Assessing Breast Cancer Risk by Combining AI for Lesion Detection and Mammographic Texture

Lauritzen, Andreas D.; von Euler-Chelpin, My C.; Lynge, Elsebeth; Vejborg, Ilse; Nielsen, Mads; Karssemeijer, Nico; Lillholm, Martin

Published in:
Radiology

DOI:
10.1148/radiol.230227

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC-ND

Citation for published version (APA):
Assessing Breast Cancer Risk by Combining AI for Lesion Detection and Mammographic Texture

Andreas D. Lauritzen, PhD • My C. von Euler-Chelpin, PhD • Elsebeth Lynge, PhD • Ilse Vejborg, MD • Mads Nielsen, PhD • Nico Karssemeijer, PhD • Martin Lilholm, PhD

From the Departments of Computer Science (A.D.L., M.N., M.L.) and Public Health (M.C.v.E.C., E.L.), University of Copenhagen, Universitetsparken 1, 2100 Copenhagen Ø, Denmark; Department of Breast Examinations, Gentofte Hospital, Gentofte, Denmark (I.V.); and Department of Radiology and Nuclear Medicine, Radboud University Medical Centre and ScreenPoint Medical, Nijmegen, the Netherlands (N.K.). Received February 3, 2023; revision requested March 17; revision received June 22; accepted June 28. Address correspondence to A.D.L. (email: al@di.ku.dk).

Supported in part by Eurostars (grant E9714 IBSCREEN).

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

See also the editorial by Poynton and Slanetz in this issue.

Background: Recent mammography-based risk models can estimate short-term or long-term breast cancer risk, but whether risk assessment may improve by combining these models has not been evaluated.

Purpose: To determine whether breast cancer risk assessment improves when combining a diagnostic artificial intelligence (AI) system for lesion detection and a mammographic texture model.

Materials and Methods: This retrospective study included Danish women consecutively screened for breast cancer at mammography from November 2012 to December 2015 who had at least 5 years of follow-up data. Examinations were evaluated for short-term risk using a commercially available diagnostic AI system for lesion detection, which produced a score to indicate the probability of cancer. A mammographic texture model, trained on a separate data set, assessed textures associated with long-term cancer risk. Area under the receiver operating characteristic curve (AUC) analysis was used to evaluate both the individual and combined performance of the AI and texture models for the prediction of future cancers in women with a negative screening mammogram, including those with interval cancers diagnosed within 2 years of screening and long-term cancers diagnosed 2 years or more after screening. AUCs were compared using the DeLong test.

Results: The Danish screening cohort included 119,650 women (median age, 59 years [IQR, 53–64 years]), of whom 320 developed interval cancers and 1,401 developed long-term cancers. The combination model achieved a higher AUC for interval and long-term cancers grouped together than either the diagnostic AI (AUC, 0.73 vs 0.70; P < .001) or the texture risk (AUC, 0.73 vs 0.66; P < .001) models. The 10% of women with the highest combined risk identified by the combination model accounted for 44.1% (141 of 320) of interval cancers and 33.7% (472 of 1401) of long-term cancers.

Conclusion: Combining a diagnostic AI system and mammographic texture model resulted in improved risk assessment for interval cancers and long-term cancers and enabled identification of women at high risk.

© RSNA, 2023

Supplemental material is available for this article.
Assessing Breast Cancer Risk with AI Lesion Detection and Mammographic Texture

Abbreviations
AI = artificial intelligence, AUC = area under the receiver operating characteristic curve

Summary
The combination of a diagnostic artificial intelligence system for lesion detection and a mammographic texture model showed improved performance for mammography-based breast cancer risk assessment in a screening cohort of Danish women.

Key Results
- In this retrospective study, a mammography-based breast cancer risk model combining an artificial intelligence (AI) system for lesion detection and a mammographic texture model was trained using examination data from 39,345 Dutch women.
- In an independent screening cohort of 119,650 Danish women, the combination model improved risk segregation for women with a negative screening mammogram and at least 5 years of follow-up compared with both the AI system (area under the receiver operating characteristic curve [AUC], 0.73 vs 0.70; \( P < .001 \)) and the texture model (AUC, 0.73 vs 0.66; \( P < .001 \)).
- Women identified by the model as having the 10% highest combined risk accounted for 44.1% (141 of 320) of interval cancers and 33.7% (472 of 1401) of long-term cancers.

