Development and Clinical Evaluation of Deep Learning Models for Personalized Breast Cancer Screening

PhD Thesis
Andreas David Lauritzen

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen on January 31st, 2023

Principal supervisor: Martin Lillholm
Co-supervisor: Mads Nielsen
Co-supervisor: Ilse Vejborg
Preface

The work for this PhD thesis was completed at the Department of Computer Science at the University of Copenhagen and Gentofte Hospital in the Capital Region of Denmark between August 2019 and January 2023. The candidate was enrolled at the PhD School of SCIENCE at the University of Copenhagen.

**Doctoral candidate and author:**
Andreas David Lauritzen

**Principal supervisor:**
Professor Martin Lillholm, PhD
Department of Computer Science,
University of Copenhagen

**Co-supervisor:**
Professor Mads Nielsen, PhD
Department of Computer Science,
University of Copenhagen

**Clinical co-supervisor:**
Head of mammography screening Ilse Vejborg, MD
Gentofte Hospital,
Capital Region of Denmark

**PhD assessment committee:**
Chair: Professor Erik Bjørnager Dam, PhD
Department of Computer Science, University of Copenhagen

Opponents: Professor Carla H. van Gils, PhD
Department of Epidemiology, Julius Center for Health Sciences and Primary Care,
University Medical Center Utrecht

Professor Sebastien Ourselin, PhD
School of Biomedical Engineering and Imaging Sciences, King’s College London

*Cover art by my fiancé Charlotte Ernstsen*
Acknowledgements

First and foremost, I would like to express my gratitude to my principle supervisor Martin Lillholm. Despite a crazy schedule, you have always made time to meet when I needed guidance for which I am thankful. You have had my back throughout the project and guided me through all the difficult times with patience and pragmatism. I appreciate our Friday morning meetings with interesting discussions, smalls lectures, and nice talks. I have been lucky to have such an excellent supervisor, mentor, and friend.

I would like to thank my co-supervisor Mads Nielsen for taking me in on the project and providing me with this opportunity. I admire your knowledge on so many disciplines which have had great value to me and this project. I have appreciated our meetings together and your invaluable feedback on the papers, abstracts, and presentations - thank you for that.

I would like to thank my co-supervisor Ilse Vejborg for giving me the opportunity to experience, first-hand, a breast cancer screening program in action. Your extensive knowledge on the clinical aspects of breast cancer and screening have been invaluable. I look forward to continuing the collaboration with you and your excellent team at Gentofte Hospital.

Furthermore, I would like to thank my colleagues and collaborators at the Department of Computer Science, Gentofte Hospital, and ScreenPoint Medical. A special thanks to Nico Karssemeijer for being so involved in our studies and for helpful feedback on manuscripts and abstracts. Thanks to My Catarina von Euler-Chelpin at the Department of Public Health for providing us with important screening and pathology data. Thanks to Elsebeth Lynge at the Department of Public Health and Nykøbing Falster Hospital for good discussions and feedback on our papers. Cheers to my fellow PhD students Mathias Perslev and Rasmus Kær Jørgensen for the chats, beers, and hangs. It would not have been the same without you.

I would like thank the PhD assessment committee for taking time to read and evaluate this thesis.

Lastly, I would like express how important my fiancé, family, and friends have been throughout this PhD. I am lucky and grateful to have you all supporting me, cheering me on, comforting me, and reminding me what is important. This would not have been possible without you.

I am pleased to present this thesis, and I hope you find it interesting.

Andreas David Lauritzen
January 31st, 2023
Abstract

Breast cancer is the single most diagnosed type of cancer among women worldwide. Fortunately, mammography screening successfully decreases breast cancer mortality, yet entails a number of practical challenges related to cumulative radiation dosage, overdiagnosis, breast radiologists’ workload, lack of breast radiologists, anxiety for screened/biopsied women, and differences in screening quality from country to country.

Personalized breast cancer screening, based on individual risk, is proposed to alleviate the mentioned practical challenges associated with large scale screening. Traditional risk models based on clinical risk factors are in use today, but are practically limited by continuous extensive data collection processes. Mammography-based risk models, that solely consider the imaging data, are more suited for clinical implementation in screening, as the mammograms are readily available. Mammography-based risk models have been studied for decades and have gradually improved, however, deep learning now offers methods of reliably estimating risk with high accuracy, robustness, and objectivity.

Mammography-based deep learning models for detection are commercially available and can detect cancer with the same accuracy as radiologists. In the first study, we employed an (artificial intelligence) AI system for lesion detection and retrospectively simulated a screening protocol in which women were categorized based on the likelihood of detecting a breast cancer. Women with likely normal mammograms were excluded from radiologist reading as healthy. Women with very high likelihood of breast cancer were recalled immediately for diagnostic tests. Women with moderate likelihood was double-read as in standard screening. The results indicated that, with AI-based screening, the radiologists’ workload could be considerably decreased while the high screening quality was preserved. This simulation study contributed to a full clinical implementation of the AI system in the Capital Region of Denmark for which prospective preliminary results indicate a safe rollout of AI in screening.

Recent studies have shown promise in estimating risk of a future breast cancer. However, current risk models does not ensure reliable risk assessment across mammographic devices from different vendors. Additionally, current risk models are trained in a conflated manner and learn features indicative of a visible breast cancers, subtle signs of developing breast cancers, and textural features indicative of a susceptibility to future breast cancer. Training for all three tasks simultaneously might lead to subpar long-term risk assessment. In the second study, we developed a texture model that was trained optimally for long-term risk, relying on features of healthy breast tissue indicative of future breast cancer. We demonstrated that the texture model could robustly estimate short- and long-term risk while generalizing across mammographic devices from different vendors.

We additionally developed a dense tissue segmentation tool to estimate planimetric percentage mammographic density (PMD), which is a known and established breast cancer risk factor.

In the third study, we combined the AI system for short-term risk and the texture model for long-term risk in a combination model with age and PMD to create a rich mammography-based risk
profile. The results indicated that training each system individually and combining them subsequently significantly improved overall risk assessment.

In the fourth study, we developed a generic machine learning framework for identifying patient groups using national health registry data. This framework was applied to identification of breast cancer relapse patients, but could be directly translated to breast cancer risk assessment and used for an even more expressive breast cancer risk profile.

These developed models might, with further rigorous validation, support clinicians in creating personalized screening protocols that could benefit patients and radiologists.
Dansk Resumé

Brystkræft er den mest diagnosticerede kæfttype blandt kvinder verden over. Heldigvis mindsker mammografi-screening brystkreftsdødeligheden, men medfører en række praktiske udfordringer relatet til stråling, overdiagnosticering, mammaradiologers arbejdsbyrde, mangel på mammaradiologer, angst for screenede/boopterede kvinder og forskelle i screeningskvalitet fra land til land.

Personaliseret brystkreftsscreening, baseret på individuel risiko, anbefales for at imødekomme de nævnte praktiske udfordringer forbundet med populationsscreening. Traditionelle risikomodeller, baseret på kliniske risikofaktorer er i anvendelse i dag, men er praktisk begrænset på grund af omfattende løbende dataindsamling. Mammografi-baserede risikomodeller, der kun tager billeddata i betragtning, er bedre egnet til klinisk implementering i screening, da mammograferne allerede er tilgængelige. Mammografi-baserede risikomodeller har været studeret i årtier og er gradvist blevet forbedret, men dyb læring giver nu mulighed for pålideligt at vurdere risiko med høj nøjagtighed, robusthed og objektivitet.


Nye studier har vist, at det er muligt at estimere risikoen for fremtidig brystkræft. Nuværende risikomodeller sikrer dog ikke pålidelig risikovurdering på tværs af mammografiske enheder fra forskellige leverandører. Derover bliver nuværende risikomodeller trænet til samtidig at lære; træk af synlige maligne lesioner, subtile tegn på udvikling af brystkræft og teksturelle træk, der kunne indikere en modtagelighed for fremtidig brystkræft. At træne modellen til alle tre opgaver samtidig kan føre til subpar risikovurdering. I det andet studie udviklede vi derfor en teksturmodell, der var trænet optimalt til at estimere risiko på lang sigt ved lære karakteristika i sundt brystvæv, der kunne indikere modtagelighed overfor brystkræft. Vi viste, at teksturmodellen pålideligt kunne estimere risiko på kort og lang sigt, alt imens modellen generaliserede på tværs af mammografiske enheder fra forskellige leverandører.

Vi udviklede desuden et værktøj til at segmentere tæt brystvæv og derved estimere den planimetriske procentvise mammografiske tæthed (PMD), hvilket er en kendt og etableret riskofaktor for brystkræft.

I det tredje studie kombinerede vi AI-systemet, der estimerede risiko på kort sigt, og teksturmodellen, der estimerede risiko på lang sigt. Kombinationsmodellen inkluderede desuden alder og
PMD for at danne en bred mammografi-baseret risikoprofil. Resultaterne indikerede en forbedret risikovurdering ved at træne hvert system for sig og herefter kombinere dem.

I det fjerde studie udviklede vi et generisk maskinlæringsværktøj til at identificere patientgrupper ved brug af nationale sundhedsregistre. Dette værktøj blev anvendt til at identificere brystkæftpatienter med tilbagefald. I fremtiden ville værktøjet kunne gentrænes til brystkæftrisikovurdering og hermed udvide den kombinerede risikoprofil.

De udviklede modeller kan, med yderligere og omfattende valideringsstudier, hjælpe klinikere med at definere personlige screeningsprotokoller, der kan komme patienter og radiologer til gavn.
**List of Papers and Abstracts**

The following list of papers, authored by the candidate, have been included in this thesis:


The following list of abstracts, authored and presented orally at international recognized conferences by the candidate, have been included in this thesis:


The following list of papers, co-authored by the candidate during the PhD project, have not been included in this thesis:


The following list of papers, authored or co-authored by the candidate, are studies that have been conducted before the PhD project:


## Contents

**Preface**  
2

**Acknowledgements**  
3

**Abstract**  
4

**Dansk Resumé**  
6

**List of Papers and Abstracts**  
8

### 1 Background  
13

#### 1.1 Breast Anatomy and Cancer  
13

#### 1.2 Breast Cancer Epidemiology  
13

#### 1.3 Clinical Breast Cancer Risk Factors  
14

- 1.3.1 General Risk Factors  
14
- 1.3.2 Reproductive Risk Factors  
14
- 1.3.3 Risk Factors Related to Previous Breast Disease  
15
- 1.3.4 Congenital Risk Factors  
15

#### 1.4 Risk Models Based on Clinical Risk Factors  
15

- 1.4.1 The Gail Model  
15
- 1.4.2 The Tyrer-Cuzick Model  
16
- 1.4.3 The BOADICEA Model  
16

#### 1.5 Preventing the Progression of Breast Cancer  
16

#### 1.6 Mammography Screening  
17

- 1.6.1 Purpose of Screening  
17
- 1.6.2 Evaluation of Screening Quality  
18
- 1.6.3 Harms and Benefits of Screening  
20
- 1.6.4 Image Acquisition  
21
- 1.6.5 Suspicious Mammographic Findings  
22

#### 1.7 Mammographic Risk Factors  
22

- 1.7.1 Mammographic Dense Tissue  
22
- 1.7.2 Wolfe, Boyd, Tabár Classification Systems  
23
- 1.7.3 Quantitative Measures of Mammographic Density  
24
- 1.7.4 Mammographic Texture  
25
- 1.7.5 Miscellaneous Mammographic Findings Associated with Risk  
25

#### 1.8 Early Computational Models for Mammography-based Breast Cancer Risk  
26

- 1.8.1 Classifying Wolfe Patterns  
27
- 1.8.2 Quantification of Mammographic Density  
27
- 1.8.3 Predicting BRCA Alterations  
28
- 1.8.4 Texture Models for Direct Risk Assessment  
28
1.9 Deep Learning Models for Mammography-based Breast Cancer Risk .................................................. 29
  1.9.1 Deep Learning in Medical Imaging ............................................................................................. 29
  1.9.2 Deep Learning Texture Model for Risk ....................................................................................... 30
  1.9.3 Conflated Breast Cancer Risk Models .......................................................................................... 31
  1.9.4 Models for Breast Cancer Detection and Short-term Risk ......................................................... 31

2 Aim of PhD project .......................................................... 33
  2.1 Challenges in Mammography Screening ...................................................................................... 33
  2.2 Personalized Screening Based on Risk .......................................................................................... 34
  2.3 Contribution Hypotheses ............................................................................................................. 35

3 Data Materials ................................................................. 37
  3.1 Screening in the Capital Region of Denmark .................................................................................. 37
     3.1.1 Breast Cancer Group Designations ......................................................................................... 37
     3.1.2 Live Collection of Raw Image Data ........................................................................................ 38
     3.1.3 Danish Screening Cohorts ..................................................................................................... 39
  3.2 Screening in Utrecht, the Netherlands ............................................................................................ 39

4 Contributions ................................................................. 41
  4.1 An Artificial Intelligence–based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload .............................................................................. 42
     4.1.1 Preliminary ............................................................................................................................ 42
  4.2 A Prospective Study of Breast Cancer Screening with AI as First Reader for Likely Normal Mammographies ........................................................................................................ 52
     4.2.1 Preliminary ............................................................................................................................ 52
  4.3 Validation of a Deep Learning-based Breast Density Estimation Tool on a Danish Screening Cohort in the Context of Personalised Risk-based Screening ...................................................... 55
     4.3.1 Preliminary ............................................................................................................................ 55
  4.4 Robust Cross-vendor Mammographic Texture Models Using Augmentation-based Domain Adaptation for Long-term Breast Cancer Risk ........................................................................ 57
     4.4.1 Preliminary ............................................................................................................................ 57
  4.5 Breast Cancer Risk Assessment Improved by Combining Artificial Intelligence for Lesion Detection and Mammographic Texture ................................................................................. 80
     4.5.1 Preliminary ............................................................................................................................ 80
  4.6 Identifying Recurrent Breast Cancer Patients in National Health Registries using Machine Learning ................................................................................................................................. 96
     4.6.1 Preliminary ............................................................................................................................ 96

5 Key Results ......................................................................... 108
  5.1 Paper 1: Simulation of AI-based screening .................................................................................... 108
  5.2 Abstract I: Preliminary results of screening with AI ..................................................................... 108
5.3 Abstract 2: Automatic dense tissue segmentation
5.4 Paper 2: Cross-vendor texture models for long-term risk
5.5 Paper 3: Combining diagnostic AI system and texture model
5.6 Paper 4: Retrospectively identifying relapse patients

6 Discussion
6.1 Advances in Breast Cancer Screening Protocols
6.2 Advances in Mammography-based Breast Cancer Risk Modelling
6.3 Advances in Patient Identification Using Health Registry Data
6.4 Warrant for Standardized Risk Evaluation Procedures
6.5 Limitations
   6.5.1 Cancer Label Noise and Lack of Longer-term Follow Up
   6.5.2 What is learned?
   6.5.3 Demographic Diversity
6.6 Paradigm Shifts In Screening for Breast Cancer
6.7 The Untold Stories and Curiosities
   6.7.1 Optimizing for Device-invariance
   6.7.2 Texture Model Training Stability
   6.7.3 Learning Density-independent Textural Features
6.8 Future work

7 Concluding Remarks

8 Appendix
8.1 Reducing the Radiologist’s Workload by Detecting Normal Mammograms with an AI System
8.3 Improving Mammographic Texture Risk Models with Long-term Cancers
8.4 An AI-based Breast Cancer Reading Protocol Evaluated Across Density Categories in a Danish Screening Cohort
1 Background

1.1 Breast Anatomy and Cancer

The female breast consists of primarily fatty, glandular, and connective tissues as displayed in Figure 1. Surrounded by fat, the glandular tissues comprise lobules which produce milk and mammary ducts which carry the milk to the nipple.

Figure 1: Cross-sectional schematic of the normal breast anatomy. (This image has been modified. Original downloaded from Wikimedia Commons: https://commons.wikimedia.org/wiki/File:Breast_anatomy_normal.jpg)

Breast cancer covers a range of diagnoses related to malignant neoplasia - a new uncontrolled growth of tissue that invades surrounding breast tissue which may disseminate to other sites in the body. Such growths can develop throughout the breast anatomy, but it most likely be located in the ducts (approximately 70% to 77% of cases), lobules (10% to 13% of cases), or both (10% of cases) [1, 2]. Breast cancers that infiltrate the surrounding tissues are referred to as invasive cancers, and cancers that are contained within their original location are referred to as in situ cancers. The latter are often characterized as a pre-stage to cancer.

Staging of breast cancers are based on the Tumor size, Nodal status, and distant Metastasis (TNM)-system [3]. The TNM-system can further be used to divide cancers into four stages: I, II, III, and IV depending on to which extent the cancer has spread. Stage I refers to a small and localized breast cancer whereas a stage IV cancer has metastasized to another part of the body. Ductal carcinoma in situ (DCIS) is generally characterized as prestage to cancer (Stage 0).

1.2 Breast Cancer Epidemiology

Breast cancer is the single most diagnosed cancer type in women worldwide [4]. Approximately one in eight women in the United States (US) and one in seven women in the European Union (EU)
receive a breast cancer diagnosis during their lifetime [1, 5]. These incidence rates have been steadily increasing over the last three decades due to more intensive breast cancer screening, societal lifestyle changes, and reproductive factors [5].

As breast cancer screening programs are being implemented more widely throughout the world, an increasing number of breast cancers are found in an early stage. The combination of early detection and effective treatments means that the mortality rates have been steadily decreasing over the last three decades. The overall five-year survival rates typically range between 80% and 90% in the US, Western EU, and the United Kingdom (UK) [1, 5, 6]. However, survival depends greatly on the stage in which the breast cancer is found. For example, in the UK five-year survival for stage I, II, III, and IV breast cancers are 98%, 90%, 72%, and 26%, respectively. Meanwhile the percentage of breast cancers diagnosed in stage I, II, III, and IV are 39%, 38%, 9%, and %5, respectively [6]. This suggests that, even with screening and efficient treatments, a lot of breast cancers are still not found in a sufficiently early stage.

1.3 Clinical Breast Cancer Risk Factors

Due to the high prevalence and incidence of female breast cancer, a large number of studies have focused on identifying underlying breast cancer risk factors. This section presents the most influential general risk factors, reproductive risk factors, risk factors related to previous breast disease, and congenital risk factors.

1.3.1 General Risk Factors

As for all types of cancer, increasing age is a risk factor for breast cancer as well. Approximately 80% of all breast cancer diagnoses are given to women over 50, and furthermore, 50-year-old women have around twice the risk of being diagnosed with a breast cancer within 10 years compared to 40-year-old women [7]. Having a high body mass index (BMI) is also a breast cancer risk factor. It has been shown that for every 5 kg/m$^2$ increase in body weight, the risk of breast cancer increases by 2% [8]. Other lifestyle choices like consuming alcohol, smoking, and lack of physical activity/exercise further increase the risk of breast cancer [9]. Lastly, hormone replacement therapy (HRT), which is usually prescribed to alleviate symptoms associated with menopause, have also been associated with elevated breast cancer risk [10].

1.3.2 Reproductive Risk Factors

It has also been shown that a number of reproductive factors can be associated with elevated breast cancer risk including, but not limited to, the age at different events such as menarche, menopause, and first childbirth. Women who experience menarche before 12 years of age and women experiencing late menopause (after 55) have elevated breast cancer risk. Having the first full-term pregnancy at a young age and the number of full-term pregnancies are inversely associated with breast cancer risk [11].
1.3.3 Risk Factors Related to Previous Breast Disease

Among women previously diagnosed with a breast cancer, between 10% and 30% develop a recurrence [12]. Women with a previous breast cancer diagnosis are also three to four times as likely to develop a new primary breast cancer compared to women without a diagnosis [13]. Furthermore, women who have atypical hyperplasia, a precancerous condition, is at higher risk [13, 14]. Women recalled for diagnostic tests yet with a negative results for breast cancer, or women who have a surgical biopsy in which the suspicious lesion was removed, are more likely to develop a breast cancer later in life [15, 16, 17, 18]. Furthermore, women with ovarian cancer are also more likely to develop breast cancer [19].

1.3.4 Congenital Risk Factors

It has been established that a familial history of breast cancer, ovarian cancer, and certain inherited genetic traits are associated with elevated breast cancer risk.

Women with a first-degree relative, e.g., a mother or a sister, with a history of breast cancer have approximately twice the risk of developing breast cancer. Women with two first-degree relatives diagnosed with breast cancer have thrice the risk of breast cancer [13]. Lastly, women with a first-degree relative diagnosed with ovarian cancer before the age of 50 are also at greater risk [13].

A lot of these familial breast cancer cases can be attributed to inherited genetic factors, some of which have been discovered. At least 68 loci, i.e., fixed positions in the human chromosomes, have been identified through genome-wide association studies, as accounting for 30% to 35% of the familial breast cancer risk [20]. The most penetrative gene mutations, in terms of breast cancer, are BRCA1 and BRCA2. Women with a BRCA1 alteration have an approximate 72% lifetime risk of developing breast cancer while women with a BRCA2 alteration have an approximate 69% lifetime risk [13]. Other moderately penetrative genes, in terms of breast cancer, include ATM, CHEK2, and PALB2 [20].

1.4 Risk Models Based on Clinical Risk Factors

In the following sections, the most influential breast cancer risk models, based on the clinical risk factors described above, are presented.

1.4.1 The Gail Model

The Gail model estimates risk using six clinical risk factors, which are usually self-reported using questionnaires [14]:

- Current age
- The number of previous biopsies
- A previous biopsy showing atypical hyperplasia
- Age at menarche
- Age at first live birth
The number of first-degree relatives with breast cancer

The relative risk of developing breast cancer was estimated using a multivariate logistic regression model and was projected to an absolute risk estimate. The relative risk could also be projected to a five-year risk estimate. The Gail model has shown moderate discriminatory accuracy compared to other risk models in the same category. It is well-suited to assess population-level risk but is less suited as a individual-level risk estimate [21, 22].

1.4.2 The Tyrer-Cuzick Model

The Tyrer-Cuzick (TC) model extends upon the Gail model to further includes genetic information, family pedigree, and a more complex statistical model. Specifically, the TC model estimates breast cancer risk by including information on BRCA1 and BRCA2 alterations in first- and second-degree relatives, i.e., grandparents, parents, aunts/uncles, siblings, and children [23]. The risk of having a genetic predisposition for breast cancer is combined with height, BMI, and the other clinical risk factors of the Gail model to estimate 10-year and absolute risk estimates. The TC model has a slightly better discriminatory accuracy than the Gail model and has more accurate risk calibration for women with history of breast cancer in the family [24].

1.4.3 The BOADICEA Model

The Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) breast cancer risk model further extends upon the idea of incorporating genetic risk factors beside BRCA1 and BRCA2 alterations. Specifically, a collection of low-penetrance genes was shown to have an additive effect explaining a larger percentage of familial breast cancers than only BRCA1 and BRCA2 [25, 26, 27].

The BOADICEA model has been updated recently to include family history, lifestyle, hormonal and reproductive risk factors, polygenic risk scores, mammographic density, and pathogenic variants. This is the most recent attempt to incorporate all known established risk factors into one model [28]. The discriminatory accuracy is better than the Gail or the TC model. However, the performance of the BOADICEA model comes with the price of an extensive data collection process. Women would need to have have a mammogram, fill out extensive questionnaires, and have their genome sequenced via blood samples for instance.

The Gail, TC and BOADICEA models all estimated breast cancer risk by using clinical risk factors. Section 1.7 will present and describe breast cancer risk factors that can be derived through mammography.

1.5 Preventing the Progression of Breast Cancer

Given the high incidence of breast cancer, poor prognosis for late stage breast cancer, and multiple breast cancer risk factors, it becomes imperative to take actions to prevent mortality on a societal level.
Breast cancer is a disease that evolves over time. In this thesis, the progression of breast cancer is defined as the time before onset, where the woman is healthy, to the time after onset of clinical symptoms. Breast cancer prevention is thus defined as the actions taken to stop the progression of breast cancer at any given phase [29]. Breast cancer prevention can be split into three phases as seen in Figure 2. Primary prevention is the measures taken to prevent the development of breast cancer in the first place. This includes reducing the above-mentioned actionable risks such as alcohol consumption and smoking while also preventing obesity, encouraging exercise, or having the first child at a young age. Secondary prevention is the actions taken to discover the breast cancer after the onset but prior to clinical symptoms occur. In this phase, mammography screening is used as a tool for secondary breast cancer prevention [29] (described in details in the next section). Tertiary prevention is the actions taken after onset of clinical symptoms or breast cancer diagnosis. In this phase, treatment consists of (neo)adjuvant therapy, surgery, radiation therapy, or a combination of these. The patient might also decline treatment and live with breast cancer in the remaining lifetime.

Figure 2: Timeline of a the three phases in breast cancer prevention including the primary, secondary, and tertiary phase. Some time the onset of breast cancer, pre-clinical manifestations become detectable on mammography. If screened using mammography, the breast cancer is likely to be detected before onset of clinical symptoms. The lead time is the time from early detection by mammography, to clinical symptoms occur or would have occurred.

1.6 Mammography Screening

The prognosis of breast cancer is largely determined by the tumor stage at diagnosis. It is therefore crucial to engage in secondary prevention and discover the breast cancer early to increase survival [30].

An established noninvasive tool for detecting clinical and pre-clinical manifestations of breast cancer is mammography. Mammography is a medical imaging modality that uses low-dose x-rays to create an image of the compressed breast tissue as seen in Figure 3.

