An Artificial Intelligence–based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload

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

Published in:
Radiology

DOI:
10.1148/radiol.210948

Publication date:
2022

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

Document license:
CC BY-NC-ND

Citation for published version (APA):
An Artificial Intelligence–based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload

Andreas D. Lauritzen, MSc • Alejandro Rodriguez-Ruiz, PhD • My Catarina von Euler-Chelpin, PhD • Elsebeth Lyng, PhD • Ilse Vejborg, MD • Mads Nielsen, PhD • Nico Karssemeijer, PhD • Martin Lillholm, PhD

From the Department 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, Denmark; ScreenPoint Medical, Nijmegen, the Netherlands (A.R.R., N.K.); Centre for Epidemiological Research, Nykøbing Falster Hospital, Nykøbing, Denmark (E.L.); Department of Radiology, Copenhagen University Hospital Herlev/Gentofte, Copenhagen, Denmark (L.V.); and Department of Medical Imaging, Radboud University Medical Centre, Nijmegen, the Netherlands (N.K.). Received May 28, 2021; revision requested July 15; revision received February 10, 2022; accepted February 16. Address correspondence to M.L. (email: groene@di.ku.dk).

Supported in part by Eurostars (grant E9714 IBSCREEN).

Background: Developments in artificial intelligence (AI) systems to assist radiologists in reading mammograms could improve breast cancer screening efficiency.

Purpose: To investigate whether an AI system could detect normal, moderate-risk, and suspicious mammograms in a screening sample to safely reduce radiologist workload and evaluate across Breast Imaging Reporting and Data System (BI-RADS) densities.

Materials and Methods: This retrospective simulation study analyzed mammographic examination data consecutively collected from January 2014 to December 2015 in the Danish Capital Region breast cancer screening program. All mammograms were scored from 0 to 10, representing the risk of malignancy, using an AI tool. During simulation, normal mammograms (score < 5) would be excluded from radiologist reading and suspicious mammograms (score ≥ recall threshold [RT]) would be recalled. Two radiologists read the remaining mammograms. The RT was fitted using another independent cohort (same institution) by matching to the radiologist sensitivity. This protocol was further applied to each BI-RADS density. Screening outcomes were measured using the sensitivity, specificity, workload, and false-positive rate. The AI-based screening was tested for noninferiority sensitivity compared with radiologist screening using the Farrington-Manning test. Specificities were compared using the McNemar test.

Results: The study sample comprised 114,421 screenings for breast cancer in 114,421 women, resulting in 791 screen-detected, 327 interval, and 1473 long-term cancers and 2107 false-positive screenings. The mean age of the women was 59 years ± 6 (SD). The AI-based screening sensitivity was 69.7% (779 of 1118; 95% CI: 66.9, 72.4) and was noninferior (P = .02) to the radiologist screening sensitivity of 70.8% (791 of 1118; 95% CI: 68.0, 73.5). The AI-based screening specificity was 98.6% (111,725 of 113,303; 95% CI: 98.5, 98.7), which was higher (P < .001) than the radiologist specificity of 98.1% (111,196 of 113,303; 95% CI: 98.1, 98.2). The radiologist workload was reduced by 62.6% (71,585 of 114,421), and 25.1% (529 of 2107) of false-positive screenings were avoided. Screening results were consistent across BI-RADS densities, although not significantly so for sensitivity.

Conclusion: Artificial intelligence (AI)–based screening could detect normal, moderate-risk, and suspicious mammograms in a breast cancer screening program, which may reduce the radiologist workload. AI-based screening performed consistently across breast densities.

© RSNA, 2022

Online supplemental material is available for this article.