Materials and Methods

Deep learning is the Dense Tissue and Early Breast Neoplasm Screening (DENSE) trial, where women with high breast density receive supplemental screening (6).

Diagnostic models are trained to detect suspicious lesions and support diagnostic assessment. These models may perform as well as radiologists (14) and can detect cancer precursors to estimate short-term risk (10,12,13). Texture models are trained to learn global features in healthy tissue indicative of breast cancer susceptibility, such as breast tissue heterogeneity, density, or both, for long-term risk (13,15,16). Conflated models rely on both global and localized features to assess risk (17); however, training for both simultaneously might yield a subpar model (13). The outcome of conflating a short-term and a long-term deep learning risk model, trained separately and subsequently combined, has not previously been reported.

In this study, we evaluated whether risk assessment improved when a commercially available diagnostic artificial intelligence (AI) system for predicting short-term risk and a texture model for predicting long-term risk, trained separately, were combined into one model. We also assessed the performance of the combination model when age and percentage mammographic density were included.

Materials and Methods

The Danish Patient Safety Authority and Danish Data Inspection Agency approved this retrospective study and waived the need for informed consent (ref. 3–3013–2118). ScreenPoint Medical provided the diagnostic AI system but did not have access to the study sample.

Study Sample
The main study sample, the Danish screening cohort, included 50–70-year-old women screened biennially in the Capital Region of Denmark breast cancer screening program from November 2012 to December 2015. Mammographic screenings were performed with Mammomat Inspiration (Siemens Healthineers) systems. The earliest examination of women with valid mediolateral oblique and craniocaudal views of each breast were used. Mammograms with visible artifacts, corrupted views, or incomplete imaging data were excluded (Fig 1A).

The Dutch training data set was independently collected and used to train the texture and combination model. This sample included 50–74-year-old women screened biennially in Utrecht, the Netherlands, from August 2003 to December 2015 using Lorad Selenia (Hologic) systems.

Before 2010, Dutch women who were identified as having high breast density had four views acquired, whereas the remaining women had two views acquired. Only examinations with four valid views, including mediolateral oblique and craniocaudal views of each breast, were included. From these four-view examinations, one examination was sampled for each woman. Women who had previous breast cancer or screenings with visible artifacts, corrupted views, or incomplete imaging data were excluded (Fig 1B). Women who had biopsy or surgical clips present in the cancer contralateral breast or both breasts were excluded as clips might indicate noncancer recall or previous breast cancer, which are associated with elevated long-term risk (18).

The Danish screening cohort of this study was reported on previously (12,19,20). These prior studies concerned simulation of stratified screening, whereas the current study concerns combining risk models.

Mammographic Evaluation and Definitions

For both cohorts, examinations were read independently by two specialized breast radiologists from a team of high-volume readers with more than 6 months of experience. Reader experience is detailed in Appendix S1. Disagreements, from the need for recall to diagnostic assessment, were addressed in a consensus meeting by the two readers or a third radiologist whose evaluation was used to break the tie. A triple test— including clinical examination, US, and eventually needle biopsy— was used to determine the diagnosis.

Screen-detected cancers were defined as cancers diagnosed within 6 months after screening. Women without a screen-detected cancer who were diagnosed with cancer within 24 months after screening via clinical symptoms or events unrelated to screening were defined as having interval cancers. Women diagnosed more than 24 months after screening were designated as having long-term cancers. Women not diagnosed with cancer throughout the follow-up period were considered healthy.

All mammograms were obtained in the study period, and the diagnostic determination was based on a follow-up period of at least 5 years from the screening visit.

Diagnostic AI System for Short-term Risk

The diagnostic AI system was Transpara (version 1.7.0; ScreenPoint Medical) (14,21,22). The system’s deep learning models were trained before this study on independent mammograms from sites in Europe and the United States to
detect lesions and calcifications suspicious for breast cancer. Findings were combined into a continuous variable, called the examination score (from 0 to 10), where 0 and 10 indicates very low and high probability of cancer, respectively. The AI system was used to score screening mammograms in the Danish screening cohort.

