Benign biopsy effect on artificial intelligence cancer detector in screening mammography – a retrospective study

Artificial intelligence cancer detector in a mammography data set

Athanasios Zouzos¹,²*, Alexandra Molovanovic³&, Karin Dembrower²,³&, Fredrik Strand¹,²¶

¹ Department of Radiology, Karolinska University Hospital, Stockholm, Sweden
² Department of Oncology and Pathology, Karolinska Institute, Stockholm, Sweden
³ Department of Radiology, S:t Görans Hospital, Stockholm, Sweden

* athanasios.zouzos@ki.se (AZ)

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Artificial intelligence (AI) cancer detectors are showing promising results and may soon be used for clinical breast cancer screening in radiology departments. Validation of such programs in different settings is mandatory.

This is a retrospective study after the application of a commercial AI cancer detector program to a cancer-enriched mammography screening dataset where 10,889 women with a median age for the study population of 56 years and a range between 40 and 74 years old where included. The program is intended for 2D mammography and it generate a prediction score for tumor presence with a decimal number between 0 and 1, where 1 represented the highest level of suspicion. The AI score median and interquartile range for women who were healthy, who had a benign biopsy, and who were diagnosed with breast cancer, was calculated. For women with a previous benign biopsy, the time between mammogram and the biopsy was stratified. The effect of increasing age was examined using stratified analysis and linear regression modelling.

The proportion of healthy women above the AI abnormality threshold (0.4) defined at our hospital was 3.5%, 11% and 84% for healthy women without a benign biopsy, healthy women with a benign biopsy, and for women with breast cancer, respectively (p<0.001) and the AI score correlated positively with increasing age of the women in the cancer group (p<0.05).

While 16% of the women who had a benign biopsy after mammography had an AI score above the cut-off point, there were 33% of those when they had their biopsy before the mammography. AI cancer detection systems can reliable be used in a screening setting. Adding information about possible prior benign biopsy and other symptoms can probably improve accuracy and subsequently reduce the workload for radiologists.
Introduction

Breast cancer is the most common cancer among women worldwide. It ranks 5th as a cause of cancer deaths because of its relatively favorable prognosis but in the last 20 years the average annual increase in breast cancer incidence rate has been 1.4% [1-3].

Well implemented screening programs have been clearly proven to reduce the mortality rate for breast cancer [4-6] and prospective studies have shown that results improve when radiologists combine mammography reading with an AI algorithm [7-9]. Furthermore, reducing reading time with the assistance of AI algorithms is possible [10, 11]. AI algorithms can be highly accurate for reading mammograms and some systems are now on a comparable level with average breast radiologists at detecting breast cancer on screening mammography [12].

In addition to the well-known risk factors of age, family history and hormonal history, there are also studies showing that benign breast disease increases breast cancer risk [13, 14]. A study that analyzed risk factors for breast cancer found that having had any prior breast procedure was associated with an increased risk of breast cancer [15]. Another study showed that women found to have false-positive mammography were more likely to develop interval cancer or cancer at the second screening compared to women not recalled [16]. Also in Gail’s risk model, prior biopsy is defined as a risk factor, and the risk is higher for women under 50 years of age [17].

In light of biopsies as a breast cancer risk factor and the increasing use of AI for screening mammography, it is important to understand if having a benign biopsy affects the AI cancer score and the ability of AI to make an accurate assessment of the screening mammogram.
Results

Initially there were 11 303 women in this retrospective case-control study. Of these 414 women were excluded. The exclusion criteria were; no mammography in conjunction with a cancer diagnosis, implants and cancer more than 12 months after mammography. The cancer group consisted in total of 917 women, the benign biopsy-group of 234 women and in the group with no cancer or biopsy there were 9 738 women, see Fig 1.

**Fig 1. Study population with exclusion criteria and subgroups.** Bx: benign biopsy; Mx: mammography.
Of the remaining 10 889 women, 8 307 had complete information on radiologist assessments (when performing data collection, we received radiologist assessments only until Dec 31, 2015) which included selections as potentially pathological by one or both radiologists and a final recall decision after consensus discussion. From those 8 307 women there where 724 women in the cancer group, 212 in the benign biopsy-group and 7 371 in the normal mammography-group.

