AI risk score on screening mammograms preceding breast cancer diagnosis

Marthe Larsen[1], Camilla F Aglen[1], Henrik W Koch[2,3], Marit A Martiniussen[4,5], Solveig R Hoff[6,7], Håkon Lund-Hanssen[8], Helene S Solli[9], Karl Øyvind Mikalsen[10,11], Steinar Auensen[12], Jan Nygård[12], Kristina Lång[13,14], Yan Chen[15], Solveig Hofvind[1,16]

[1] Section for Breast Cancer Screening, Cancer Registry of Norway, Oslo, Norway

[2] Department of Radiology, Stavanger University Hospital, Stavanger, Norway

[3] Faculty of Health Sciences, University of Stavanger, Stavanger, Norway

[4]Department of Radiology, Østfold Hospital Trust, Kalnes, Norway

[5] Institute of Clinical Medicine, University of Oslo, Oslo, Norway

[6] Department of Radiology, Ålesund Hospital, Møre og Romsdal Hospital Trust, Ålesund, Norway

[7] Department of Circulation and Medical Imaging, Faculty of Medicine and Health Sciences, National University for Science and Technology, Trondheim, Norway

[8] Department of Radiology and Nuclear Medicine, St Olavs University Hospital, Trondheim, Norway

[9] Department of Radiology, Hospital of Southern Norway, Kristiansand, Norway

[10] SPKI – the Norwegian Centre for clinical artificial intelligence, University Hospital of North Norway, Tromsø, Norway

[11] Department of Clinical Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway

[11] Department of Register Informatics, Cancer Registry of Norway, Oslo, Norway

[12] Department of Translational Medicine, Diagnostic Radiology, Lund University, Lund, Sweden

[13] Unilabs Mammography Unit, Skåne University Hospital, Malmø, Sweden

[15] School of Medicine, University of Nottingham, Clinical Science Building, Nottingham City Hospital, Nottingham, United Kingdom

[16] Department of Health and Care Sciences, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway

Corresponding author: Solveig Hofvind, sshh@kreftregisteret.no, +47 22 92 88 28

Data sharing statement: Research 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.

Abbreviations:

AI - artificial intelligence

Summary Statement: Over 38% of both screen-detected and interval cancers were assigned the highest artificial intelligence risk score, on screening mammograms preceding breast cancer diagnosis.

Key results:

- In this retrospective study of 1602 patients with breast cancer, 38.3% (389/1016) of screen-detected cancers and 39.4% (231/586) of interval cancers had the highest malignancy risk score assigned by artificial intelligence (AI) on the screening mammogram prior to diagnosis.
- Mammographic features were associated with AI score 10 (high risk) versus 1-7 (low risk) on prior mammograms for screen-detected invasive cancers ($p<0.001$).
- Among invasive screen-detected cancers with AI score 10 and 1-7 at prior mammograms, density with calcifications were observed for 13.6% (43/317) and 4.7% (15/322), respectively.

Abstract

Background: Few studies have evaluated the role of artificial intelligence (AI) being used in prior screening mammograms.

Purpose: To examine AI risk scores assigned to screening mammograms of women who were later diagnosed with breast cancer.

Methods: Image data and screening information of examinations performed from January 2004 to December 2019 as part of BreastScreen Norway were used in this retrospective study. Prior screening examinations from women later diagnosed with cancer were assigned an AI risk score by a commercially available AI system (1–7=low risk of malignancy, 8–9=intermediate risk, 10=high risk of malignancy). Mammographic features of the cancers based on the AI score were also assessed. The association between AI score and mammographic features were tested with bivariate test.

Results: A total of 2787 prior screening examinations from 1602 women (mean age 59 years, standard deviation 5.1) with screen-detected (n=1016) or interval cancer (n=586) showed an AI risk score of 10 for 389 (38.3%) and 231 (39.4%) respectively, on the mammograms in the screening round prior to diagnosis. Among the screen-detected cancers with AI scores available two screening rounds (four years) before diagnosis, 23.0% (122/531) had a score of 10. Mammographic features were associated with AI score for invasive screen-detected cancers ($p<0.001$). Density with calcifications was registered for 13.6% (43/317) of screen-detected cases with a score of 10 and 4.7% (15/322) for those with a score of 1-7.

Conclusion: More than one in three screen-detected and interval cancer cases had the highest AI risk score at prior screening suggesting the use of AI in mammography screening may lead to earlier detection of breast cancers.

Introduction

Breast cancer is the most common cancer type in women worldwide, accounting for 2.3 million new cancers in 2020 and a predicted 3 million new cancers in 2040 (1). Despite reduced disease-specific mortality due to the implementation of screening and improved treatment, breast cancer is the second most common cause of cancer death among women in developed countries (2). Standardized mammographic screening is recommended by several international health authorities (3).

More than 99% of screening examinations are determined to have a negative outcome (4, 5). Due to the low prevalence of the disease, the interpretation of the screening mammograms requires trained and experienced breast radiologists to keep sensitivity and specificity at acceptable levels. Retrospective review studies have reported that 20-30% of screen-detected and interval cancers were classified as false negatives at prior screening (6, 7). Furthermore, a recent publication reported that approximately 10% of screen-detected and interval cancers were discussed at a consensus meeting in the prior screening round but were dismissed and women were not recalled for further assessments (8).

Artificial intelligence (AI) has shown great potential in the field of radiology to reduce workload and increase diagnostic accuracy (9). In mammographic screening, AI can be used as a standalone technique to triage examinations and/or as support for radiologists in their interpretations. The performance of different AI algorithms is being evaluated in both retrospective and prospective studies (10-15). Several retrospective studies have reported AI risk scores for screen-detected cancers (16-19) or have reported the AI risk score of interval cancers for different AI systems (17, 19-24). For example, a retrospective study reviewing interval cancers showed that 19% of 429 interval cancers had the highest AI score and that AI was able to flag with the correct suspicious location (20). However, few studies have assessed the AI risk scores on the prior mammograms of screen-detected cancers (17, 24).

With the aim of exploring the potential for earlier detection of breast cancer, imaging data collected in BreastScreen Norway were used, and AI scores on the mammograms from screening examinations preceding breast cancer diagnosis were analyzed. Furthermore, to determine whether the cases with a high AI score on prior mammograms were of clinical

relevance and had the potential to be diagnosed earlier, prognostic histopathological tumor characteristics and mammographic features were analyzed.

Materials and methods

This retrospective registry study included imaging data and screening information from BreastScreen Norway which is administered by the Cancer Registry of Norway (4). The study was approved by the Regional Committees for Medical and Health Research Ethics (#13294) and had a legal basis in accordance with Articles 6 (1) (e) and 9 (2) (j) of the GDPR. Pursuant to Section 35 of the Health Research Act, the Regional Committees for Medical Research Ethics has granted the project exemption from the requirement of consent (25).