Population-based mammographic screening reduces breast cancer mortality by detecting early signs of disease (1–3). Such screening programs generate a substantial number of mammograms that require reading by radiologists. Considering the low prevalence of breast cancers in a screening cohort, most screening mammograms are normal with no visible malignancy. Radiologist readings in low-prevalence conditions are more likely to miss disease signs, producing false-negative test results (4). Conversely, recalling women for further examination exposes them to potentially unnecessary discomfort and anxiety (5) and increases workload in the clinic. Furthermore, the need for specialized radiologists is growing due to an increase in breast cancer incidence and widely implemented screening programs (6,7), although shortages of radiologists have been reported worldwide (8,9).

In 2007, Fenton et al (10) found that in 43 clinics in the United States, the use of computer-aided detection systems to assist radiologists in reading screening mammograms yielded a higher recall rate but did not improve the detection of invasive breast cancer. However, current computer-aided detection systems are based on artificial intelligence (AI) and now achieve cancer detection accuracies, in stand-alone setups, equal to the average accuracy of multiple (up to 101) radiologists (11–13). Such systems may also improve the performance and productivity of radiologists when used to support decisions (14–16).
An Artificial Intelligence–based Mammography Screening Protocol for Breast Cancer

Abbreviations
AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, BI-RADS = Breast Imaging Reporting and Data System, RT = recall threshold

Summary
In a simulation study of breast cancer screenings in more than 114,000 women, an artificial intelligence–based mammography screening protocol had similar sensitivity to that of radiologist screening and may reduce the radiologist workload.

Key Results
- In a retrospective simulation study of 114,421 women who underwent screening mammography, artificial intelligence (AI)–based screening had a comparable sensitivity to (69.7% vs 70.8%, respectively; P = .02) and higher specificity than (98.6% vs 98.1%, P < .001) the radiologists.
- Mammograms from 71,585 of 114,421 screenings were read only by the AI system and the radiologist workload was reduced by 63%.
- With AI-based screening, 529 of 2107 false-positive screenings were avoided, a 25% decrease.

These circumstances motivate investigations into making population-based screening programs more effective using AI and potentially improving screening outcomes, while maintaining an equally high safety level for screened women.

We aimed to evaluate whether implementing a reading protocol based on an AI system in the clinic could improve screening outcomes and reduce the number of mammograms needing to be read by radiologists. We retrospectively examined whether an AI system could detect normal, moderate-risk, and suspicious mammograms. With this categorization, an exclusion strategy to relieve radiologists of reading normal and suspicious mammograms failed to be processed by the AI system.

Materials and Methods
The Danish Data Inspection Agency and Danish Patient Safety Authority approved this retrospective study with the use of relevant data and waived the need for informed consent (reference no. 3–3013–2118). ScreenPoint Medical provided the AI software, but no employee had access to data from our study.

Study Sample
Biennial breast cancer screening using Mammomat Inspiration systems (Siemens Healthineers) is offered to asymptomatic women 50–69 years of age in the Capital Region of Denmark.

For our study, two screening samples were used. The development sample included women consecutively screened from November 2012 to December 2013. All 54,997 women in the development sample were previously reported (17,18). Prior studies investigated screening according to mammographic density and texture, whereas our study reports on a reading protocol. The testing sample included women consecutively screened from January 2014 to December 2015.

During retrospective collection of raw image data, women with incomplete image data were excluded (Fig 1). It could be assumed that for AI-based screening implemented on site in clinics, no image data would be lost. None of the available mammograms failed to be processed by the AI system.

Imaging Protocol and Diagnosis
The screening workflow is depicted in Figure 2A. A radiographer captures four full-field digital mammograms, including a craniocaudal and mediolateral oblique view for each breast. Usually, within 10 days, every mammogram is read independently by two specialized radiologists. From 2012 to 2015, seven radiologists were employed with an average of 7.1 years of high-volume experience. See Appendix E1 (online) for details. If disagreement occurs on a recall decision, a consensus meeting is held with a third radiologist. Women with suspicious findings (positive screening result) are invited to a triple test consisting of a clinical examination, imaging with mammography and US, a biopsy, if relevant, and/or further breast imaging (eg, MRI). Women without suspicious findings (negative screening result) are reinvited for screening 2 years later. Radiologists assign a breast density to each woman using the fourth edition of BI-RADS (19). In cases of disagreement, the highest BI-RADS density is used.

