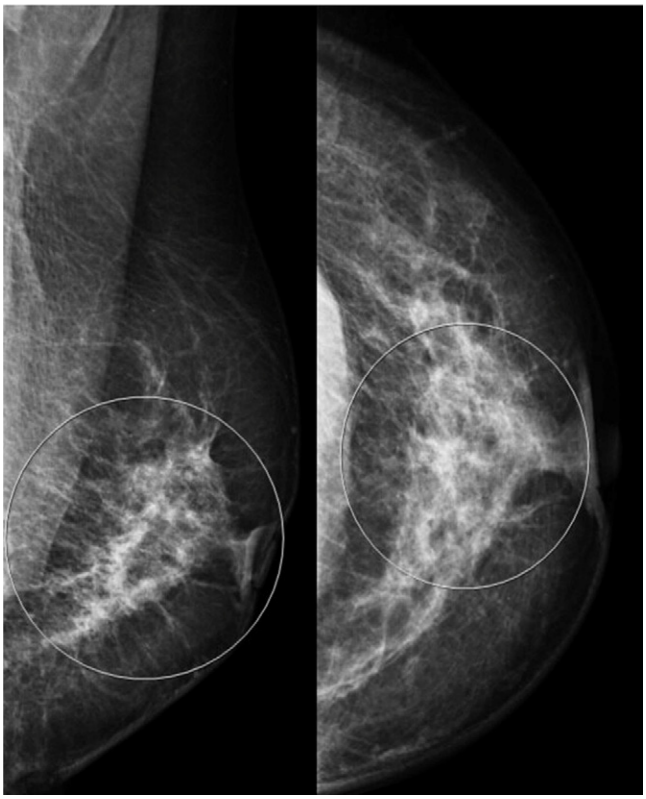


Figure 2: Left craniocaudal and mediolateral oblique mammograms in a 66-year-old female patient with an invasive interval cancer, 14 mm, histologic grade 2, lymph node negative, estrogen receptor positive, progesterone receptor positive, and human epidermal growth factor receptor 2 negative, with Ki67 less than 30%. The white circles illustrate the location of the tumor. The examination was classified as positive by the artificial intelligence system when the top 5% risk scores were defined as positive and 95% were defined as negative (risk score, 88.1).

Table 2: Screen-detected and Interval Cancers Classified as Positive by AI System at Different Thresholds for Positive Examinations and False-Positive Screening Results and Cases Dismissed at Consensus Classified as Negative at Different Thresholds for Negative Examinations

Triage Scenario	Cutoff Values[#]	SDC Classified Positive by AI	IC Classified Positive by AI	FP Cases Classified Negative by AI	Cases Dismissed at Consensus, Classified Negative by AI
1% positive (*n* = 6619) 99% negative (*n* = 655 076)	93.0	2484 (65.3)	128 (11.5)	15 302 (95.4)	33 540 (97.6)
5% positive (*n* = 33 090) 95% negative (*n* = 628 605)	62.4	3276 (86.1)	333 (30.0)	12 944 (80.7)	29 950 (87.1)
10% positive (*n* = 66 172) 90% negative (*n* = 595 523)	39.3	3502 (92.0)	495 (44.6)	10 987 (68.5)	26 563 (77.3)
30% positive (*n* = 198 522) 70% negative (*n* = 463 173)	9.8	3727 (97.9)	796 (71.7)	5815 (36.3)	16 273 (47.3)
50% positive (*n* = 330 905) 50% negative (*n* = 330 790)	3.1	3781 (99.3)	946 (85.2)	2725 (17.0)	8973 (26.1)

Note.—Unless otherwise noted, data are numbers, with percentages in parentheses. Examinations with the same risk score as *n* = p_x* 661 695 were included in the pool of positive cases for all cutoff values (p_x* = 0.01, 0.05, 0.1, 0.3, and 0.5). The study sample included 3807 screen-detected cancers and 1110 interval breast cancers. False-positive (*n* =16 040) screening results were recalled for further assessment after the consensus meeting but concluded negative after being recalled for further assessment. Cases dismissed at consensus (*n* = 34 379) were selected for consensus by either or both radiologists but concluded negative in consensus and not recalled for further assessment. AI = artificial intelligence, FP = false-positive, IC = interval cancers, SDC = screen-detected cancers.
[#] An AI score above the cutoff value of the continuous AI risk score given by the AI system was defined as positive.