There was a significant difference (p<0.001) in AI score between the cancer-group, the benign biopsy-group and the normal mammography-group (Table 1).

<table>
<thead>
<tr>
<th>Sub groups</th>
<th>n</th>
<th>Age, years</th>
<th>% above cut-off point</th>
<th>AI-score</th>
<th>P-value</th>
<th>P-value vs. Cancer</th>
<th>P-value vs. Normal w/o biopsy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal with benign biopsy</td>
<td>234</td>
<td>51.2</td>
<td>11% (26/234)</td>
<td>0.051</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Cancer</td>
<td>917</td>
<td>60.6</td>
<td>83% (768/917)</td>
<td>0.933</td>
<td>N/A</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Normal without biopsy</td>
<td>9738</td>
<td>55.5</td>
<td>3.5% (345/9738)</td>
<td>0.018</td>
<td>&lt;0.001</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>

Data are presented in median and interquartile range (p25-p75).

The percentage above the cut-off point, for AI assessments, in the group with normal mammography was 3.5%, and 83% in the cancer-group. In the benign biopsy-group 11% of the AI assessments were above the cut-off point. The distribution of AI scores for women diagnosed with breast cancer is shown in Fig 2, for healthy women with benign biopsy in Fig...
3, and for healthy women without benign biopsy who remained healthy in Fig 4.

**Fig 2.** N=917. Median 0.933, p25-p75 (0.666-0.983). Cut-off point 0.40

**Fig 3.** N=234. Median 0.051, p25-p75 (0.016-0.174). Cut-off point 0.40

**Fig 4.** N=9738. Median=0.018. p25-p75 (0.005-0.065). Cut-off point 0.40

In Table 2, we present how the AI score is associated with the age of the woman. There was a significant increase of the AI score in relation to age category in the cancer group (p<0.05). There was no significant increase in the AI score in relation to age in the group with normal mammography or in the group with benign biopsies. The median age for the study population
was 56 years and the median AI score was 0.023. The median age for the cancer group was 61 years and the median AI score was 0.933. The median age for the group with previous benign biopsies was 49 years and the median AI score 0.051. The median AI score for healthy women was 0.018 and the median age for that group was 59 years.

Table 2. AI-score, age order

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Normal mammography (No bx, no cancer)</th>
<th>Benign biopsy-group (No cancer)</th>
<th>Cancer-group</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>9738 0.018 (0.005–0.065)</td>
<td>234 0.051 (0.016–0.174)</td>
<td>64 0.069 (0.018–0.179)</td>
</tr>
<tr>
<td>40–49</td>
<td>3152 0.021 (0.006–0.068)</td>
<td>144 0.048 (0.016–0.155)</td>
<td>43 0.074 (0.019–0.142)</td>
</tr>
<tr>
<td>50–59</td>
<td>2834 0.015 (0.005–0.058)</td>
<td>68 0.051 (0.017–0.260)</td>
<td>16 0.039 (0.014–0.113)</td>
</tr>
<tr>
<td>60–69</td>
<td>2577 0.018 (0.005–0.064)</td>
<td>22 0.078 (0.012–0.174)</td>
<td>5 0.173 (0.090–0.449)</td>
</tr>
<tr>
<td>70+</td>
<td>1175 0.020 (0.006–0.077)</td>
<td>0 0 0 0</td>
<td>97 0.958 (0.820–0.985)</td>
</tr>
</tbody>
</table>

Data are presented in median with interquartile range (p25-p75). Bx= benign biopsy, Mx=mammography.

The benign biopsy-group was stratified by time between biopsy and mammography into three categories: 0-6 months, 6-24 months and more than 24 months. There was no significant difference between the time-stratified categories (Table 3). In Table 3, we describe the AI score
related to the time between biopsy and mammography. Within 6 months after mammography, 104 of 234 had had benign biopsy. The proportion of women with AI scores above the threshold, was, 16% for those with a benign biopsy within 6 months from mammography, and 33% for those with a benign biopsy 6 months before mammography.