Radiologist recall decisions, diagnostic assessment outcomes, and BI-RADS density data used in our simulation study were extracted from the original screening reports.

Diagnosis data were retrieved from the Danish Pathology Register, which is linked with the screening program, in terms of biopsies and results, by personal identification numbers. Women with a positive screening who were diagnosed with breast cancer or ductal carcinoma in situ within 6 months of being screened were labeled as having screen-detected cancers. Women with a positive screening but without a screen-detected cancer had false-positive results.
Cancers in women with a negative screening or a negative recall examination diagnosed within 24 months after screening (or before the next screening) were labeled as interval cancers. Interval cancers were discovered outside screening (e.g., self-palpation or general practitioner consultation) and confirmed with a triple test. Long-term cancer information was collected for breast cancers diagnosed 2–5 years after screening.

**AI System**

The AI system was Transpara (version 1.7.0, ScreenPoint Medical) (11,14,20–24). This commercially available system received Conformité Européenne (CE) mark approval and was cleared by the U.S. Food and Drug Administration. It uses deep convolutional neural networks, trained on more than 1 million mammograms from sites in Europe and the United States, to detect lesions (soft tissue and calcifications) suspicious for breast cancer on full-field digital mammograms captured on machines from different vendors. Analysis results comprise an examination score ranging from 0 to 10, indicating risk of the presence of visible cancer, calibrated such that 10% of a population falls within each of the 10 categories. Transpara was used with default settings, and our study data were completely independent of the development of the AI system. Information on image processing is in Appendix E2 (online).

**Simulation of AI-based Screening**

In the default screening workflow (radiologist screening), depicted in Figure 2A, two radiologists read all mammograms without any interaction with the AI system.

To establish an AI system baseline, we initially investigated AI-only screening where the AI system substituted for both readers.

In the main retrospective simulation study of AI-based screening, depicted in Figure 2B, mammograms were processed by the AI system yielding an examination score. Two thresholds were used to categorize normal, moderate-risk, and suspicious mammograms. The exclusion threshold of 5 was chosen prior to analysis, meaning that approximately 50% would be categorized as normal, in accordance with the literature (23,25–28). The recall threshold (RT) was used to determine when a mammogram was categorized as suspicious.
The categories and actions were as follows. An examination score less than 5 was categorized as normal and was excluded from reading by a radiologist (negative screening result). An examination score of 5 or greater and less than or equal to the RT was categorized as moderate risk and was read by two radiologists (the radiologist recall decisions were extracted from the original screening reports). A mammogram with an examination score greater than the RT was categorized as suspicious and was excluded from reading by a radiologist; the woman was recalled directly.

The RT was derived by applying AI-based screening to the development sample and was fitted such that the number of missed screen-detected cancers (by AI) equaled the number of suspicious screening examinations diagnosed later as interval cancer. Consequently, the radiologist and AI-based screenings were matched by sensitivity in the development sample. The AI-based screening was then validated in the testing sample with the fitted RT.

The AI-based and radiologist screening performances were compared by using a series of metrics, collectively referred to as the screening outcome: sensitivity, specificity, workload, and false-positive rate.

Women with screen-detected or interval cancer constituted the group with breast cancer. Interval cancers were assumed to be false-negative findings. Workload reduction was the percentage of mammograms read by the AI system only, corresponding to normal or suspicious mammograms (radiologists would read moderate-risk mammograms only).

AI-based and radiologist screening were additionally evaluated for the four BI-RADS densities separately.