1.6.1 Purpose of Screening

Mammography is sensitive to small non-palpable lesions in the breast tissue and can therefore be used as a secondary prevention tool to screen asymptomatic women for early signs of breast cancer. The aim is to screen women at regular intervals, e.g., once every one or two years, such that the screening
is likely to fall within the detectable time, as seen in Figure 2. That is, after the time pre-clinical manifestations start to occur and before the clinical symptoms occur or diagnosis. The time from the screening mammogram reveals a malignant lesion to the clinical symptoms occur or would occur is called the lead time. The longer the lead time is, the earlier the breast cancer have been detected. The purpose of mammography screening is to discover women who are at high risk of having a breast cancer and is therefore not a diagnostic test. Diagnostic tests will take place if the radiologists determine that the screening mammogram shows suspicious/abnormal findings indicative of breast cancer and decide to recall.

### 1.6.2 Evaluation of Screening Quality

It is important to measure the screening quality and effectiveness. Although screening is not a diagnostic test, many of the metrics and analyses used for diagnostic tests may be used to evaluate screening quality as well [29]. Given a binary screening recall decision and a binary result of the diagnostic test for breast cancer, a two by two contingency table can be created as shown in Table 1.

<table>
<thead>
<tr>
<th>Screening decision</th>
<th>Breast cancer</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Present</td>
<td>Not present</td>
<td></td>
</tr>
<tr>
<td>Recalled</td>
<td>True positives (TP)</td>
<td>False positives (FP)</td>
<td></td>
</tr>
<tr>
<td>Not recalled</td>
<td>False negative (FN)</td>
<td>True negatives (TN)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Two by two contingency table of the screening decisions (recalled vs. not recalled) and the results diagnostic test for breast cancer (present or not present)

Women in the TP group are generally referred to as having screen-detected cancers (SDC). Women in the FN group are generally referred to as having interval cancers (IC), as the cancer was found in the interval between screenings due to other events or clinical symptoms developed. ICs are either cancers that was over-looked at screening, which in many cases can attributed to masking of lesions in dense breasts [31], or cancers that was not yet detectable. Women in the FP group are sometimes referred to as noncancer recalls. Furthermore, women who had an SDC or IC in the subsequent screening round are also considered in this thesis. These cancers are referred to long-term cancers.
(LTC). All cancer groups, and how they are derived, have been further described in Section 3.1.1 and Figure 9.

From Table 1, the most common screening quality indicators can be calculated and used to quantify screening quality. Most of these quality indicators are considered in this thesis and especially in Paper I.

**Cancer detection rate (CDR)** is the fraction of women with breast cancer who was recalled at screening out of the total number of screened women \((N)\). 

\[ CDR = \frac{TP}{N}. \]

**False positive rate** is the fraction of women recalled at screening yet did not receive a breast cancer diagnosis out all women without breast cancer. 

\[ FP rate = \frac{FP}{FP+TN}. \]

Sometime the FP rate is calculated as the number of FPs out of the total population. 

\[ FP rate = \frac{FP}{N}. \]

**False negative/interval cancer rate** is the fraction of women with breast cancer who were not recalled at screening out of all with breast cancer. 

\[ IC rate = \frac{FN}{FN+TP}. \]

**Screening sensitivity** is the fraction of women with breast cancer who were recalled at screening out of all the women who had breast cancer. 

\[ sensitivity = \frac{TP}{TP+FN}. \]

A high sensitivity indicates that radiologists correctly recall a high portion of the women who had a breast cancer.

**Screening specificity** is the fraction of healthy women who were not recalled at screening out of all women without breast cancer. 

\[ specificity = \frac{TN}{TN+FP}. \]

A high specificity indicates that radiologists are correctly identifying normal or benign mammograms that does not warrant recall or diagnostic tests.

**Positive predictive value (PPV)** is the fraction of women with breast cancer who were recalled at screening out of all women who were recalled. 

\[ PPV = \frac{TP}{TP+FP}. \]

A high PPV indicates that a large portion of the recalled women end up receiving a breast cancer diagnosis.

**Negative predictive value (NPV)** is the fraction of women without breast cancer not recalled at screening out of all women who were not recalled. 

\[ NPV = \frac{TN}{TN+FN}. \]

A high NPV indicates that a large portion of women not recalled also did not receive a breast cancer diagnosis.

Besides the mentioned indicators, a series of other indicators are used the measures screening quality. The tumor size of an invasive breast cancer, as measured after surgery, is an indicator of early detection. A small tumor size would generally suggest that the cancer have been detected early. Lymph nodal status is also an indicator of early detection. Generally, if the breast cancer is detected early, spread to the surrounding lymph nodes is less likely. These two indicators of lesion size and spread to lymph nodes are not explored in this thesis.
It is important to note that all the mentioned indicators depend highly on the screening interval and the length of follow up. In Europe, most screening rounds are an average of two years in length, whereas in the US, most screening rounds are one year in length [32]. This should be taken into account when reporting screening quality using the mentioned methods.

1.6.3 Harms and Benefits of Screening

The first large trials on using mammography in screening for breast cancer was carried out between the 1960s and the 1980s. The studies showed a reduction in breast cancer mortality by up to 30% for women between the age of 50 and 69 who were screened compared to women who were not screened [33]. In the early 1980s, 1990s, and early 2000s, the American Cancer Society, the UK National Health Service, and the European Council all started recommending recurring mammography screening for early detection of breast cancer in asymptomatic women [34, 35, 36]. After these recommendations were made, screening mammography has continued to be rolled out in large parts of the western world. Today, there are a substantial amount of recent evidence and studies that find recurring breast cancer screening using mammography decreases mortality [33, 37, 38, 39, 40]. In a review by the EUROSCREEN Working Group on several European breast cancer screening studies, the pooled estimates of mortality reduction were 25% in incidence-based mortality studies and 31% in case-control studies. The pooled estimates of mortality reduction for women who chose to participate in screening were 38% and 48% in incidence-based an case-control studies, respectively.

A general concern in screening is overdiagnosis. Overdiagnosis occurs when a breast cancer might not progress to produce clinical symptoms or become life threatening within the remaining lifetime. Equivalently, the breast cancer would not have been diagnosed within the remaining lifetime if not for screening. The diagnosed woman would have been overdiagnosed and may have received unwarranted treatment. Although it is difficult to measure the rate of overdiagnosis, it is estimated that 1% to 10% of breast cancer cases are overdiagnosed in Europe and in the UK [41, 42].

X-rays are ionizing radiation that might, in very few women, induce a breast cancer. It has been estimated that in 100,000 women screened annually from the age of 40 to 55 and biannually from 56 to 74, approximately 86 women would develop a radiation-induced breast cancer [43]. Other studies that simulated radiation risk found that 27 and 49 women per 100,000 screened would develop a radiation-induced breast cancer in biennial and annual screening, respectively [44, 45]. Yet, the benefit of screening still outweighs the risk of developing radiation-induced breast cancer [43]. Nonetheless, screening should be moderated so that women do not receive a higher cumulative x-ray dosage than necessary.

For a large number of women recalled at screening due to suspicious/abnormal findings, the diagnostic imaging or biopsy will reveal normal or benign conditions. In Denmark, in a period between 2014 and 2016, only one in four recalled women received a diagnosis [46]. A lot of diagnosed women will experience discomfort and anxiety in the process of diagnostic testing and waiting for the result.
Lastly, there exists a number of statistical biases that clinicians and researches should be wary of when reporting results. These biases might obscure the underlying benefits and harms by screening. Lead time bias occurs when survival seems to improve due to screening, however, it was the earlier detection and diagnosis that falsely seemed to increase lifespan. Length time bias occurs as very aggressive and fast growing cancers are more likely to be detected after the onset of clinical symptoms while slow growing cancers are more likely to be detected with pre-clinical manifestations in screening. As a consequence, the estimates of survival/mortality might be inaccurate. Selection bias occurs as women who are participating in screening might differ from women who decline screening in terms of risk and this might obscure the estimated survival/mortality.

### 1.6.4 Image Acquisition

A screening mammogram typically comprise four standard projections, referred to as views: A mediolateral oblique (MLO) view and a craniocaudal (CC) view for each breast. During screening, radiographers capture an MLO view by compressing the breast at a 45° angle, where the breast and armpit are compressed uniformly. To capture a CC view, the radiographer compresses the breast directly from the top down (see Figure 3). This process generates four views shown in Figure 4.

![Figure 4: Screening mammography of a 51 year old asymptomatic women. No suspicious findings were made by radiologists. Top-left: L-MLO, top-right: R-MLO, bottom-left: L-CC, and bottom-right: R-CC. It can be clearly observed that different tissue types within the breast attenuate x-rays to varying degrees. The resulting views contain characteristics of the breast tissue which can be examined by radiologists.](image)

In recent years, most screening programs in western countries have switched, or are planning to switch, from screen-film (SFM) to full-field digital mammography (FFDM) [47]. Opposed to SFM, FFDMs can more easily be stored digitally in large quantities, be transferred to other clinics, be retrieved by radiologists who wish to look at prior mammograms, support new technology such as
computer aided diagnostic (CAD) systems, and allow for custom post-processing.

The raw/untouched output from the FFDM systems is not suitable for human interpretation. Vendors of FFDM systems usually provide a post-processing algorithm that, through non-linear image transformations, enhances global and local contrasts. The resulting processed views are interpretable by radiologists. Post-processing algorithms vary from vendor to vendor, and the transformation parameters also depend on the screening site's/clinic's preferences.

1.6.5 Suspicious Mammographic Findings

A malignancy can manifest in the mammography in many different ways from very large, dense, and high-contrast lesions to very small lesions that are almost indistinguishable from the background tissue. Suspicious lesions are usually categorized as either soft-tissue lesions or calcifications or a combination. Soft-tissue lesions are characterized by radiologists by shape, density (pixel intensity), orientation, and location among other characteristics. Soft-tissue lesions appearing spiculated, irregular, and with high density are likely malignant [48]. Calcifications are calcium deposits in the breast which are very common, more prevalent in postmenopausal women, and usually benign [49]. However, pleomorphic calcifications arranged in certain segmental/branching patterns with irregular shapes or borders can indicate a malignancy or a pre-stage to breast cancer (DCIS).

1.7 Mammographic Risk Factors

It turns out that the mammography in itself contains information about the screened individual including age, race, body weight, and most importantly, breast cancer risk [50, 51, 52]. Specifically, the mammography contains independent information about immediate, short-, and long-term breast cancer risk. This thesis will focus primarily on risk factors derived via mammography, i.e., the imaging data.

1.7.1 Mammographic Dense Tissue

By far the most breast cancers develop in epithelial or stromal tissues of the breast. Epithelial cells form a protective lining on the surface of all internal structures in the breast including glands and ducts. The stroma is defined as the non-functional connective/structural parts of the breast that support normal function. It is therefore reasonable to deduce that a woman with a larger amount of epithelial or stromal tissues is more likely to develop a breast cancer [4]. Epithelial or stromal tissues are together referred to as dense breast tissue, which becomes apparent on a mammography. Dense breast tissue attenuates x-rays and thus absorbs more photons than the fatty tissue. Consequently, dense and fatty tissues appear distinct from each other in the mammogram. The amount of mammographic dense tissue can be assessed and measured in various ways and used as a breast cancer risk factor. However, having high breast density can also create a masking effect that lowers screening sensitivity. Radiologists are more likely to miss a malignant lesion in women with high breast density as the lesion might be partially or fully masked by the surrounding dense tissue during compression.
Conversely, screening sensitivity is higher for women with low breast density as lesions are less likely to be masked.

It should be noted that breast density varies considerably between women and is affected by multiple factors including age, race, BMI, and hormonal treatments [54, 55, 56].

### 1.7.2 Wolfe, Boyd, Tabár Classification Systems

The first attempts to measure mammography-based breast cancer risk were the Wolfe, Boyd, and Tabár classifications systems. Radiologists would use these systems to manually assign women into distinct groups based on the mammographic appearance of dense tissue. The Wolfe, Boyd, and Tabár classification systems have been instrumental to the development and improvement of later risk models.

Note that in early literature, breast density would refer to both the amount and appearance of mammographic dense tissue. Today, and in this thesis, a clear distinction is made between breast density, which usually refers to the amount of dense tissue, and the appearance of dense tissue which concerns the distribution throughout the breast (mammographic texture, see Section 1.7.4).

J. Wolfe recognized that the amount and appearance of dense tissue could be associated with breast cancer risk. In 1976, these findings resulted in a manual classification system with four groups of increasing risk [57, 58]:

- N1: mainly fatty breasts
- P1: mainly fatty breasts with linear densities covering less than 25% of the breast
- P2: linear densities covering more than 25% of the breast
- Dy: Radio-dense breasts

The original study reported a 22-fold increase in risk between women categorized as Dy compared to N1. In a later review from 1984, relative risks between 0.06 and 8.02 were reported when using the Wolfe classification system, however, the most credible studies of the review proved an associating between certain breast parenchyma patterns and breast cancer risk [59].

In 1995, N. Boyd published a study in which women were categorized by percentage mammographic density (PMD). PMD is defined as:

\[
PMD = 100\% \cdot \frac{\text{Dense tissue area}}{\text{Total breast area}}
\]

The Boyd classification consisted of six PMD groups: 0%, <10%, 10%-25%, 25%-50%, 50%-75%, and >75% [60]. Women in the highest PMD group had a relative risk between 4.04 and 6.05 compared to women in the lowest PMD group. This made PMD one of the strongest predictors of breast cancer.

However, this level of relative risks using breast density have not been replicated in recent studies of large screening cohorts. A recent study found that women with extremely dense breasts had a 2.5 age-adjusted relative risk of breast cancer compared to women with almost entirely fatty breasts [61].

In 1997, L. Tabár created yet another classification system based on appearance of four components: nodular densities, linear densities, homogeneous fibrous tissue, fat tissue [62, 58]:

...
• I: balanced proportion of all components with a slight predominance of fibrous tissue.
• II: mainly fatty tissue
• III: mainly fatty tissue with retroareolar residual fibrous tissue
• IV: mainly nodular densities
• V: mainly fibrous tissue

Category IV and V were considered high-risk and the remaining categories were considered low-risk. The Tabár classification has been shown to correlate highly with the Wolfe and Boyd classification systems [63].

1.7.3 Quantitative Measures of Mammographic Density

A large number of more recent methods to quantitatively or semi-quantitatively measure breast density have been proposed and used in literature.

The Boyd classification system was later standardized as a part of the Breast Imaging-Reporting and Data System (BI-RADS 4th edition). Using BI-RADS density descriptors, women are assigned to one of four groups based on their breast density: (1) Almost entirely fatty, (2) scattered areas of fibroglandular density, (3) heterogeneously dense, and (4) extremely dense [64, 65]. Most women have low to moderate breast density. For instance, in Denmark approximately 70% of screened women have BI-RADS density 1 or 2 while less than 5% have BI-RADS density 4 [61]. The BI-RADS density scale is currently in use in some breast cancer screening programs, usually for research purposes [66]. Examples of each BI-RADS density category are shown in Figure 5.

![Figure 5: Examples of four MLO views in BI-RADS (4th edition) density category 1, 2, 3, and 4, respectively. The left-most breast appear almost entirely fatty (PMD < 25%) whereas the right-most breast is extremely dense (PMD > 75%).](image)

BI-RADS density remains a subjective measure and can vary to a large degree between radiologists [67]. Yet, breast density may also be measured automatically or semi-automatically using computational models. An example of a semi-automatic method is the Cumulus tool which allows a radiologist to select an intensity threshold that separates dense from fatty tissue [68]. Modern fully automatic breast density estimation tools rely on deep learning models [69, 70]. PMD may in some
cases also be calculated with approximations of tissue volumes instead of areas [71, 72, 73]. Examples of computational models for breast density estimation are presented and described in Section 1.8.2

1.7.4 Mammographic Texture

It turns out that the distribution of dense tissue within the breast, and not solely the amount, is a breast cancer risk factor as well. Specifically, the distribution of the breast parenchyma, defined as the functional anatomy of the breast such as lobules and ducts, is associated with risk [74].

The above-mentioned Wolfe and Tabár classification systems were among the first to associate distinct parenchymal patterns with elevated breast cancer risk [57, 62]. These systems described distinct appearances including the prominence of ducts, nodular densities, linear structures/densities, fatty tissue appearance [62, 58]. However, these classification systems relied on the radiologist’s subjective decisions on mammographic appearance.

Computational models are capable of automatically measuring tissue patterns associated with elevated risk in an objective fashion. Such models may rely on manually selected features or features derived automatically using deep learning. Common for both is that the data-driven measure correlate visually with Wolfe and Tabár patterns. Specifically, high-risk women seem to stand out from low-risk women by their breast tissue appearing more heterogeneous, having high-contrast structures and linearities, and larger dense tissue areas, whereas low-risk women have more homogeneous tissue and less high-contrast structures. Features derived by computational models do not necessarily correspond to known biological processes or anomalous breast anatomy but rather corresponds to mammographic features that are statistically relevant for the task, i.e, risk assessment. In this thesis, data-derived mammographic features of breast tissue heterogeneity are referred to as mammographic texture [74]. Figure 6 displays four mammographic views of four women with low and high texture (measured by a deep learning model) vs. women with low and high BI-RADS density.

1.7.5 Miscellaneous Mammographic Findings Associated with Risk

Mammographic views may contain other artifacts that are associated with elevated breast cancer risk.

As mentioned in Section 1.3.3, screening false-positive women have elevated breast cancer risk [18, 15, 16, 17]. Recalled women will, as part the diagnostic tests, have a biopsy taken of the suspicious lesion(s) given diagnostic imaging deems it necessary. During a biopsy, a tiny metal marker (clip) may be left at the site of the tested lesion. In case of a positive outcome for breast cancer, these clips can help radiologists to locate the malignant lesion preoperatively. In case of negative outcome for breast cancer, the clips are left in the breast and will appear as small, bright, and distinct objects in future mammograms.

Women that have had breast conserving surgery will most likely also receive radiation therapy. Surgical clips will be left in the breast that locate the area of the removed lesion such that radiation can be accurately delivered. After treatment, the women will likely return to screening where the surgical clips will be visible in the mammogram.
The presence of clips, indicating a previous false-positive biopsy or a previous breast cancer, may therefore entail an elevated breast cancer risk. However, not all false-positive women will have clips in their breast and should not be considered a true breast cancer risk factor. Yet, presence of clips should be considered when developing mammography-based risk models.

Implanted medical devices such as pacemakers, cardioverter-defibrillators, and loop recorders may appear as bright objects in a mammography as well. Having an implanted medical device might indicate a cardiovascular disease which is known to share a number of risk factors with breast cancer [75]. Furthermore, medical implants might leading to reduced contrast in mammographic views [76]. In this thesis, the breast cancer risk associated with implanted medical devices is not explored.

1.8 Early Computational Models for Mammography-based Breast Cancer Risk

Computational mammography-based breast cancer risk models derive risk estimates directly from the mammography through image analysis techniques in a data-driven approach. These models do not rely on clinical risk factors such as medical history, personal/familial/reproductive history, self-reported information, or genetics. Neither do they rely on manual classification by radiologists or experts. The following sections will present early and more recent mammography-based computational models for breast cancer risk.
1.8.1 Classifying Wolfe Patterns

Very early mammography-based risk models, developed between 1985 and 1995, primarily focused on manually selecting simple image features that could be extracted from digitized SFMs [77, 78, 79]. The feature space would be grouped or linearly combined to align with Wolfe patterns. Wolfe’s classification was the predominant model for breast cancer risk at the time. Manually selected features would include the dynamic ranges of intensities, variance, histogram skewness, kurtosis, entropy, and similar image descriptors. More advanced analyses were used as well in attempts to create more expressive features spaces such as fractal dimensions and frequency (Fourier) features [78, 79].

It was shown that such feature spaces could be constructed and, to some extent, grouped/aligned to Wolfe patterns. However, the data samples were very limited (less than 120 mammograms), and the agreement with radiologists were occasionally poor [78, 79]. However, these studies laid the ground work for future studies.

1.8.2 Quantification of Mammographic Density

Between 1994 and 2006, several studies were published on assessing breast cancer risk through quantitative measurement of mammographic breast density.

The first studies offered a tool for experts, e.g., radiologists, that allowed them to interactively choose an intensity threshold in digitized SFMs that separated dense from fatty tissue. Specifically, an expert would, on a computer workstation, manually increase the threshold until all the radio-dense tissue was covered. This produced a segmentation mask with pixel-wise classification of dense and fatty tissue [80, 68]. It was confirmed, through these studies, that breast density was a risk factor [68]. The intra- and inter-reader agreements were higher than semi-quantitative classifications such as Boyd’s density classification. Even a naive reader would have a high agreement with an experienced radiologist. Additionally, the density estimates seem to correlate highly with Boyd’s density classification [80]. Due to these reason, interactive thresholding (later known Cumulus) became the gold-standard for breast density estimation for many years.

Besides the semi-automatic thresholding methods, several studies focused on automatically estimating breast cancer risk either via segmentation or classification into known scales, e.g., BI-RADS density.

N. Karssemeijer developed a model for automatic classification of mammograms into four density categories (PMD: <5%, 5–25%, 25–75%, and 75–100%) by automatically detecting the border to the pectoral muscle, computing variance, skewness, distance to the skin line, and other descriptors. These image descriptors were combined using a K-nearest neighbors (KNN) classifier. The accuracy of the automatically estimated PMDs, compared to the manually quantified PMDs, was a modest 67%, but increased to 80% in mammograms obtained on a mammographic device that was newer at the time [81]. Other studies on (near-)automatic density estimation, based on various segmentation or classification methods, followed with decent results. However, the studies were small with less than 200 subjects. Nonetheless, the studies proved that density could be estimated objectively and reliably [82, 83, 84].
Later, studies focused on estimating the dense tissue volume as opposed to area. It was hypothesized, that area-based models could not adequately capture complex fatty/dense tissue interactions over time, and that volume-based models would be more expressive \[71\]. Since x-rays are attenuated differently through fatty and dense tissues, a physical model can be used to approximate the dense tissue volume \[72\]. However, such models might need extensive calibration using phantoms and was not directly clinical usable. Still, the studies provided the groundwork for volume-based models which are in use today.

After FFDMs became more available, volume-based models became more reliable as information on tube voltage, anode material, filtration, and compressed breast thickness could be directly extracted from the image header. S. van Engeland et al. provided a method for mapping intensity to dense tissue thickness using image acquisition parameters \[73\]. This volume-based model was compared to density estimated in MRI, considered the gold-standard, which yielded a very high correlation coefficient of 0.97. However, it later turned out that the volumetric and Cumulus estimated density related equally to clinical breast cancer risk factors such as age, BMI, nulliparity, and more \[85\].

### 1.8.3 Predicting BRCA Alterations

Having BRCA1 or BRCA2 alterations has been associated with elevated breast cancer risk as mentioned in Section 1.3.4.

In 1999 and 2002, two studies were published on using textural features, derived through classical image analysis techniques, to assess risk via BRCA alterations. The authors extracted a single large region of interest (ROI) from behind the nipple and measured minimum/maximum/average intensities, histogram balance, skewness, contrast, coarseness, and frequency features and combined them using linear discriminant analysis (LDA) to predict BRCA alterations from the mammogram \[86, 87\].

This model was, with high accuracy, able to identify women with BRCA alterations \[87\]. The results showed that high-risk women tended to have dense breasts and their breast tissue appeared more coarse than low-risk women. However, the model’s output only correlated to a low degree with risk, as measured by the Gail model \[86\].

### 1.8.4 Texture Models for Direct Risk Assessment

From 2008 to 2016 a series of studies were published on developing mammographic texture models for direct breast cancer risk assessment without estimating specific physical effects such as those measured by Wolfe patterns, density, or BRCA alterations.

The theoretical background for the texture models was published in a study from 2008 on developing a generic machine learning (ML) framework for assessing specific biological effects \[88\]. This framework was applied to the effect of HRT and age on mammographic appearance. As with many of the previously described computational risk models, the framework relied on extraction of selected image features using image analysis techniques. Specifically, the Hessian matrix was computed for every breast tissue pixel at three scales (1, 2, and 4 mm) to measure local curvature. The eigenvalues $e_1$ and $e_2$ of the corresponding two first eigenvectors, that align with the second-order structures in
the breast, were computed. The *stripiness* features, as described by the authors, are then measured by the follow ratio:

$$q_s = \frac{|e_1| - |e_2|}{|e_1| - |e_2| + \epsilon},$$

where $\epsilon$ is a small constant for numerical stability and $q_s$ is the ratio as scale $s$. Intuitively, this measure searches for elongated structures in the breast tissue at the location of the pixel. The $q_s$ feature vector was extracted from 10,000 randomly chosen locations in the breast tissue from the mammograms. A KNN classifier was trained, on a subset of mammograms, to identify pixels from mammograms of women with HRT vs. placebo. During inference on a testing mammogram, 10,000 features vectors would be extracted. Using the KNN model, each vector was given a probability of belonging to a mammogram of a women receiving HRT. All probabilities were averaged to produce a final continuous measure interpreted as likelihood for HRT.

The described model, with additional scales and distance features, was used to train for and assess breast cancer risk in a sample of 495 women. 245 women developed a breast cancer two to four years from the baseline screening and 250 women remained healthy for at least four years after the baseline screening. All mammograms used for training and validation were from the baseline screening visit. The results on a small sample of SFMs showed that the mammographic texture resemblance (MTR) marker achieved a higher AUC than when using BI-RADS density or PMD [89].

Essentially, a ML model was trained to directly measure breast cancer risk and was a better predictor than than BI-RADS density or PMD. The MTR marker was, to some degree, shown to be independent from breast density [89, 90].