Table 3. Study population - Benign biopsies

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Age, years</th>
<th>% above the cut-off point</th>
<th>AI-score</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Benign biopsy before mammography</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-6 months</td>
<td>9</td>
<td>49.1</td>
<td>33% (3/9)</td>
<td>0.150</td>
<td>ref</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(44.5-54.9)</td>
<td></td>
<td>(0.099–0.449)</td>
<td></td>
</tr>
<tr>
<td>6-24 months</td>
<td>39</td>
<td>48.1</td>
<td>5.1% (2/39)</td>
<td>0.062</td>
<td>0.123</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(46.7–52.7)</td>
<td></td>
<td>(0.016–0.150)</td>
<td></td>
</tr>
<tr>
<td>&gt;24 months</td>
<td>16</td>
<td>41.0</td>
<td>0% (0)</td>
<td>0.025</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(40.3–48.4)</td>
<td></td>
<td>(0.015–0.099)</td>
<td></td>
</tr>
<tr>
<td><strong>Benign biopsy after mammography</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-6 months</td>
<td>104</td>
<td>48.3</td>
<td>16% (17/104)</td>
<td>0.065</td>
<td>ref</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(44.4–52.5)</td>
<td></td>
<td>(0.017–0.274)</td>
<td></td>
</tr>
<tr>
<td>6-24 months</td>
<td>44</td>
<td>49.0</td>
<td>4.5% (2/44)</td>
<td>0.037</td>
<td>0.362</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(45.6–55.3)</td>
<td></td>
<td>(0.008–0.109)</td>
<td></td>
</tr>
<tr>
<td>&gt;24 months</td>
<td>22</td>
<td>51.2</td>
<td>9% (2/22)</td>
<td>0.031</td>
<td>0.378</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(45.6–53.0)</td>
<td></td>
<td>(0.017–0.165)</td>
<td></td>
</tr>
</tbody>
</table>

Data are presented in median and interquartile range (p25-p75).

In the group with benign biopsy after mammography the abnormal assessment by AI, above cut-off point, was 15% while the radiologists had a recall rate up to 57% in this group (Table 4).
The radiologists had a recall rate of 2% and the rate for abnormal assessments by AI was 3.8% in the group with normal mammograms and with benign biopsy (Table 4). For the group with only normal mammograms the recall rates were 1% and 3.6% respectively. Radiologists and the AI program had similar rates of recall for the total study population as a whole.

Table 4. Recall rate and abnormal assessments

<table>
<thead>
<tr>
<th></th>
<th>Radiologist 1</th>
<th>Radiologist 2</th>
<th>Consensus</th>
<th>AI-% above cut-off point</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>10% (831/8307)</td>
<td>10% (827/8307)</td>
<td>9.2% (767/8307)</td>
<td>11% (880/8307)</td>
</tr>
<tr>
<td>Normal and benign biopsy</td>
<td>3.6% (274/7583)</td>
<td>3.1% (234/7583)</td>
<td>2% (154/7583)</td>
<td>3.8% (290/7583)</td>
</tr>
<tr>
<td>Normal</td>
<td>2.8% (203/7371)</td>
<td>2.2% (162/7371)</td>
<td>1% (73/7371)</td>
<td>3.6% (265/7371)</td>
</tr>
<tr>
<td>Benign biopsy</td>
<td>33% (71/212)</td>
<td>34% (72/212)</td>
<td>38% (81/212)</td>
<td>12% (25/212)</td>
</tr>
<tr>
<td>- Biopsy before Mx</td>
<td>8.5% (5/59)</td>
<td>6.8% (4/59)</td>
<td>6.8% (4/59)</td>
<td>8.5% (5/59)</td>
</tr>
<tr>
<td>- Biopsy after Mx</td>
<td>49% (66/135)</td>
<td>50% (68/135)</td>
<td>57% (77/135)</td>
<td>15% (20/135)</td>
</tr>
<tr>
<td>Cancer</td>
<td>77% (557/724)</td>
<td>82% (593/724)</td>
<td>85% (613/724)</td>
<td>81% (590/724)</td>
</tr>
<tr>
<td>With benign biopsy</td>
<td>75% (12/16)</td>
<td>88% (14/16)</td>
<td>88% (14/16)</td>
<td>81% (13/16)</td>
</tr>
<tr>
<td>W/o benign biopsy</td>
<td>77% (545/708)</td>
<td>82% (579/708)</td>
<td>85% (599/708)</td>
<td>82% (577/708)</td>
</tr>
</tbody>
</table>

Recall rate by radiologist and abnormal assessments by AI. Data are presented in percentage and fraction. Mx: mammography.