Study sample

Examinations were identified from the Cancer Registry of Norway where information from all breast centers and screening units in BreastScreen Norway are stored. Digital Imaging and Communications in Medicine (DICOM) imaging data from all 372 580 digital screening examinations performed during the period from January 2004 to December 2019 in five breast centers in BreastScreen Norway were analyzed with an AI system (Figure 1). Results from two breast centers, have been included in previous studies (23, 26). A total of 122 969 examinations were included in these studies, but the aim of these studies was overall AI performance, and to explore different clinical workflow for AI and radiologists.

After excluding examinations performed after breast cancer diagnosis and examinations where less than four images were processed with the AI system, the overall study sample, sample A, included 344 337 examinations, 1929 screen-detected and 586 interval cancers (Figure 1). Results for all examinations were presented to give an overview of the AI performance. In the analysis of AI scores on prior examinations for cancer cases, examinations among women without breast cancer, screen-detected cancers with priors outside the study period or without priors, examinations not following the biennial screening scheme were excluded. We excluded mammograms from examinations four or further rounds back in time (eight years or more), since the data file included only digital images, and thus a limited number of prior screening examinations. 5From study sample A, 338 958 examinations from women without a screen-detected or an interval cancer were excluded. Examinations were also excluded based on the following: 756 screen-detected cancers with

prior screening examination outside the study period or detected at first attendance in BreastScreen Norway, 639 examinations not following the biennial screening scheme, and 181 examinations with four or more screening rounds prior to diagnosis. Ultimately, study sample B included 2787 prior screening examinations, including 1733 prior screening examinations from 1016 women with screen-detected cancer and 1054 prior screening examinations from 586 women with interval cancer.

Imaging and reading procedure

BreastScreen Norway invites women aged 50-69 to two-view (craniocaudal and mediolateral oblique view) mammography screening of each breast biennially (4). From 2017 to 2021, the rate of attendance was 75%, recalls 3.3%, screen-detected cancer 0.64% and interval cancer 0.18% (21). Screening examinations are independently read by two breast radiologists, and both radiologists assign each breast a score from 1 to 5 (4) . 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 score 2 or higher, the examination is discussed in a consensus meeting by the same or other radiologists. Consensus decide whether to recall the woman. Recall assessment might include clinical examination, additional imaging (mammography, ultrasound, and eventually MRI), and needle biopsy.

In this study, the radiologists did not have AI results available. All data on AI-assessments were collected retrospectively.

Variables of interest

A screen-detected cancer was defined as a histologically verified ductal carcinoma in situ (DCIS) or invasive breast cancer diagnosed after a recall for further assessment due to mammographic findings and within 6 months after screening (4). An interval cancer was defined as breast cancer detected after a negative screening result or more than 6 months after being recalled with a negative outcome and within 24 months after screening. Screen-detected and interval cancer were considered the reference standard.

Prognostic histopathological tumor characteristics included histological type (DCIS or invasive), tumor diameter, histological grade 1–3, lymph node involvement and immunohistochemical subtypes for invasive cancers. The subtypes were classified as luminal A–like, luminal B–like Her2−, luminal B–like Her2+, Her2+, and triple negative and were

determined based on estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (Her2) status (27). Information about mammographic features was reported by the radiologists and classified as mass, spiculated mass, architectural distortion, asymmetric density, density with calcifications, and calcifications alone.

AI system

All women included in the study were screened with MAMMOMAT Inspiration, Siemens Healthcare. Examinations were analyzed with Transpara v. 1.7.0 (ScreenPoint Medical). For each examination, a malignancy risk score (AI score) from 1 to 10 was determined for each examination, where 1-7 indicated a low risk of malignancy, 8-9 intermediate risk and 10 high risk of malignancy. The AI score indicates risk of malignancy at the given screening examination. We have previously shown that 4.4% of the screen-detected and 30.2% of the interval cancers have an AI score 1-7, 8.8% and 24.9% for AI score 8 and 9, and 86.6% and 44.9% an AI score of 10 (23). The AI system aims to assign approximately 10% of the examinations to each score. The AI system uses convolutional neural networks to identify calcifications and soft tissue lesions, and is trained, validated and tested on mammograms from four different vendors (28). Transpara v. 1.7.0 does not include prior examinations in the risk score assessment.

Statistical analysis

Descriptive analyses were performed for study samples A and B separately. Frequencies and percentages are presented to summarize AI findings. Histopathological tumor characteristics and mammographic features are presented only for invasive cancers with an AI score of 10 (high risk) and 1-7 (low risk) at P1, the screening examination 2 years or less prior to diagnosis of breast cancer, and percentages were calculated from nonmissing values. Associations were tested with bivariate tests with a significance level of 0.05. All analyses were performed by M.L. with Stata (StataCorp. 2021. *Stata Statistical Software: Release 17*).

Results

Characteristics of study sample A – all examinations

After excluding 3740 examinations performed after breast cancer diagnosis and 24 492 examinations where less than four images were processed with the AI system, study sample A, included 344 337 examinations (Figure 1). Mean age was 59.7 years (Standard deviation, SD=5.7) and 13.4% (46 087/344 337) were prevalent examinations (Table 1).

A total of 21.9% (75 547/344 337) of the examinations had an AI score of 1, while 9.0% (31 025/344 337) had an AI score of 10 (Figure 2, Supplementary Figure 1). A total of 88.0% (1697/1929) of the screen-detected cancers and 39.4% (231/586) of the interval cancers had a score of 10 (Figure 2). The probability of screen-detected cancer for examinations with AI score 1 was 0.01% (9/75 547) and 5.5% (1697/31 025) for AI score 10. The percentage of screen-detected cancers with an AI score of 10 ranged from 86.4% (248/287) to 91.7% (341/372) among the five breast centers, whereas the percentage of interval cancers with an AI score of 10 ranged from 35.4% (52/147) to 53.3% (32/60) (Supplementary Table 1).

Characteristics of study sample B - prior examinations of cancer cases

For women with screen-detected cancer, 1016 examinations were performed at the two years prior to diagnosis (P1), 531 were performed four years prior to diagnosis (P2), and 186 were performed six years prior to diagnosis (P3). For women with interval cancers, the numbers were 586 at P1, 308 at P2 and 160 at P3. Mean age at diagnosis was 62.4 years (SD=5.0) for screen-detected cancers and 61.2 years (SD=5.9) for interval cancers (Table 1).