Two other studies were published in the same period on models for direct breast cancer risk assessment but did not seem to have the same discriminatory abilities [91, 92].

### 1.9 Deep Learning Models for Mammography-based Breast Cancer Risk

Modern mammography-based risk models are based on deep neural networks trained on large cohorts of mammograms. In the following sections, I provide background information on deep learning in medical imaging and how such models may be used to estimate breast cancer risk.

#### 1.9.1 Deep Learning in Medical Imaging

Deep learning (DL) is a collection of ML methods that allow a computational model to automatically learn highly complex patterns in the input data and make accurate predictions about that data. DL models are based on deep artificial neural networks (DNN) which consist of several layers of processing units, called neurons, which is inspired by biological networks in the brain. The layers-wise architecture of DNNs allows the model to learn task-relevant features in the input data at different levels of abstraction in a completely automatic fashion [93]. As opposed to conventional ML methods, there is no need to manually select image features for analysis as they are derived automatically with DL models.
Since DNNs, the backpropagation algorithm, and stochastic gradient descent could be efficiently implemented on computationally fast hardware, DL has presented effective solutions to many medical tasks including detection of breast cancers and assessment of breast cancer risk in mammography [94, 69, 95, 96].

### 1.9.2 Deep Learning Texture Model for Risk

The first attempts to directly assess breast cancer risk using DL models was carried out by K. Petersen, M. Kallenberg et al. between 2014 and 2016 [97, 69]. Essentially, the already developed framework for breast cancer risk, described in Section 1.8.4, was used as inspiration for a DL texture model.

Specifically, the authors defined a convolutional sparse auto-encoder (CSAE) network that was pre-trained in an unsupervised fashion and later fine-tuned for risk assessment. The network was pre-trained using a large number of image patches extracted at multiple scales from the mammograms. Each scale patch was passed through the network to obtain latent feature maps which were combined for each scale patch. The multi-scale feature map was then passed through a decoder to reconstruct the original input patch. This allowed the network to be trained in an unsupervised fashion and learn mammographic features without any labels. The network architecture is shown in Figure 7.

![Figure 7: The convolutional sparse auto-encoder (CSAE) network architecture used for breast cancer risk and density estimation. This image have been modified from the original publication with permission from the last author. Citation: M. Kallenberg et al., “Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring,” in IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1322-1331, May 2016, doi: 10.1109/TMI.2016.2532122.](image_url)

The pre-trained CSAE, now with a logistic regression layer at the end instead of the decoder, was fine-tuned to assess breast cancer risk using cancer contra-lateral views or for dense tissue segmentation. For risk assessment, 500 patches from the breast tissue were extracted, each patch passed through the network to obtain 500 posterior probabilities of cancer. The patches used for training were extracted solely from the cancer contra-lateral view to exclude diagnosed malignant lesions. The 500 posterior probabilities were averaged to produce a single continuous risk estimate per view. The CSAE performed better than the previous texture model based on the stripiness features [69]. For dense tissue segmentation, a patch for each pixel in the mammogram was extracted such that the model could be applied in a sliding window to produce a full segmentation mask.
1.9.3 Conflated Breast Cancer Risk Models

Since 2019, a series of studies have been published on DL models for assessing breast cancer risk using mammograms. These studies differ from the DL texture model, described previously, in many aspects. However, most important is the choice of training data. Specifically, how future cancer are defined in terms time from the baseline screening visit, and whether diagnosed or potentially present malignancies should be included.

A. Yala et al. collected a training sample of 31,806 women screened between 2009 and 2012 in the US. All mammograms used to train and validate the DL ResNet model were obtained in this period and are referred to as the baseline mammograms. For each woman, the authors collected information on follow-up and whether the women remained healthy for five years (considered a negative label) or developed cancer within five years (considered a positive label). The model was then trained using these binary labels and the baseline mammograms [96]. Clearly, the model will, during training, encounter positive samples that have visible cancers, samples with minimal signs of a developing cancer (precursors), and samples for which the cancer has not yet manifested in the mammogram. Consequently, the training procedure was conflated and the model was optimized to detect present, developing, and future cancers simultaneously [96, 98]. Women who developed cancer within one year from the baseline mammogram were excluded from the validation dataset where the model achieved an AUC of 0.68. In later studies, the model was extended with further modules to combine multiple latent features maps and to predict risk at multiple intervals after the baseline mammogram. The AUCs achieved by the newer model were approximately 0.70 on average in seven demographically different samples, however, the training procedure remained conflated [99].

Dembrower et al. also trained a DL ResNet model to assess breast cancer risk using a Swedish screening sample of 9,969 women. However, women diagnosed with a breast cancer within one year were excluded from the training and validation datasets [100]. Consequently, the model will have encountered positive samples with precursors and positive samples without any pre-clinical manifestations during training. However, it is a lot less likely that the model will have encountered visible cancers although it might still be possible. The training procedure was therefore also conflated to rely on more than healthy breast tissue, although to a lesser degree than the risk model developed by A. Yala et al.. In a testing dataset with women that remained healthy for at least one year, the model achieved an AUC of 0.65 in women with a mean follow-up time of 3.6 years.

After these two studies were published, it was suggested that training models for breast cancer risk using a conflated training procedure might inhibit the models from learning risk-predictive features of healthy tissue [101].

1.9.4 Models for Breast Cancer Detection and Short-term Risk

The intended purpose of a AI-based system for breast cancer detection is to present visual material that support radiologists and clinicians during screening or diagnostic testing. As with the previously mentioned DL models, modern AI systems typically consist of one or several convolutional neural
networks (CNN). However, the models have been trained to delineate or mark suspicious lesions in the input mammogram. A map with markings is generated using a CNN that segments lesions or a model that automatically detect ROIs with a possible lesion [102]. The AI system might also employ a classification model that produces a score for each of the marked lesions reflecting the likelihood of malignancy [103, 104]. The markings and likelihoods are overlayed the mammogram and shown to the radiologist.

AI systems for breast cancer detection now perform, in a stand-alone setup, on par with average radiologists and are therefore useful in screening as an extra reader, as decision support, or to triage mammograms as likely normal or warranting recall [103, 105, 106, 107].

Although such AI systems are developed for detection, they may be repurposed for short-term breast cancer risk [108, 109]. It has been suggested that many screen-detected cancer are visible at the previous screening round one to two years prior [110]. These cancers might have been present as minimal signs of developing cancer (precursors), masked cancers in dense breasts, or benign-looking findings that did not warrant recall. A sensitive AI system, able to detect precursors before radiologists deem them likely malignant, can therefore be used as a proxy for short-term risk.

It has been shown that AI systems for lesion detection are excellent at identifying women with a high likelihood of cancer within two years from screening with AUCs ranging between 0.68 and 0.79 [108, 95].
2 Aim of PhD project

This section will present the aim of this PhD project by describing some of the current challenges in mammography screening and how personalized screening, based on modern risk models, might alleviate some of these challenges.

2.1 Challenges in Mammography Screening

Despite the fact that breast cancer screening successfully decreases the overall breast cancer mortality, implementing such a population-based screening program might entail a number of practical limitations and challenges.

Firstly, as mentioned in Section 1.6.3, x-rays are ionizing radiation that might cause a radiation-induced cancer in few women [43, 44, 45]. Thus the cumulative x-ray dosage should be moderated.

Secondly, as mentioned in Section 1.6.3, some women are overdiagnosed as they receive a breast cancer diagnosis and treatment which was not beneficial in the remaining lifetime [41, 42].

Thirdly, population-based breast cancer screening entails a considerable workload for radiologists in two aspects: During reading of screening mammography and during diagnostic testing of recalled women. In Europe, mammograms are usually read blindly by two radiologists and recalled women undergo clinical examination, mammography/tomosynthesis, ultrasound, and a needle biopsy in case of suspicious and/or palpable findings. In the Capital Region of Denmark alone, the target group is about 220,000 women and all participating women are read independently by at least two full time breast radiologists. Between 2014 and 2016, less than 3% were recalled due to suspicious findings and approximately one in four recalled women tested positive for breast cancer [46]. Consequently, radiologists spent large amounts of time reading screening mammograms where only a very small fraction of the mammograms had suspicious findings and even fewer led to a diagnosis.

Fourth, there is a general lack of radiologists worldwide including specialized high-volume breast radiologists [111, 112, 113, 114]. This might be due breast cancer screening being implemented more widely, an aging population, and the general high requirements for radiologists.

Fifth, most women going through screening, and especially women recalled for diagnostic tests, experience discomfort and anxiety when waiting for the results of a screening or a biopsy [115]. Considering the large number of false positive screenings, there might be potential to prevent some of the discomfort and anxiety for women.

Sixth, there is no worldwide standard way of doing breast cancer screening. For instance, most European programs are population/region-based, screen women biennially, and usually have at least two independent readers. Women in the US are generally screened annually in smaller breast clinics by a single radiologist [32]. The European guidelines for quality assurance in breast cancer screening and diagnosis, developed by the European Commission, specify comprehensive guidelines and requirements that breast cancer screening programs should follow. In Denmark, the regional screening programs adhere rigorously to the European guidelines, but adherence varies throughout Europe [116]. It has been suggested, with moderate certainty, that adhering to these guidelines was associated with improved survival [117].
2.2  Personalized Screening Based on Risk

All the above-mentioned challenges warrant a new approach to more effective screening without compromising on screening quality and safety for women. It is well-known that women have very different individual risks of breast cancer as described in section 1.3 and 1.7. Given a woman’s individual risk could be measured reliably and accurately, clinicians might be able to create safe personalized screening strategies as outlined in Figure 8.

![Figure 8: Schematic of how a mammography-based risk model could be used to categorize women by individual risk and used for different clinical interventions.](image)

Women that are categorized as low-risk might enter breast screening with longer screening intervals, e.g., every third or fourth year. Low-risk women might also be read by a single radiologist instead of two. Women categorized as high-risk may be directly recalled for diagnostic tests or supplemental screening. Supplemental screening might involve tomosynthesis or MRI, both of which are more sensitive than mammography but also more costly. High-risk women might further enter screening with shorter screening intervals, e.g., every year. In any case, all screened women could be read by an AI system as an additional reader and provide decision support to radiologists.

Breast cancer screening can be viewed as a zero-sum game, especially in national/regional-funded screening in Europe. Usually, screening programs have to operate with a limited amount of resources including time, budget, clinicians, radiologists, radiographers, modalities, and equipment. Risk models might help redirect the limited resources from low- to high-risk women. If a sufficient amount of resources can be safely saved from screening low-risk women differently, the resources could be spent for supplemental imaging of a few high-risk women or women where masking is a concern due to breast density. This is achievable since most women are not at high risk of developing breast cancer or have extremely dense breasts.

These reading/workflow optimization should at least maintain screening quality and safety for the participating women. However, such protocols might even facilitate increased detection rates, earlier detection, less false positives, and reduce discomfort and anxiety for low-risk women without increasing total cost.

In turns out that reliable diagnostic AI systems are available for breast cancer detection. Such systems might be directly implementable in screening sites or clinics. DL provides powerful frameworks to measure likelihood of specific biological effects such as assessing risk of a future breast
cancer. Both diagnostic and risk models might support safe personalized risk-stratified screening.

2.3 Contribution Hypotheses

This thesis sought to develop, advance, and evaluate tools to reliably assess breast cancer risk. The main focus was to develop clinically usable risk models to support clinicians in realizing personalized breast cancer screening using only information available at screening, e.g., the screening mammogram and age. Specifically, four papers and six abstracts are included in this thesis, contributing to three aspects of breast cancer screening, risk, and recurrence with the following hypotheses:


We used a commercially available AI system to detect normal, moderate-risk, and suspicious mammograms. Women with normal mammograms were excluded from radiologist reading, moderate-risk mammograms were double-read by breast radiologists, and women with suspicious mammograms were recalled directly. We hypothesized that this screening protocol could decrease workload without compromising screening quality. Additionally, **Abstract III** and **Abstract VI** presented the initial results that led to Paper I.

**Abstract I: A Prospective Study of Breast Cancer Screening with AI as First Reader for Likely Normal Mammographies** *(Accepted and presented orally at RSNA 2022)*

Following the simulation study, presented in Paper I, the AI system was clinically implemented in the breast cancer screening program in the Capital Region of Denmark. In this abstract, we presented the preliminary results following the clinical implementation. Specifically, women with mammograms classified as likely normal by the AI system are only read by a single radiologist. The remaining women are double-read by two radiologists.

We hypothesized that the preliminary results would indicate that the high screening quality was preserved, yet with reduced reading workload for radiologists. Note, that the results of the abstract have changed as the study period was extended and more data have been collected. The most recent results have been summarized in the key results in Section 5.2.

**Abstract II: Validation of a Deep Learning-based Breast Density Estimation Tool on a Danish Screening Cohort in the Context of Personalised Risk-based Screening** *(Accepted and presented orally at ECR 2020)*

We trained a U-Net model for automatic segmentation of dense breast tissue, non-dense tissue, and the background/pectoral muscle in mammographic views using 500 radiologists annotated view/mask pairs. We hypothesized that this segmentation model could robustly estimate planimetric PMD associated with elevated breast cancer risk. This segmentation model was later used in Paper II and Paper III.

We developed a mammographic texture model trained to estimate long-term breast cancer risk while adapting to mammographic devices from other vendors than the vendor used for training. We hypothesized that this texture model would be able to assess risk five years after screening in an independent validation screening cohort and identify women at high-risk that might receive supplemental screening. Additionally, Abstract V presented initial results of training the texture model which led to Paper II.

**Paper III: Breast Cancer Risk Assessment Improved by Combining Artificial Intelligence for Lesion Detection and Mammographic Texture** *(submitted to Radiology)*

Conflated risk models trained to simultaneously estimate short- and long-term breast cancer risk might yield subpar performance. We hypothesized that two models, a diagnostic AI system relying on subtle local signs of developing cancer for short-term risk, and a mammographic texture model relying on global features for long-term risk, both trained separately and optimally would complement each other in increasing overall risk assessment. We aggregated the two components using a combination model with age and PMD to identify clinically relevant high-risk women eligible for supplemental screening. Additionally, Abstract IV, was a preliminary study on combining an AI system and a texture model that led to Paper III.

**Paper IV: Identifying Recurrent Breast Cancer Patients in National Health Registries using Machine Learning** *(Recommended for publication in Acta Oncologica - revisions pending)*

Approximately 10% to 30% of breast cancer patients experience a recurrence but are often not flagged as such in national health registries. We developed a general machine learning framework using a simplistic encoding scheme and a machine learning model trained using patient-specific events in health registry data. We hypothesized that this model could retrospectively identify patients at high risk for breast cancer recurrence. The general framework might be useful in settings where clinicians are manually identifying patients and might further be used prospectively to flag low- or high-risk recurrence patients in real time.
3 Data Materials

The following sections present data materials used in the papers and abstracts of this thesis. Preliminary methods, relevant for each of the contributions, are described prior to the contributions in Section 4.

3.1 Screening in the Capital Region of Denmark

DL models require large amounts of training data to successfully converge and generalize to unseen data. Additionally, models meant to be clinically usable must be validated in large cohorts of women to demonstrate that requirements have been met based on sufficient and proper evidence. Consequently, to develop and validate diagnostic/risk DL models, large screening cohorts must be collected.

We used the screening population in the Capital Region of Denmark throughout the contributions. This breast cancer screening program is regional and high-volume with a target group of approximately 220,000 women screened at five clinics. Asymptomatic women between the age of 50 and 69 are invited to biennial mammography screening. This includes both women with and without a previous breast cancer diagnosis. In 2016, women with a previous breast cancer diagnosis between the age of 70 and 79 were invited to screening as well. However, women between 70 and 79 with a previously diagnosed breast cancer are not considered in Paper I, II, or III, as the study samples was gathered before 2016.

Radiographers capture at least four mammographic views: A craniocaudal and a medio-lateral oblique view for each breast on a Mammomat Inspiration or Revelation system from Siemens. Two radiologists read every mammogram independently, usually within 10 days, and determine whether to recall the women or not. All readers are specialized breast radiologists, and at least one of the readers is a senior high-volume breast radiologists reading at least 5,000 screening mammograms per year. If the readers disagree on a recall decision, a third radiologist breaks the tie, a consensus meeting is held, or one of the two senior radiologists forces a recall. A recall for diagnostic tests entails palpation, clinical examination, diagnostic mammography/tomosynthesis, and ultrasound. If the radiologist deem that the diagnostic imaging indicates a possible malignancy, a needle biopsy of the lesion is performed. Thus, all breast cancers are diagnosed based on the so-called triple-test of palpation, diagnostic imaging, and needle biopsy [66]. If the triple-test is positive for invasive breast cancer or DCIS, the woman will be offered appropriate treatment such as breast conserving surgery or mastectomy, axillary dissection, adjuvant or neoadjuvant treatment depending on the TNM-staging. If the triple-test is negative, the woman return to screening.

3.1.1 Breast Cancer Group Designations

To measure the quality of a screening program data on sensitivity, specificity, detection rate, false-positive rate and interval cancer rate is crucial. Furthermore, during training and validating of texture risk models, it is important to know the time from screening to diagnosis. To meet these two requirements, in Paper I, II, and III, we consider three distinct breast cancer groups based on time
from screening to diagnosis in the screening workflow: screen-detected cancers (SDC), interval cancers (IC), and long-term cancers (LTC). The designation process for each cancer group within the screening workflow is shown in Figure 9. Specifically, women who were recalled at screening and received a pathologically proven breast cancer diagnosis within six months from screening had an SDC. Women who received a breast cancer diagnosis within 24 months from screening (or before the next screening invitation if this comes first), that was not screen-detected, had an IC. Women who remained healthy for 24 months and subsequently developed a breast cancer within the given follow-up period had a LTC. The length of the follow-up period was, in this thesis, five years after the last screening visit date in the study sample.

![Flow diagram of breast cancer screening workflow and designation of three distinct breast cancer groups: screen-detected, interval, and long-term cancers.](image)

Figure 9: Flow diagram of breast cancer screening workflow and designation of three distinct breast cancer groups: screen-detected, interval, and long-term cancers.

### 3.1.2 Live Collection of Raw Image Data

Since 2012, a data collection system has been in place to collect raw image data in real time from the screening program in the Capital Region of Denmark. Whenever a woman has a mammogram in one of the five clinics, the images are sent directly, in an anonymized format, to a secure location accessible from the Department of Computer Science at the University of Copenhagen.

Images are stored in Digital Imaging and Communications in Medicine (DICOM) format. The DICOMs are "for processing", which is the raw data from the imaging modality directly measuring x-ray attenuation through the breast.

Every night the received images are indexed in a database where parts of the DICOM header are archived for fast and easy access to information on the mammograms. This includes, but is
not limited to, the acquisition parameters, imaging parameters, patient age, view positioning, image laterality, and study dates. At the time of writing, 715,532 studies have been collected and indexed in our database. Collected mammograms from the Capital Region of Denmark screening program were used in Paper I, II, and III.

3.1.3 Danish Screening Cohorts

In paper I, II, and III, women screened in the Capital Region of Denmark from November 2012 to December 2015 were used for training and/or validation.

These two datasets were collected over two periods as seen in Figure 10. The Danish screening cohort 1 was collected from November 1st, 2012, to December 31st, 2013, which included 54,977 women. Pathological outcomes were extracted from the Danish Pathology Register for these women with at least five years of follow-up [118]. Cancer groups were defined (SDC/IC/LTC/healthy) as described above in Section 3.1.1.

At a later point, pathological outcomes were extracted, and cancer groups were defined in Danish screening cohort 2 for women screened from January 1st, 2014, to December 31st, 2015, with at least five year of follow-up as well. Furthermore, women screened in both Danish screening cohorts were marked as such to facilitate exclusion.

![Figure 10: Diagram of the two Danish screening cohorts including start and end dates.](image)

In Paper I, the Danish screening cohort 1 was used as a development sample and the Danish screening cohort 2 was used as an independent validation sample. In Paper II, the Danish screening cohort 1 was used as training sample and Danish screening cohort 2 was used as an independent hold-out testing sample. In Paper III, the Danish screening cohort 1 and 2 were used in combination as a validation sample.

3.2 Screening in Utrecht, the Netherlands

To be capable of validating developed risk models in a completely independent population screened on a different scanner, a Dutch population of women screened in Utrecht, the Netherlands was used in Paper II, and III. The breast cancer screening programs resembles the screening program in the Capital Region of Denmark, comply to the same European guidelines, and was described extensively in [119].

Dutch women are invited to biennial screening between the age of 50 and 75. Before 2003, all women were screened using SFM. Between 2003 and 2007, digital mammography was partly implemented in the screening clinics with Lorad Selenia systems from Hologic. Before 2010, radiographers always captured an MLO view for each breast but only captured CC views in the first screening visit,
for women with dense breasts, and for women with visible abnormalities. Around 2010, and in the following years, the screening routines changed such that both MLO and CC views were obtained at all screening visits.

Generally, imaging data was collected in raw DICOM format, however a subset of women had corresponding Hologic-processed views available as well. This subset of women with both raw and corresponding processed views from the same compression was used in Paper II. In Paper II, the Dutch screening cohort was used as both a training and a validation dataset. The texture model trained on the Danish screening cohort 1 was validated in the Dutch screening cohort, meanwhile the texture model trained on the Dutch screening cohort was validated in the Danish screening cohort 2. In Paper III, the Dutch screening cohort was only used to train the combination model which was validated in both Danish screening cohorts. Consequently, the texture and combination models were always validated in a training-independent sample of women from another population screened on a different mammographic device.
4 Contributions

This section presents all contributions included in this thesis in their original (published or submitted) format and in full length. Prior to each contribution there will be a preliminary section describing relevant aims or methods not already described previously in the thesis or in the contribution itself.
4.1 An Artificial Intelligence–based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload

4.1.1 Preliminary

The overall aim of this study was to develop an AI-based screening protocol that yielded a non-inferior screening quality while considerably reducing the radiologists workload in a retrospective simulation study. For this purpose the AI-system Transpara version 1.7 was employed and used. The AI-system is described in details in the contribution below and in [105, 107, 103, 120, 95, 120].
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 Lyge, PhD • Ilse Vejborg, MD • Mads Nielsen, PhD • Nico Karssemeijer, PhD • Martin Lilholm, 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, Nyköbing Falster Hospital, Nyköbing, Denmark (E.L.); Department of Radiology, Copenhagen University Hospital Herlev/Gentofte, Copenhagen, Denmark (I.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: grayng@fhi.dk).

Supported in part by Eurostars (grant E9714 IBSCREEN). *Current address: Department of Image Guided Therapy Systems, Philips Healthcare, Utrecht, the Netherlands.

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

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 (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).

This copy is for personal use only. To order printed copies, contact reprints@rsna.org
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 could be simulated and assessed. In particular, we evaluated whether the AI-based screening sensitivity was noninferior and screening specificity was higher compared with those of radiologist screening. Finally, we investigated the quality of AI-based and radiologist screening across Breast Imaging Reporting and Data System (BI-RADS) densities.

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 cranio-caudal 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.

Figure 1: Flow diagrams of the data collection process for the development and testing samples. The exclusion criteria and the number of breast cancer screenings at each stage are shown. “Screen” refers to a four-view mammographic examination.
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 $\alpha = .05$ (31). To test non-inferiority for sensitivities, a one-sided Farrington-Manning test was used with the null hypothesis $H_0: S_A - S_R \geq \delta$, where $S_A$ is the radiologist screening sensitivity and $S_R$ the AI-based screening sensitivity. $H_0$ was tested at a significance level of $\alpha = .05$. If $H_0$ was rejected, the alternative hypothesis, $H_1: S_A - S_R < \delta$, 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 118,039 women who underwent screening examinations, of whom 3618 (3%) were excluded; thus, 114,421 women with a mean age of 59 years ± 6 (SD) were included. Of the 114,421 women, 7911 had screen-detected cancers, 327 had interval cancers, 1473 had long-term cancers, and 111,830 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 114,421). The false-positive rate in the development sample was 2.41% (1299 of 53,951), which was higher ($P < .001$) than that
Radiology:
healthy women in the testing sample. Figure 3B depicts the BI-RADS densities 3 and 4.

ties 1 and 2 was higher (BI-RADS density; the AUC for women with BI-RADS densities
healthy women, the performance was reduced with increasing
ually, the AUC remained relatively stable across BI-RADS densities.

development sample are shown in Table E2 (online).
pairs of age groups,
ference in AUC across age groups (for all possible
dence of a di/\text{T}_h (95\% CI: 0.67, 0.70) for long-term cancers.
/\text{T}_h (95\% CI: 0.74, 0.77) for interval cancers, and 0.68
/\text{T}_h (95\% CI: 0.77, 0.80) for screen-detected can-
ds of the testing sample of 1.84% (2107 of 114421). 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\% (107478 of 113303; 95\% CI: 94.7, 95.0), which was lower (\( P < .001 \)) than the radiologist specificity of 98.1\% (111196 of 113303; 95\% CI: 98.1, 98.2). The number of false-positive screenings increased by 276.5\% (5825 of 2107) compared with radiologist screening.
Evaluation of Al-based Screening

In the development sample, the RT was fitted to be 9.989, such that the number of screen-detected cancers missed in Al-based screening equalled the number of women with interval cancers.

In the development sample, the RT was fitted to be 9.989, such that the number of screen-detected cancers missed in Al-based screening equaled the number of women with interval cancers. In the testing sample, the RT was fitted to be 9.989, such that the number of screen-detected cancers missed in Al-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 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% (111196 of 113303; 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% (111725 of 113303; 95% CI: 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 (P < .001). Screening outcomes for the development sample 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% (111196 of 113303; 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% (111725 of 113303; 95% CI: 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 (P < .001). Screening outcomes for the development sample...
are shown in Table E4 (online).