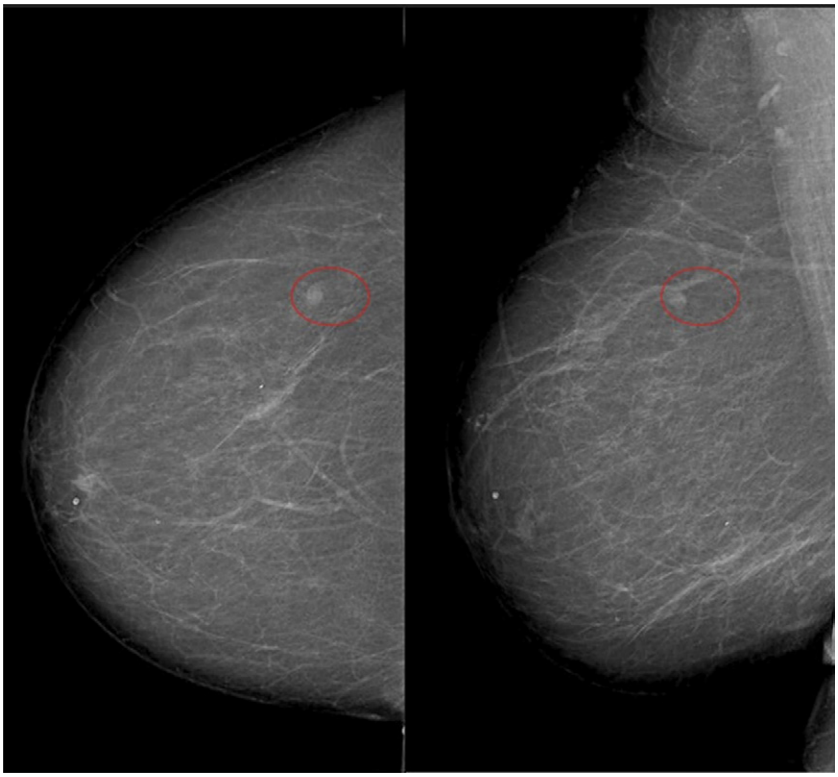


Figure 3: Right craniocaudal and mediolateral oblique mammograms in a 65-year-old female patient with invasive screen-detected breast cancer, 7 mm, lymph node negative, estrogen receptor negative, progesterone receptor negative, and human epidermal growth factor receptor 2 negative. The red circles illustrate the location of the tumor in the patient's right breast. The examination was classified as negative by the artificial intelligence (AI) system when the top 50% risk scores were defined as positive and 50% were defined as negative. No markings on the mammograms were available by the AI system because of low risk score.

Table 3: Histopathologic Tumor Characteristics of Invasive Screen-detected and Invasive Interval Cancers Classified as Positive and Negative when 5% of Examinations with Highest AI Risk Score Were Defined as Positive and 95% with Lowest Scores as Negative by AI System

	Screen-detected Cancers			Interval Cancers		
Variable	Positive by AI (n = 2762)	Negative by AI (n = 448)	P Value[#]	Positive by AI (n = 314)	Negative by AI (n = 723)	P Value[#]
Tumor diameter (mm)[†]	13 (9–19)	9 (7–13)	<.001	20 (13–30)	17 (12–25)	<.01
Information not available	45	4		18	69	
Histologic grade						
Grade 1	777 (28.3)	164 (37.1)	<.001	52 (16.7)	96 (13.6)	.15
Grade 2	1391 (50.7)	212 (48.0)		160 (51.5)	343 (48.7)	
Grade 3	578 (21.0)	66 (14.9)		99 (31.8)	266 (37.7)	
Information not available	16	6		3	18	
Lymph node positive	595 (21.9)	46 (10.3)	<.001	123 (40.6)	221 (32.1)	.009
Information not available	42	3		11	34	
Immunohistochemical subtypes			.02			.005
Luminal A–like	1065 (46.6)	193 (52.6)		80 (29.5)	191 (30.1)	
Luminal B–like, HER2–	638 (27.9)	96 (26.2)		88 (32.5)	180 (28.4)	
Luminal B–like, HER2+	347 (15.2)	43 (11.7)		62 (22.9)	112 (17.7)	
HER2+	93 (4.1)	6 (1.6)		22 (8.1)	51 (8.0)	
Triple negative	142 (6.2)	29 (7.9)		19 (7.0)	100 (15.8)	
Information not available	477	81		43	89	

Note.—Unless otherwise indicated, data are numbers or numbers with percentages in parentheses. AI = artificial intelligence, HER2 = human epidermal growth factor receptor 2.
[#] Nonparametric test for tumor diameter and χ^2 test for categorical tumor characteristics variables.
[†] Data are medians (IQRs).