AI scores on prior screening mammograms of screen-detected cancers

A total of 38.3% (389/1016) of the examinations at P1 had an AI score of 10 and 23.7% (241/1016) had a score of 8 or 9 (Table 2). Among the 389 screen-detected cancers with an AI score of 10 at P1, 11 (2.8%) had an AI score <10 at diagnosis. A total of 27.3% (106/389) of those with score 10 at P1 were discussed at the consensus meeting, whereas 22.6% (88/389) were concluded to be normal after the consensus meeting and not recalled and 4.6% (18/389) of underwent recall assessment with a negative outcome. In comparison, 11.1% (43/386) of the examinations with score 1-7 at P1 were discussed at consensus. Among the consensus cases with an AI score of 10, 83.0% (88/106) were dismissed at the consensus meeting, whereas 80.7% (71/88) of the dismissed cases were selected for consensus (an interpretation score of 2 or higher) by only one of the two radiologists.

For the 531 women with screen-detected cancer and an AI score available at P2, 23.0% (122/531) had an AI score of 10, and 90 women had an AI score of 10 at both P1 and P2 (Table 2, Figure 3). At P3, 16.7% (31/186) had an AI score of 10.

AI scores on prior screening mammograms of interval cancers

For interval cancers, 39.4% (231/586) had an AI score of 10, and 24.9% (146/586) had an AI score of 8 or 9 at P1 (Table 2). A total of 35.5% (82/231) of the interval cancers with an AI score of 10 at P1 were discussed at consensus, and among these cases, 41.5% (34/82) underwent recall assessment with a negative outcome, and 58.5% (48/82) were dismissed. Among the dismissed cases, 85.4% (41/48) were selected for consensus by only one of the two radiologists. At P2 and P3, 23.4% (72/308) and 23.1% (37/160) of the interval cancers had an AI score of 10, respectively (Table 2).

Prognostic histopathological tumor characteristics and mammographic features

Among the screen-detected cancers with an AI score of 10 and 1-7 at P1, 85.4% (332/389) and 89.9% (347/386) were invasive, respectively. Median tumor diameter for screen-detected cancers with an AI score of 10 at P1 was 13 mm (IQR: 9-18) and 11 mm (IQR: 7-17) for cancers with score 1-7 ($p<0.05$) (Table 3). Histological grade 3 was observed for 17.0.1% (56/329) of the cancer with score 10 at P1 and for 24.8% (85/343) of cancers with score 1-7 at P1 ($p<0.05$). The most frequent mammographic feature of the invasive cases was spiculated mass, found in 41.3% (131/317) of those with a score of 10 at P1 and 43.5% (140/322) of those with a score of 1-7 (Table 4). The association between mammographic feature and AI score 10 versus 1-7 at P1 was statistically significant for screen-detected cancers ($p<0.001$). Density with calcifications was registered for 13.6% (43/317) of screen-detected cases with a score of 10 and 4.7% (15/322) for those with a score of 1-7.

Among the interval cancers with an AI score of 10 at P1, 94.8% (219/231) were invasive, while 94.7% (198/209) of those with an AI score of 1-7 were invasive. No statistically significant associations were observed for tumor characteristics for invasive interval cancers with AI score 10 at P1 versus score 1-7 (Table 3). However, 30.5% (65/213) of the interval cancers with score 10 at P1 were histological grade 3 and 37.3% (79/212) were lymph node positive versus 39.5% (75/190) and 30.3% (57/188) for cancers with score 1-7. For invasive interval cancers, the association between mammographic feature and AI score 10 versus 1-7

at P1 was not statistically significant (p=0.216). However, calcifications alone were observed for 6.9% (6/87) of those with an AI score of 10 and for none of the cases with an AI score of 1-7 (Table 4).

Discussion

In our study, including 344 337 examinations, we found that 88.0% of the screen-detected and 39.4% of the interval cancers had a score of 10. When considering prior examinations for cancer cases, we found that the prior screening examinations (P1) were classified as high risk by the AI system (score 10) in 38.3% of screen-detected cancers and 39.4% of interval cancers. In two screening rounds prior to diagnosis (P2), 23.0% and 23.4% of the screen-detected and interval cancers had an AI score of 10, respectively. Mammographic features were associated with AI score (10 versus 1-7) at P1, 2 years prior to diagnosis for invasive screen-detected cancers (p<0.001). Density with calcifications was registered for 13.6% (43/317) of screen-detected cases with a score of 10 at P1 and 4.7% (15/322) for cancers with an AI score of 1-7.

Prior mammograms for 745 screen-detected cancers in a cancer-enriched sample were analyzed with the same commercially available AI system as in our study (29). Of these cases, 41.9% had an AI score of 10 at P1. In our study, the corresponding percentage was 38.3%. We observed a lower proportion of examinations with an AI score of 10 than the expected proportion of 10%, and in the enriched sample (26), a higher proportion was observed. This might explain the lower proportion of screen-detected cancers with an AI score of 10 in our study. In a study in which results from another AI system were analyzed, it was reported that in the top 10% with the highest risk score, 45% of screen-detected cancers were selected by the AI system at the prior screening (17).

For interval cancers, we observed a larger proportion of cases with an AI score of 10 than what was reported in a study from Sweden (39.4% versus 33.3%) (20). An updated version of the AI system was used in our study (version 1.7.0 versus 1.5.0). In a study in which version 1.6.0 was used, it was reported that 37.5% of interval cancers were identified at 90% specificity (21).

Review studies have reported that screen-detected and interval cancers classified as missed had histopathological favorable tumor characteristics compared with those classified as true

negatives (6, 7). Missed cases might have suspicious visible findings on prior examinations and have the potential to be detected earlier. In our study, a higher proportion of grade 3 tumors (overall p-value=0.030) for those with AI score 1-7 versus 10 at P1 indicates less favorable tumor characteristics and might thus be true negatives cases. On the other hand, larger tumor diameter (p=0.016) for AI score 10 versus 1-7 at P1 indicated the opposite. The clinical relevance of earlier detection for cases with a score of 10 at P1 thus remains unclear when taking histopathological tumor characteristics into account.

The results for mammographic features for invasive screen-detected cancers indicate calcifications alone or in combination with density to be more common for screen-detected cancers with an AI score of 10 versus 1-7 at P1 (overall p-value<0.001). Poorer survival for women with small (<15 mm) screen-detected invasive cancers presenting as calcification and large (>=15 mm) tumors presenting as a density with calcifications has been reported (30). It might thus be of importance to recall women with calcifications and score 10 for assessment as it might lead to earlier detection of relevant cancers. However, the proportion of calcifications and AI score 10 among disease free women also must be explored in this context as is might influence the rate of false positive screening results (recalled with a negative outcome).