**Mammographic Texture Model for Long-term Risk**

The texture model was developed by the authors of this article to identify mammographic texture patterns associated with long-term breast cancer risk (release 1.0; available at https://github.com/Andefar/BreastTextureModel). The texture model architecture, consisting of the deep learning encoder SE-ResNet18 followed by two fully connected layers, outputs a single risk estimate (23). Five texture models were trained separately on five folds of examinations, each fold consisting of a training and validation set, curated from the Dutch training data set. These five folds were curated by age matching each woman with breast cancer with 20 women without a breast cancer diagnosis. Matches were randomly split into five folds stratified according to the time from screening to diagnosis and age. In the five training sets, only cancer contralateral views were included to exclude diagnosed or potential malignancies. Training views were processed into six formats, each of which approximates vendor-specific processing algorithms to simulate mammograms acquired on different devices. Each texture model was trained to identify cancers until convergence, indicated by the highest area under the receiver operating characteristic curve (AUC) in the corresponding validation set. The final texture score was calculated as the average of the risk estimates from the five texture models. The training procedure is detailed in Appendix S1 and Figure S1. Women in the Danish screening cohort were scored for texture.

**Estimation of Percentage Mammographic Density**

Percentage mammographic density values were obtained using a dense tissue segmentation tool, developed by the authors of this article, trained on the independent data of 500 Dutch mammograms with the corresponding radiologist-annotated masks. The tool is described in Appendix S1.
Combining Short-term and Long-term Risk Models

Measures of risk (examination score, texture risk, age, percentage mammographic density), which we refer to as risk covariates, were combined using a three-layer neural network, called the combination model (Fig 2), to adequately capture nonlinear interactions. The combination model was trained on risk covariates derived exclusively from the Dutch training data set. To optimally learn risk-associated features, the model was trained to differentiate women at future risk for interval cancers or long-term cancers from those women without a cancer diagnosis. The combination model was trained on 80% of examinations in the Dutch training data set until convergence was indicated by the highest AUC in the validation set of the remaining 20%. The combination model’s output was referred to as the combined risk. The training procedure is detailed in Appendix S1 and Figure S2.

Two combination model variations were trained such that one included the examination score and texture risk, and another included those features in addition to age and percentage mammographic density. Both models were applied to the Danish screening cohort.

Note that examinations with surgical or biopsy clips were omitted from the texture model training to learn texture-based risk. To reinclude risk related to the presence of clips (see Appendix S1), texture scores were combined with the presence of clips, using the combination model, and together referred to as the texture risk. Clips were detected using an SE-ResNet18 model trained on independent data to detect mammographic views with clips.

Statistical Analysis

Statistical analyses were performed by one author (A.D.L.) in R (version 4.1.0; The R Foundation). Risk assessment performances were measured in the Danish screening cohort for age and percentage mammographic density combined as a baseline model, texture risk, examination score, and the two combination model variations. AUC and sensitivity are reported at 90% specificity for interval and long-term cancers as one group to adhere to previous literature (24). The DeLong method was used to calculate 95% CIs and compare AUCs between models, and \( P < .05 \) was considered indicative of a statistically significant difference (25). As subanalyses, AUCs for interval and long-term cancers, separately, and percentages of interval and long-term cancers in the 10% of women with the highest combined risk are reported. For all outcomes, screen-detected cancers were excluded as they are not related to future risk estimation.

Results

Demographics of the Study Samples

A total of 161,628 examinations were initially acquired from the Danish screening program. For 40,498 women screened twice...
in the study period, only the earliest examination was included. A total of 1480 examinations were excluded due to artifacts, corruption, or missing views (Fig 1A). Thus, 119650 women (median age, 59 years [IQR, 53–64 years]) were included in this study. Of these women, 320 of 119650 (0.3%) had interval cancers, with a median follow-up time of 17.3 months (IQR, 12.0–21.2 months), and 1401 of 119650 women (1.2%) had long-term cancers, with a median follow-up time of 37.9 months (IQR, 28.0–50.9 months) (Table 1). Throughout the 5-year follow-up period, 117030 of 119550 women (97.9%) received no breast cancer diagnosis. The median percentage mammographic density was 9.8% (IQR, 3.0%–19.5%; minimum = 0.0%, maximum = 60.1%).