Table 4: Number of Examinations and Screen-detected and Interval Cancer Rates for Each Density Category, 1–10, Classified by the AI System

Mammographic Density Category Given by AI System	No. of Examinations[#]	Screen-detected Cancer[†]	Interval Cancer[†]
1	35 441 (5.4)	124/3.5	27/0.8
2	185 706 (28.1)	850/4.6	163/0.9
3	147 606 (22.3)	874/5.9	204/1.4
4	125 391 (19.0)	825/6.6	247/2.0
5	60 332 (9.1)	441/7.3	119/2.0
6	29 734 (4.5)	206/6.9	92/3.1
7	27 917 (4.2)	189/6.8	88/3.2
8	14 420 (2.2)	115/8.0	45/3.1
9	19 079 (2.9)	113/5.9	71/3.7
10	16 069 (2.4)	70/4.4	54/3.4
Total	661 695 (100)	3807/5.8	1110/1.7

Note.—AI = artificial intelligence.
[#] Data in parentheses are percentages.
[†] Data are numbers/numbers per 1000 examinations.

Mammographic Density and AI Scores

A total of 28.1% (185 706 of 661 695) of the examinations were classified into density category 2, and 22.3% (147 606 of 661 695) were classified into category 3 (Table 4). Category 8 had the lowest proportion of examinations, 2.2% (14 420 of 661 695). The mean rate of screen-detected cancer was 5.8 per 1000 examinations (3807 of 661 695) for the entire study sample and 4.4 per 1000 examinations (70 of 16 069) for density category 10. The rate of screen-detected cancer was highest for density category 8, with a rate of 8.0 per 1000 examinations (115 of 14 420). The mean interval cancer rate was 1.7 per 1000 examinations (1110 of 661 695). The lowest rate, 0.8 per 1000 (27 of 35 441), was observed for density category 1 and the highest rate, 3.7 per 1000 (71 of 19 079), was observed for density category 9. When including screen-detected and interval cancers, the AUC was statistically significantly higher for density category 2 compared with category 6 and 10 (Fig 4).

Discussion

In this retrospective study of more than 660 000 mammography screening examinations, including close to 4000 screen-detected and 1000 interval cancers, we found an AUC of 0.93 for a commercially available AI system with inclusion of breast cancers detected within 2 years after screening (screen-detected and interval cancers). When we divided the screening examinations based on AI risk score (50% negative and 50% positive), 99.3% of the screen-detected and 85.2% of the interval cancers were classified as positive.

An external evaluation of three commercially available AI systems performed in Sweden showed AUCs of 0.916, 0.859, and 0.877 for the different systems with inclusion of cancers detected within 2 years (23). In another study from Sweden that also evaluated Lunit's AI system, no screen-detected cancers were observed among the 50% of examinations with the lowest AI score, and 14 (4.0%) screen-detected cancers were observed among the 90% of examinations with the lowest AI score (24). Corresponding results in our study were 0.7% and 8.0%, respectively. Among examinations in the top 1% of AI scores, interval cancers were detected in 12.0% in the Swedish study and 12.2% in our study. Different cutoff points of the continuous risk score for the highest 1% (69.6 in the Swedish study vs 92.9 in our study), different versions of the AI system, and the low number of cases classified as negative in the study from Sweden might be the reasons for the differences.

According to Table 2, a scenario in which the 50% of screening examinations with the lowest AI score was read only by AI, 40% by one reader, and 10% by two readers, the screen reading volume could be reduced by 70%. This scenario would miss less than 1% of the screen-detected cancers but include a potential of adding 85% of the interval cancers and reducing the consensus rate by 26% and false-positive rate by 17%. For a scenario with a threshold of 10% positive, 69% of the false positives would be considered negative by AI and consensus could be reduced by 77%. These are benefits that need to be considered in the light of missing 8% of the screen-detected cancers, but with the possibility of detecting up to 45% of the interval cancers. The scenarios described above have several assumptions; all cancer cases should be scored positive in the initial interpretation, selected for consensus, and recalled for further assessment, and all the cancers should be detected. These assumptions might be unrealistic because among the 50% of screening examinations with the highest risk score (n = 330 848), the cancer cases contributed only to 1.1% (3781 of 330 848) of the examinations. In addition, we do not know how availability of information about AI risk score would affect the radiologists' decision at consensus and recall assessment. To understand this, prospective studies are needed. Different screening scenarios and possible consequences on cancer detection and false-positive screening results must be discussed (25,26).