There are several limitations of the study. First, the use of mammograms from a single vendor and only including women from Norway limit the generalizability of the findings. Second, there is a lack of knowledge about the correlation between the location of the AI markings and the location of the cancer. This limitation may lead to an overestimation of our findings in favor of the AI system. A review of the hotspot versus the location of the cancer is needed to understand this issue.

In conclusion, we found that more than one in three screen-detected and interval cancer cases had an AI risk score of 10 at prior screening. This indicates a potential of AI to detect breast cancer earlier, which could lead to less harmful treatment for the affected women. Review studies and prospective studies comparing location of AI markings versus the location of the cancer is needed to understand this issue further. Furthermore, high AI score on mammograms from women not diagnosed with breast cancer represent a challenge which is important to consider.

Acknowledgments

We thank all personnel involved in the collection of imaging data at the breast centers included in the study.

References

1. Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, et al. Current and future burden of breast cancer: Global statistics for 2020 and 2040. The Breast. 2022;66:15-23.
2. Lauby-Secretan B, Scoccianti C, Loomis D, Benbrahim-Tallaa L, Bouvard V, Bianchini F, et al. Breast-cancer screening--viewpoint of the IARC Working Group. N Engl J Med. 2015;372(24):2353-8.
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, Larsen M, Thy J, Mangerud G, et al. BreastScreen Norway: 25 years of organized screening. Cancer Registry of Norway; 2022. Report No.: ISBN 978-82-93804-03-1.
5. Blanks RG, Wallis MG, Alison RJ, Given-Wilson RM. An analysis of screen-detected invasive cancers by grade in the English breast cancer screening programme: are we failing to detect sufficient small grade 3 cancers? Eur Radiol. 2021;31(4):2548-58.
6. Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. NPJ Breast Cancer. 2017;3:12.
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.
8. Martiniussen MA, Sagstad S, Larsen M, Larsen ASF, Hovda T, Lee CI, et al. Screen-detected and interval breast cancer after concordant and discordant interpretations in a population based screening program using independent double reading. Eur Radiol. 2022;32(9):5974-85.
9. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18(8):500-10.
10. Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ. 2021;374:n1872.
11. Mammography Screening With Artificial Intelligence (MASAI): ClinicalTrials.gov; [28.03.2023]. Available from: https://ClinicalTrials.gov/show/NCT04838756.
12. Artificial Intelligence in Large-scale Breast Cancer Screening (ScreenTrustCad): ClinicalTrials.gov; [28.03.2023]. Available from: https://ClinicalTrials.gov/show/NCT04778670.
13. Lång K, Josefsson V, Larsson AM, Larsson S, Högberg C, Sartor H, et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol. 2023;24(8):936-44.
14. Yoon JH, Strand F, Baltzer PAT, Conant EF, Gilbert FJ, Lehman CD, et al. Standalone AI for Breast Cancer Detection at Screening Digital Mammography and Digital Breast Tomosynthesis: A Systematic Review and Meta-Analysis. Radiology. 2023;307(5):e222639.
15. Hickman SE, Woitek R, Le EPV, Im YR, Mouritsen Luxhøj C, Aviles-Rivero AI, et al. Machine Learning for Workflow Applications in Screening Mammography: Systematic Review and Meta-Analysis. Radiology. 2022;302(1):88-104.
16. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Teuwen J, Broeders M, Gennaro G, 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.

17. Dembrower K, Wåhlin E, Liu Y, Salim M, Smith K, Lindholm P, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. Lancet Digital Health. 2020;2(9):e468-e74.
18. 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.
19. Lauritzen AD, Rodríguez-Ruiz A, von Euler-Chelpin MC, Lynge E, Vejborg I, Nielsen M, et al. An Artificial Intelligence-based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload. Radiology. 2022:210948.
20. 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):5940-7.
21. Wanders AJT, Mees W, Bun PAM, Janssen N, Rodríguez-Ruiz A, Dalmış MU, et al. Interval Cancer Detection Using a Neural Network and Breast Density in Women with Negative Screening Mammograms. Radiology. 2022;303(2):269-75.
22. Byng D, Strauch B, Gnas L, Leibig C, Stephan O, Bunk S, et al. AI-based prevention of interval cancers in a national mammography screening program. Eur J Radiol. 2022;152:110321.
23. Larsen M, Aglen CF, Lee CI, Hoff SR, Lund-Hanssen H, Lång K, et al. Artificial Intelligence Evaluation of 122 969 Mammography Examinations from a Population-based Screening Program. Radiology. 2022;303(3):502-11.
24. Park GE, Kang BJ, Kim SH, Lee J. Retrospective Review of Missed Cancer Detection and Its Mammography Findings with Artificial-Intelligence-Based, Computer-Aided Diagnosis. Diagnostics. 2022;12(2):387.
25. Lov om helseregistre og behandling av helseopplysninger (helseregisterloven) [21.04.2023]. Available from: https://lovdata.no/dokument/NL/lov/2014-06-20-43.
26. 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):8238-46.
27. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thürlimann B, 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.
28. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, 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.
29. 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.
30. Tabar L, Tony Chen HH, Amy Yen MF, Tot T, Tung TH, Chen LS, et al. Mammographic tumor features can predict long-term outcomes reliably in women with 1-14-mm invasive breast carcinoma. Cancer. 2004;101(8):1745-59.

Table 1. Characteristics related to all examinations, Study sample A, and the study sample including cancers and screening examinations prior to cancer diagnosis, Study sample B.

	Study sample A, n=344 337	Study sample B, n=2787 prior examinations and 1602 cancer cases
Age at screening, mean (SD) years	59.7 (5.7)	59.7 (5.1)
Age at diagnosis, screen-detected cancers, mean (SD) years	60.9 (5.8)	62.4 (5.0)
Age at diagnosis, interval cancers, mean (SD) years	61.2 (5.9)	61.2 (5.9)
Prevalent screening examinations, n (%)	46 087 (13.4%)	281 (7.4%)*

SD, standard deviation

*Prior examinations can be prevalent screening examinations only for interval cancer cases.

Table 2. AI score on prior screening examinations for women with screen-detected or interval cancer.