### Statistical Analysis

Statistical analyses (A.D.L.) were conducted in R (version 4.1.0, The R Foundation). CIs were computed at a level of 95% using 1000 bootstrap samples. Detection performance was measured using the area under the receiver operating characteristic curve (AUC). The DeLong method was used to determine CIs and significant differences between AUCs (29). The method of David Collett was used for CIs for sensitivities and specificities (30). The McNemar exact test was used to test for difference in specificities, with a significance level of \( \alpha = 0.05 \) (31). To test non-inferiority for sensitivities, a one-sided Farrington-Manning test was used with the null hypothesis \( H_0: S_R - S_{AI} \leq \delta \), where \( S_R \) is the radiologist screening sensitivity and \( S_{AI} \) the AI-based screening sensitivity. \( H_0 \) was tested at a significance level of \( \alpha = 0.05 \). If \( H_0 \) was rejected, the alternative hypothesis, \( H_1: S_R - S_{AI} < \delta \), was accepted. The inferiority margin, \( \delta \), was set to 0.05 in accordance with clinicians and recent literature (24,25,32).

### Results

The development sample was used only to fit the RT. The following demographics and results are based on the analysis of the testing sample alone, and corresponding results for the development sample are in Tables E1–E4 (online).

### Study Sample Characteristics

The testing sample comprised 118039 women who underwent screening examinations, of whom 3618 (3%) were excluded; thus, 114421 women with a mean age of 59 years ± 6 (SD) were included. Of the 114421 women, 791 had screen-detected cancers, 327 had interval cancers, 1473 had long-term cancers, and 111830 were healthy in the 2-year follow-up period; 2107 women had false-positive screenings. The recall rate in the development sample was 3.18% (1717 of 53 951), which was higher \( (P < .001) \) than that of the testing sample of 2.53% (2898 of 114421). The false-positive rate in the development sample was 2.41% (1299 of 53951), which was higher \( (P < .001) \) than that
of the testing sample of 1.84% (2107 of 114,421). Demographics are presented in Table 1 for the testing sample and Table E1 (online) for the development sample.

### Cancer Detection Performance of a Stand-Alone AI System

The results of using the AI system in a stand-alone setup for breast cancer diagnosis are presented in Table 2. Detection performance was measured using the AUC for screen-detected, interval, and long-term cancers using the examination score. The AI system achieved an AUC of 0.97 (95% CI: 0.97, 0.98) for screen-detected cancers, 0.74 (95% CI: 0.71, 0.77) for interval cancers, and 0.68 (95% CI: 0.67, 0.70) for long-term cancers. There was no evidence of a difference in AUC across age groups (for all possible pairs of age groups, \( P < .05 \)). Corresponding results in the development sample are shown in Table E2 (online).

For screen-detected, interval, and long-term cancers individually, the AUC remained relatively stable across BI-RADS densities. When segregating women with any breast cancer type from healthy women, the performance was reduced with increasing BI-RADS density; the AUC for women with BI-RADS densities 1 and 2 was higher (\( P = .001 \)) than that for women with BI-RADS densities 3 and 4.

Figure 3A depicts the distribution of examination scores of healthy women in the testing sample. Figure 3B depicts the distribution of examination scores for screen-detected cancers in the testing sample and shows that 89.8% of women (710 of 791) had an examination score of 10 and 94.6% of women (748 of 791) had an examination score of 9 or 10.

When simulating AI-only screening, the sensitivity was matched to the radiologist screening sensitivity in the development sample. The workload reduction was 100%, as both radiologists’ recall decisions were substituted. The corresponding specificity in the testing sample was 94.9% (107,478 of 113,303; 95% CI: 94.7, 95.0), which was lower (\( P < .001 \)) than the radiologist specificity of 98.1% (111,196 of 113,303; 95% CI: 98.1, 98.2). The number of false-positive screenings increased by 276.5% (5825 of 2107) compared with radiologist screening.