The following two screening mammograms have been assessed by radiologists and the AI cancer detection program, see Figs 5 and 6. These examples illustrate concordant and discordant assessments.
**Fig 5.** Woman in her 50s, selected by radiologists for potential pathology in the left breast. High AI-score. The biopsy result showed hyperplastic breast epithelial cells that could represent a degenerated fibroadenoma.
Fig 6. Woman in her 50s, selected by radiologists for potential pathology in the right breast.

Low AI-score. The biopsy showed lymph node.

Material and method

Study population

This project is a retrospective study based on a case-control subset from the dataset Cohort of Screen-Aged Women (CSAW). CSAW is a complete population-based cohort of women 40 to
74 years of age invited to screening in the Stockholm region, Sweden, between 2008 and 2015 [18]. The exclusion criteria in CSAW were: prior history of breast cancer, diagnosis outside the screening range and incomplete mammography examinations. From this large CSAW-cohort a case-control subset was separately defined to contain all women from Karolinska University Hospital, Stockholm, who were diagnosed with breast cancer (n=1303), at screening or clinically during the interval before the next planned screening, and 10 000 randomly selected healthy controls [18]. The purpose of the case-control subset is to make evaluation more efficient by not having to process an unnecessary amount of healthy controls while preserving the representability of the CSAW screening cohort in which it is nested. Additional exclusion criteria for the current study were: implants and cancer diagnosis more than 12 months after mammography. The study population was divided into three groups based on their status: cancer, benign biopsy, and normal.

The cancer-group was defined as “biopsy verified breast cancer within 12 months from mammography”. For those women, the most recent screening mammography prior to diagnosis was selected for analysis.

The benign biopsy-group was defined as “benign biopsy without ever having breast cancer”. The screening mammography closest in time to the biopsy, before or after was chosen. The group was stratified by time between biopsy and mammography, both from biopsy to mammography and from mammography to biopsy.

The normal-group had had neither breast cancer nor prior benign biopsy. The most recent screening mammography was selected for this group. Women in the screening program with lesions that were previously deemed benign and were not selected for further diagnostic work, were also included in this group. Only biopsy verified lesions were included in the benign biopsy-group.
Mammography assessments

Screening mammograms were prospectively assessed independently by two radiologists working at a certified screening center, with more than five years of screening experience for each of them. The following screening decision data was collected: flagging of abnormal screening by one or both radiologists and the final recall decision after consensus discussion. Screening decisions and clinical outcome data was collected by linking to regional cancer center registers. Those data were collected so to be compared with the assessment results from AI cancer detector.

The AI cancer detector algorithm used for detection of tumor signs was provided by a commercial vendor (Lunit, Seoul, South Korea), version 5.5.0.16. The reason for choosing Lunit's algorithm for this study was the results of the retrospective analysis published 2020 [9] where a comparison of three of the most known algorithms showed that best of those was AI (Lunit) with a sensitivity and specificity comparable to Breast Cancer Surveillance Consortium benchmarks [19].

The algorithm was originally trained on 170 230 mammograms, from 36 468 women diagnosed with breast cancer and 133 762 healthy controls. The mammograms in the original training were sourced from five institutions: three from South Korea, one from the USA, and one from the UK. The mammograms were acquired on mammography equipment from GE Healthcare, Hologic, and Siemens, and there were both screening and diagnostic mammograms. In the clinical version, the AI cancer-detector software creates visual prompts in the mammograms where the suspicion of tumor is above a certain threshold. In this study, we did not use the images, but instead used the underlying continuous prediction score of the algorithm. The generated prediction score for tumor presence was a decimal number between 0 and 1, where 1 represented the highest level of suspicion. The program assessed both breasts and the highest
score among the four images was selected as representing the exam. The cut-off point (0.4) (AI abnormality threshold) defining a screening abnormality was determined in a prior study having the AI system operate at the first radiologists’ specificity level [9].