From the Dutch screening program, a total of 120506 examinations were initially acquired. Of these, 81161 examinations were excluded for women who underwent screening mammography more than once or due to missing or corrupted views, artifacts, previous breast cancer, or clips in both or in the cancer contralateral breast (Fig 1B). Thus, the Dutch training data set included 39345 women (median age, 56 years [IQR, 52–63 years]), of whom 152 (0.4%) had interval cancers, with a median follow-up of 15 months (IQR, 10–19 months), and 808 (2.1%) had long-term cancers, with a median follow-up of 47 months (IQR, 34.3–50.0 months) (Table 1). No breast cancer was diagnosed in 37922 of 39345 women (96.4%) throughout the 5-year follow-up period. The median percentage mammographic density was 12.7% (IQR, 4.7%–23.7%; minimum = 0.0%, maximum = 77.0%), which was higher than that observed in the Danish screening cohort (P < .001).

**Performance of the Risk Assessment Models**

Using the Danish screening cohort, the baseline model combining age and percentage mammographic density achieved an AUC of 0.62 [95% CI: 0.60, 0.63]; P < .001. The AI system’s examination score achieved an AUC of 0.70 [95% CI: 0.69, 0.72] for interval and long-term cancers together, which was higher than that of age and percentage mammographic density (AUC, 0.62 [95% CI: 0.60, 0.63]; P < .001). The combination model with texture risk and the examination score achieved a higher AUC (0.73 [95% CI: 0.72, 0.74]) for predicting interval and long-term cancers together than either texture risk (AUC, 0.66 [95% CI: 0.65, 0.68]; P < .001) or the examination score (AUC, 0.70 [95% CI: 0.69, 0.72]; P < .001) alone.

With age and percent mammographic density additionally included in the combination model, the AUC was 0.72 (95% CI: 0.71, 0.74) for predicting interval and long-term cancers together, which was not different from the combination model without these patient characteristics (AUC, 0.73 [95% CI: 0.72, 0.74]; P = .24).

**Identification of Women at High Risk Using the Combination Model**

At 90% specificity, the baseline model combining age and percentage mammographic density had a sensitivity of 17.1% (294 of 1721 patients [95% CI: 15.3, 18.9]) for interval and long-term cancers together (Table 2). Texture risk had a sensitivity of 24.5% (421 of 1721 patients [95% CI: 22.4, 26.6]) for interval and long-term cancers together, which was higher than the 17.1% sensitivity for age and percentage mammographic density (P < .001). The examination score had a sensitivity of 33.9% (584 of 1721 patients [95% CI: 31.7, 36.2]) for interval and long-term cancers together, which was higher than the 17.1% sensitivity for age and percentage mammographic density (P < .001).

The combination model with texture risk and the examination score achieved a sensitivity of 36.5% (629 of 1721 patients [95% CI: 34.3, 38.9]) for interval and long-term cancers together, which was higher than that of texture risk alone (sensitivity, 24.5% [95% CI: 22.4, 26.6]; P < .001) and the examination score alone (sensitivity, 33.9% [95% CI: 31.7, 36.2]; P < .01). The combination