**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 specificities 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 cohort of mammographic examinations, 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 (Transpara 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 Southern Denmark for providing screening mammograms. We also thank the Support for attending meetings and/or travel from EU research, clinical studies, statistical analysis, and manuscript editing, A.D.L., M.C.v.E.C., M.N., M.L.; and manuscript editing, A.D.L., M.C.v.E.C., E.L., I.V., N.K., M.L.


References
4.2 A Prospective Study of Breast Cancer Screening with AI as First Reader for Likely Normal Mammographies

4.2.1 Preliminary

After Paper I was recommended for publication, the AI system (Transpara version 1.7) for lesion detection and reader stratification was implemented into screening in the Capital Region of Denmark as of November 18th, 2021. The implemented screening protocol is slightly more conservative than the protocol presented in Paper I, and is shown alongside the standard screening protocol in Figure 11.

The protocol for screening with AI consists of the following steps: Firstly, the AI system scores all screened women on a scale from 1 to 10 based on the likelihood of breast cancer. Mammograms with an exam score below a certain threshold are considered likely normal. Secondly, likely normal women are read by a single senior radiologist whereas the remaining women are read by two radiologists (at least one senior). Thirdly, upon disagreement between any two readers, a third reader will step in to break a tie on a recall decision or a consensus meeting is held. Additionally, all readers have decision support, provided by the AI system, available during reading. Only after the mammogram has been examined for suspicious findings, the reader will look at the report generated by the AI system. This report contains markings of lesion and corresponding likelihoods of malignancy. The reader then reevaluates their own findings against the AI system’s finding and make a final recall decision.

The threshold for likely normal mammograms has been decided by the managerial team to reflect a desired reading workload reduction while considering the safety for screened women as well. Before May 3rd, 2022, the threshold was 5 which entailed an overall reading workload reduction of approximately 25% as 50% of women would classify as having likely normal mammograms. After
May 3rd, 2022, the threshold was increased to 7 such that approximately 70% of mammograms would classify as likely normal and the overall reading workload reduction would be 35%.

To study the outcomes of screening with AI, three screening cohorts were collected: a screening cohort before AI, a screening cohort with AI before the threshold was increased, and a screening cohort with AI after the threshold was increased. The three cohorts are shown in Figure 12. In the three cohorts, we measure the following outcomes: Workload reduction, recall rate, consensus meeting rate (indicative of disagreement), and detection rate in the group of likely normal women. Note, that the results reported in the submitted abstract (show below) differs from the results presented at RSNA 2022 and in this thesis. The most current results are presented in Key Results Section 5.2.

Figure 12: Timeline of screening before and with AI. Three screening cohorts were collected. The cohort "Before AI" was defined as women screened from October 1st, 2020, to October 31st, 2021. The cohort "With AI (before threshold increase)" was defined as women screened from November 18th, 2021, to May 2nd, 2022. The cohort "With AI (after threshold increase)" was defined as women screened from May 3rd, 2022, to September 30th, 2022.
A Prospective Study of Breast Cancer Screening with AI as First Reader for Likely Normal Mammographies

Andreas D. Lauritzen, Ilse Vejborg, Martin Lillholm

Purpose: The effect of employing artificial intelligence (AI) systems into breast cancer screening programs to safely reduce radiologist’s workload has been investigated in retrospective studies. In Denmark, an AI system has been fully implemented into a large regional screening program and the direct effects can be measured prospectively. We aimed to investigate how initial results affect workload reduction and recall rate after AI implementation.

Methods and Materials: Women in the Capital Region of Denmark aging between 50-69 are screened biennially. Each full-field digital mammography (FFDM) is independently read by two radiologists that decide whether to recall for diagnostic mammography, ultrasound, and eventually needle biopsy. Upon disagreement, a consensus conference is held with a third radiologist. In November 2021, the AI system Transpara (ScreenPoint Medical), was taken into use in this screening program. The AI system assigns a score on a scale of 1-10 to each exam with increasing likelihood of cancer. FFDMs with an exam score less than or equal to 5 are considered likely normal and are therefore read by the AI system as first reader and by a senior radiologist as second reader. The remaining FFDMs are doubly read by radiologists as normally. Data on radiologists’ decisions was extracted from October 2020 - October 2021 and from January 2022 - February 2022 which we refer to as baseline screening and AI screening, respectively.

Results: The baseline screening sample consisted of 59,325 women (1816 recalled). The AI screening sample consisted of 11,205 women (312 recalled). With AI screening 58% of FFDMs were likely normal and were therefore read by AI and only one radiologist, which corresponded to a 29% workload reduction. The recall rate was 3.06% (95% CI: 2.92%, 3.20%) at baseline screening and 2.78% (95% CI: 2.48%, 3.09%) with AI screening, which was lower but not significantly different (p=0.12).

Conclusions: Initial results of screening with AI as first reader, in cases of likely normal FFDMs, reduced the reader workload by 29% and resulted in a lower recall rate, however not significantly so. More time is needed to collect additional data and to detect whether recall rate will safely decrease without sacrificing cancer detection rate. An ongoing study for future publication is currently monitoring rate of consensus conferences, level of reader agreement, interval cancer rate, and cancer detection rate.

Clinical Relevance/Application: It is important to monitor AI screening performance to ensure a continuously high safety level. It might enable clinicians to work at a higher threshold for likely normal mammographies and decrease workload and recall rate even further.
4.3 Validation of a Deep Learning-based Breast Density Estimation Tool on a Danish Screening Cohort in the Context of Personalised Risk-based Screening

4.3.1 Preliminary

Mammographic density is an established breast cancer risk factor as described previously in Section 1.7.3. It has furthermore been shown that combining risk models with mammographic density improves risk assessment [121, 122, 28].

To quantitatively measure breast density and PMD, we developed and trained a U-Net model to segment mammographic views (MLO and CC) in three pixel-wise categories: background/pectoral muscle, non-dense tissue, and dense tissue. The used U-net model architecture is shown in Figure 13.

![Figure 13: U-net model for segmenting breast tissues in mammographic views. In this example, an L-MLO view is given as input to the model which generates a segmentation mask where each pixel has been classified as either background/pectoral muscle, non-dense tissue, or dense tissue. In this example, the purple area denotes the non-dense tissue while the white area denotes the dense tissue. The background and pectoral muscle has no overlay.](image)

The model was trained using 500 pairs of views (collected from the Dutch screening population) and corresponding radiologist-annotated masks (obtained with a drawing tool and interactive thresholding) using a soft DICE loss [123]. The planimetric PMD was calculated percentage of dense tissue pixels out of to the total number of dense and non-dense tissue pixels. The PMDs for each view were averaged to obtain an exam-level PMD. To validate the segmentation model, it was applied to all women in Danish screening cohort 1 with valid imaging data. The results of this validation study are presented in the abstract below.

In Paper II and III, we aimed to investigate whether planimetric PMD could contribute to a richer risk profile. The segmentation model was therefore applied to Danish screening cohort 2 and the Dutch screening cohort as well.
Validation of a Deep Learning-based Breast Density Estimation Tool on a Danish Screening Cohort in the Context of Personalised Risk-based Screening

Andreas D. Lauritzen, My Catarina von Euler-Chelpin, Elsebeth Lynge, Ilse Vejborg, Mads Nielsen, Martin Lillholm

**Purpose:**
To validate a fully automatic density estimation tool on a screening cohort in terms of agreement with radiologists’ BI-RADS and cancer risk segregation.

**Methods and materials:**
This study was based on the Danish Capital Region breast cancer screening program from November 1st, 2012 to December 31st, 2013. 4-view FFDMs were available for 53956 women. The cohort’s median age (IQR) was 59 (54-65), and the cohort comprised 568 cancers. Radiologist’s BI-RADS, 4th edition scores were available from two readers. Using a deep learning-based fully automatic tool developed by the University of Copenhagen, all FFDMs were scored for planimetric percent mammographic density (PMD). The correspondence between two-reader consensus BI-RADS and PMD was evaluated in terms of Spearman correlation and, after categorisation of PMD, with weighted kappa statistics (WKS). The latter was compared to the readers’ inter-observer WKS. In terms of cancer risk segregation, the area under the ROC-curve (AUC) was compared for PMD and consensus BI-RADS; both with age as a covariate in a logistic regression model.

**Results:**
The correlation between PMD and consensus BI-RADS was 0.85. The WKS between PMD and consensus BI-RADS was 0.693, and the radiologist inter-observer WKS was 0.692. The AUC for cancer risk was 0.60 (0.57-0.62) for PMD+age and 0.59 (0.56-0.61) for consensus BI-RADS+age.

**Conclusion:**
PMD matched radiologists’ BI-RADS in terms of agreement between categorised PMD & consensus BI-RADS and the radiologists’ inter-observer agreement; both were substantial. For cancer risk segregation, there was no significant difference between consensus BI-RADS and PMD. Regarding personalised screening using mammographic density as a risk factor, the results suggest that automated PMD would work as well as consensus BI-RADS of two radiologists.

**Limitations:**
The two readers work at the same clinic.
4.4 Robust Cross-vendor Mammographic Texture Models Using Augmentation-based Domain Adaptation for Long-term Breast Cancer Risk

4.4.1 Preliminary

The overall aim of this study was to develop a DL risk model that learns and relies on mammographic textural features indicative of an increased susceptibility to breast cancer. For this purpose, we chose a single DL network architecture based on multiple criteria. During the study we observed that clips had an effect on texture model training. We therefore further developed a DL model to detect visible clips, implanted medical devices, and views with disruptive image features.

**Residual Network Encoder for Texture Analysis** For the purpose of extracting risk predictive features from mammographic views, a single DL network architecture and training procedure was chosen. The criteria for the network were that the model should be fairly recently published, be parameterized sufficiently to learn mammographic features, be well-validated in multiple image domains, e.g., natural and medical images, and be easily implementable using standard libraries. The aim of this thesis was, not to develop a new non-reproducible network for marginally better performance, but rather to systematically collect and design a case-control dataset for training the model and carry out proper large-scale clinical validation. Keeping the network and training procedure fixed allowed for more accurate analyses of other effects like changing training sample size, data augmentation, and removal of noisy training samples as described in Paper II and III. The choice fell on the squeeze-and-excitation residual network model with 18 blocks (SE-ResNet18) which fulfilled all the above-mentioned requirements. This model significantly improved performance over other state-of-the-art DL network with only very little additional computational cost [124]. The model won the ImageNet classification challenge which also meant that the pre-trained weights could be used in Paper II and III for quicker and more stable convergence during training [125]. During inference, a mammographic view (MLO or CC) is passed through the SE-ResNet18 encoder, and the output feature map is global average pooled and passed through two fully-connected layers both with dropout and batch normalization [126, 127]. The output is interpreted as a measure of risk. The full model architecture is shown in Figure 14. Throughout this thesis and included contribution, this model is referred to as the texture model. The full training procedure for the texture model is described in Paper II.

**Residual Network Encoder for Clip Detection** Women with a previous breast cancer or a noncancer recall are known to have elevated breast cancer risk as described in Section 1.3.3 [13, 17, 18, 15, 16]. During a biopsy or prior to radiation therapy a clip may be left in the breast which appear as a bright object in the mammogram. DL models are highly sensitive to easily distinguishable objects with high intensities and clearly defined boarders [128]. The presence of clips may therefore prevent the texture model from learning true risk-predictive tissue features [128, 101]. To obtain a true mammographic texture model, we aimed to completely exclude views with clips from training (not validation). A classifier was therefore trained to detect clips in views, and the texture model training dataset could then be constructed accordingly without views with clips. The clip detection
model was the SE-ResNet18 network which is the same network as described above. This model was trained using 10,000 manually annotated views of randomly sampled mammograms collected from the Capital Region of Denmark breast cancer screening program from the 1st of January 2012 to 31st of October 2012. The 10,000 views were manually annotated (by the candidate) into four different classes: normal-looking views without artifacts, views with one or several clips, views with implanted medical device(s), and views with severe disruptive imaging features/artifacts, e.g., wrong cropping with missing breast tissue, phantoms, corruptions, and alike. Women with breast implants (breast augmentation), were classified as belonging to the class of normal-looking views. The training procedure was the same as described in Paper II, except the model was optimized to distinguish between the four classes. The clip detection model achieved AUCs of 0.99 for views with clips, 0.99 for views implanted medical devices, and 0.91 for views with severe disruptive features/artifacts measured using 5-fold cross validation. The clip detection model was deemed adequate to support construction of training datasets for the texture model. The clip detection model was applied to Danish screening cohort 1 and 2 and the Dutch screening cohort. Figure 15 shows three views in the Dutch screening cohort with clips, a implanted pacemaker, and disruptive imaging features (cropping and wrong intensity scale). In the Danish screening cohort 1, the AUC for identification of ICs
using the clip measure was 0.56 (95% CI: 0.53, 0.59) and 0.52 (95% CI: 0.52, 0.53) for LTCs. This suggested that presence of clips was a weak breast cancer risk factor.

These two models for measuring mammographic texture and for detecting clips, respectively, were used in the Paper II and Paper III.
Robust Cross-vendor Mammographic Texture Models Using Augmentation-based Domain Adaptation for Long-term Breast Cancer Risk

Andreas D. Lauritzen*, My Catarina von Euler-Chelpin, Elsebeth Lyngé, Ilse Vejborg, Mads Nielsen, Nico Karssemeijer, and Martin Lillholm

Abstract

Purpose: Risk-stratified breast cancer screening might improve early detection and efficiency without comprising quality. However, modern mammography-based risk models do not ensure adaptation across vendor-domains and rely on cancer precursors, associated with short-term risk, which might limit long-term risk assessment. We report a cross-vendor mammographic texture model for long-term risk.

Approach: The texture model was robustly trained using two systematically designed case-control datasets. Textural features, indicative of future breast cancer, were learned by excluding samples with diagnosed/potential malignancies from training. An augmentation-based domain adaption technique, based on flavorization of mammographic views, ensured generalization across vendor-domains. The model was validated in 66,607 consecutively screened Danish women with flavorized Siemens views and 25,706 Dutch women with Hologic-processed views. Performances were evaluated for interval cancers (IC) within two years from screening and long-term cancers (LTC) from two years after screening. The texture model was combined with established risk factors to flag 10% of women with the highest risk.

Results: In Danish women, the texture model achieved an area under the receiver operating characteristic (AUC) of 0.71 and 0.65 for ICs and LTCs, respectively. In Dutch women with Hologic-processed views, the AUCs were not different from AUCs in Danish women with flavorized views. The AUC for texture combined with established risk factors increased to 0.68 for LTCs. The 10% of women flagged as high-risk accounted for 25.5% of ICs and 24.8% of LTCs.

Conclusions: The texture model robustly estimated long-term breast cancer risk while adapting to an unseen processed vendor-domain and identified a clinically relevant high-risk subgroup.

Keywords: breast cancer risk, mammography, domain adaptation, data augmentation, noisy labels

*First Author, E-mail: al@di.ku.dk

1 Introduction

Breast cancer is the leading cause of death in women worldwide. Approximately 13% of women in the E.U. and U.S. will receive a breast cancer diagnosis in their lifetime. Population-based screening by mammography, an x-ray of the compressed breast, succeeds in decreasing mortality by detecting early breast cancer signs. Screening programs often employ a one-size-fits-all approach where all women are screened under equal conditions. This approach is inefficient
considering the low breast cancer incidence and general lack of specialized radiologists. Personalized screening has the potential to improve screening by reallocating radiologists’ resources efficiently to high-risk women without increasing the total cost or compromising screening quality.\textsuperscript{5,6}

M. Gail et al. showed that breast cancer risk could be estimated using established clinical risk factors such as family history of breast cancer, reproductive history, and history of previous biopsies.\textsuperscript{7} The Tyrer-Cuzick (TC) model, extended the Gail model to incorporate genetic information, e.g., BRCA mutations in first- and second-degree relatives.\textsuperscript{8} Whereas models based on clinical risk factors can estimate risk reasonably well, the extensive data collection and potentially inconsistent self-reporting makes implementation in population-based screening programs laborious and complex.

Conversely, mammography-based risk models derive information directly from the already available screening mammogram. Mammography-based classification schemes like the Wolfe and Boyd models showed that the amount of radio-dense tissue, e.g., lobules and ducts, relative to the amount of fatty tissue could successfully categorize high-risk women.\textsuperscript{9,10} This ratio between dense and fatty tissue, the percentage mammographic density (PMD), became an established risk factor. L. Tabár et al. later showed that the distribution of dense/parenchymal breast tissue, not solely the amount, was a risk factor as well.\textsuperscript{11} The study showed that mammographic features, representing breast heterogeneity, could be manually scored to identify high-risk women. Such features are often referred to in literature as mammographic texture. Mammographic density and texture are global features of the mammogram associated with long-term breast cancer risk. This is contrary to detection of actual cancerous lesions, minimal signs of developing cancer (precursors), or other localized findings which are associated with short-term breast cancer risk. Furthermore, women recalled at screening with suspicious findings, but with a negative result for breast cancer, have an elevated long-term risk as well.\textsuperscript{12}

Modern mammography-based risk models automatically derive breast cancer predictive features such as precursors, texture, or density, and rely on supervised deep learning (DL) methods trained on large datasets of labeled digital mammograms. DL risk models are more compatible with population-based screening programs as digital mammograms can be made available for analysis immediately after or during screening. DL risk models relying on precursors or breast density are currently being evaluated in real and simulated clinical settings.\textsuperscript{5,6} However, long-term risk models have not been evaluated in terms of clinical utility.

We distinguish between three approaches to estimate mammography-based breast cancer risk by DL methods. Firstly, a diagnostic model, trained to detect or segment suspicious local findings, can be used to estimate short-term risk.\textsuperscript{13,14} Secondly, a texture model can be trained to estimate long-term risk by learning systemic differences in tissue composition between high- and low-risk women, e.g., mammographic texture or density patterns indicative of elevated risk.\textsuperscript{15,16} Thirdly, a conflated model can be trained to simultaneously estimate immediate, short-, and long-term risk by learning both features of visible cancers/precursors, and systemic/global differences in mammographic texture.\textsuperscript{17,18} It has, however, been suggested that a visible lesion is an easily
distinguishable feature but is weakly predictive of long-term risk and therefore might prevent an
DL model from learning strongly predictive features of long-term risk.\textsuperscript{16,19}

Some of the previous studies were performed with mammograms in raw format, i.e., the
untouched output from a mammographic device directly representing x-ray attenuation.\textsuperscript{6,15} Raw
data is well-defined and mostly homogeneous across vendors and mammographic devices, which
makes it easier to develop robust AI algorithms. However, models that rely on availability of raw
data have limited clinical value, because clinics use processed mammograms prepared by the
mammography device with a contrast suitable for human visual interpretation. Raw mammograms
are normally not archived whereas processed mammograms are routinely archived. The
appearance of processed views depends highly on the vendors image post-processing pipelines
and even depends on the screening clinics' own preferences/settings. Consequently, DL risk
models must not only estimate risk in processed views but also robustly across mammographic
devices from different vendors to remain clinically relevant at a large scale. Most of the relevant
work we compared to considered only processed mammograms from a single vendor.

Modern DL models often suffer from domain shift problems and will in many cases of medical
images not adapt to unseen domains, e.g., to processed mammograms obtained on a
mammographic device from another vendor than the one used for training. The ability to adapt to
unseen vendor-domains would allow retrospective external validation at multiple international
clinics and minimize clinical implementation overhead. Device-domain adaptation for DL breast
cancer risk models have not been investigated in depth yet.

In this study, we developed a clinically relevant, robust, cross-vendor texture model for long-
term breast cancer risk. The main contributions of this study were as follows:

1. We trained an DL model, based on a well-validated architecture, for texture long-term
breast cancer risk using screen-available information only. To ensure optimal training and
not depressing a potentially weak long-term risk signal, we train exclusively on views with
no visible or potentially visible diagnosed cancers.

2. We showed that training for long-term risk can produce erratic convergence which we
alleviated by excluding noise-inducing training samples and systematically designing a
case-control training dataset for an ensemble texture model.

3. To ensure adaptation across vendor-domains, we trained the texture model using a data
augmentation technique converting raw to processed views in multiple flavors. The texture
model was successfully applied to domains of flavor-processed and vendor-processed
views.

4. The texture model was shown to generalize well across populations and vendors in two
large independent screening datasets and identified clinically relevant high-risk women.

2 Related Work

Breast cancer risk is used in different contexts throughout the literature and is often intertwined or
confused with detection/segmentation. It is important to consistently distinguish between models
created purely for detection, risk assessment, or a combination and report results accordingly. In
the following sections, we describe existing methods for estimating breast cancer risk: a diagnostic-based, a texture-based, and two conflated models.

2.1 Diagnostic Models for Risk Estimation

A. Gastounioti et al. employed a diagnostic model, the ProFound AI tool (iCAD), which detects/segments suspicious lesions and was compatible with Hologic, GE, and Siemens devices. The model included age, masses, microcalcifications, asymmetry, and breast density to provide a short-term risk estimate in a cohort of screen negative women with two-year follow-up. The model achieved an area under the receiver operating characteristic curve (AUC) of 0.68.13 No long-term risk estimates were presented. There exist other studies using a similar approach to estimate short-term risk.14 We focused on a long-term risk model by excluding localized findings from training as suggested by Liu et al.16

2.2 Mammographic Texture Model for Risk Estimation

Kallenberg et al. trained an autoencoder on multi-scale image patches to learn mammographic features in an unsupervised fashion. The model was finetuned to measure texture risk, given only cancer contra-lateral views, assuming tissue patterns persisted across left and right breasts.15 The model achieved an area under the receiver operating characteristic curve (AUC) of 0.61 in a dataset with two years of follow-up. The training and validation datasets were small, from the same population, and single-vendor raw views only.

2.3 Conflated Models for Risk Estimation

Dembrower et al. trained an Inception-ResNet-v2 model to estimate risk on multi-scale image patches of mammograms. The model included age and device acquisition parameters.17 The model was trained and validated on women from one population with Hologic-processed views only who was cancer-free for 12 months and achieved an AUC of 0.65. The training views might include early or slow-growing malignancies thus yielding a conflated model. There was no evaluation of robustness regarding early stopping criteria.

Yala et al. trained a conflated model, named Mirai, to estimate one to five-year risk.18 The model consisted of a ResNet encoder, a transformer model to aggregate features, a risk factor predictor, and finally an additive-hazard layer to estimate risk at multiple timepoints. Mirai was trained using an adversarial approach to learn device-invariant features, thus being able to generalize across mammographic devices. However, only two systems from the same vendor (Hologic) were considered. Mirai achieved an AUC between 0.68 and 0.73 in seven independent datasets, all with Hologic mammograms. However, during validation, 42 to 82% cancers were diagnosed with one year and 54 to 91% within two years which might not be adequate to measure the long-term risk estimation performance and favoring Mirai's conflated approach.20

None of the above-mentioned studies investigated ability to adapt to unseen vendors, e.g., from Siemens to Hologic. As presented above, the problem of learning features of long-term risk in a
data-driven fashion is inherently difficult and noise-filled. When approached as a supervised learning task with per-study labels, a modern DL model it is very likely to overfit by learning long-term risk-irrelevant features due to weak long-term risk signal compared to other visible features and model overparameterization.

Additional noise is present as high-risk women, e.g., high density and texture-risk women, might not develop a breast cancer within follow-up or at all. This can lead to erratic convergence and have not been discussed in literature in this context. Furthermore, existing methods are conflated models in terms training, but also mix detection of future/current cancers at different intervals during validation. We aimed to accommodate these above-mentioned limitations.

3 Materials and Methods

In the following sections, we describe our methods for developing and testing the texture model. Initially, we define four datasets – two for training and two for testing. Next, we describe our efforts to stabilize training. Then, we describe the augmentation-based domain adaptation technique to enable a cross-vendor texture model. Finally, we describe the training procedure and the ensemble texture model.

This study was approved by The Danish Data Inspection Agency and Danish Patient Safety Authority including collection and analysis of relevant data. The need for informed consent was waived.

3.1 Screening Populations and Datasets

The Danish screening population was from the breast cancer screening program in the Capital Region of Denmark, where women between the age of 50 and 70 were screened biennially on Mammomat Inspiration systems (Siemens Healthineers).

The Dutch screening population was from the breast cancer screening program in Utrecht, the Netherlands, where women between the age of 50 and 74 were screened biennially on Lorad Selenia systems (Hologic). Screened women had four mammographic views captured, two for each breast. The mammograms were read by two specialized radiologists who determined whether to recall the women for diagnostic tests. All breast cancers were diagnosed as invasive or ductal carcinoma in situ based on clinical examination consisting of mammography, ultrasound, and needle biopsy.

Breast cancers were separated into three distinct groups: Screen-detected cancers (SDC), interval cancers (IC), and long-term cancers (LTC). Recalled women who received a breast cancer diagnosis within six months from screening had SDCs. Women who did not have an SDC yet received a breast cancer diagnosis within 24 months from screening had ICs. LTCs are breast cancers diagnosed at least two years from screening regardless of recall. In practice, LTCs were SDCs or ICs in subsequent screening rounds.

From the two screening populations, two training datasets and two testing datasets were collected. Fig. 1a and 1b depict the collection processes in a flow diagram including the
inclusion/exclusion process for women with visible artifacts, corrupted views, previous breast cancer diagnoses, and biopsy clips in the cancer contra-lateral breast or both breasts.

Fig. 1 (a) Flow diagram of selection process for the Danish training and testing dataset, (b) Selection process for the Dutch training and processed testing dataset, and (c) curation of views for training and validation in the filtering stage. *Visible artifacts include implanted medical devices, foreign unknown objects, and other disruptive image artifacts.