	Screen-detected cancers			Interval cancer		
AI score	P1 n (%)	P2 n (%)	P3 n (%)	P1 n (%)	P2 n (%)	P3 n (%)
1	62 (6.1%)	60 (11.3%)	19 (10.2%)	40 (6.8%)	30 (9.7%)	17 (10.6%)
2	40 (3.9%)	25 (4.7%)	9 (4.8%)	21 (3.6%)	10 (3.3%)	6 (3.8%)
3	43 (4.2%)	33 (6.2%)	13 (7.0%)	24 (4.1%)	15 (4.9%)	8 (5.0%)
4	50 (4.9%)	37 (7.0%)	16 (8.6%)	23 (3.9%)	20 (6.5%)	14 (8.8%)
5	57 (5.6%)	29 (5.5%)	12 (6.5%)	33 (5.6%)	31 (10.1%)	10 (6.3%)
6	62 (6.1%)	42 (7.9%)	20 (10.8%)	27 (4.6%)	19 (6.2%)	17 (10.6%)
7	72 (7.1%)	45 (8.5%)	19 (10.2%)	41 (7.0%)	26 (8.4%)	13 (8.1%)
8	96 (9.5%)	52 (9.8%)	18 (9.7%)	40 (6.8%)	30 (9.7%)	20 (12.5%)
9	145 (14.3%)	86 (16.2%)	29 (15.6%)	106 (18.1%)	55 (17.9%)	18 (11.3%)
10	389 (38.3%)	122 (23.0%)	31 (16.7%)	231 (39.4%)	72 (23.4%)	37 (23.1%)
Total	1016 (100%)	531 (100%)	186 (100%)	586 (100%)	308 (100%)	160 (100%)

AI, artificial intelligence; P1, the screening examination prior to diagnosis (≤2 years prior to diagnosis); P2, the screening examination two examinations prior to diagnosis (≤4 years prior to diagnosis); P3, the screening examination three examinations prior to diagnosis (≤6 years prior to diagnosis).

Table 3. Histopathological tumor characteristics of invasive tumors stratified by an AI score of 10 and an AI score of 1-7 at P1.

	Screen-detected cancers			Interval cancers		
Tumor characteristics of invasive cancers	AI score 10, n=332	AI score 1-7, n=347	p-value*	AI score 10, n=219	AI score 1-7, n=198	p-value*
Diameter, median (IQR) mm	13 (9-18)	11 (7-17)	0.016	18 (12-27)	16 (11-25)	0.405
NA, n	5	4		12	22	
Histological grade, n (%)			0.030			0.662
1	103 (31.3%)	87 (25.4%)		38 (17.8%)	28 (14.7%)	
2	170 (51.7%)	171 (49.9%)		110 (51.6%)	87 (45.8%)	
3	56 (17.0%)	85 (24.8%)		65 (30.5%)	75 (39.5%)	
NA, n	3	4		6	8	
Lymph node involvement, n (%)	66 (20.1%)	65 (19.0%)	0.702	79 (37.3%)	57 (30.3%)	0.748
NA, n	4	4		7	10	
Immunohistochemical subtypes, n (%)			0.060			0.106
Luminal A-like	154 (51.3%)	143 (45.7%)		65 (34.4%)	42 (25.5%)	
Luminal B-like, Her2-	89 (29.7%)	93 (29.7%)		53 (28.0%)	49 (29.7%)	
Luminal B-like, Her2+	35 (11.7%)	45 (14.4%)		37 (19.6%)	25 (15.2%)	
Her2+	11 (3.7%)	6 (1.9%)		17 (9.0%)	16 (9.7%)	
Triple negative	11 (3.7%)	26 (8.3%)		17 (9.0%)	33 (20.0%)	
NA, n	32	34		30	33	

AI, artificial intelligence; P1, the screening examination prior to diagnosis, ≤2 years prior to diagnosis; Her2, human epidermal growth factor receptor. An AI score of 10 indicates high risk of malignancy and 1-7 indicates low risk; NA, Information Not Available

*Overall association between each tumor characteristic variable and AI score (10 versus 1-7) were tested with bivariate test for continuous or categorical outcome as appropriate.

Table 4. Mammographic features for invasive screen-detected and interval cancers stratified by an AI score of 10 and an AI score of 1-7 at P1.

AI, artificial intelligence; P1, the screening examination prior to diagnosis, ≤2 years prior to diagnosis.

	Invasive screen-detected cancers			Invasive interval cancers		
	AI score 10, n=332	AI score 1-7, n=347	p-value*	AI score 10, n=219	AI score 1-7, n=198	p-value*
Mammographic feature, n (%)			<0.001			0.216
Mass	23 (7.3%)	78 (24.2%)		13 (14.9%)	17 (21.8%)	
Spiculated mass	131 (41.3%)	140 (43.5%)		38 (43.7%)	25 (32.1%)	
Architectural distortion	12 (3.8%)	11 (3.4%)		5 (5.8%)	3 (3.9%)	
Asymmetric density	58 (18.3%)	53 (16.5%)		17 (19.5%)	30 (38.5%)	
Density with calcifications	43 (13.6%)	15 (4.7%)		8 (9.2%)	3 (3.9%)	
Calcifications alone	50 (15.8%)	25 (7.8%)		6 (6.9%)	0 (0%)	
Information not available	15	25		132	120	

An AI score of 10 indicates high risk of malignancy and 1-7 indicates low risk.

*Overall association between AI score (10 versus 1-7) and mammographic feature tested with bivariate test.

Figure Legends

Figure 1. Flow chart of exclusions and the final study sample A (all examinations) and B (cancer cases and prior examination of cancer cases). *The artificial intelligence (AI) system can process more and less images than the standard of four images. However, due to storage format of the mammograms, we had technical issues with some images from mainly one breast center.

Figure 2. Distribution of artificial intelligence (AI) scores for all examinations (n=344 337), screen-detected cancers (n=1929), and interval cancers (n=586) for the full dataset. An AI score of 1 indicates low risk of malignancy and an AI score of 10 indicates high risk of malignancy.

Figure 3. A) Left craniocaudal (L-CC) and mediolateral oblique (L-MLO) mammography views from diagnosis for a 57-year-old woman with an invasive screen-detected cancer. The tumor was histologic grade 2, lymph node negative, estrogen receptor positive, and progesterone receptor positive. The white arrows point to the location of the tumor. B) L-CC and L-MLO from P1 (the screening examination prior to diagnosis, ≤2 years prior to diagnosis). C) L-CC and L-MLO from P2 (the screening examination two examinations prior to diagnosis, ≤4 years prior to diagnosis). The artificial intelligence system gave a score of 10 (high risk of malignancy) at diagnosis (3A), P1 (3B) and P2 (3C).

Supplemental Material

Supplementary Table 1. Screen-detected (SDC) and interval cancers (IC) stratified by artificial intelligence (AI) score 1-10 for breast center Agder, Troms og Finnmark, Østfold, Trøndelag, and Møre og Romsdal. Percentages are calculated based on the total number of SDC or IC for each breast center.