### Table 2: Model Performance for Breast Cancer Risk in the Danish Screening Cohort

<table>
<thead>
<tr>
<th>Model</th>
<th>Combined Interval and Long-term Cancers AUC</th>
<th>Interval Cancers AUC</th>
<th>Long-term Cancers AUC</th>
<th>Sensitivity at 90% Specificity for Combined Interval and Long-term Cancers*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient data†, texture risk</td>
<td>0.62 (0.60, 0.63)</td>
<td>0.65 (0.62, 0.68)</td>
<td>0.61 (0.59, 0.62)</td>
<td>17.1 [294/1721] (15.3, 18.9)</td>
</tr>
<tr>
<td>Examination score</td>
<td>0.70 (0.69, 0.72)</td>
<td>0.75 (0.72, 0.78)</td>
<td>0.69 (0.68, 0.71)</td>
<td>24.5 [421/1721] (22.4, 26.6)</td>
</tr>
<tr>
<td>Texture risk and examination score</td>
<td>0.73 (0.72, 0.74)</td>
<td>0.78 (0.75, 0.80)</td>
<td>0.72 (0.70, 0.73)</td>
<td>36.5 [629/1721] (34.3, 38.9)</td>
</tr>
<tr>
<td>Patient data†, texture risk, and examination score</td>
<td>0.72 (0.71, 0.74)</td>
<td>0.77 (0.74, 0.80)</td>
<td>0.71 (0.70, 0.73)</td>
<td>36.1 [621/1721] (33.8, 38.4)</td>
</tr>
</tbody>
</table>

Note.—Data in parentheses are 95% CIs. Risk estimation for age and percentage mammographic density as a baseline model, texture risk, the artificial intelligence system’s examination score, and combination models with different combinations of risk covariates were assessed in the Danish screening cohort using AUC analysis. All patients with screen-detected cancers were excluded from this analysis, and patients with interval (n = 320) and long-term (n = 1401) cancers were included. AUC = area under the receiver operating characteristic curve.

* Data are percentages, with numbers of patients in brackets.
† Patient data include age and percentage mammographic density.
model that additionally included age and percentage mammographic density had a sensitivity of 36.1% (621 of 1721 patients [95% CI: 33.8, 38.4]), which was not different from the combination model without these patient characteristics (sensitivity, 36.5% [95% CI: 34.3, 38.9]; \( P = .66 \)).

Women identified by the texture risk and the examination score combination model as having the 10% highest combined risk (\( n = 11875 \)) accounted for 35.7% (615 of 1721) of interval cancers and long-term cancers together, 44.1% (141 of 320) of interval cancers alone, and 33.7% (472 of 1401) of long-term cancers alone (Figs 3–6).

Discussion

Current mammography-based breast cancer risk models can estimate short-term and long-term risk. We determined whether risk assessment improved when combining a commercially available diagnostic artificial intelligence (AI) system for lesion detection to estimate short-term risk and a mammographic texture model for long-term risk, trained separately and combined subsequently, which has not been previously reported. In a Danish screening cohort of 119,650 women, the risk assessment of the combination model was improved for interval and long-term cancers together, with an area under the receiver operating characteristic curve (AUC) of 0.73 (95% CI: 0.72, 0.74), when compared with either the AI system’s examination score (AUC, 0.70 [95% CI: 0.69, 0.72]; \( P < .001 \)) or texture risk (AUC, 0.66 [95% CI: 0.65, 0.68]; \( P < .001 \)). The combination model identified women with the 10% highest combined risk, accounting for 44.1% (141 of 320) of interval cancers and 33.7% (472 of 1401) of long-term cancers. At 90% specificity, the combination model predicted future cancers with a sensitivity of 36.5% (629 of 1721 patients [95% CI: 34.3, 38.9]). The combination model that additionally included age and percentage mammographic density yielded an AUC of 0.72 (95% CI: 0.71, 0.74), which was not different from the combination model without these patient characteristics (AUC, 0.73 [95% CI: 0.72, 0.74]; \( P = .24 \)).

McCarthy et al (26) investigated the Gail and Tyrer-Cuzick models in a sample of 35,921 women with 6 years of follow-up. The Gail and Tyrer-Cuzick models achieved AUCs of 0.64 (95%
CI: 0.61, 0.65) and 0.62 (95% CI: 0.60, 0.64), respectively. The combination model in our study, using the examination score and texture risk, achieved a higher AUC (0.73 [95% CI: 0.72, 0.74]) than the Gail or Tyrer-Cuzick models, within a comparable follow-up period. In a study by Yang et al (9), breast cancer risk was assessed in 5693 women with a 5-year follow-up period by using The Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm, which yielded an AUC of 0.70. The combination model in our study had a slightly higher AUC of 0.73, which is of note given that the model developed by Yang et al included extensive patient data and our model was built using only examination-based information.