In deciding the cutoff for defining positive and negative examinations, false negatives versus false positives must be considered. By defining a large proportion of examinations as negative and not to be interpreted by radiologists, the number of false positives could be reduced substantially, but the risk of defining a cancer case as negative will increase compared with keeping the proportion of negative examinations low. We chose a threshold of 5% for positive versus negative screenings to describe histopathologic findings and mammographic features because of the selection rate of the average reader in the double reading setting. Other thresholds may have better fit in other screening programs and need to be considered.

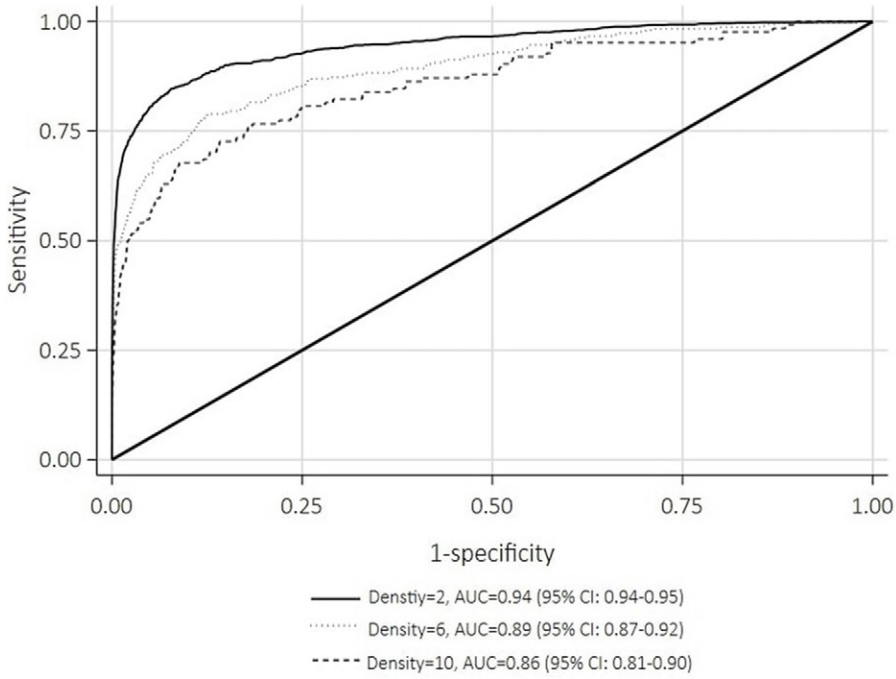


Figure 4: Receiver operating characteristic (ROC) curve for breast density category 2, 6, and 10 provided by the artificial intelligence system. Area under the ROC curve (AUC) values are presented with 95% CIs. Both screen-detected and interval cancers were included in the analysis.

According to publications reporting cancer detection rates by mammographic density (11,29,30), we expected increasing rates of cancer by increasing mammographic density provided by the AI system. The recent report with data from BreastScreen Norway showed an interval cancer rate of 4.1 per 1000 for the 2% of examinations with highest density and 4.6 per 1000 for the top 3% (29). In this study, we observed the highest interval cancer rates for density categories 9 (3.7 per 1000) and 10 (3.4 per 1000). Category 10 included 2.4% of the examinations. For screen-detected cancers, we found the highest rates for density category 8: 8.0 per 1000. The rate was 4.4 per 1000 screened for category 10. The lowest rate was observed for density category 1. We consider the results related to mammographic density in this study somewhat uncertain and that more studies are needed to understand the results and eventually be able to put the measurement method into clinical use.

Results on tumor diameter, histologic grade, lymph node status, and subtypes indicate that invasive screen-detected cancers classified as positive when defining the top 5% of examinations with the highest AI scores as positive had less favorable tumor characteristics than those classified as negative. A limited number of studies have reported on AI scores and tumor characteristics, but our findings support results in a study using another AI system (27). For interval cancers, lymph node status indicated less favorable characteristics among those classified as positive, but the percentage of histologic grade 3 and triple-negative tumors was lower for those classified as positive. The complexity of interval cancers, including tumors with different growth patterns and mammographic features, makes this issue challenging.