AI score	Agder, n (%)		Troms og Finnmark, n (%)		Østfold, n (%)		Trøndelag, n (%)		Møre og Romsdal, n (%)	
	SDC	IC	SDC	IC	SDC	IC	SDC	IC	SDC	IC
1	0 (0)	11 (9.2)	1 (0.3)	8 (6.7)	3 (0.6)	12 (8.2)	4 (1.4)	2 (3.3)	1 (0.2)	7 (5.0)
2	2 (0.5)	6 (5.0)	1 (0.3)	7 (5.9)	1 (0.2)	7 (4.8)	0 (0)	0 (0)	0 (0)	1 (0.7)
3	1 (0.3)	6 (5.0)	0 (0)	3 (2.5)	5 (0.9)	6 (4.1)	0 (0)	4 (6.7)	0 (0)	5 (3.6)
4	0 (0)	3 (2.5)	1 (0.3)	6 (5.0)	0 (0)	5 (3.4)	3 (1.1)	1 (1.7)	4 (0.9)	8 (5.7)
5	1 (0.3)	8 (6.7)	5 (1.7)	7 (5.9)	1 (0.2)	7 (4.8)	2 (0.7)	5 (8.3)	2 (0.5)	6 (4.3)
6	3 (0.8)	8 (6.7)	2 (0.7)	3 (2.5)	8 (1.5)	7 (4.8)	3 (1.1)	4 (6.7)	6 (1.4)	5 (3.6)
7	0 (0)	8 (6.7)	5 (1.7)	7 (5.9)	5 (0.9)	13 (8.8)	5 (1.7)	3 (5.0)	2 (0.5)	10 (7.1)
8	6 (1.6)	10 (8.3)	6 (2.0)	9 (7.6)	18 (3.3)	7 (4.8)	7 (2.4)	2 (3.3)	10 (2.3)	12 (8.6)
9	18 (4.8)	15 (12.5)	16 (5.4)	25 (21.0)	29 (5.4)	31 (21.1)	15 (5.2)	7 (11.7)	30 (6.9)	28 (20.0)
10	341 (91.7)	45 (37.5)	261 (87.6)	44 (37.0)	469 (87.0)	52 (35.4)	248 (86.4)	32 (53.3)	378 (87.3)	58 (41.4)
Total	372 (100)	120 (100)	298 (100)	119 (100)	539 (100)	147 (100)	287 (100)	60 (100)	433 (100)	140 (100)

AI, artificial intelligence; SDC, screen-detected cancer; IC, interval cancer. An AI score of 1 indicates low risk of malignancy and an AI score of 10 indicates high risk of malignancy.

Figure legend

Supplementary Figure 1. Distribution of artificial intelligence (AI) score 1-10 for all screening examinations for breast center. An AI score of 1 indicates low risk of malignancy and an AI score of 10 indicates high risk of malignancy.

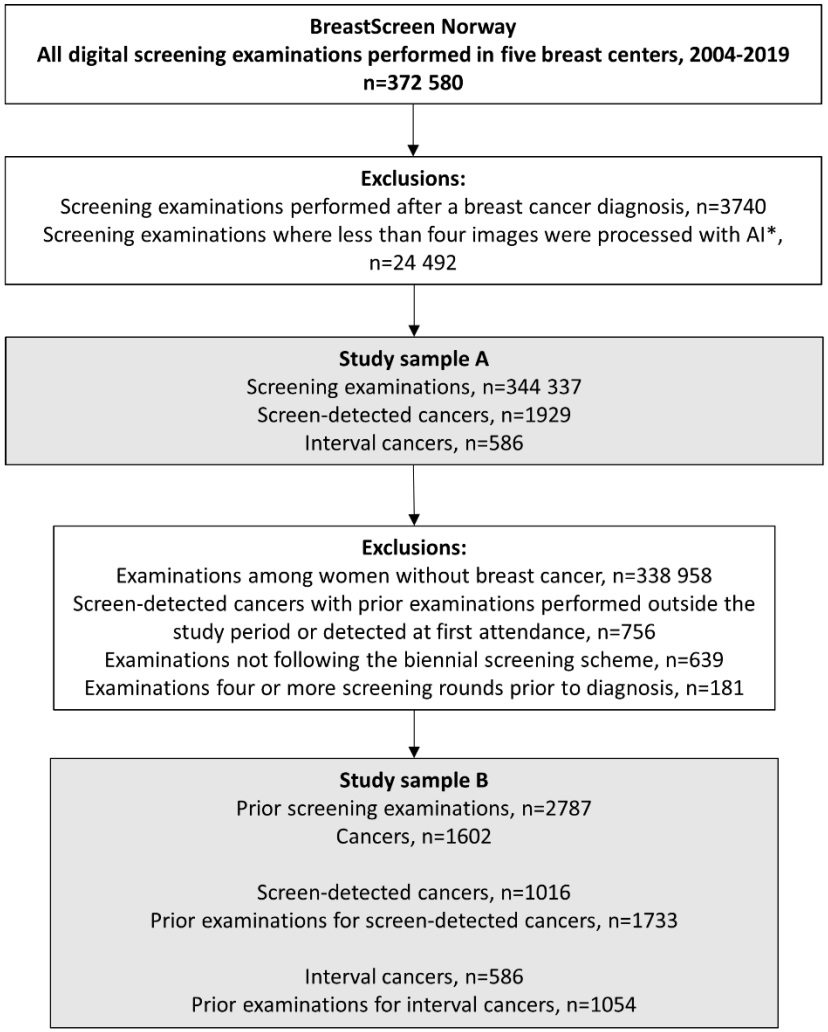

BreastScreen Norway
All digital screening examinations performed in five breast centers, 2004-2019
n=372 580
Exclusions:
Screening examinations performed after a breast cancer diagnosis, n=3740
Screening examinations where less than four images were processed with AI*, n=24 492
Study sample A
Screening examinations, n=344 337
Screen-detected cancers, n=1929
Interval cancers, n=586
Exclusions:
Examinations among women without breast cancer, n=338 958
Screen-detected cancers with prior examinations performed outside the study period or detected at first attendance, n=756
Examinations not following the biennial screening scheme, n=639
Examinations four or more screening rounds prior to diagnosis, n=181
Study sample B
Prior screening examinations, n=2787
Cancers, n=1602
Screen-detected cancers, n=1016
Prior examinations for screen-detected cancers, n=1733
Interval cancers, n=586
Prior examinations for interval cancers, n=1054