The Danish training dataset was collected from the Danish screening population as screened women from November 1st, 2012, to December 31st, 2013. The Dutch training dataset was collected from the Dutch screening population by randomly sampling a single screening for each woman between August 1st, 2003, and January 1st, 2015. After exclusions, the maximum amount of SDCs, ICs, LTCs, and healthy women were sampled from the Danish and Dutch training population, such that each group of women were equal in size. Both datasets consisted of raw mammographic views.

The Danish testing dataset was collected from the Danish screening population as screened women from January 1st, 2014, to December 31st, 2015. The Danish testing dataset consisted of raw mammographic views. This dataset was a hold-out testing dataset to measure performance for models trained on the Danish or Dutch training datasets.

The Dutch processed testing dataset was collected from the Dutch screening population between August 1st, 2003, and January 1st, 2015, and was defined as women for which raw and corresponding Hologic-processed views were available. The two Dutch datasets were not independent, but this was not a limitation because the Dutch processed testing dataset was only used for evaluation of texture models trained on Danish data.

3.2 Curation of Training Data to Stabilize Convergence

Due to noisy labels and the inherent difficulty of the task, the reference training of the texture model was erratic with unstable convergence (see Fig. 5). To attain reliable risk estimates, we
curated a systematically designed case-control dataset for robust training. This curation process consisted of two stages: The noise-identification stage and the filtering stage.

Literature on training DL models, with noisy or corrupted labels, suggests training a mentor/teacher network to filter samples or generate pseudo-labels for a student network to improve classification.\textsuperscript{21} We partly employed the same strategy by training a reference model in the noise-identification stage to obtain reference risk scores. These reference risk scores were used, in the filtering stage, to exclude noise-inducing samples and create risk-stratified ensemble folds for the ensemble texture model. The noise-identification stage consisted of four steps:

1. Cancers were randomly matched with 20 healthy women on age (± one year age difference).
2. Cancers were then randomly split into five ensemble folds. In practice, we sampled the folds using a cross validation method stratified by age and cancer groups (SDC/IC/LTC/healthy). Note that the model performance was not cross validated per se as the folds were only used for training an ensemble model. The matched healthy women were added to the fold of their matched cancer.
3. Training views with a diagnosed cancer (SDC, IC, and LTCs) were excluded. For the matched healthy women, the same views were excluded, e.g., given a woman with cancer in the left breast, the two left views were excluded for both the woman with cancer and the matched healthy women. Unmatched healthy women were randomly split into the five validation sets. No validation views were excluded. Consequently, the reference texture model was trained on two cancer contra-lateral views and validated on all four available views.

The reference risk scores were extracted after training the reference model for 40 epochs. The training procedure is described in section E below. This stage was applied to both the Dutch and Danish training datasets such that reference risk scores were available for all women in the two datasets.

The filtering stage, shown in Fig. 1c, was fundamentally like the noise-identification stage except for the sample filtering and ensemble fold sampling. Initially (before step 1), 10% of women healthy throughout follow-up with the highest reference risk were excluded. These were likely noisy samples as they might have a high PMD or texture risk and did not develop a breast cancer within follow-up. The SDCs, ICs, and LTCs with the lowest 4% reference risk, in each cancer group, were excluded as well because long-term risk estimation is inherently difficult hence a subset of women with a future cancer will be difficult to classify. Women with a future cancer could have low-risk tissue patterns that correlate highly with tissues of healthy women or have a cancer subtype that does not correlate with texture. In step 2, the ensemble folds were resampled and stratified on reference risk, cancer group, and age.

Unmatched/excluded women were split randomly, stratified on cancer group and reference risk, into the five validation sets. This filtering stage produced five folds for training a robust and better converging ensemble texture model (see Fig. 5).
3.3 Augmentation-based Domain Adaptation

A clinically relevant texture model must adapt to new target domains of unseen processing types. There are several methods for domain adaptation including data augmentation and adversarial trainings.\textsuperscript{22,23} Adversarial approaches can be difficult to train in practice and often leads to failed convergence.\textsuperscript{18,24} When the target domain is well defined, in our case by classical image transformations, adaptation can be successfully achieved by data augmentation which outperform adversarial approaches in certain cases.\textsuperscript{22} To realize augmentation of a single raw view to multiple processed views of different formats, we developed a tool consisting of a series of four parameterized image transformations steps inspired by previous literature.\textsuperscript{25,26}

In the first step, seen in Fig. 2a, a mask $M$ delineating the breast tissue was generated along with a distance map $D$ measuring distance to the skin-air-boundary. The background defined by $M$ was set to 0 which also removes burned-in view annotations.

In the second step, pixel values in the breast tissue in view $I$ at position $i, j$ were transformed by a logarithmic and an inverse sigmoid function

$$I_{i,j} = \frac{I_{max}}{1 + \exp\left(\alpha \log_{10}(\frac{I_{i,j}}{\bar{I}_{breast}})^2 - 1\right) + \beta},$$

where $\bar{I}_{breast}$ was the mean pixel value in the fully compressed breast tissue. The parameters $\alpha$ and $\beta$, were manually chosen parameters between values 4 to 5.5, and 0.7 to 1.2, respectively. The max value of a processed view $I_{max}$ was 4095.

![Fig. 2. (a) Example of the breast mask M and distance map D. Note that the rest of the grid has been omitted for viewing purposes and the grid and lines are not to scale. The tissue between the blue and orange lines is the edge tissue. (b) Example of a raw and a Hologic-processed view. Below is an example of a flavor 1 and 6 view. Flavor 2 to 5 views have been omitted in this example.](image)
In the third step, the edge tissue thickness was corrected for, and contrast was enhanced close to the skin-air-boundary as described by Tortajada et al.\textsuperscript{25}

In the fourth step, the fully compressed breast tissue was contrast enhanced by addition of an enhancement mask as described by Panetta et al.\textsuperscript{26}

\[ I_{i,j} = \gamma \cdot I_{i,j} + \delta \cdot \frac{F_{i,j}}{I_{\text{max}}} \cdot I_{i,j}, \quad (2) \]

where \( F_{i,j} \) was the enhancement mask, which in practice was the difference between the original view and the low pass filtered view. Scaling parameters \( \gamma \) and \( \delta \), were manually tuned to approximate processed views from different vendors. A set of parameters is referred to as a flavor profile. The output of the tool, defined by the flavor profile, is a processed view which we refer to as a flavor. We manually constructed seven flavor profiles producing seven different flavors. The Danish and Dutch training datasets were both processed into six different flavors and the Danish testing dataset was processed into a seventh flavor. Consequently, the texture model was validated in a flavor not seen during training. Fig. 2b shows an example of a raw view, the corresponding Hologic-processed view, and examples of flavors created with the image conversion tool.

### 3.4 Training Procedure for the Texture Model

The texture model was generally trained using the same procedure and only the augmentation scheme and training/validation dataset varied. The texture model takes as input a single mammographic view, passes it through a ResNet18 encoder\textsuperscript{27}, initialized with pre-trained weights from the ImageNet challenge dataset\textsuperscript{28}, followed by two fully connected layers that output a single continuous confidence measure. The model architecture is shown in Fig. 3.

![Fig. 3. The texture model architecture. The input view was repeated thrice in the channel axis and passed through ResNet18. The output feature map was global-average-pooled to a feature map of size 512, which was passed through a dropout (DO) layer with a probability of dropout of 0.5, a batch normalization (BN) layer, and a fully connected layer (FC) using, ReLU activation, to 512 hidden neurons. The last layers were a DO layer with a probability of 0.25, BN, and a FC layer with sigmoid activation to a single output neuron.](image)

The model was trained to distinguish between cancer contra-lateral views of women with a future cancer and views from healthy women, by minimizing the binary cross entropy (BCE) in batch \( B \) of size 10.
\[ \text{BCE} = -\frac{1}{|B|} \sum_{i=1}^{B} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i), \]  

where \( y \) is the binary ground-truth label and \( \hat{y} \) is the scaled confidence generated by the model.

The model was trained using Adam optimizer \(^{29}\) with a learning rate of \(10^{-5}\). During training right views are flipped across the vertical axis. Training views were randomly augmented to mimic difference in scanner positions by randomly rotating (by -15° to 15°), scaling (by a factor -0.2 to 0.2), or shearing in the horizontal direction (by a factor of -0.15 to 0.15).

The final texture model was an ensemble of five instances of the above-mentioned model trained using the same parameters. These five instances were trained on the corresponding five splits defined in the given curated training dataset. Each instance of the model was trained until convergence in the corresponding validation set indicated by the highest validation AUC where all cancer groups were considered as positives. All views, regardless of training or inference were scaled to a resolution of 0.255 mm per pixel and a size of padded to a size of 1194 by 938 pixels.

During inference on a testing dataset, the view-level texture risk score is an average of five confidence measures provided by each instance of model. Screened women had four views captured so to obtain a study-level risk score the four view-level risk scores were averaged. During inference, views in the given testing dataset were standardized to have the same mean and standard deviation as the used training dataset based on 1,000 randomly sampled views from the given testing dataset. In some experiments, the flavor augmentation was used. In these cases, all views in the given training dataset were processed into six flavors and randomly sampled into each batch.

An epoch was defined as an iteration through 1/6 of the full training dataset and such that the model was trained on all flavor-augmented training samples by 6 epochs. The experiment section specifies in which experiment the flavor augmentation was used. All models have been trained and tested using Python (version 3.6.8) and TensorFlow (version 2.1.0) on Nvidia TITAN RTX GPUs.

4 Experiments

As described in the data curation section (Sec. 3.2), two reference texture models were initially trained to reduce noise and stabilize training. Meanwhile, this also provided an opportunity to analyze convergence before during and after the noise-identification and filtering stages. These convergence plots are shown in Fig. 5.

Afterwards, four series of experiments were conducted, referred to as A, B, C, and D, to establish a baseline performance using raw views, and assess whether the flavor augmentation ensured proper adaptation to domains of flavorized and Hologic-processed views. We estimated whether additional training data was needed and whether age, PMD, and presence of clips (a surrogate for a previous false-positive screen), could improve risk estimation. Fig. 4 depicts these fours series of experiments in a flow diagram. Results were denoted \( R_{\text{train-format--test-format}} \), e.g., \( R_{\text{Dan--Nl-flav--proc}} \) denoted the results of applying the texture model, trained on Danish training dataset with flavor views, to the NL processed testing dataset with Hologic-processed views.
Fig. 4. Flow diagram of the four series of experiments. Boxes denote trained models, the used training dataset, and view formats. The oval endpoints denote the testing dataset used and view format. DK test=Danish testing dataset, NL proc. test=Dutch processed testing dataset, RF=risk factors.

The first series of experiments (see Fig. 4, series A) quantified a baseline performance in the Danish testing dataset of raw Siemens views using a texture model trained on raw Hologic or Siemens views. We also assessed whether the model trained on raw views, without flavor augmentation, would be able to generalize to flavor views.

In second series of experiments (see Fig. 4, series B), we quantified how well a texture model, now trained with flavor augmentation, would generalize to an unseen flavor (flavor 7) using the Dutch and the Danish training dataset of flavor views (flavor 1 to 6), respectively. Essentially, we simulated an unseen scanner.

In the third series of experiments (see Fig. 4, series C), baseline performance in the Dutch processed testing dataset was measured in raw views with the Danish trained model. We compared the baseline to the performance of the texture model trained with and without flavor augmentation to assess the performance on Hologic-processed views.

In the fourth series of experiments (see Fig. 4, series D), a texture model was trained on both the Dutch and Danish training dataset simultaneously to assess whether the training dataset size was sufficient. A significantly higher AUC would indicate the model was not saturated in terms of data. We concatenated the five ensemble sets from each of the training datasets and trained five new instances of the model using flavor views (flavor 1 to 6). This model was applied to the Danish testing dataset of flavor views (flavor 7). To assess whether risk factors could contribute to a full risk model, we separately trained a small neural network of three fully connected layers to combine the texture, age, presence of clips, and PMD.

Finally, we assessed whether the texture model was able to identify high-risk subsets of women in the Danish testing dataset. First, we reported how many women with ICs and LTCs were in the
group of women with the 10% highest risks using $R_{\text{flav-flav-D\text{-}tr+RF\text{-}D}}$. Second, we reported sensitivities for ICs and LTCs at 90% specificity using $R_{\text{flav-flav-D\text{-}tst}}$ as well. Third, the Danish testing dataset was split by texture risk (using $R_{\text{flav-flav-D\text{-}tr}}$) and PMD quantiles into 16 bins (4x4 quantiles) to assess the distribution and which of these bins might be flagged as high-risk. Quantiles were computed using healthy women only.

5 Results

In the following sections, the results are presented. Initially, the convergence of the reference texture model was investigated before and after the filtering stage. Next, the results of the four series of experiments are presented along with appropriate significance testing. Lastly, the percentages of ICs and LTCs in the high-risk groups were presented along with sensitivities at 90% specificity and the relationship between texture and PMD quantiles and odds ratios (OR) for each of the 16 bins.

5.1 Performance of the Reference Texture Model

Fig. 5 shows the convergence of the Danish reference model trained for 40 epochs during the noise-identification stage (top row) and after the filtering stage (bottom row).

In the top row, the validation performances depended highly on when the training is stopped. Before the filtering stage, the folds converged, in terms of ICs, between epoch 14 and 39. The difference in AUC(IC) at epoch 28, the mean convergence point (CP), was 0.10. After the filtering
stage, the largest difference in AUC(IC) at CP was 0.04. Before the filtering stage, the folds converged, in terms of LTCs, at epoch 21, 27, 31, 40, and 29, respectively. The difference in AUC(LTC) at CP was 0.05. After the filtering stage, fold one to five converged, in terms of LTCs, at epoch 33, 20, 20, 20, and 20, respectively. When evaluated at CP, the largest difference was 0.01 between folds. For both ICs and LTCs the validation curves were much closer together and converged largely at the same time after the two stages. Even with this curation of a training data, the folds overfit and does not guarantee equal convergence points, however, this effect was lessened by averaging over five instances of the ensemble texture model during inference.

5.2 Future Breast Cancer Detection Performance

Risk estimation performances, presented in Table 1, was measured using AUC.

| Experiment | AUC (IC) | AUC (LTC) | AUC (IC | LTC) |
|------------|----------|-----------|----------|
| **Series A** |          |           |          |
| N-tr raw→D-tr | 0.68 (0.65, 0.72) | 0.64 (0.62, 0.66) | 0.65 (0.63, 0.67) |
| N-tr raw→D-tr | 0.70 (0.66, 0.74) | 0.65 (0.62, 0.67) | 0.66 (0.64, 0.68) |
| N-tr raw→D-tr | 0.55 (0.50, 0.59) | 0.51 (0.49, 0.53) | 0.52 (0.50, 0.54) |
| N-tr raw→D-tr | 0.52 (0.48, 0.57) | 0.52 (0.50, 0.55) | 0.51 (0.49, 0.53) |
| **Series B** |          |           |          |
| N-tr flav→D-tr | 0.69 (0.65, 0.73) | 0.64 (0.62, 0.66) | 0.65 (0.63, 0.67) |
| N-tr flav→D-tr | 0.70 (0.66, 0.74) | 0.64 (0.62, 0.66) | 0.65 (0.64, 0.67) |
| **Series C** |          |           |          |
| raw→N-tr raw | 0.66 (0.61, 0.72) | 0.60 (0.58, 0.62) | 0.61 (0.59, 0.63) |
| raw→N-tr proc | 0.53 (0.47, 0.59) | 0.55 (0.53, 0.57) | 0.54 (0.52, 0.56) |
| flav→proc | 0.66 (0.60, 0.71) | 0.63 (0.61, 0.65) | 0.63 (0.61, 0.65) |
| **Series D** |          |           |          |
| flav→D-tr flav | 0.71 (0.67, 0.75) | 0.65 (0.63, 0.67) | 0.66 (0.64, 0.68) |
| flav→D-tr flav | 0.71 (0.67, 0.75) | 0.65 (0.63, 0.67) | 0.68 (0.66, 0.70) |

For AUC(LTC) only LTCs were considered positive. For AUC(IC) only ICs were considered positive. For completeness, we included AUC(IC | LTC) where both ICs and LTCs were considered positives. All women not considered positive were negative. A paired (unless stated otherwise) two-sided DeLong test was used to determine whether two AUCs were significantly different and used to compute 95% confidence intervals (CI).30

In series A, the baseline results $R^N_{raw→D-tr}$ and $R^N_{raw→D-tr}$ yielded AUCs(IC) of 0.68 and 0.70, respectively. The AUCs(LTC) were 0.65 and 0.64, respectively. There were no significant differences between the AUCs(IC) or AUCs(LTC) with $p(IC)=0.24$ and $p(LTC)=0.35$, which indicated good adaptation across raw views and robust model training. $R^N_{raw→N-tr}$ and $R^N_{raw→N-tr}$ yielded
near random performance that was significantly worse compared to baseline $p<0.001$ for ICs and LTCs.

In series B, the results $R_{\text{flav} \rightarrow \text{flav}}^{N \rightarrow D}$ and $R_{\text{flav} \rightarrow \text{flav}}^{D \rightarrow D}$ yielded AUCs(IC) of 0.69 and 0.70, respectively. The AUCs(LTCs) were both 0.64. There was no difference in AUC between these two results with $p(\text{IC})=0.24$ and $p(\text{LTC})=0.9$. $R_{\text{flav} \rightarrow \text{flav}}^{N \rightarrow D}$ and $R_{\text{flav} \rightarrow \text{flav}}^{D \rightarrow D}$ were not different from the corresponding baselines $R_{\text{raw} \rightarrow \text{raw}}^{N \rightarrow D}$ and $R_{\text{raw} \rightarrow \text{raw}}^{D \rightarrow D}$ with $p(\text{IC})=0.57$/$p(\text{LTC})=0.64$ and $p(\text{IC})=0.76$/$p(\text{LTC})=0.39$, respectively. This suggested that the texture models, trained with flavor augmentation, generalized to an unseen flavor domain without any loss of risk assessment performance.

In series C, the baseline results $R_{\text{raw} \rightarrow \text{raw}}^{D \rightarrow N}$ yielded an AUC(IC) of 0.66 and AUC(LTC) of 0.60, which is low due to difference in curation of datasets. $R_{\text{flav} \rightarrow \text{proc}}^{D \rightarrow N}$ yielded an AUC(IC) of 0.66 as well, but a significantly higher AUC(LTC) of 0.63 with $p(\text{LTC})<0.001$, which indicated that the flavor model adapted better to unseen processed views than the baseline on raw views. $R_{\text{raw} \rightarrow \text{proc}}^{D \rightarrow N}$ yielded near random performance which was significantly worse than the baseline with $p<0.001$ for ICs and LTCs. $R_{\text{flav} \rightarrow \text{proc}}^{D \rightarrow N}$ was not different from $R_{\text{raw} \rightarrow \text{proc}}^{D \rightarrow N}$ with $p(\text{IC})=0.17$ and $p(\text{LTC})=0.42$, using an unpaired test. This suggested that the texture model, trained with flavor augmentation, performed equally well when validated on flavorized Siemens views of Hologic-processed views.

In series D, $R_{\text{flav} \rightarrow \text{flav}}^{D \rightarrow N}$ did not yield significantly higher AUCs than $R_{\text{flav} \rightarrow \text{flav}}^{N \rightarrow D}$, with $p(\text{IC})=0.06$ and $p(\text{LTC})=0.14$. However, $R_{\text{flav} \rightarrow \text{flav}}^{D \rightarrow N}$ gave a higher AUC(LTC) than $R_{\text{flav} \rightarrow \text{flav}}^{N \rightarrow D}$ with $p(\text{LTC})=0.02$. There was only partial evidence that doubling the training dataset size increased risk estimation performance. $R_{\text{flav} \rightarrow \text{flav}}^{D \rightarrow N}$ gave a significantly higher AUC(LTC) of 0.68 compared to $R_{\text{flav} \rightarrow \text{flav}}^{D \rightarrow N}$, with $p(\text{LTC})<0.001$ suggesting that established risk-factors, combined with texture in a small neural network, contributed with additional information on long-term risk.

5.3 High-risk Subgroups and Distribution of Cancers Texture and PMD quantiles

In the group of women in the Danish testing dataset with the highest risk from $R_{\text{flav} \rightarrow \text{flav}}^{D \rightarrow N}$, 25.5% of women had an IC and 24.8% had an LTC.

Using risk from $R_{\text{flav} \rightarrow \text{flav}}^{D \rightarrow N}$, the sensitivities at 90% specificity was 25.5% for ICs and 25.5% for LTCs.

Using texture risk from $R_{\text{flav} \rightarrow \text{flav}}^{D \rightarrow N}$ and PMD quantiles of healthy women, three matrices were created using and shown in Fig. 6.
Fig. 6. Matrix plots of (a) the distribution of all women in the Danish testing dataset, (b) distribution of ICs, and (c) distribution of LTCs in texture and PMD quantiles. Texture scores were from $R_{flav\rightarrow flavD\&N-tr\rightarrow D-trt}$. Quantiles were calculated in healthy women only. Underneath each percentage in (a) and (b), the odds ratios (OR) are specified when compared to the reference group and exact Fisher p-values are in parentheses.

The leftmost matrix (a) shows the distribution of all women, the middle matrix (b) shows the distribution of ICs, and the rightmost matrix (c) show the distribution of LTCs. Texture and PMD were correlated as most cancers were found near the diagonal and in the top right corner in all matrices. The highest ORs found in both matrices were in the top rows. This indicated that, at any PMD quantile, being in the fourth texture quantile yielded a significantly higher OR of having an IC or LTC compared to the reference group. Women in (D3, T4) corresponded to 8.3% of women and accounted for 20.9% of ICs and 14.2% of LTCs. However, simply selecting the top 8.3% of women with highest risks, using $R_{flav\rightarrow flavD\&N-tr\rightarrow RF\rightarrow D-trt}$, accounted for 24.2% and 22.1% of ICs and LTCs, respectively.

Notably, the highest ORs for each row are not in D4 column, which indicate that high-density women are not at the highest risk when simultaneously considering texture. This effect seems most pronounced for LTCs where the highest ORs are located in the D1 or D2 column.

6 Discussion

In this study, we developed a clinically usable, cross-vendor, and robust texture model for long-term breast cancer risk. By employing a flavor augmentation-based domain adaptation technique, we enabled the texture model to adapt to the domain of unseen Hologic-processed views in a large independent dataset. The texture model was trained to rely on features of healthy breast tissue to optimally learn long-term risk features. The texture model was trained as an ensemble model on a curated training dataset that prevented erratic convergence and yielded a robust risk estimate.

6.1 The Texture Model Domain Adaptation Abilities

In the series A experiments, results suggested that the texture model adapted well across domains of raw views from different devices, and that the ensemble training gave consistent results when
populations and devices changed. $R_{\text{raw} \leftrightarrow \text{flav}}$, $R_{\text{flav} \leftrightarrow \text{tst}}$, $R_{\text{tst} \leftrightarrow \text{proc}}$ failed to adapt to processed views which indicated that a domain adaptation technique was needed to support a clinical workflow with processed views.

In the series B experiments, we simulated a new scanner by using a seventh and unseen flavor during testing. In this case the flavor texture models adapted equally well to an unseen flavor and performed equal to their raw baseline counterparts. This indicated that the augmentation-based domain adaptation technique succeeded in enabling the model to learn risk features in vendor-agnostic flavor views and might enable integration with the typical clinical workflow.

The series C experiments showed that the flavor augmentation enabled to train a texture model that adapted better to Hologic-processed views than its corresponding raw baseline. It also confirmed possible integration in a clinical workflow. $R_{\text{raw} \leftrightarrow \text{raw}}$ suggested that there was a bias in which women from the Dutch population, for which both a raw and a Hologic-processed view were available. Series B and C experiment highlighted the gain of training on flavorized raw views. There was no need to collect large datasets of processed views across several sites and scanners, as flavor augmentation successfully enabled training a cross-vendor texture model using augmented single site data only.

The series D experiment showed only little evidence that doubling the dataset increased performance. These results indicated that the performance is close to the upper limit for this particular model. However, adding age, clips, and PMD yielded better long-term risk assessment. If used in a clinical workflow this best performing full risk model should be used.

### 6.2 Comparison to Existing Risk Models

We compared the texture model to traditional models for estimating risk. The Gail model achieved an AUC of 0.60 (95% CI 0.58 to 0.62) across 29 datasets in 5 to 10-year risk estimation. The TC model achieved, in 35,921 women with six years follow-up, an AUC of 0.62 (95% CI 0.60 to 0.64).

The BOADICEA model included all established risk factors: personal, familial, reproductive, hormonal, and lifestyle factors along with polygenic risk scores, breast density, and pathogenic variants. It achieved an AUC of 0.70 (95% CI 0.66 to 0.73) for five-year risk in 5,693 women. Our texture model achieved a considerably higher AUC than the Gail and TC model, but the BOADICEA model achieved a slightly higher five-year AUC than our model that yielded 0.68 (95% CI 0.66 to 0.70), however likely not significantly so, and the study sample was less than a tenth the size of ours. The texture model, however, has the advantage using screen-available information only as opposed to collecting patient history, gene data, and questionnaires.

A. Gastounioti et al. used iCAD to achieve an AUC of 0.68 (95% CI 0.64-0.72) in detecting cancers for 5,139 screen negative women with up to two years after screening. Comparably, we achieved an AUC(IC) of 0.71 (95% CI 0.67 to 0.75) for the same interval after screening (IC).