In a recent publication, mammographic density provided by Lunit was compared with radiologists (BI-RADS 5th edition) and an automated software measuring mammographic density, Volpara, version 3.4.1 (Volpara Health) (28). Lunit's density category 1–2 was defined as density a, 3–5 as b, 6–8 as c, and 9–10 as d. Mammographic density measures given by Lunit showed similar agreement with radiologists as seen between Volpara and Lunit ($\kappa = 0.52$ and 0.50, respectively). In our study, Lunit's density category 9–10 corresponded to 5.3% of the examinations and category 3–5 corresponded to 50.4% of the examinations. This is in line with findings from a recent study using data from BreastScreen Norway, in which 6.0% of examinations were classified as d (extremely dense) and 53.9% were classified as BI-RADS b by Volpara (29). However, 33.5% of the examinations in our study were category 1–2 and 10.9% were category 6–8. This is not in line with 13.7% with category a and 26.4% with c in the recently published study (29).

Strengths of this study include the large study sample, with more than 660 000 mammography examinations from a regular screening setting representing eight breast centers in three of the four health regions in Norway. Furthermore, the images were not used to train the AI algorithm, making our study an independent external test of the AI system.

Our study also had limitations. Our mammograms are from only one vendor, Siemens. In addition, the study did not include location of marking by the AI system or comparison with actual cancer location. Finally, no radiologic review was performed for interval cancers (missed vs true), and this is an important aspect to consider in the evaluation of AI regarding sensitivity.

In conclusion, results of our study support the growing literature indicating that AI can accurately identify cancers on digital mammograms and potentially help triage low-risk mammograms away from busy radiologists. The exact proportion of examinations defined as negative must be discussed and decided based on agreements about acceptable levels of potentially false-negative and false-positive screening results and workload reduction in specific screening settings. Future research should prospectively evaluate AI performance and AUC with different AI–radiologist combinations, including AI as a support in the interpretation and as a stand-alone approach, in addition to the cost-effectiveness of using AI in screen reading.

Acknowledgments: The study was financially supported by the Research Council of Norway, The Norwegian Breast Cancer Society, and the Norwegian Cancer Society.

Author contributions: Guarantors of integrity of entire study, **M.L., H.S.S., S.H.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual con-

tent, 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.O., S.R.H., S.H.**; clinical studies, **S.A.**; statistical analysis, **M.L., S.H.**; and manuscript editing, **M.L., C.F.O., C.I.L., T.H., S.R.H., M.A.M., K.Ø.M., H.L.H., H.S.S., J.F.N., S.H.**

Data sharing: Research data from the Cancer Registry of Norway, not the image data, used in the analyses can be made available on request to *https://helsedata.no/*, given legal basis in Articles 6 and 9 of the GDPR and that the processing is in accordance with Article 5 of the GDPR.

Disclosures of conflicts of interest: M.L. No relevant relationships. **C.F.O.** No relevant relationships. **C.I.L.** Grant or contract from the National Cancer Institute; textbook royalties from McGraw Hill, Oxford University Press, and UpToDate. **T.H.** No relevant relationships. **S.R.H.** No relevant relationships. **M.A.M.** No relevant relationships. **K.Ø.M.** No relevant relationships. **H.L.H.** Payment ($450) from Pfizer for lecturing on AI in mammography at a local oncology meeting (Trondheim, Norway, October 2023), payment covered preparation as well as the actual presentation. **H.S.S.** No relevant relationships. **M.S.** No relevant relationships. **Å.Ø.S.** No relevant relationships. **S.A.** No relevant relationships. **J.F.N.** No relevant relationships. **S.H.** No relevant relationships.