Figure 1

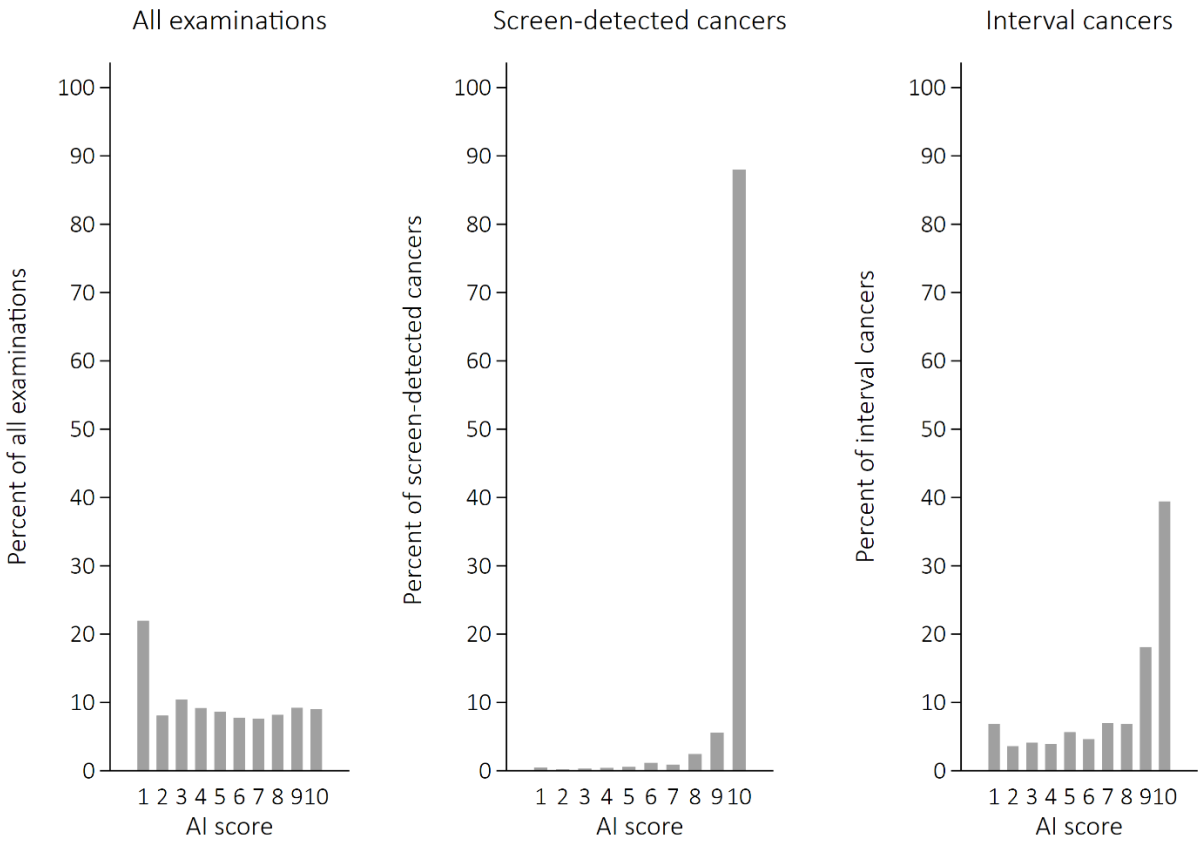

All examinations
Screen-detected cancers
Interval cancers
Percent of all examinations
Percent of screen-detected cancers
Percent of interval cancers
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10
AI score