Kallenberg et. al achieved an AUC of 0.61 (95% CI 0.57 to 0.66) using a texture model. Our texture model was inspired by this study and results indicate a clear improvement over the previous results.
Dembrower et al. achieved an AUC of 0.65 (95% CI 0.63 to 0.66) one- to six-year risk estimation in a dataset of 2,283 women.\textsuperscript{17} This AUC corresponded well with our results: 0.65 (95% CI 0.63 to 0.67). However, cancers diagnosed close to screening are generally easier to detect, as indicated by our texture model’s AUC(IC) of 0.71. Consequently, the AUC for our texture model could be hypothesized to increase if we evaluated risk after one year after screening as well.

Mirai obtained AUCs between 0.68 to 0.73 in 6 months to five-year risk estimation in seven validation datasets.\textsuperscript{18} These results are not comparable to ours, as the authors considers all cancers six months to five years after screening as positives and 42% to 82% of cancers were diagnosed in the first year. This distribution of time from screening to diagnosis might not be representative of a population-based screening program.\textsuperscript{20}

Neither of the four studies on DL risk models assessed adaptation to unseen vendor-processed views.

6.3 Clinical Relevance

For both ICs and LTCs it was observed, in Fig. 6, that being in the fourth texture quantile, at any PMD quantiles, yielded the largest ORs compared to the reference group. This result showed that texture consistently contributed to better risk assessment besides PMD. Women in third PMD quantile and fourth texture quantile could be flagged as high-risk.

It turned out that when selecting 8.3% of the women with the highest risk, using the texture risk combined with established risk factors, 24.2% of ICs and 22.1% of LTCs were detected. These percentages were higher than the percentages when choosing women in (D3, T4) which also corresponded to 8.3% of the women, but only accounted for 20.9% of ICs and 14.2% of LTCs. This suggest that, in a clinical workflow, simply choosing women with the highest risk is more effective that using texture/PMD quantiles to flag women as high-risk. Using texture/PMD quantiles, we observed that the highest ORs were not found in the D4 column indicating that supplementary screening strictly for high-density women\textsuperscript{5} may not be optimal if texture was considered as well. A large proportion of persistently healthy women, with low texture risk, will unnecessarily receive supplemental screening while some high-risk high-texture women will be missed. However, it is well-known that screening sensitivity is lower for women with high breast density due to a masking effect.\textsuperscript{5} It may therefore still be valuable for clinicians to consider PMD along with texture risk.

6.4 Limitations

Although we successfully developed a texture model, our study was limited by using only one testing dataset of vendor-processed view, which was inherently different from the Danish consecutive screening cohort due to differences in curation. We cannot accurately assess whether six flavors for training was sufficient to generalize to other vendors though results suggest it. The study was limited by the similarity of the populations, as they had the same screening intervals, number of readers, and were demographically comparable.
6.5 Conclusion and Future Perspectives

We successfully identified a clinically relevant high-risk cohort using a mammographic texture model for long-term risk. We would like to investigate how to incorporate it into screening workflow in collaboration with radiologists, and whether texture correlates with certain subtypes of breast cancer. The robust cross-vendor texture model should further be validated in a cohort with longer follow-up, more demographic diversity, and with different screening intervals to confirm measured performance and robustness. Lastly, we want to supplement the texture model, that estimates long-term risk, with a detection model, that estimate immediate/short-term risk using localized findings or precursors. Combining two such risk models, after being trained optimally and separately, might increase short- and long-term risk estimation, as no weak risk signals are suppressed.

Disclosures

None of the authors have anything to disclose.

Acknowledgements

This work was partly supported by Eurostars (grant E9714 IBSCREEN) and was approved by the Danish Data Inspection Agency and Danish Patient Safety Authority (reference number 3–3013–2118).

References


4.5 Breast Cancer Risk Assessment Improved by Combining Artificial Intelligence for Lesion Detection and Mammographic Texture

4.5.1 Preliminary

The aim of this study was to investigate whether the AI-system, used in Paper I, in combination with the texture model, developed and validated in Paper II, could be combined to improve overall risk assessment. All the relevant methods and components of this study have been described in the end of the paper below in the supplemental material section.
Breast Cancer Risk Assessment Improved by Combining Artificial Intelligence for Lesion Detection and Mammographic Texture

Andreas D. Lauritzen. MSc. Department of Computer Science, University of Copenhagen
My C. von Euler-Chelpin. PhD. Department of Public Health, University of Copenhagen
Elsebeth Lynge. PhD. Department of Public Health, University of Copenhagen
Ilse Vejborg. PhD. Department of Breast Examinations, Gentofte Hospital
Mads Nielsen. PhD. Department of Computer Science, University of Copenhagen
Nico Karssemeijer. PhD. ScreenPoint Medical
Martin Lillholm. PhD. Department of Computer Science, University of Copenhagen

A diagnostic AI system and a mammographic texture model for short- and long-term risk, respectively, were trained separately and combined subsequently to improve overall mammography-based breast cancer risk assessment.

Key points:
- In this retrospective study, a mammography-based breast cancer risk model, combining an artificial intelligence system for lesion detection and a mammographic texture model, was trained using 39,345 exams of Dutch women.
- In an independent screening cohort of 119,650 Danish women, the combination model improved risk segregation compared to both the artificial intelligence system and texture model.
- 10% of women with the highest combined risk accounted for 44.7% of interval cancers and 33.7% of long-term cancers.

Abbreviations
Artificial intelligence (AI)
Deep learning (DL)
Screen-detected cancers (SDC)
Interval cancers (IC)
Long-term cancers (LTC)
Area under the receiver operating characteristic curve (AUC)
Tyrer-Cuzick (TC)
Percentage mammographic density (PMD)
Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA)
Abstract

Background
Safe clinical implementation of personalized breast cancer screening requires reliable risk models. Recent mammography-based models can estimate short- or long-term risk, yet risk assessment may improve by combining short- and long-term risk models.

Purpose
To evaluate whether risk assessment improved when combining a diagnostic artificial intelligence (AI) system and mammographic texture model.

Materials and Methods
This retrospective study examined a Danish screening sample of consecutively screened women in the study period from November 2012 to December 2015 with at least five-year follow up.
A commercially available diagnostic AI system for lesion detection produced exam scores, indicating likelihood of cancer, as a surrogate for short-term risk in the Danish screening sample.
A mammographic texture model, trained on a separate Dutch training sample, assessed long-term risk in the Danish study sample.
Exam scores and texture risks were evaluated individually and combined using a neural network to determine joint short- and long-term risk in mammograms from the study period. Model performances were evaluated for interval cancers (IC) within two years from screening and long-term cancers (LTC) diagnosed two years after screening using area under the receiver operating characteristic curve (AUC) and compared with DeLong's test.

Results
The Danish screening sample comprised 119,650 women (median age: 59, IQR: 53, 64) of whom 320 developed IC and 1401 developed LTC. The combination model achieved a higher AUC for ICs than the exam score (0.78 vs. 0.75, P<0.001) and texture risk (0.78 vs. 0.71, P<0.001), and a higher AUC for LTCs than the exam score (0.71 vs. 0.69, P<0.001) and texture risk (0.71 vs. 0.65, P<0.001). The 10% of women with highest combined risk accounted for 44.7% (143/320) of ICs and 33.7% (472/1401) of LTCs.

Conclusion
Combining a diagnostic AI system and mammographic texture model improved risk assessment for ICs and LTCs, enabled identification of high-risk women, and could support personalized screening.
Introduction

Breast cancer risk differs among women (1–3). Yet, most breast cancer screening programs are one-size-fits-all where all women are screened with the same protocol. Risk-stratified screening is proposed to optimize screening for individual women and to better exploit resources, an important factor given the increasing lack of specialized radiologists (4–6).

A variety of risk models are used in practice. The Gail and Tyrer-Cuzick (TC) models demonstrated that five-year/ten-year/lifetime risk could be assessed using clinical risk factors and pathogenic variants (7,8). The BOADICEA model further included breast density, lifestyle factors, and complex genetic information for improved risk assessment (9). However, extensive data collection processes associated with these models complicates implementation in screening practices. Deriving risk using screen-available data such as mammograms minimizes overhead and maximizes clinical utility.

Mammography-based models characterize parenchymal/dense tissue patterns associated with risk. The Wolfe, Boyd, and Tabár classifications demonstrated that mammograms could be manually categorized by percentage mammographic density (PMD) and tissue heterogeneity (often called mammographic texture) to identify high-risk women (1–3). Texture and PMD are global mammographic features associated with long-term risk. Conversely, minimal localized signs of developing cancer (precursors) are associated with short-term risk (10). Moreover, screening false-positive women have increased long-term risk (11).

Mammography-based deep learning (DL) models estimate risk robustly and objectively without the need for questionnaires or genetic workup (10,12–14), and are therefore better suited for breast screening practices than traditional risk models.

Diagnostic models are trained to detect suspicious lesions and support diagnostic assessment. These models may perform as well as radiologists (15) and can be reused for short-term risk by detecting localized precursors of cancer (10,13,14).

Texture models are trained to learn global features in healthy tissue indicative of breast cancer susceptibility, e.g., breast tissue heterogeneity, density, or both, for long-term risk (14,16,17).

Conflated models rely on both global and localized features to assess risk (18). However, training for both simultaneously might yield a subpar model (14).

The outcome of conflating a short-term and a long-term DL risk model, trained separately and optimally and subsequently combined, has not previously been reported.

We combined a short-term and a long-term DL risk model complementing each other to obtain a state-of-the-art mammography-based risk estimate. A commercially available diagnostic AI system was used for short-term risk and combined with a texture model for long-term risk. The models were trained separately and combined afterwards using a combination model. We assessed the effect of adding age and PMD to the combination model and ability to identify high-risk women.
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).

Screening Data Collection and Study Sample

The Danish screening population was 50–70-year-old women screened biennially in the Capital Region of Denmark breast cancer screening program between November 2012 and December 2015 using Mammomat Inspiration systems (Siemens Healthineers).

The Dutch screening population was 50–74-year-old women screened biennially in Utrecht, the Netherlands between August 2003 and December 2015 on Lorad Selenia systems (Hologic).

For both populations, exams were read independently by two specialized breast radiologists from a team of high-volume readers described in supplemental material S1. Upon disagreement on need for recall and diagnostic assessment, a consensus meeting was held, or a third radiologist broke the tie. Diagnostic assessment was performed with a triple test, including a clinical mammography, ultrasound, and biopsy if imaging deemed it necessary. Breast cancers were invasive or ductal carcinoma in situ diagnosed by a triple test.

Screen-detected cancers (SDC) were defined as cancers diagnosed within six months after recall. Women without a SDC yet diagnosed within 24 months from screening had interval cancers (IC), usually detected through clinical symptoms or screening-unrelated events. Women diagnosed after 24 months from screening had long-term cancers (LTC). LTCs were SDCs or ICs in subsequent screening rounds. Women without a diagnosis throughout follow up were considered healthy.

All mammograms were obtained in the study period, and their labels (SDC/IC/LTC/healthy) were based on a follow-up period of at least five years from the screening visit in the study period.

The main study sample, the Danish screening sample, was collected from the Danish screening population as one screening per woman with valid mediolateral oblique (MLO) and craniocaudal (CC) views for each breast (Figure 1a). All results were reported using this Danish screening sample.

A Dutch training sample was collected and used only to train the texture and combination model. The Dutch training sample was collected from the Dutch screening population as one screening per women with valid MLO and CC views (Figure 1b). Before 2010, Dutch women only had two views captured unless their breast density was high. Consequently, the median breast density was expected to be exaggeratedly high compared to the Danish screening sample. Women with prior breast cancer or with biopsy clips in both or in the cancer contra-lateral breast were excluded. Clips might indicate a noncancer recall thus elevated long-term risk. This effect was removed from texture model training by excluding views with clips. The Dutch training sample was independent from the Danish screening sample. The Danish screening population was reported previously (13,19,20).
Figure 1: Flow diagrams of the data collection processes for (a) the Danish screening sample and (b) the Dutch training sample.

*Visible artifacts cover foreign objects, implanted medical devices, and miscellaneous disruptive image artifacts. SDC=screen-detected cancer. IC=interval cancer. LTC=long-term cancer.

Diagnostic AI System for Short-term Risk

The diagnostic AI system was Transpara® (v1.7.0, ScreenPoint Medical) (15,21,22). The system’s DL models were trained on mammograms from a large number of mostly European and US sites to detect lesions and calcifications suspicious for breast cancer. Findings were combined into a continuous exam score between 0-10, where 10 indicates a high likelihood of cancer. The AI system scored women in the Danish screening sample. The AI system was trained prior to this study on independent data.

Mammographic Texture Model for Long-term Risk

The texture model was trained solely on the Dutch training sample to learn mammographic texture patterns associated with long-term risk. The training procedure is detailed in supplemental material S2. The output texture scores were combined with presence of clips, using the combination model described below, and together referred to as the texture risk. Women in the Danish screening sample were scored for texture risk using the trained model.

Combining Short- and Long-term Risk Models

The above-mentioned risk co-variates, i.e., texture risk, exam score, age, and PMD, were combined using a neural network called the combination model, see Figure 2. The combination model was trained using risk co-variates derived exclusively from the Dutch training sample to identify future cancers (ICs/LTCs) from other women (healthy/SDCs). The combination model’s output was referred to as the combined risk. The training procedure and PMD estimation is detailed in the supplemental material S3. Two combinations models were trained: one including the AI system’s exam score and texture risk, and another with all risk co-variates. Both combination models were applied to the Danish screening sample.
Figure 2: Schematic of the combination model. A mammogram is scored individually for texture risk, lesions (exam score), and percentage mammographic density. These three risk co-variates are then given as input, along with age, to the combination model that produces a single continuous combined risk.

Statistical Analysis

Statistical analyses (A.D.L.) were done in R (v4.1.0, The R Foundation). Risk assessment performances for the AI system’s exam score, texture risk, and combination models were measured using area under receiver operating characteristic curve (AUC). DeLong's method was used to calculate 95% confidence intervals (CI) and testing AUCs differences with a 0.05 significance level (23). Using the Dutch training sample, the superior of the two combination models was found by the highest AUC for ICs and LTCs. Using the superior model on the Danish screening sample, we reported sensitivities at 90% specificity, percentage of ICs and LTCs in the 10% and 20% of women with the highest combined risk, and number of SDCs in the 10% and 20% of women with the lowest combined risk.

Results

In the following sections, we present sample demographics, risk assessment performances in the Danish screening sample, sensitivity at 90% specificity, and identification of clinically relevant high-risk subgroups using the combination model.

Demographic of the Study Samples

The demographics of the Danish screening sample and the Dutch training sample are presented in Table 1. The Danish training sample comprised 119,650 women after inclusions and exclusions as described in Figure 1a. 320 women had ICs with a median follow-up time of 17.3 months. 1401 women had LTCs with a median follow-up time of 37.9 months. 117,030 women were healthy throughout the five-year follow-up period. The median age was 59 years (IQR: 53, 64) and the median PMD was 9.8%.

The Dutch training sample comprised 39,345 women after inclusions and exclusions as described in Figure 1b. 152 women had ICs with a median follow-up time of 15 months. 808 women had LTCs with a median follow-up time of 47 months. 37,922 women had no breast cancer diagnosis throughout the five-year follow-up period. The median age was 56 years (IQR: 52, 63) and the median PMD was 12.7%. The median PMD in the Dutch training sample was, as expected, higher than in the Danish screening sample due to the differences in screening protocols.
Table 1: Danish screening sample and Dutch training sample demographics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Danish screening sample</th>
<th>Dutch training sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of women</td>
<td>119,650</td>
<td>39,345</td>
</tr>
<tr>
<td>SDCs</td>
<td>899</td>
<td>463</td>
</tr>
<tr>
<td>Median SDC follow-up in months* (IQR)</td>
<td>0.9 (0.7, 1.4)</td>
<td>-</td>
</tr>
<tr>
<td>ICs</td>
<td>320</td>
<td>152</td>
</tr>
<tr>
<td>Median IC follow-up in months* (IQR)</td>
<td>17.3 (12, 21.2)</td>
<td>15 (10, 19)</td>
</tr>
<tr>
<td>LTCs</td>
<td>1401</td>
<td>808</td>
</tr>
<tr>
<td>Median LTC follow-up in months* (IQR)</td>
<td>37.9 (28, 50.9)</td>
<td>47 (34.3, 50)</td>
</tr>
<tr>
<td>Healthy women</td>
<td>117,030</td>
<td>37,922</td>
</tr>
<tr>
<td>Median age (IQR)</td>
<td>59 (53, 64)</td>
<td>56 (52, 63)</td>
</tr>
<tr>
<td>Women age 50-54 (% of sample)</td>
<td>37288 (31.2%)</td>
<td>16,963 (43.1%)</td>
</tr>
<tr>
<td>Women age 55-59 (% of sample)</td>
<td>27296 (22.8%)</td>
<td>7,464 (19%)</td>
</tr>
<tr>
<td>Women age 60-64 (% of sample)</td>
<td>25584 (21.4%)</td>
<td>6,811 (17.3%)</td>
</tr>
<tr>
<td>Women age 65- (% of sample)</td>
<td>29482 (24.6%)</td>
<td>8,107 (20.6%)</td>
</tr>
<tr>
<td>Median PMD (IQR)</td>
<td>9.8% (3%, 19.5%)</td>
<td>12.7% (4.7%, 23.7%)†</td>
</tr>
<tr>
<td>Clips (% of sample)</td>
<td>2271 (1.9%)</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1: * Follow-up time in the NL training sample were given in whole months. Consequently, all SDC follow-up times were 0 months, and follow-up times for ICs and LTCs were integers. †The observed median PMD in the Dutch training sample was higher than the observed PMD in the Danish screening sample as we expected due to differences in curation of data in the two populations.

SDC=screen-detected cancer. IC=interval cancer. LTC=long-term cancer. IQR=interquartile range.

Risk Assessment Performances

The risk assessment performances for the AI system, the texture model, and the combination model with different combination of risk co-variates, were assessed using AUC in the Danish screening sample. In these analyses, all SDCs detected in the study period have been excluded. The detection performance for ICs vs. healthy was denoted AUC(IC). The detection performance for LTCs vs. healthy was denoted AUC(LTC). The detection performance for women who, in the follow up period, developed IC or LTC vs. healthy was denoted AUC(IC | LTC). All AUCs are presented in Table 2.

Table 2. Detection performance in the Danish screening sample.

| Co-variates                  | AUC(IC)         | AUC(LTC)         | AUC(IC | LTC)          |
|------------------------------|-----------------|------------------|------------------|
| a. Texture risk              | 0.71 (0.68, 0.74) | 0.65 (0.64, 0.67) | 0.66 (0.65, 0.67) |
| b. Exam score                | 0.75 (0.72, 0.78) | 0.69 (0.68, 0.71) | 0.70 (0.69, 0.72) |
| c. Texture risk + Exam score | 0.78 (0.75, 0.80) | 0.71 (0.70, 0.73) | 0.73 (0.71, 0.74) |
| d. Texture risk + Exam score + Biomarkers* | 0.77 (0.75, 0.80) | 0.72 (0.71, 0.73) | 0.73 (0.72, 0.74) |

Table 2: Area under the receiver operating characteristic curve (AUC) for detection of interval cancers (IC), long-term cancers as one group (IC | LTC). The 95% confidence intervals are shown in parentheses. * Biomarkers = Age and percentage mammographic density (PMD).

The texture risk achieved an AUC(IC) of 0.71 (95% CI: 0.68, 0.74), an AUC(LTC) of 0.65 (95% CI: 0.64, 0.67), and an AUC(IC | LTC) of 0.66 (95% CI: 0.65, 0.67).

The exam score of the AI system achieved an AUC(IC) of 0.75 (95% CI: 0.72, 0.78), an AUC(LTC) of 0.69 (95% CI: 0.68, 0.71), and an AUC(IC | LTC) of 0.70 (95% CI: 0.69, 0.72).

The first combination model with two risk co-variates, texture risk and exam score, achieved an AUC(IC) of 0.78 (95% CI: 0.75, 0.80), an AUC(LTC) of 0.71 (95% CI: 0.70, 0.73), and an AUC(IC | LTC) of 0.73 (95% CI: 0.71, 0.74). This first combination model yielded a higher AUC(IC) than both texture risk alone with p<0.001 and the AI system’s exam score alone with p<0.001. The first combination model yielded a higher AUC(LTC)
than both texture risk alone with p<0.001 and the AI system’s exam score alone with p<0.001. This first combination model also yielded a higher AUC(IC | LTC) than both the texture risk alone with p<0.001 and the AI system’s exam score alone with p<0.001.

When including all risk co-variates into the combination model (texture risk, exam score, age and PMD), the AUC(IC) was 0.77 (95% CI: 0.75, 0.80), AUC(LTC) was 0.72 (95% CI: 0.71, 0.73), and AUC(IC | LTC) was 0.73 (95% CI: 0.72, 0.74). This full combination model with age and PMD did not achieve a significantly higher AUC(IC) or AUC(IC | LTC) than the combination model without age and PMD with P=0.65 and P=0.08, respectively. However, this full combination model achieved a significantly higher AUC(LTC) than the combination model without age and PMD with P=0.03.

**Clinically Relevant High-risk Subgroups**

The two combination model’s discriminatory abilities were equal when compared to each other in the Dutch training sample (using an 80%/20% training/validation split, see supplemental material S3). For AUC(IC), AUC(LTC), and AUC(IC | LTC) the tests yielded P=0.17, P=0.16, and P=0.42, respectively. As there were no difference in overall risk assessment, the first (and the simplest) combination model was chosen to identify a high-risk subgroup of women. In these analyses, SDCs were excluded as well. At 90% specificity, the first combination model achieved a sensitivity of 45.3% (145/320, 95% CI: 39.8%, 50.9%) for ICs, 34.6% (485/1401, 95% CI: 32.1%, 37.2%) for LTCs, and 36.6% (630/1721, 95% CI: 34.3%, 38.9%) for ICs and LTCs combined.

Using the first combination model, the subgroup of women with the 10% highest combined risk accounted for 44.7% (143/320) of ICs, 33.7% (472/1401) of LTCs, and 35.7% (615/1721) of ICs and LTCs combined. Using the first combination model, the subgroup of women with the 20% highest combined risk accounted for 61.3% (196/320) of ICs, 51.4% (720/1401) of LTCs, and 53.2% (916/1721) of ICs and LTCs combined. There were no SDCs in the group of women with the lowest 10% combined risk, and one SDC (0.1% of 899 SDCs) in the group of women with the lowest 20% combined risk.

**Discussion**

To obtain a state-of-the-art mammography-based breast cancer risk estimate, we trained a combination model comprising a diagnostic artificial intelligence (AI) system for short-term risk, relying on detection of localized subtle signs of developing cancers, and a texture model for long-term risk, relying on global texture features.

The combination risk model was validated retrospectively in an independent study sample of 119,650 consecutively screened Danish women. The combination model’s risk assessment was improved for interval cancers (IC) up to two years after screening, long-term cancers (LTC) two years after screening, and ICs and LTCs combined, when compared to either the AI system’s exam score or texture risk alone (P<0.001 in all cases).

The full combination model with the biomarkers age and percentage mammographic density (PMD) yielded a higher AUC than the combination model without biomarkers for LTCs (P=0.03). For ICs alone and ICs and LTCs combined, the risk assessment performances were unchanged (P=0.65 and P=0.08, respectively). There was therefore only partial evidence that including age and PMD increased overall risk segregation. The combination model predicted ICs and LTCs with a sensitivity indicating it may support clinicians in moving towards safe and efficient personalized breast cancer screening.
The Dutch training sample, used to train the combination and texture model, had a higher median PMD (12.7%) than the Danish screening sample (9.8%), potentially due to differences in screening protocols and data curation. However, the texture model still generalized across populations and mammographic devices.

Clinical Relevance
The combination model identified a high-risk subgroup of women with the 10% highest combined risk, accounting for 44.7% (143/320) of ICs and 33.7% (472/1401) of LTCs, who may be eligible for supplemental screening. The combination model simultaneously identified a high-risk subgroup comprising 20% of women with the highest combined risk, and a low-risk subgroup comprising 20% of women with the lowest combined risk. The high-risk subgroup accounted for 61.3% (196/320) of ICs and 51.4% (720/1401) of LTCs, whereas only one SDC was in the low-risk subgroup in the study period. Consequently, custom risk-stratified screening strategies could be composed using the combination model depending on clinical needs.

Comparison to Other Risk Models
We compared the combination model to clinical risk models, and another conflated deep learning (DL) model for joint short- and long-term risk. Recently, McCarthy et al. investigated the Gail and Tyrer-Cuzick (TC) model in a validation sample of 35,921 women with six years follow-up. The Gail and TC models achieved AUCs of 0.64 (95% CI: 0.61, 0.65) and 0.62 (95% CI: 0.60, 0.64), respectively (24). The combination model achieved higher AUCs than the Gail or TC models within a comparable follow-up period. The combination model's 10-year/lifetime risk assessment is still to be determined. We hypothesize that with longer follow up, cancer labels would be more accurate as more high-risk women would develop breast cancer.

Recently, Yang et al. studied the BOADICEA model in a sample of 5,693 women with five-year follow up, which yielded an AUC of 0.70 (95% CI 0.66, 0.73) (9). Results suggest that the combination model estimates risk with a slightly higher AUC or at least as high as the BOADICEA model, using only screen-available information at five years from screening.