References

1. Arnold M, Morgan E, Rumgay H, et al. Current and future burden of breast cancer: Global statistics for 2020 and 2040. Breast 2022;66:15–23.
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):2353–2358.
3. Ren W, Chen M, Qiao Y, Zhao F. Global guidelines for breast cancer screening: a systematic review. Breast 2022;64:85–99.
4. Bjørnson E, Holen ÅS, Sagstad S, et al. BreastScreen Norway: 25 years of organized screening. Report No.: ISBN 978-82-93804-03-1. Cancer Registry of Norway; 2022.
5. Hovda T, Tsuruda K, Hoff SR, Sahlberg KK, Hofvind S. Radiological review of prior screening mammograms of screen-detected breast cancer. Eur Radiol 2021;31(4):2568–2579.
6. 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 2022;29(Suppl 1):S180–S191.
7. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 2021;31(6):3797–3804.
8. Bahl M. Artificial intelligence in clinical practice: implementation considerations and barriers. J Breast Imaging 2022;4(6):632–639.
9. Mann RM, Athanasiou A, Baltzer PAT, et al. Breast cancer screening in women with extremely dense breasts recommendations of the European Society of Breast Imaging (EUSOBI). Eur Radiol 2022;32(6):4036–4045.
10. Koch HW, Larsen M, Bartsch H, Kurz KD, Hofvind S. Artificial intelligence in BreastScreen Norway: a retrospective analysis of a cancer-enriched sample including 1254 breast cancer cases. Eur Radiol 2023;33(5):3735–3743.
11. Vachon CM, Scott CG, Norman AD, et al. Impact of artificial intelligence system and volumetric density on risk prediction of interval, screen-detected, and advanced breast cancer. J Clin Oncol 2023;41(17):3172–3183.
12. Lovdata. Kreftregisterforskriften. https://lovdata.no/dokument/SF/forskrift/2001-12-21-1477. Published 2001. Accessed February 18, 2024.
13. Lov om helseregistre og behandling av helseopplysninger (helseregisterloven). https://lovdata.no/dokument/NL/lov/2014-06-20-43. Accessed February 18, 2024.
14. 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.
15. Yoon JH, Han K, Suh HJ, Youk JH, Lee SE, Kim EK. Artificial intelligence-based computer-assisted detection/diagnosis (AI-CAD) for screening mammography: outcomes of AI-CAD in the mammographic interpretation workflow. Eur J Radiol Open 2023;11:100509.
16. Dembrower K, Salim M, Eklund M, Lindholm P, Strand F. Implications for downstream workload based on calibrating an artificial intelligence detection algorithm by standalone-reader or combined-reader sensitivity matching. J Med Imaging (Bellingham) 2023;10(S2 Suppl 2):S22405.
17. Lunit INSIGHT MMG. https://www.lunit.io/en/products/mmg. Accessed February 18, 2024.
18. 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.
19. D'Orsi CJ, Newell MS. BI-RADS decoded: detailed guidance on potentially confusing issues. Radiol Clin North Am 2007;45(5):751–763, v.
20. Barazi H, Gunduru M. Mammography BI RADS Grading. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK539816/. Accessed September 4, 2023.
21. Sickles E, D'Orsi C, Bassett L, et al. ACR BI-RADS mammography. In: ACR BI-RADS Atlas, Breast Imaging Reporting and Data System. 5th ed. Reston, Va: American College of Radiology, 2013.
22. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29–36.
23. 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):1581–1588.
24. 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.
25. Larsen M, Aglen CF, Hoff SR, Lund-Hanssen H, Hofvind S. Possible strategies for use of artificial intelligence in screen-reading of mammograms, based on retrospective data from 122,969 screening examinations. Eur Radiol 2022;32(12):8238–8246.
26. Dahlblom V, Dustler M, Tingberg A, Zackrisson S. Breast cancer screening with digital breast tomosynthesis: comparison of different reading strategies implementing artificial intelligence. Eur Radiol 2023;33(5):3754–3765.
27. Larsen M, Aglen CF, Lee CI, et al. Artificial intelligence evaluation of 122 969 mammography examinations from a population-based screening program. Radiology 2022;303(3):502–511.
28. Lee SE, Son NH, Kim MH, Kim EK. Mammographic density assessment by artificial intelligence-based computer-assisted diagnosis: a comparison with automated volumetric assessment. J Digit Imaging 2022;35(2):173–179.
29. Larsen M, Lynge E, Lee CI, Lång K, Hofvind S. Mammographic density and interval cancers in mammographic screening: moving towards more personalized screening. Breast 2023;69:306–311.
30. Bodewes FTH, van Asselt AA, Dorrius MD, Greuter MJW, de Bock GH. Mammographic breast density and the risk of breast cancer: a systematic review and meta-analysis. Breast 2022;66:62–68.