### Table 2: Cancer Detection Performance of the AI System in the Testing Sample

<table>
<thead>
<tr>
<th>Variable</th>
<th>All Cancers</th>
<th>Screen-detected and Interval Cancers</th>
<th>Screen-detected Cancer</th>
<th>Interval Cancer</th>
<th>Long-term Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of women*</td>
<td>114,421</td>
<td>114,421 (2,591/111,830)</td>
<td>114,421 (1,118/113,303)</td>
<td>114,421 (791/113,630)</td>
<td>114,421 (1,473/112,948)</td>
</tr>
<tr>
<td>Sample</td>
<td>0.78 (0.77, 0.79)</td>
<td>0.91 (0.89, 0.92)</td>
<td>0.97 (0.97, 0.98)</td>
<td>0.74 (0.71, 0.77)</td>
<td>0.68 (0.67, 0.70)</td>
</tr>
<tr>
<td>Age range (y)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50–54</td>
<td>0.78 (0.76, 0.80)</td>
<td>0.91 (0.88, 0.93)</td>
<td>0.97 (0.96, 0.98)</td>
<td>0.71 (0.64, 0.78)</td>
<td>0.68 (0.65, 0.71)</td>
</tr>
<tr>
<td>55–59</td>
<td>0.76 (0.73, 0.78)</td>
<td>0.90 (0.87, 0.92)</td>
<td>0.97 (0.95, 0.98)</td>
<td>0.76 (0.70, 0.82)</td>
<td>0.67 (0.64, 0.70)</td>
</tr>
<tr>
<td>60–64</td>
<td>0.78 (0.76, 0.80)</td>
<td>0.91 (0.89, 0.93)</td>
<td>0.97 (0.96, 0.98)</td>
<td>0.72 (0.66, 0.79)</td>
<td>0.68 (0.65, 0.71)</td>
</tr>
<tr>
<td>65–69</td>
<td>0.80 (0.78, 0.81)</td>
<td>0.91 (0.89, 0.93)</td>
<td>0.97 (0.96, 0.98)</td>
<td>0.74 (0.69, 0.79)</td>
<td>0.69 (0.66, 0.72)</td>
</tr>
<tr>
<td>≥70</td>
<td>0.85 (0.74, 0.96)</td>
<td>0.93 (0.84, 1.00)</td>
<td>0.95 (0.85, 1.00)</td>
<td>0.85 (0.72, 0.98)</td>
<td>0.75 (0.53, 0.96)</td>
</tr>
<tr>
<td>BI-RADS density</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.79 (0.77, 0.81)</td>
<td>0.93 (0.91, 0.95)</td>
<td>0.97 (0.96, 0.98)</td>
<td>0.77 (0.71, 0.83)</td>
<td>0.69 (0.66, 0.71)</td>
</tr>
<tr>
<td>2</td>
<td>0.78 (0.77, 0.80)</td>
<td>0.90 (0.88, 0.91)</td>
<td>0.97 (0.96, 0.98)</td>
<td>0.71 (0.66, 0.75)</td>
<td>0.68 (0.65, 0.70)</td>
</tr>
<tr>
<td>3</td>
<td>0.75 (0.73, 0.77)</td>
<td>0.89 (0.87, 0.92)</td>
<td>0.96 (0.95, 0.98)</td>
<td>0.72 (0.66, 0.78)</td>
<td>0.66 (0.63, 0.69)</td>
</tr>
<tr>
<td>4</td>
<td>0.75 (0.71, 0.80)</td>
<td>0.85 (0.79, 0.92)</td>
<td>0.97 (0.95, 0.99)</td>
<td>0.74 (0.62, 0.85)</td>
<td>0.69 (0.64, 0.75)</td>
</tr>
</tbody>
</table>

Note.—Except where indicated, data are areas under the receiver operating characteristic curve, with 95% CIs in parentheses. AI = artificial intelligence, BI-RADS = Breast Imaging Reporting and Data System.