Mirai is a conflated DL model for risk estimation at multiple timepoints from screening up to five years (25). Mirai was validated in seven datasets from different locations measuring risk assessment performance by AUC for cancers diagnosed within five years from screening and including only screening negative women. Mirai achieved AUCs between 0.68 and 0.73 (median 0.71), which corresponded to the combination model's AUC for ICs and LTCs combined. However, in that study 42%-82% and 54%-92% of cancers were diagnosed in the first and second year, respectively (26). This overrepresentation of cancers in the first two years heavily weighs detection over risk assessment, which will lead to higher AUCs. No separate AUC were reported for LTCs alone.

Study Limitations
Our study was limited by using only one single-institution study sample. However, the combination model's training data was completely independent from the Danish screening sample and was obtained with a different mammographic device. We did not validate our model in screening protocols with different
Future Work
In future work, we will validate the combination model in samples from US clinics with demographic diversity and determine whether the combination model adapts sufficiently well to other vendors. We will study how combined risk can be translated to lifetime/absolute risk for direct comparison with traditional models in the same sample. Further including clinical risk factors into the combination model may improve risk assessment. Lastly, results suggested that, with five-year follow up, minimal signs of developing cancer, precursors, or masked cancers outweigh textural features in this study. The combination model will be validated on a sample with longer follow-up.

References


Supplemental Materials

S1 - Reader experience
In the Capital Region of Denmark breast cancer screening program from November 2012 to December 2015, seven radiologists were employed as full-time high-volume readers of screening mammography. Being a high-volume radiologist entails reading at least 5,000 screening mammograms per year. In the beginning of the Danish screening sample (November 1st, 2012) the radiologists each had 0.5, 4.5, 5.5, 6.5, 7, 7, and 18.5 years of experience, respectively, with an average of 7.1 years.

All screening mammograms in the study were read by two readers of whom at least one was a senior radiologist. Newly employed radiologists or radiologists with only clinical/symptomatic mammography experience in training can step in as first reader. The senior radiologists decide when the training is sufficient for the new radiologists to become second reader.

The radiologists in the breast cancer screening program in Utrecht are full-time high-volume readers of screening mammography as well. However, individual years of experience could not be retrieved. The screening program before 2009 has been described in (1).

S2 – Texture Model Training
The texture model was generally trained following the method described by Lauritzen et al. to ensure optimal and robust training for long-term risk (2).

Curation of Training and Validation Sets
The texture model was trained on the Dutch training sample. Specifically, five training and validation sets were curated from the Dutch training sample, and the curation process can be seen in Figure S1, and are defined by the steps below:

1. A reference texture model was trained on the full Dutch training sample in a five-fold cross validation setup to obtain reference texture scores for all women in the dataset. The healthy women with the 10% highest risk were excluded. SDCs, ICs, and LTCs with the 4% lowest reference risk were excluded as well. These exclusions mitigated erratic convergence by removing noise-inducing samples from training.
2. All cancers were matched with 20 healthy women based on age at screening.
3. Afterwards, cancers were randomly split into five folds using a cross validation strategy stratified on reference texture score, cancer group (healthy/SDC/IC/LTC), and age. The matched healthy women were allocated to the same fold as their corresponding matched cancer. Each fold now consisted of a training and a validation set.
4. From each of the five training sets only the two cancer contra-lateral views were included. For instance, given a woman with a cancer in the left breast, the model will only be trained on views of
the right breast. For the 20 matched healthy women, the same views were included for training – in this case the views of the right breast. The five validation sets consisted of all four available views. Unmatched and excluded women were randomly split in five sets, based on reference risk and cancer group, and appended to the allocated to one of the five validation sets.

Figure S1: Flow diagram of the collection process of (left) the Dutch training sample, and (right) curation of the five training and validation sets used for training the texture model. SDC=screen-detected cancer. IC=interval cancer. LTC=long-term cancer.

In curation of the five ensemble folds from the Dutch training sample, as shown in Figure 1b, each fold contained approximate 355 SDCs, 116 ICs, 621 LTC, and 21856 healthy women. These number might vary from fold to fold due to random sampling.

Training of the five instances of the texture model

After curation of the training and validations sets, raw mammographic views were converted to a processed format to ensure generalization processed mammograms from different vendors. Specifically, the full Dutch training sample, consisting of raw views, was converted into six difference formats of processed views. This was done using an image processing tool, capable of approximating vendor-specific processing pipelines, also described by Lauritzen et al. in (2).

Each training and validation set therefore contained six different processed formats of the same image.

Five texture models were trained separately on each of the five training sets, with the six processed formats, to classify cancer-contra lateral views from women with cancers (SDC/IC/LTC) from healthy women. Training was defined as minimizing the binary cross entropy using the Adam optimizer with a learning rate of $10^{-5}$ (3). During training views were further randomly augmented by rotation, scaling, or shearing. All views were scaled to a resolution of 0.255 mm per pixel and padded to a size of 1194 by 938 pixels. Training was stopped at convergence, indicated by the highest AUC, considering all cancers positive, in the validation dataset. This yielded five trained instances of the texture model. The final texture score was obtained by averaging the output from all five texture models.

The Danish screening sample was processed into one of the six processed formats described above.
S3 – Combination Model Training

The combination model was trained solely on risk co-variates derived from the Dutch training sample. To obtain exam scores for training the combination model, the AI system was applied to Dutch training sample. Texture scores used for training the combination model were obtained from each of the validations sets described in supplemental material S1. PMDs were obtained by applying a dense tissue segmentation tool developed by the University of Copenhagen trained on independent data. The segmentation tool was published and presented at ECR 2020 under the study “Validation of a Deep Learning-based Breast Density Estimation Tool on a Danish Screening Cohort in the Context of Personalised Risk-based Screening” (4). Ages were readily available in the DICOM header of the view.

The architecture of the combination model is shown in Figure S2. The combination model receives five inputs: the AI system’s exam score, the texture score, presence of clips, age, and PMD. Note that the number of inputs might vary in this study for different experiments.

The input is passed by two layers consisting of a fully connected layer, batch normalization and a dropout layer. The third and final layer outputs a single continuous measure between 0 and 1, which we refer to as the combined risk.

![Figure S2: Schematic of the full architecture of the combination model. In this example, all five risk- covariates are given as input to the model. Through three layers of fully connected, batch normalization, and dropout layers, the input is processed into a single continuous risk estimate between 0 and 1. PMD=percentage mammographic density. ReLU=rectified linear unit.](image)

The Dutch training sample was randomly split into a training set comprising 80% of the women and a validation set comprising the remaining 20% of the women. The training and validation set remained constant throughout this study. The combination model was trained on the training set until convergence indicated by the highest AUC for ICs and LTCs combined. The model was trained to classify ICs and LTCs from healthy women and SDCs by minimizing the binary cross entropy using the Adam optimizer with a learning rate of $10^{-4}$ (3).

The combination model was thus only trained on risk- covariates derived from the Dutch training sample and subsequently applied to the Danish screening sample.
Reference for Supplemental Materials


4.6 Identifying Recurrent Breast Cancer Patients in National Health Registries using Machine Learning

4.6.1 Preliminary

The main aim of this study was to develop a general framework for identifying specific patient endpoints in national health registries. The framework was trained and applied to retrospectively identifying breast cancer recurrence patients. In Denmark, patients with breast cancer is followed for up to 10 years, but patients with recurrence are inconsistently and manually coded in national registries. Information on recurrence status in the The Danish Breast Cancer Cooperative Group (DBCG) clinical database is therefore far from complete and cannot be used to directly query the recurrence patients, for research, monitoring, or quality assurance.

Prior to this study, clinicians would manually review patient’s clinical history which was extensively laborious. In Paper IV, we aimed to automate flagging of patients with recurrence based on their medical history which is archived in three different national health registries. The three registries were the DBCG clinical database, the Danish National Patient Register (NPR), and the Danish Pathology Registry (PDB) [129, 130, 118].

Given that a machine learning framework reliably can estimate likelihood of a patient having a recurrence, patients may be sorted with regards to the likelihood of recurrence in decreasing order. This of value to clinicians that are manually reviewing patients as they can start from the top and work their way down the list. This will increase the data completeness faster compared to having a list in random order.
Identifying recurrent breast cancer patients in national health registries using machine learning

Andreas David Lauritzen¹ & Tobias Berg²,³ Maj-Britt Jensen³, Martin Lillholm¹ and Ann Søegaard Knoop²

¹Department of Computer Science, University of Copenhagen
²Department of Oncology, University Hospital Copenhagen, Rigshospitalet
³Danish Breast Cancer Group, Department of Oncology, University Hospital Copenhagen, Rigshospitalet

Abstract

Background

More than 4,500 women are diagnosed with breast cancer each year in Denmark, however, despite adequate treatment 10-30% of patients will experience a recurrence. The Danish Breast Cancer Group (DBCG) stores information on breast cancer recurrence but to improve data completeness automated identification of patients with recurrence is needed.

Methods

We included patient data from the DBCG, the National Pathology Database and the National Patient Registry for patients with an invasive breast cancer diagnosis after 1999. In total, relevant features of 79,483 patients with a definitive surgery were extracted. A machine learning (ML) model was trained, using a simplistic encoding scheme of features, on a development sample covering 5,333 patients with known recurrence and three times as many non-recurrent women. The model was validated in a validation sample consisting of 1,006 patients with unknown recurrence status.

Results

The ML model identified patients with recurrence with AUC-ROC of 0.93 (95% CI: 0.93-0.94) in the development, and an AUC-ROC of 0.86 (95% CI: 0.83-0.88) in the validation sample.

Conclusion

An off-the-shelf ML model, trained using the simplistic encoding scheme, could successfully identify recurrence patients across multiple national registries with high performance. This approach might potentially enable researchers and clinicians to better and faster identify patients with recurrence and reduce manual patient data interpretation.
Introduction
More than 4,500 Danish women are diagnosed with breast cancer each year making it one of the most frequent cancers in Denmark(1). Due to improvements in diagnostics, treatments, and follow-up the survival after a breast cancer diagnosis has significantly improved over the last 20 years(2). However, recurrence still occurs in 10-30 % of patients worldwide despite surgery and adjuvant treatment(3). Specifically, patients with high lymph node burden, large tumors, and estrogen receptor (ER)-positives are at risk of late recurrence up to 32 years after primary diagnosis(4).
It is crucial to follow patients with recurrent breast cancer over time to optimize treatments, decrease mortality, and to monitor the epidemiology of recurrent breast cancer(5).
The Danish Breast Cancer Group (DBCG) maintains a large clinical database with detailed information on diagnoses, treatments, and outcomes for patients diagnosed with breast cancer(6). To incorporate information for all patients with recurrence in the clinical database, an automatic identification method is needed, considering the extensive number of breast cancer patients.
Methods for identification of patients will typically rely on use of national health registries of diagnoses or treatments or on manual input of involved clinicians. For other cancer types, identification is easily solvable by simple rule-based coding across national health registries when treatment options are specific to recurrence or when relapses occur close to the primary diagnosis date. This is, however, not the case for breast cancer recurrence.
It is therefore a non-trivial task to identify patients with breast cancer recurrence based on information on surgery, pathological findings, and anti-cancer treatments. Consequently, there is a need for more advanced methods to identify recurrence patients, which is the focus of this study.
To identify patients with recurrent breast cancer in Denmark, an algorithm using data from multiple national health registries was created. This study's main contributions are:

1. To pool and compile information on recurrent breast cancer patients across multiple national health registries, including information on diagnosis and treatments, into a simplistic data structure.
2. Apply three off-the-shelf computational models to identify recurrent breast cancer patients using this data structure.
3. Validate the performance of the superior of the three models in an independent validation sample.

This model should enable clinicians and researchers to retrospectively identify women with recurrent breast cancer in health registries in an automated fashion and allow for follow-up on a national level.

Methods and Materials
Data Sources
In this study, three different data sources were linked to collect patient information: the DBCG clinical database, the Danish National Patient Register (NPR), and the Danish National Pathology Registry (PDB)(6–8).
The DBCG clinical database holds information on all breast cancer patients in Denmark and is valid for large epidemiological studies of breast cancer treatment(9). Patients with a first-time record of an invasive breast tumour in PDB are included in the DBCG with >95%completeness(10). Data on demographics, diagnostics, treatment, and follow-up are registered via electronic case report forms. Clinical follow-up of up to ten years are registered for patients with early breast cancer included in a protocolled treatment programme (90% of
all patients. Clinical follow-up for recurrence or new malignant disease is collected from the hospital department responsible for the individual patients. A yearly reminder is sent to these departments to make sure clinicians report new patient-specific events(11). However, the reporting related to recurrence is not uniformly handled across departments resulting in partial or completely missing follow-up. The DBCG clinical database is further updated with information on vital status(12).

The NPR is a Danish register on all hospital discharges, visits, and outpatient treatments. NPR stores partial information on surgical procedures and oncological treatments including infusion dates. Date, hospital, department, disease and treatment codes on all hospital discharges and outpatient visits for the study cohort were retrieved(7).

PDB is a register with detailed pathological information on all cell and tissue samples in Denmark. Dates, hospitals, departments, pathology codes, and free-text on all benign and malignant cell and tissue samples were retrieved(8).

All data sources were linked using the unique personal identification number (PIN) assigned to all Danish citizens. Data from all registries were extracted September 2019.

Data Linking and Encoding

NPR included patient entries until June 27th, 2019. PDB contained patient entries until September 17th, 2019. We identified patients in the DBCG clinical database from January 1st, 1999, to June 6th, 2019, which defined the set of patients diagnosed with breast cancer and were used to look up patient-specific events in NPR and PDB. Patients who did not receive a definitive surgery after primary diagnosis, typically patients with primary metastatic breast cancer, were excluded as these patients do not undergo protocolled treatment and follow-up. Figure 1 shows a diagram of the selection process.

Figure 1: Top 15 contributing features of a logistic regression model fitted to the whole development sample. Importance is estimated as the coefficient(weight) divided by the corresponding standard error. NPR=Danish National Patient Register. DBCG=Danish Breast Cancer Group. PDB=Danish National Pathology Registry. BC=breast cancer. PIN=personal identification number.

Data from NPR and PDB are not depended on the last date of follow-up from the DBCG, specifically we do not censor or exclude women at their last follow-up in the DBCG.
In total, 220 different codes, describing patient-specific events, and their corresponding dates were included. These codes were chosen as being clinically relevant when monitoring breast cancer patients and were derived with involved clinicians. From the NPR, 73 codes related to a wide range of treatments (including surgery). From the PDB, 147 codes related to pathology. To distance relapse-related events from events related to the primary cancer, all registrations up to 210 days after the primary diagnosis of breast cancer were excluded from analysis. This exclusion limit was chosen prior to development of the models based on the length a typical breast cancer treatment programme.

Finally, each patient was assigned a binary feature vector of length 220, encoding presence (1) or absence (0) of each of the different codes. Consequently, the data structure used to encode patient history was a binary matrix of patient cohort size by 220.

Study Samples

In this study two study samples were collected from a subset of the entire study cohort. Firstly, a development sample which were used for development and fitting of the models, and secondly a blinded validation sample. The development sample was constructed to include all patients with known relapse and three times as many randomly chosen patients known to be relapse-free within the registered clinical follow-up time in DBCG).

The validation sample was constructed as a random sample of patients with unknown relapse-status and yielded 1,006 patients in total. These patients differ from the development sample as patients who are completely unknown – despite receiving a definitive surgery – may consist of patients who afterwards do not receive a protocolled treatment. The relapse status of these patients was manually retrieved by an external team, with medical background, without knowledge of features of the models. The relapse status of the patients was not known at the time of development of the models.

Prediction Models

The task of identifying patients with uncoded relapse was approached as a classification problem in which a model predicts recurrence from the binary feature vector. The posterior probability, provided by the model, of having relapse is referred to as the recurrence score.

Three models were evaluated for predicting relapse: vector magnitude, a logistic regression (LR) model, and a random forest (RF) model. Vector magnitude was measured as the Euclidean norm of each patient’s binary feature vector. The LR and RF models were developed and evaluated in a 10-fold cross-validation scheme stratified by relapse status, using the development sample. The final generalization performance was measured in the validation sample.

The LR and RF models were fitted with NPR and PDB features alone and combined. Likewise, the vector magnitude was measured for the same combinations of NPR and PDB features. The superior model, in terms of cross-validated classification performance on the development sample, was applied to the validation sample. We further considered two different cut-off points and evaluated sensitivity and specificity of the model in terms of identifying relapse patients.

Statistical Analysis

All computational model fitting, statistical analysis, and plotting were conducted in R(v4.1.0). LR models were fitted using the glm function from the stats namespace using the iteratively reweighted least squares method until convergence, which is defined as no significant increase in the log-likelihood between the predictions and true labels. RF models were fitted using the randomForest package ported from Breiman and Cutler’s original implementation of the algorithm with default parameter settings in the R package(13). 500 trees are constructed using random bootstrap samples, with replacements, of the input data. Approximately
two thirds of the input data are used to grow the trees and the remaining is used for testing. The area under the receiver operating characteristic curve (AUC-ROC) was measured with the pROC package. The optimal threshold of the ROC is chosen as closest top-left operating point. DeLong's method was used to obtain 95% confidence intervals (CI) and determining significance difference between two paired AUC-ROCs (14).

Balanced accuracy was calculated as the mean of the sensitivity and specificity. The function varImp from the caret package was used to determine feature importance for both LR and RF models.

Results

83,040 unique breast cancer patients diagnosed after 1999 were identified in the DBCG clinical database. 3,557 patients were excluded as they did not receive definitive surgery. The cohort, defined of merged PINs in NPR, PDB and the DBCG clinical database, after exclusions, consisted of 79,461 patients of whom 5,333 had a known relapse, 59,919 patients were relapse-free patients, and 14,209 patients had an unknown relapse status.

The development sample comprised all 5,333 patients with known relapse and 15,999 patients known to be relapse-free within follow-up. The median follow-up time for patients in the development sample was 4.70 years.

In the validation sample, 1,006 patients were randomly sampled from the 14,209 patients with unknown relapse status. Recurrence status for 89 patients could not be retrieved and were therefore excluded. Finally, the validation sample comprised 917 patients of whom 306 had a confirmed recurrence. The median follow-up time for patients in the validation sample was 3.64 years.

Figure 1 shows the full selection process for the development and validation samples.

Cross-validated Relapse Detection Performance in the Development Sample

The development sample contained 5,333 patients with known relapse and 15,999 relapse-free patients. Segregation performance of the different model are shown in Table 1.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC-ROC (NPR)</th>
<th>AUC-ROC (PDB)</th>
<th>AUC-ROC (NPR &amp; PDB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector magnitude*</td>
<td>0.83 (0.82 - 0.83)</td>
<td>0.74 (0.73 - 0.74)</td>
<td>0.81 (0.81 - 0.82)</td>
</tr>
<tr>
<td>Logistic regression, recurrence score</td>
<td>0.89 (0.88 - 0.90)</td>
<td>0.85 (0.84 - 0.86)</td>
<td>0.92 (0.92 - 0.93)</td>
</tr>
<tr>
<td>Random forest, recurrence score</td>
<td>0.90 (0.89 - 0.90)</td>
<td>0.86 (0.85 - 0.87)</td>
<td>0.93 (0.93 - 0.94)</td>
</tr>
</tbody>
</table>

Table 1: Segregation performance for different models across registries measured by AUC-ROC with 95% CI. * Cross validation does not apply for vector magnitude

Using all 220 features, the vector magnitude yields an AUC-ROC of 0.81 (95% CI: 0.81 - 0.82). Fitting an LR and a RF model using all features yielded an AUC-ROC of 0.92 (95% CI: 0.92 - 0.93) and 0.93 (95% CI: 0.92 - 0.93), respectively. The LR and RF model achieved balanced accuracies of 87.0% and 88.1% respectively.

Regardless of using NPR features alone, PDB features alone, or combining NPR and PDB the RF model performed better than the LR model (p<0.01, p<0.001, and p<0.001, respectively).

Feature Importance

The 15 most important features in an LR model, fitted on the development sample, are shown in Figure 2a. From NPR, features include general cytotoxic treatment, bisphosphonate, breast surgery and radiation therapy, but also more specific treatments such as capecitabine, and fulvestrant.

From PDB, important pathological features were included. The event of having a cytologic screening or a
malignant positive biopsy (invasive ductal carcinoma or carcinoma) contributes to the LR model.

Figure 2b shows the 15 most important features for the RF model. 11 out of 15 features are shared among the models. No features related to breast surgeries are present in the RF model's top 15 features. Using the top 15 most important features of the RF model both the RF and LR model yields an AUC-ROC of 0.90 (0.90 - 0.91).

Validation Sample Performance

The RF model trained on all 220 features from both NPR and PDB achieved an AUC-ROC 0.86 (95% CI: 0.83 - 0.88) on the validation sample. Figure 3a shows the score distribution for the confirmed recurrence patients in the validation sample.

Figure 3: (a) Histogram of the recurrence scores for the confirmed recurrence patients in (a) the validation sample and (b) the development sample, respectively.
48 (15.7%, 48/306) patients with a confirmed recurrence had a recurrence score of less than 0.1. For comparison, the score distribution for patients with known recurrence in the development sample is shown in Figure 3b, where 320 (6.0%, 320/5,333) patients are in the category with a score less than 0.1. The score distribution for relapse-free patients in the validation and development samples are shown in the appendix Figure 5a and 5b.

The ROC curve of using the recurrence score of the RF model on the validation sample to identify relapse patients can be seen in Figure 4. The optimal threshold of 0.348 yielded a specificity of 83.5% and a sensitivity of 74.8%. The manually chosen threshold of 0.1 yielded a specificity of 69.7% and a sensitivity of 84.3%. Additional metrics are shown in Table 2.

Figure 4: Receiver operating characteristic curve illustrating the diagnostic ability of the RF model in the validation sample. The two highlighted points correspond to the optimal (closest top-left) operating point and the operating point corresponding to choosing a threshold of 0.1. Inside the corresponding parentheses, the specificity and sensitivity are displayed, in that order.

<table>
<thead>
<tr>
<th>Relapse status</th>
<th>RF model Predictions</th>
<th>True</th>
<th>False</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>229</td>
<td>101</td>
<td>330</td>
<td></td>
</tr>
<tr>
<td>False</td>
<td>77</td>
<td>510</td>
<td>587</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>306</td>
<td>611</td>
<td>917</td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity = 74.8%
Specificity = 83.5%
Balanced accuracy = 79.2%

Table 2: Contingency table of relapse status and RF model predictions in the validation sample at the optimal threshold for the recurrence score at 0.348. Below are diagnostic performance statistics.

Discussion

In this study we have demonstrated a model capable of predicting relapse using data on 79,461 patients previously diagnosed with early breast cancer. A random forest (RF) model achieved an AUC-ROC of 0.93 (95% CI: 0.93-0.94) when separating relapse patients from healthy patients in a 10-fold cross validation scheme. Furthermore, in a manual validation of 917 randomly sampled patients, the RF model yielded an AUC-ROC of 0.86 (95% CI: 0.83 - 0.88). Choosing a threshold of the recurrence score of 0.348, the RF model correctly identified 74.8% of relapse patients and 83.5% of the healthy patients.
Clinical Relevance

The results of this study suggest that automatic patient identification is possible in large data pools such as health record registries with a simplistic encoding scheme and off-the-shelf computational models. This might enable researchers and clinicians to identify patients with specific, but unregistered, endpoints (such as relapse), and thus solve problems otherwise time consuming or deemed practically infeasible. It might also help triaging, by flagging low- and high-risk patients, when clinicians are manually identifying relapse patients.

The used models could assist clinicians in additionally two ways. Firstly, a set of important features, given a downstream task or endpoint e.g., relapse, can be discovered among a larger group of features without extensive prior knowledge. The models will be able to weigh each feature individually in terms of contribution to correct predictions. This is demonstrated by the high AUC of 0.90 using only the 15 most contributing features. Secondly, with extensive domain-specific knowledge of important features, the model can include a larger set of features to detect relapse patients with high performance.

We initiated the current study believing that we needed both the PDB and NPR to identify patients with relapse, but results showed that either of the two registries perform comparably well but with slightly decreased performance than when combined. In some cases, it might only be possible or feasible to include a single registry. In the case of this study, using only NPR yielded an AUC-ROC of 0.90, which is likely still a clinically usable model. With an initial detection performance using a single registry, evaluation can determine whether performance is adequate or if additional features are needed from other registries.

The RF model’s top 15 variables included several features that are clinically relevant. This covers both chemotherapies as vinorelbine and capecitabine and endocrine therapy as fulvestrant. All of which are primarily used as treatment in the recurrent setting. We also see the general code for basic cytotoxics which means the model can be used even if specific treatments are changed. Additionally, having an assessment of estrogen receptor (ER) or human epidermal growth factor receptor 2 (HER2) code contributes highly to the RF model. This is clinically translatable to the fact that having a new malignant biopsy taken after an initial breast cancer diagnosis is most likely due to a relapse.

An important factor when implementing a recurrence-patient identification model in a clinical/research practice is to contemplate whether it can stand alone or contribute to existing solutions. This decision might be influenced by what is already in place. In our case, the DBCG clinical data base, being an already long existing database, might implement the model to improve detection of late recurrences, however, only serve as supplement to the current clinical follow-up in the first ten years after diagnosis.

The Models and Encoding Scheme

Results showed that the performance of the RF model was superior to the performance of the LR model regardless of training on NPR codes and PDB codes separately or combined. This suggest that the RF model can capture more meaningful interactions when using higher dimensional training data. Using codes from both NPR and PDB yielded significantly better results than using codes from each register individually.