Figure 2

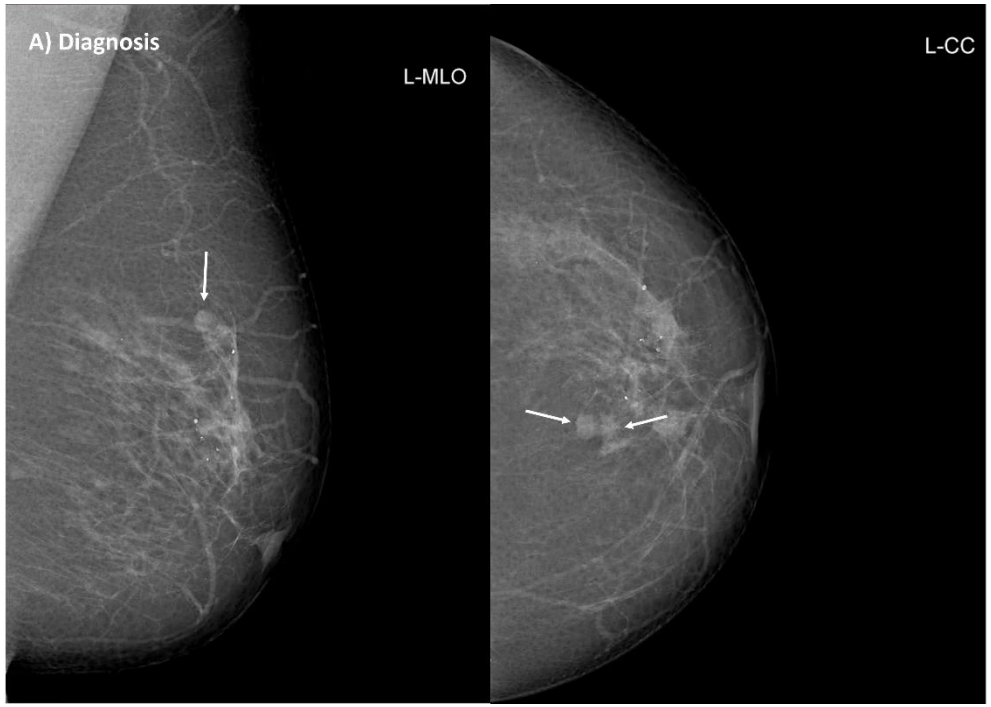


Figure 3A

Figure 3B

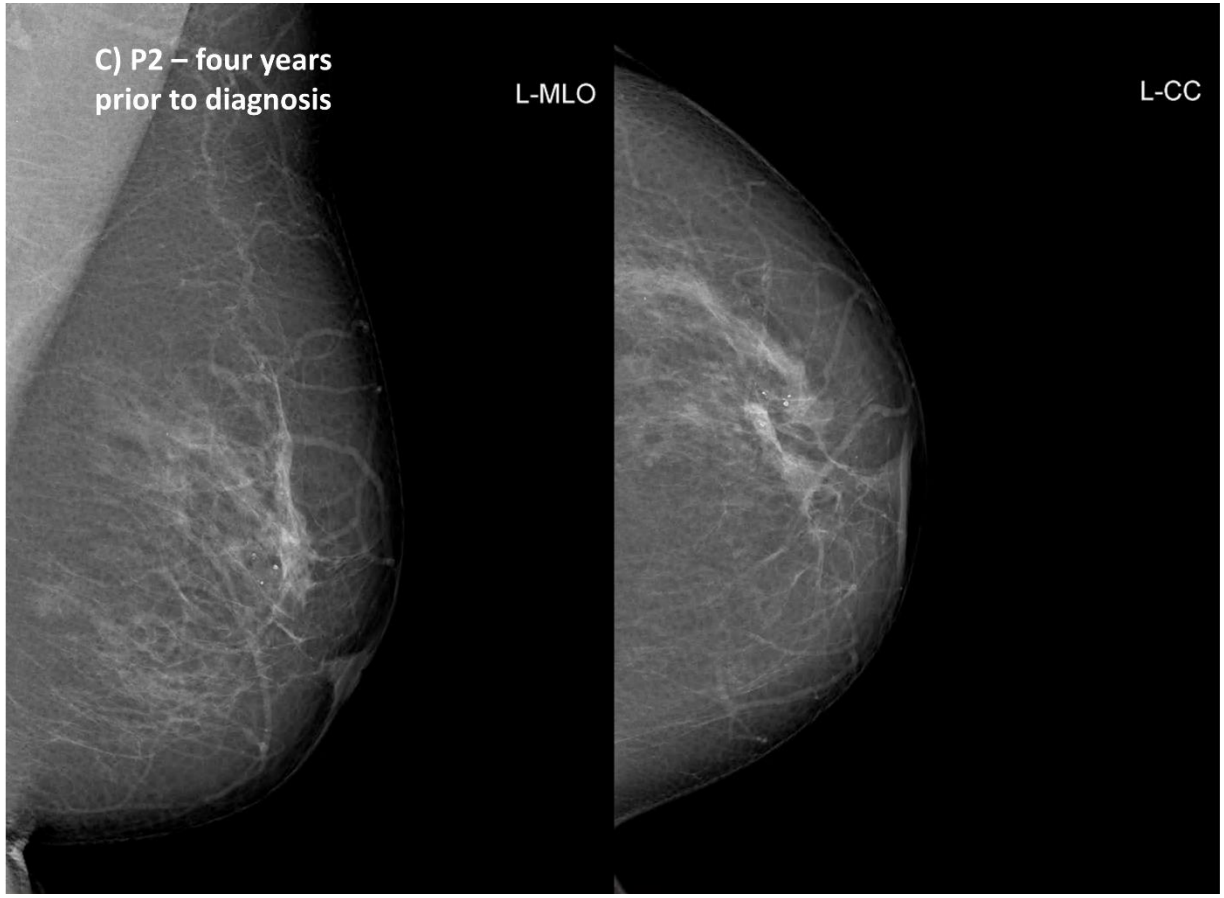
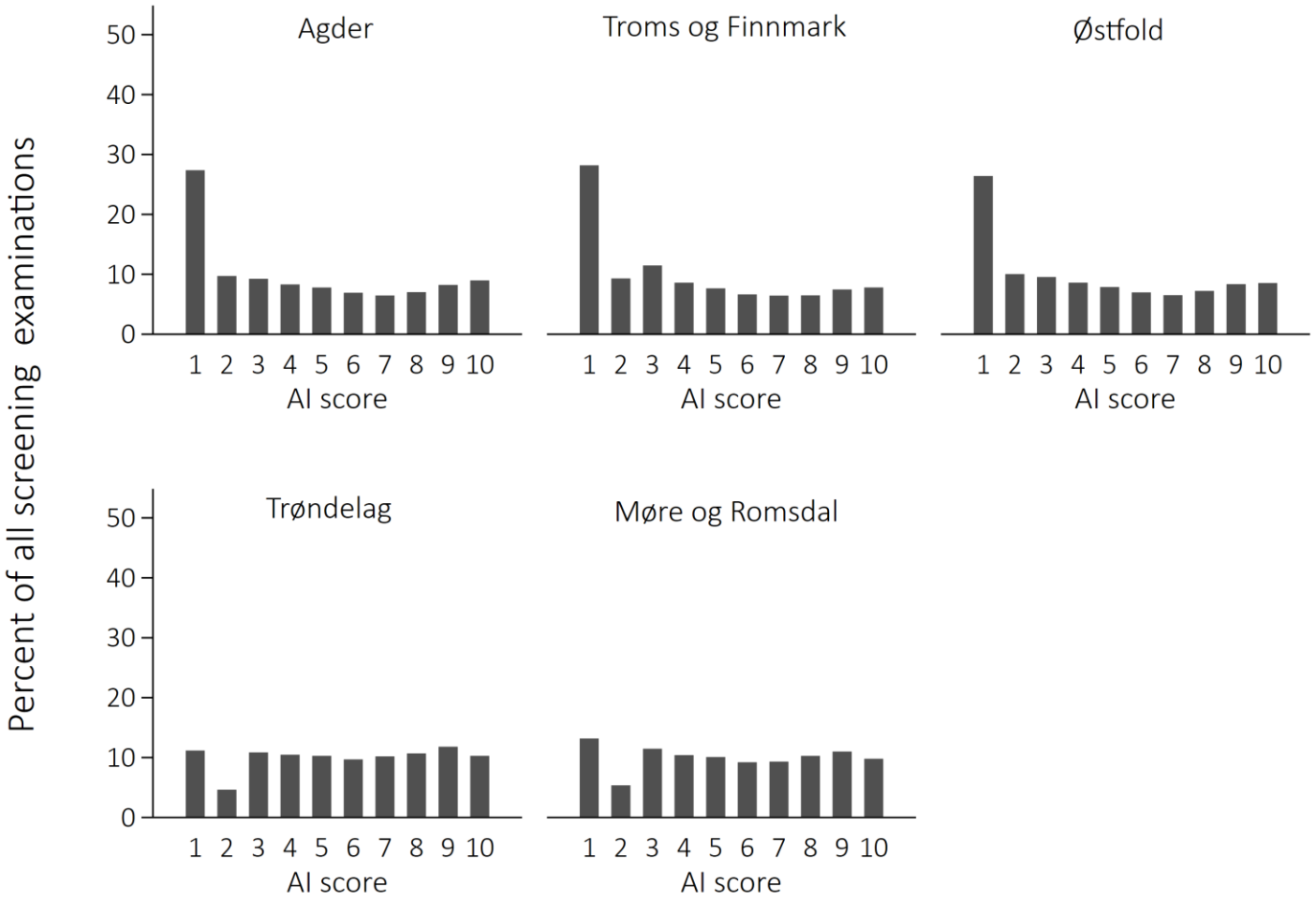


Figure 3C



Supplementary Figure 1