* Data in parentheses are the number of positive screening examinations/number of negative screening examinations.
Applying AI-based screening on the testing sample, using the thresholds 5 and 9.989 for exclusion and recall, respectively, 71,499 screening examinations would have been labeled normal, 42,836 would have been labeled as moderate risk and to be read by radiologists, and 86 would have been suspicious and the women would have been recalled automatically. Of these 86 women, 70 had screen-detected cancers, none had interval cancers, two had long-term cancers, and 14 were cancer free; 1.5% (12 of 791) of screen-detected cancers were categorized as normal and, therefore, would have been missed. Categorizations are shown in Table 3 for the testing sample and Table E3 (online) for the development sample. Example mammograms are shown in Figure 4.

The measured screening outcomes of the radiologist and AI-based screening in the testing sample are presented in Table 4. In the radiologist screening, radiologists read all mammograms without interference from the AI system and achieved a sensitivity of 70.8% (791 of 1118; 95% CI: 68.0, 73.5) and a specificity of 98.1% (111,196 of 113,303; 95% CI: 98.1, 98.2). Simulating AI-based screening on the testing sample achieved a sensitivity of 69.7% (779 of 1118; 95% CI: 66.9, 72.4) and a specificity of 98.6% (98.5, 98.7). A total of 529 false-positive screenings may have been avoided, corresponding to a 25.1% (529 of 2107) reduction from the number of false-positive screenings with radiologist screening. Radiologists would have avoided reading images from 71,585 screenings due to the exclusion of normal or suspicious mammograms, which corresponds to a 62.6% (71,585 of 114,421) workload reduction. AI-based screening sensitivity was noninferior to radiologist screening (P = .02). The AI-based screening specificity was higher than that of the radiologist screening.

### Table 3: AI System Examination Score Characteristics of the Testing Sample

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sample</th>
<th>Screen-detected Cancer</th>
<th>Interval Cancer</th>
<th>Long-term Cancer</th>
<th>No Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of examinations</td>
<td>114,421 (100)</td>
<td>791 (1)</td>
<td>327 (0)</td>
<td>1473 (1)</td>
<td>111,830 (98)</td>
</tr>
<tr>
<td>Mean examination score*</td>
<td>4.16 (2.95)</td>
<td>9.60 (1.04)</td>
<td>6.71 (3.02)</td>
<td>6.12 (3.13)</td>
<td>4.08 (2.90)</td>
</tr>
<tr>
<td>Median examination score†</td>
<td>3.67 (1.46–6.61)</td>
<td>9.94 (9.78–9.97)</td>
<td>7.67 (4.31–9.51)</td>
<td>6.77 (3.30–9.09)</td>
<td>3.58 (1.44–6.50)</td>
</tr>
</tbody>
</table>

#### Category

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Moderate Risk</th>
<th>Suspicious</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of examinations</td>
<td>71,499 (62)</td>
<td>42,836 (37)</td>
<td>86 (0)</td>
</tr>
<tr>
<td>Mean examination score</td>
<td>4.16 (2.95)</td>
<td>9.60 (1.04)</td>
<td>9.94 (9.78–9.97)</td>
</tr>
<tr>
<td>Median examination score</td>
<td>3.67 (1.46–6.61)</td>
<td>9.94 (9.78–9.97)</td>
<td>7.67 (4.31–9.51)</td>
</tr>
</tbody>
</table>

Note.—Except where indicated, data are numbers of examinations, with percentages in parentheses. AI = artificial intelligence.

* Data are means, with SDs in parentheses.
† Data are medians, with IQRs in parentheses.

### Table 4: AI–based and Radiologist Screening Outcomes in the Testing Sample

<table>
<thead>
<tr>
<th>Variable</th>
<th>Radiologist Screening</th>
<th>AI-based Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity (%)</td>
<td>70.8 (68.0, 73.5)</td>
<td>69.7 (66.9, 72.4)</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>98.1 (98.1, 98.2)</td>
<td>98.6 (98.5, 98.7)</td>
</tr>
<tr>
<td>No. of screenings</td>
<td>114,421</td>
<td>114,421</td>
</tr>
<tr>
<td>True positive</td>
<td>791</td>
<td>779</td>
</tr>
<tr>
<td>True negative</td>
<td>111,196</td>
<td>111,725</td>
</tr>
<tr>
<td>False positive</td>
<td>2107</td>
<td>1578</td>
</tr>
<tr>
<td>False negative</td>
<td>327</td>
<td>339</td>
</tr>
<tr>
<td>Workload reduction (%)</td>
<td>…</td>
<td>62.6</td>
</tr>
</tbody>
</table>

Note.—Numbers in parentheses are 95% CIs. AI = artificial intelligence.