Mirai is another model for breast cancer risk estimation at multiple time points from screening to up to 5 years (27). Mirai was validated in seven data sets from different countries that included only women with negative screening findings and achieved a median AUC of 0.71 for cancers diagnosed within 5 years from screening. This AUC is approximately equal to the AUC of 0.73 (95% CI: 0.72, 0.74) for our combination model; however, in that study, 42%–82% of cancers were diagnosed in the 1st year and 54%–92% of cancers were diagnosed in the 2nd year (28). This overrepresentation of cancers diagnosed in the 1st 2 years may not be adequate to measure risk assessment performance and may lead to higher AUCs. Unlike in our study, no separate AUC value was reported for long-term cancers alone.

Our study had limitations. First, the study sample was from a single institution and, therefore, our findings require validation in external data sets from institutions with different screening intervals (1 or 3 years), single reading, and with more demographic diversity to ensure generalization to screening in other countries. Second, the training and testing data sets had different median percentages of mammographic density (12.7% vs 9.8%, \( P < .001 \)), likely due to differences in data curation between the two samples regarding women with high breast density. Specifically, in the Dutch sample, there is an overrepresentation of women with high breast density as we only sampled women with four views. This may impact the calibration of risk estimates in the combination model. Third, our study is limited by the lack of longer follow-up. The combination model requires validation in samples with follow-up of 10 years or more from screening. Fourth, a few examinations were excluded due to artifacts (pacemakers, etc) or incomplete imaging data and could not be assessed for risk. The nature of these examinations should be determined, and the implication of the exclusion assessed.

Finally, the density bias between women who developed cancer and healthy women was not corrected for in-texture model training. Therefore, the texture score will, indirectly, include breast density as a risk factor and approximate density in the combined risk models, so that density does not contribute further.

**Figure 5:** Full-field digital mammograms (left mediolateral oblique view) in a 60-year-old woman with an interval cancer show (A) the screening mammogram at the area of interest and (B) the same area of interest on a clinical mammogram obtained during diagnostic testing a year later, whereby the blue circle indicates the diagnosed lesion. The woman was not recalled as a result of screening. The screening mammogram (A) had a very high combined risk score (highest 99%) as determined by the combination model with texture risk and the examination score.

**Figure 6:** Full-field digital mammograms (right mediolateral oblique view) in a 59-year-old woman with a long-term cancer show (A) the screening mammogram at the area of interest and (B) the same area of interest on a clinical mammogram obtained during diagnostic testing 3 years later, whereby the blue circle indicates the diagnosed lesion. The woman was not recalled as a result of screening. The screening mammogram (A) had a very high combined risk score (highest 99.9%) as determined by the combination model with texture risk and the examination score. The four objects are surgical clips left in the breast from previous breast-conserving surgery.
In conclusion, our findings indicate that mammography-based breast cancer risk assessment is improved when combining an artificial intelligence system for lesion detection and a mammographic texture model. In early experiments, logistic regression analysis yielded significantly worse performance than the neural network. Thus, future work should focus on investigating the combination model architecture. We should further determine whether the combination model adapts sufficiently to other mammographic devices and institutions. Additional research should focus on translating combined risk to lifetime or absolute risk for comparison with traditional models.

Author contributions: Guarantors of integrity of entire study, A.D.L., M.L.; 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, A.D.L., I.V., M.N.; experimental studies, A.D.L., M.N., N.K., M.L.; statistical analysis, A.D.L., M.C.V.E.C., M.N., M.L.; and manuscript editing, all authors

Disclosures of conflicts of interest: A.D.L. The artificial intelligence lesion detection software was provided by ScreenPoint Medical. M.C.V.E.C. No relevant relationships. E.L. No relevant relationships. I.V. No relevant relationships. M.N. Stockholder in Biomediq A/S. N.K. Patents planned, issued, or pending with ScreenPoint Medical; stockholder in ScreenPoint Medical, Volpara HealthCare, and QView Medical. M.L. Board member for Biomediq A/S; stockholder in Biomediq A/S and Cerebrit A/S.

References