Collection of data

The Stockholm-Gotland regional cancer center (RCC) provided personal identification numbers for all women who fulfilled the inclusion criteria for CSAW. The identification numbers were linked to the local breast cancer quality register, “Regional Cancercentrum Stockholm-Gotland Kvalitetsregister för Bröstcancer” to collect data about breast cancer diagnosis. All diagnoses of breast cancer were biopsy verified. Benign diagnoses were collected from the medical record “Take Care” which is hospitals data-based system that provides detailed health information about every registered patient.

All images were 2D full-field digital mammograms acquired on Hologic mammography equipment. The personal identification numbers were also linked to the radiological image repository to extract all digital mammograms from PACS. The dataset has been stored securely following the requirements of the ethical approval.

Statistical analyses

The statistical analysis was done per patient and not per lesion. Stata version 14 or later was used for statistical analyses. Wilcoxon rank sum-test and quantile regression were used to examine differences between groups. These statistical tests of differences in median were chosen due to the skewed distribution of AI scores.
Compliance with ethical standards

The collection and use of the dataset for the purpose of AI has been approved by the Swedish Ethical Review Board 2017-02-08, and the need for informed consent was waived. The Diary number was 2016/2600-31.

Discussion

The strength of this study is the large number of cancer cases and that all women were sampled from a screening cohort. A limitation is the relatively small number of benign biopsies which makes it difficult to consider different types of benign lesions. Another limitation is the retrospective setting: since the AI program did not have the opportunity to make recalls and choose women for further diagnostic biopsy it could not affect who got biopsied. All information about benign biopsies was based on radiologists’ decisions. In contrast to radiologists, the AI-program calculates a score for the likelihood of breast cancer based on the image alone and does not consider any information about symptoms given by the woman at screening.

Furthermore, in this study, we did not consider the exact location of the presumed abnormality where the AI showed a high AI score. We did not use the images, but instead we used the underlying predictions score of the algorithm as a per patient analysis. AI scores above the threshold for women having benign biopsy was 16% if the biopsy was done 6 months after the mammography and 33% if the biopsy was done 6 months before the mammography examination. This can cause questions about the probability that AI is affected by alternations on mammography because of the biopsy. Further analysis of the data can be valuable to evaluate whether the lesions that AI showed responded to the actual finding that the patient was recalled...
Although we used a specific AI cancer-detector software for our study which may reduce the external validity of the study, the mammograms were acquired on mammography equipment from GE Healthcare, Hologic, and Siemens, and there were both screening and diagnostic mammograms which covers a broad amount of systems with diverse imaging characteristics.

The data that confirm the external validation of the AI program is the significant difference in AI score between the normal group and the cancer group as well as the association between higher AI score and increasing age, in the cancer-group. Like radiologists, AI detects more cancer in the screening mammograms of older women. No difference was shown between AI assessment and radiologist assessment when selecting for cancer in accordance with results from prior studies.

Another interesting finding is the recall rate for the group with mammography prior to benign biopsy. It was 57% for the radiologist and 15% for the AI program. Those figures are more likely to be explained by the study design where the AI program didn’t have access to the clinical evaluation of the patients while the radiologists selected all the patients with symptoms as in routine praxis almost independently from imaging.

The most intriguing finding this study presents is that even if the proportion of (false positive) AI scores increased from 3.6% to 8.5% with a previous benign biopsy. In other words, there was a significant difference in AI scores between the normal group and the benign biopsy-group despite both groups consisting of healthy women. This is probably an indication that incorporation of information about prior biopsy in AI cancer detector programs would further improve their accuracy. Prospective studies regarding that matter are required.

These results amplify the indications from previous studies, that AI cancer detectors can be reliable enough to incorporate in a clinical screening assessment. If AI is to make more accurate
assessments it will probably require clinical information (previous biopsy, symptoms) in addition to simply analyzing mammograms. Based on the observation that only 15% of the women who had a biopsy after mammography had an AI score above the cut-off point, there might be a role for AI in reducing the number of unnecessary biopsies.

References