### Evaluation of AI-based Screening

In the development sample, the RT was fitted to be 9.989, such that the number of screen-detected cancers missed in AI-based screening equaled the number of women with interval cancers that would have been additionally recalled. Only one screen-detected cancer was missed; therefore, only one woman with interval cancer would have to be recalled.

Applying AI-based screening on the testing sample, using the thresholds 5 and 9.989 for exclusion and recall, respectively, 71,499 screening examinations would have been labeled normal, 42,836 would have been labeled as moderate risk and to be read by radiologists, and 86 would have been suspicious and the women would have been recalled automatically. Of these 86 women, 70 had screen-detected cancers, none had interval cancers, two had long-term cancers, and 14 were cancer free; 1.5% (12 of 791) of screen-detected cancers were categorized as normal and, therefore, would have been missed. Categorizations are shown in Table 3 for the testing sample and Table E3 (online) for the development sample. Example mammograms are shown in Figure 4.

The measured screening outcomes of the radiologist and AI-based screening in the testing sample are presented in Table 4. In the radiologist screening, radiologists read all mammograms without interference from the AI system and achieved a sensitivity of 70.8% (791 of 1118; 95% CI: 68.0, 73.5) and a specificity of 98.1% (111,196 of 113,303; 95% CI: 98.1, 98.2). Simulating AI-based screening on the testing sample achieved a sensitivity of 69.7% (779 of 1118; 95% CI: 66.9, 72.4) and a specificity of 98.6% (98.5, 98.7). A total of 529 false-positive screenings may have been avoided, corresponding to a 25.1% (529 of 2107) reduction from the number of false-positive screenings with radiologist screening. Radiologists would have avoided reading images from 71,585 screenings due to the exclusion of normal or suspicious mammograms, which corresponds to a 62.6% (71,585 of 114,421) workload reduction. AI-based screening sensitivity was noninferior to radiologist screening (P = .02). The AI-based screening specificity was higher than that of the radiologist screening.
AI-based Screening Outcome across BI-RADS Densities

The measured sensitivities and specificities are shown in Figure 5A and 5B, respectively, according to BI-RADS density and the screening protocol. The AI-based screening sensitivity was reduced between 0.6 and 1.8 percentage points for BI-RADS densities 1–3 compared with the radiologist screening sensitivity. The sensitivities were equal for BI-RADS density 4. Noninferiority could not be established for the individual BI-RADS density groups due to smaller sample sizes. AI-based screening specificities increased between 0.3 and 0.6 percentage points across all BI-RADS densities. All measured specificities were higher than the radiologist screening specificities (P < .001 for BI-RADS densities 1–3, P < .01 for BI-RADS density 4).

Discussion

To efficiently manage resources during a time when more radiologists are needed worldwide (8,9), we proposed a screening protocol based on an artificial intelligence (AI) system and investigated whether it could detect normal, moderate-risk, and suspicious mammograms, thereby improving screening outcomes and reducing the radiologist workload. In a retrospective simulation study of AI-based screening applied to an independent testing sample, 71,499 mammographic examinations were determined to be normal and 86 were suspicious, leading to a 62.6% (71,585 of 114,421) radiologist workload reduction. AI-based screening sensitivity was noninferior to radiologist screening sensitivity (P = .02), and AI-based screening specificity was higher than radiologist screening specificity (P < .001). Using AI-based screening, the number of false-positive screenings was reduced by 529 compared with radiologist screening (25.1% reduction); 1.5% of screen-detected cancers were missed. AI-based screening performed consistently across Breast Imaging Reporting and Data System densities.

Similar detection performances of the AI system have been found in cohorts from other vendors and countries (11,14). In previous studies, several AI systems (Tranpara and other products) were used to safely reduce the radiologist workload. These studies showed noninferior sensitivities and a low number of missed screen-detected cancers (23,25,32); however, the workload reduction was substantially less (17%–19%) than in our study. In two studies with comparable workload reduction, the study samples of 15,986 and 7,364 women were considerably smaller than ours and without independent validation (24,27). One study showed that preselecting mammograms for single-reading instead of double-reading safely yielded a 33% workload reduction, but in a smaller sample of 18,015 screenings (20).

With AI-based screening on the testing sample, no women with interval cancers were recalled due to the high RT of 9.989. The optimal RT in the testing sample was 9.954. The high RT is likely due to differences in the development and testing samples. From the period of the development sample (2012–2013) to that of the testing sample (2014–2015), the recall and false-positive rate declined (P < .001 for both) (6). Managerial factors, such as when to recall, could influence the RT.

The cancer detection performance of the AI system was reduced with increasing BI-RADS density, which was expected due to masking by fibroglandular tissue (17). This effect was also partially explained by women with interval cancers making up a larger proportion of those with high BI-RADS densities. Moreover, interval cancers are harder to detect and, therefore, AUC is reduced as BI-RADS density grows.

Our study has limitations. First, with AI-based screening, radiologists would read a subgroup of mammograms of the sample, where the likelihood of encountering a cancer might be greater than with standard screening. Second, our study examined the effect of replacing both radiologists’ readings in the case of normal or suspicious mammograms. Replacing only the first reader would be clinically more realistic, as it is safer
and would raise fewer ethical and legal concerns in a prospective study. Third, 14 healthy women would have been recalled with the AI-based screening due to their mammograms being categorized as suspicious. One would have been recalled with radiologist screening as well. Consequently, 13 persistently healthy women with negative screenings would have been recalled. Fourth, for the individual BI-RADS density groups, noninferior sensitivities could not be established due to small sample sizes in the density strata. Visual inspection of the AI-based screening reveals that sensitivities are similar to that of the radiologist screening and the CIs have a high degree of overlap. Fifth, potential cancers not detected within the 2-year follow-up period might have affected the screening outcomes for the considered screening round. We assess the number of such cancers to be low and believe this did not affect screening quality across multiple rounds. Finally, our study included data from a single European institution and AI system, meaning that we cannot guarantee that the results of our screening protocol can be generalized to clinics with other screening regimens (including different screening intervals), management styles, or AI systems.

In conclusion, the incorporation of an artificial intelligence (AI) system in population-based breast cancer screening programs could potentially improve screening outcomes and may considerably reduce the workload of radiologists. AI-based screening would likely support most of these screening programs by increasing efficiency. The protocol is adjustable, by changing the thresholds, to yield more conservative screening or to match a desired recall rate. This might support resource-limited and/or single-reader programs or programs with non-specialized radiologists. For highly specialized radiologists, it might increase the reading time for high-risk mammograms. For institutions with comparable management regimens, the results of our AI-based screening protocol are likely generalizable, but further validation is needed. Additional examination of thresholds is needed, along with an investigation of misclassified cancers based on findings by the AI system, diagnostic and pathologic reports, and previous screenings. A prospective trial is needed to determine the impact of AI-based screening on radiologist performance before the clinical implementation of this protocol.

Acknowledgments: We thank Esther Smits, MSc, Freeke Porte, MSc, and colleagues at ScreenPoint Medical for assisting in the installation and support of Transpara, and the Capital Region of Denmark for providing screening mammograms. We also thank Aksel Karl Georg Jensen, PhD, from the Department of Public Health at the University of Copenhagen for biostatistical assistance.


References