Simply measuring vector magnitude of the binary vector yielded a relatively high detection performance, although much lower than the LR or RF model. This suggests that the binary encoding scheme could potentially generalize to other identification problems in alike health registries without extensive knowledge about complex interactions between patient-specific events.
Performance Difference in the Development and Validation Sample
An evident flaw of the model was that 48 of 306 patients with confirmed relapse had a recurrence score of less than 0.1 in the validation sample. This was also partially expressed in the difference in AUC-ROC as it decreases from 0.93 in the development sample to 0.86 in the validation sample. There might be several reasons for the considerable amount of missed relapse patients.

The patients with unknown relapse-status have clinical features excluding them from protocolled treatment and are known to be at a higher risk of recurrence than patients who are allocated to protocolled treatment.

When clinically evaluating each of the 48 low recurrence score patients several groups of patients form. 16 patients had primary metastatic breast cancer, which were not marked as such. 16 patients died within 100 days of their primary diagnosis. For 8 patients, no treatment was administered after relapse. A significant amount of the patients had simple antineoplastic treatment courses. For most of the patients in these groups, no or very few codes were found in either registries. This suggest that there are relapse patients with less features than average due to a short disease course, simple treatments courses, or no treatment at all. This also applies to the development cohort, however the fraction of such patients is overrepresented in the validation sample, as can be seen in Figure 3a and 3b. Regardless of the choice of model, relapse patients with a low feature count are difficult to distinguish from healthy patients.

Our model identifies patients with recurrent breast cancer with high performance in terms of AUC-ROC yet still misses some relapse patients. However, in a clinical setting the model might still support existing tools for identifying relapse patients.

Related Work
A similar study from 2021 by Valachis et al. on identification and characterization of metastatic breast cancer were able to achieve a balanced accuracy of 94.1% in a validation set using a support vector machine model(15). The study’s training and validation sets were from the same regional registry and included 282 and 70 patients with metastatic disease, respectively. In our study, balanced accuracies of 88.1% and 79.2% were achieved by the RF model in the development and validation sample, respectively. However, our development sample included 5,333 patients with metastatic disease from 12 different clinics and predictive values were estimated using cross-validation. Our validation sample, which was independent in terms of model development, included 306 patients with metastatic disease. Thus, our samples were considerably larger, not only in the number of cases but also controls, and our models were rigorously validated across multiple sites.

Study Limitations
Our study has a limitation caused by the complex interactions between multiple patient registries. Such registries are frequently subject to large changes in terms of how events are coded, and this complicates the data collection and evaluation process. This applies especially when collecting data over several years and across several versions. This study was also limited by population differences between the development and validation samples. Ideally, the sampling process should use the same pool for validation and development
A source of error in fitting the models is caused by the follow up of patients with relapse-free status and their individual follow-up dates. Despite having a relapse-free status, these patients could experience a relapse after the DBCG follow-up period, and with information from the recurrence being retrieved from NPR and PDB. Thus, a limited number of patients, marked as relapse-free, might have a relapse after follow-up and this would be reflected in their feature vectors. These patients increase the false-negative rate of the model.

**Future Perspectives**

Future perspectives include encoding time explicitly for each feature during encoding. In this study, the model is agnostic to when certain events occur except for the 210 days limit. Including time and order of occurrence should increase detection performance and increase explainability of the model’s behavior. Letting the model know when a patient changes relapse status at different timepoints, might further enhance performance and would increase interpretability – e.g., a patient at first being relapse-free during of follow-up and then unknown status afterwards.

Additionally, features regarding early death, no treatment, and comorbidities could be included into the model in order to improve classification of patients with relapses within 210 days. A future model should further be able to provide a date or time interval in which the relapse occurred, possible by retrospectively evaluating features and their respective interactions.

**Approval**

This study was approved by the Knowledge Center for Data Review, Capital Region (VD-2019-201).

**References**


Appendix

Figure 5: Histograms of the recurrence scores for relapse-free patients in (a) the validation sample and (b) in the development sample.
5 Key Results

In this section, the most important findings are presented for each of the contributions.

5.1 Paper 1: Simulation of AI-based screening

• In a cohort of 114,421 consecutively screened women in the Capital Region of Denmark breast cancer screening program, a retrospective simulation of an AI-based screening protocol yielded a screening sensitivity comparable to standard screening with double reading, and a higher screening specificity.
• 71,585 of the screened women were classified as having likely normal mammograms and were excluded from double reading. Consequently, the reading workload reduced by 63%.
• 529 women recalled in standard screening were not recalled in the simulation of AI-based screening. With a total of 2107 false positive women in standard screening, this yielded a 25% reduction in false positives.
• Results suggested that AI-based screening performed consistently across BI-RADS density categories although sample sizes were too small to determine non-inferiority in screening sensitivities.
• The study contributed to full clinical implementation of AI in the Capital Region of Denmark breast cancer screening program

5.2 Abstract 1: Preliminary results of screening with AI

• Following Paper I and clinical implementation of AI in the Capital Region of Denmark breast cancer screening program, three screening cohorts were collected with a total of 118,845 screenings. On May 3rd, 2022, the threshold for likely normal mammograms was increased from 5 to 7. The overall reductions in reading workload were 27.5% and 37.5% before and after the threshold was increased, respectively, compared to before AI. The reading workload reduction of 37.5% was estimated to be approximately equal to 21,000 less screening reads per year.
• The recall rate decreased from 3.04% before AI to 2.20% with AI after the threshold was increased. It was estimated that this is equivalent to 476 less recalls or 23,000 less screening reads per year compared to standard screening.
• The detection rates in the group of women with likely normal mammograms were 10.89 per 100,000 before the threshold was increased and 13.39 per 100,000 after the threshold was increased. These detection rates were not significantly different. We assume the detection rate for double-read women is constant, and consequently, the preliminary results suggest that the threshold was increased safely from 5 to 7.

5.3 Abstract 2: Automatic dense tissue segmentation

• A dense tissue segmentation model was trained using 500 pairs of mammographic views and radiologist-annotated masks. The DICE overlap between the true and predicted breast tissue
masks was 0.97 and 0.69 for non-dense and dense tissue, respectively. The segmentation model was validated in a cohort of 53,956 women in terms of agreement with radiologists’ BI-RADS density and risk assessment.

- The continuous PMD was grouped into four categories. The agreement between the automatic density categories and consensus (max) BI-RADS density of two radiologists were substantial using the kappa statistic in the validation cohort.
- PMD combined with age in a multivariate logistic regression model yielded an AUC of 0.59 in detection of future cancers (ICs and LTCs) in the validation cohort. BI-RADS density combined with age yielded an AUC of 0.60 in detection of future cancers. The two AUCs were not significantly different.
- Results suggested that PMD, measured by the segmentation tool, reliably segmented breast and dense tissue, and could be used for breast cancer risk.

### 5.4 Paper 2: Cross-vendor texture models for long-term risk

- A mammography-based breast cancer risk model was developed to assess mammographic textural features indicative of long-term risk and trained using an augmentation-based domain adaptation technique, based on flavorization of raw to processed views.
- When the texture model was trained with flavor augmentation on 82,484 Dutch and Danish women and validated in an independent screening cohort, with views of an unseen processed format, the texture model achieved AUCs of 0.71 and 0.65 for ICs and LTCs, respectively.
- The texture model, trained solely on Danish women screened on a Siemens device, was validated in an independent sample of 25,706 Dutch women screened on a Hologic device. Long-term risk assessment improved when training the texture model with flavor augmentation and validating on Hologic-processed views compared to both training and validating the texture model with raw views.
- The overall risk assessment ability of the texture model, trained with flavor augmentation, was not different in the independent sample of Dutch women with Hologic-processed views and the independent sample of Danish women with flavor-processed views.
- In the independent Danish validation cohort, women with the 10% highest risk, as determined by the texture model combined with age and PMD, accounted for 25.5% of women who developed IC and 24.8% of women who developed LTC.

### 5.5 Paper 3: Combining diagnostic AI system and texture model

- An AI system for lesion detection, used as a surrogate for short-term breast cancer risk, and a mammographic texture model for long-term breast cancer risk were aggregated in a combination model which was trained using exams of 39,345 Dutch women.
- Using an independent cohort of 119,650 consecutively screened Danish women, the combination model was better, in term of risk assessment, than both the AI system and texture model for women who developed IC or LTC.
• Women in the independent Danish cohort with the 10% highest combined risk accounted for 44.7% of ICs and 33.7% of LTCs. These women might be eligible for supplemental screening.

5.6 Paper 4: Retrospectively identifying relapse patients

• Using health registry data of 21,332 patients of whom 5,333 developed a breast cancer recurrence, a general machine learning framework was developed and trained to identify patients with high risk of recurrence. The model achieved an AUC of 0.93 in 10-fold cross validation and 0.86 in an independent validation sample of 1,006 patients.
• Using the optimal risk threshold (top-left on the operating characteristic curve), the sensitivity was 74.8% and the specificity was 83.5% in the validation sample.
• The model is currently being used by clinicians to retrospectively identify recurrence patients on a national level. The purpose is to have a near-complete clinical database on recurrence patients for large-scale studies, monitoring, and quality assurance.
6 Discussion

The aim of this thesis was to develop, advance, and evaluate tools supporting clinicians in creating personalized breast cancer screening strategies. To do so, modern machine learning models were investigated for detection, risk, and recurrence of breast cancer. Specifically, an AI system was employed for lesion detection and reader stratification. Models for dense tissue segmentation, for textural features indicative of long-term risk, for combining multiple mammography-based risk factors, and for identifying relapse patients were developed and validated.

An AI-based screening protocol was successfully simulated and implemented clinically for which results indicated preserved screening quality while reducing the radiologists’ workload. A texture model was developed to assess risk in mammograms of different devices and was trained optimally for long-term risk by relying on features of healthy breast tissue. A combination model comprising the AI system and the texture model showed a synergistic effect in which overall risk assessment was improved. For women who already had breast cancer, a model was developed and trained to assess risk of relapse based on national health registry data and is now in clinical use.

6.1 Advances in Breast Cancer Screening Protocols

Paper I was a simulation of an AI-based breast cancer screening protocol. Similar simulation studies have been published in recent years for different populations and diagnostic AI systems [131, 132, 133, 134, 135]. Our study was therefore not novel in terms of aim and hypothesis, yet at time of publication, it was one of the largest studies including 114,421 women in the validation sample. Our study confirmed the other studies’ findings and increased the scientific evidence for reliable breast cancer screening reading triage. Furthermore, our study provided the evidence needed prior to clinical implementation in the Capital Region of Denmark breast cancer screening program. To the candidates knowledge, no other region-based breast cancer screening program have fully implemented AI for triaging likely normal mammograms for single reading. In the Hospital Universitario Reina Sofia in Córdoba, Spain, the AI system studied in this thesis, is also in use prospectively [136].

The preliminary prospective results presented in Abstract I indicated a safe rollout of the screening protocol with AI as first reader. However, the evidence is still incomplete. At the time of writing, the AI system has been in use for more than a year. This allows for cohort studies on screening before and after AI was implemented where we may measure outcomes such as cancer detection rate, lymph node status, and tumor size. However, the follow up period is currently not long enough to report the true screening sensitivity, specificity, and IC rate.

We could still calculate proxy quality indicators, e.g, one-year IC rates or one-year FP rates, in both cohorts from before and after implementation of AI. These proxy quality indicators would be comparable across the two cohorts and might approximate the standard quality indicators.
6.2 Advances in Mammography-based Breast Cancer Risk Modelling

Mammography-based breast cancer risk modelling was advanced in three aspects: A modern texture model for long-term risk was developed, the texture model generalized across mammographic devices, and by combing short- and long-term risk models, risk assessment improved to levels not reported previously in large cohorts of consecutively screened women.

The idea behind our texture model have been partly derived from the work by Kallenberg et al. from 2016 [69]. The results presented in Paper II demonstrate a large increase in risk assessment abilities, clinical usability, and reliability. To the candidate’s knowledge, no other study have attempted on using mammography-based DL model to approximate breast cancer risk in women before clinical or pre-clinical symptoms or manifestations. The rationale is that long-term risk cannot be approximated by developing a model that relies on features of visible signs of breast cancers. The hope is that, with longer follow up and more rigorous validation, we can known whether that texture models are better predictors of risk in truly healthy women than diagnostic or conflated risk models.

To the candidate’s knowledge, no other study has been published on mammography-based DL risk models that work reliably across mammographic devices from different vendors. Yala et al. demonstrated that it was possible to train a risk model in an adversarial setup to generalize across two mammographic devices of the same vendor. Although, we cannot guarantee adaptation to other vendors, e.g. GE, Philips, or Fujifilm, the contribution provides methods and preliminary results that advances the current research on this topic.

Research was also advanced in terms of improving overall risk assessment by aggregating short- and long-term risk predictors for a richer risk profile. There exists studies on risk models, trained using a conflated training procedure, that are comparable to the combination model. The AUC obtained by Mirai in a sample of screened women in Barretos, Brazil was 0.73 which was the same as the combination model validated in the Danish screening sample. However, the Barretos validation sample only included 2,057 women of whom 70 developed breast cancer within five years [99]. Furthermore, about 68.5% of cancers in the full Barretos sample were diagnosed within two years from baseline screening. The AUCs were lower in the six other validation samples [99]. These AUCs could possibly indicate that it was more feasible to train for the tasks of short- and long-term risk assessment separately and then subsequently aggregate the two risk factors.

The current level of risk assessment by the combination risk model should be disseminated to clinicians in order to initiate prospective studies on supplemental screening for high-risk women.

Additionally, a tool for automatically segmenting dense tissue in mammograms was developed. Results indicated that the segmentation tool reliably estimated PMD that was in high agreement with radiologists’ BI-RADS density. The automatically-estimated PMD was indicative of short- and long-term breast cancer risk. The segmentation model was trained on a relatively small single-institution sample of 500 mammograms, annotated by a single reader, and captured on a single type of mammographic device. It is therefore unlikely that the model will adapt to processed views of other mammographic devices. The model was developed primarily for research purposes for which we
investigated the relationship between PMD and texture. The PMD was further used to create a richer mammography-based risk profile described in Paper II and III. However, the flavor augmentation technique could be directly transferred to the training procedure of the U-net which might allow the model to reliably estimate PMD in other populations scanned on unseen devices.

Lastly, it was demonstrated that mammography-based risk models can assess risk equally well as (or better than) risk models based on clinical risk factors such as the BOADICEA model. This is important when considered the increasing implementation of breast cancer screening worldwide. As the target groups increase, risk has to be automatically and reliably estimated to support risk-stratified clinical workflows. Risk models based on clinical risk factors are inherently laborious to maintain due to the extensive and continuous collection process. This would require a large amount of resources in terms of time, budget, and staff. The mammography-based risk models studied in this thesis have been developed such that risk can be derived directly from screen-available data, i.e, the screening mammography and age. There might be an implementation cost, however when in place, the continuous costs would be marginal.

6.3 Advances in Patient Identification Using Health Registry Data

The work on relapse patient identification resulted in a general framework to identify patient endpoints using health registries. The general task of predicting patient endpoints using electronic health records or registries has been investigated many times prior and have recently been studied with different DL methodologies [137, 138, 139]. In our study, and a study by Valachis et al., it was demonstrated that breast cancer relapse could be partially predicted using health register data using simpler machine learning models such as random forests and support vector machines [140]. Even multivariate logistic regression performed almost as well as random forest. This might indicate that encoding is the limiting factor rather than choice of classifier or that the used registries does not include a sufficient amount of information related to the task.

Our study was preliminary in the sense that the encoding scheme was simple and only parsed register events to a binary feature vector. This encoding scheme comes with practical advantages during implementation, however it is likely the limiting factor in achieving accurate identification of patient endpoints.

6.4 Warrant for Standardized Risk Evaluation Procedures

Throughout the literature on mammography-based breast cancer risk models, there has never been an accepted and standardized way of reporting risk assessment abilities and even a consensus definition of risk.

In this thesis, three types of risk were considered: Immediate risk, short-term risk, and long-term risk. Immediate risk, being the risk of detecting a breast cancer in the mammogram at the time of screening, can be considered equivalent to detection. Short-term risk is the risk of detecting a breast cancer within a short period of time after screening and usually within the next screening round. In the case of biannual screening, it is within two years. Long-term risk is the risk of detecting a breast
cancer in the next screening round, i.e., at least two years from screening. If the follow up period is longer than five years or if the screening interval changes, a more nuanced and established definition of risk is needed. There is also a need for a more consistent definition of lifetime risk and how it is derived. Lifetime risk might be estimated up to the age of 70, 80, or until death, and methods to convert five or ten year risk to lifetime risk differ as well.

There are two important clinical aspects regarding training and evaluation of mammography-based risk models. Firstly, a differentiation between texture risk models and conflated risk models should be made. It is important for clinical decision-making to know whether a model was trained for detection of breast cancer (immediate risk), risk of breast cancer (short- or long-term), or a combination. Secondly, the validation sample should be representative of a screening cohort to best simulate clinical use. For instance, a conflated model might be evaluated in a sample of screening negative women where most of the cancers are found within one or two years. Detection is inherently an easier task to solve than risk assessment and the conflated model will therefore yield a higher AUC. If the conflated model is presented as a risk model and shown to yield high AUCs, this is not necessarily related to risk assessment (long-term risk) but rather high detection accuracy (immediate or short-term risk), as the model was trained for detection and the validated sample favored detection. This is a problem in current literature that should be considered in the future [141]. However, if the problem is to identify women who are eligible for supplemental screening given a high five-year risk, the conflated training and validation approach might still be valid.

As in many other areas of machine learning, a challenge dataset with predefined labels and outcomes could be created to set a standard for the research community. Some attempts have been made to establish a breast cancer risk benchmark/challenge dataset [142, 143]. However, these benchmarks/challenges have not gained sufficient traction yet. As increasingly more studies on breast cancer risk are published, discussions should happen across institutions on how to standardize risk assessment quality. A solution might be a public benchmark cohort that at least fulfills the following seven criteria:

- Is anonymized and easy to download
- Comprise both annually and biennially screened women
- Comprise consecutively screened women to best simulate a real screening setting
- Comprise women screened on mammographic devices from different vendors
- Have sufficient follow-up. Preferably 10 years.
- Be demographically diverse
- Have predefined outcomes and metrics interpretable by clinicians

6.5 Limitations

This thesis and the included studies share common limitations that must be addressed to advance research on the topic of mammography-based breast cancer risk.
6.5.1 Cancer Label Noise and Lack of Longer-term Follow Up

In Paper I, II, and III, screened women were followed for at least five years before cancer group designations were made. In few cases, the follow up time was up to six years.

Five years of follow up is sufficient to demonstrate clinical usability. Five year follow up includes, in biennial screening, two and a half screening rounds. However, it has been estimated that around 11% to 22% of screen-detected cancers in Norway were visible and warranted recall in the previous screening round two years prior \[110, 144\]. Consequently, a small fraction of the women who developed a long-term cancer (LTC) should actually have been labeled screen-detected cancer (SDC) in the previous round assuming perfect detection. Additionally, high-risk women who developed a breast cancer after follow up ended are also wrongly labeled as healthy. This might be problematic and noise-inducing when training risk models, which was highlighted in Paper II, but it is also a problem during validation. Radiologists’ labels are considered the gold-standard, but we might never know the true labels of all mammograms.

In Paper III, we discovered that the diagnostic AI system was equal or better than the texture model in identifying LTCs. This might indicate that for women who developed LTC, minimal/subtle signs of cancer or actual cancers were actually already present at least two years prior.

With increasing length of follow up, the probability of encountering a cancer, say ten years from screening, that should have been screen-detected decreases. The ideal scenario would be 20 years of follow up, i.e., from the beginning of screening at the age of 50 to the age of 70 when women usually exit the screening program. Consequently, we could estimate risk throughout the screening period.

6.5.2 What is learned?

DL has recently demonstrated solutions to many medical imaging problems. However, features extracted by these DL models are inherently latent/hidden and are largely uninterpretable. In health care, it is important to understand and ensure the quality of used tools. This paradigm stands completely opposed to the innate nature of DL models.

In this thesis, no guarantees can be made about the learned features and how the models make predictions based on these features. This is a limitation, as we cannot explain to clinicians a complete cause-effect relationship between mammography and risk. Nonetheless, DL models are powerful statistical frameworks that are fitted to the data presented, and we can control which biases are introduced. For instance, by training on age-matched data, the texture model is not likely to learn age relevant features. By excluding clips from training, the texture model model is less likely to consider them during inference. By excluding, to the greatest possible extent, visible or potentially visible malignancies from texture model training, the model is prevented from relying mostly on features relevant for detection. The model is more likely to learn textural features associated with long-term risk. Nonetheless, we can still visually inspect high- and low-risk women, categorized by the texture model, as seen in Figure 6 and observe that tissue heterogeneity is a risk predictive feature of the mammogram.

Many studies have tried to circumvent such issues under the topic of explainable AI with mixed
results. It has been shown that even explainable DL models have failure cases, thus the promise of explainable AI is not likely to be solved soon [145]. Many studies have shown that DL models are clinically relevant and useful. DL models should therefore not be dismissed solely due not being explainable. Instead extensive and rigorous validation should be done and shortcomings examined and mitigated.

### 6.5.3 Demographic Diversity

All of the studies of this thesis are limited by the lack of demographic diversity. The models have been developed and evaluated using exams of only Danish or Dutch women, which may be considered to be representative for most of Europe. However, no guarantee can be made on whether our models will generalize other populations. e.g., women in Asia, Africa, the Americas, or other regions. This is problematic for two reasons: Firstly, the models should be equally reliable for all women. Secondly, if the models are to be clinically usable at a large scale, they should generalize across populations.

### 6.6 Paradigm Shifts In Screening for Breast Cancer

As discussed in Section 6.5.2, DL models are largely uninterpretable. However, throughout this thesis, we have demonstrated that DL models are still clinically relevant and usable. One could argue that such powerful tools should not be ruled out from clinical use solely because we don’t understand them fully [145].

There is a paradigm shift happening around the world in which clinicians are becoming more aware and trusting in AI applications and can directly measure the benefits. In the Capital Region of Denmark, AI has been rolled out in breast cancer screening as one of the first region-based programs in the world to fully incorporate AI. This thesis and its contributions did not only advance the state-of-the-art, but did also present scientific evidence that hopefully can support the paradigm shift towards general trust in AI in health care.

### 6.7 The Untold Stories and Curiosities

During this PhD project several research trails have been followed that was never published or disseminated to the public. This section will briefly describe these methods and results.

#### 6.7.1 Optimizing for Device-invariance

Regarding training device-invariant texture models, the best solution was to augment mammograms by processing raw views into multiple flavors which approximated vendor-specific algorithms. Typically, vendor-specific algorithms are based on a series of well-defined classical image transformations, but are also proprietary and kept secret. However, there exists published literature on algorithms that approximates the vendors’ algorithms. We utilized these existing public methods to create our own processing algorithm and manually tuned it to our needs. This was a data-specific solution, yet there might exists a methodical solution to such problems. As mentioned in Paper II, adversarial
approaches might be difficult to train and only occasionally converges. However, multistage training or other workarounds might succeed [98].

Another solution could be learning device-invariant representations using an information-theoretic objective that can be optimized for directly using backpropagation and gradient descent. A conditional-marginal divergence loss term has been derived and published that, given a constant $c$, can construct a latent variable $z$ that does not encode information about $c$. This seemed like an elegant solution and it was proven to work with the MNIST dataset [146].

This loss term was implemented and used to train a variational autoencoder (VAE) to compress the input mammograms to a device-invariant latent representation and reconstruct the input. The VAE was trained on a limited sample size of 1,000 mammograms obtained on two different devices and downscaled by a factor of six of the original resolution. The complete loss consisted of a reconstruction loss term, a Gaussian prior loss term, and the conditional-marginal divergence loss term. This training procedure succeeded in encoding the mammogram to a latent variable that could not be used to classify (using a multi-layer perceptron) the original mammographic device (AUC of 0.53). However, the latent representation was too general and could not be used for risk assessment. The AUC for detection of LTCs was 0.59 using the latent variable which was approximately equal to the AUC of the dense tissue segmentation model in Abstract II. Increasing image resolution or sample size within feasible limits did not alleviate the problem. The network architecture, an example input view, and the corresponding reconstructed view are shown in Figure 16.

![Variational Autoencoder Network](image)

**Figure 16**: The variational autoencoder network trained to learn latent representations that do not encode information about the mammographic device. Three experiments were conducted. First, we tried classifying device, using a multi-layer perceptron (MLP), from a random vector which, as expected, yielded an AUC of 0.5. When introducing a reconstruction loss term and a Gaussian prior loss term, the latent space was sampled using a mammographic view as input, the AUC of device classification was 0.74. When further including the conditional-marginal divergence loss term for invariance, the AUC was 0.53, which indicated that the MLP was not able to identify the mammographic device from the latent space.

<table>
<thead>
<tr>
<th>VAE Loss</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.50</td>
</tr>
<tr>
<td>With prior term (KLD)</td>
<td>0.74</td>
</tr>
<tr>
<td>With invariance term</td>
<td>0.53</td>
</tr>
</tbody>
</table>

### 6.7.2 Texture Model Training Stability

Large parts of the PhD project have been focused on improving texture model training stability. Training DL models with noisy cancer labels, cancer contra-lateral views, and FFDMs for an inherently
difficult task initially led to poor training stability. In early iterations of the texture model, the performance depended almost completely on the stopping criteria as described in Paper II. Consequently, we could not say with certainty that the texture model would generalize to unseen data.

We found that removing noise-inducing samples, removing the age bias, and matching cases and controls, removing views with visible clips and medical devices, data augmentation, and using ensemble models to aggregate several texture risk scores made the texture model more reliable and robust. However, these are domain-specific workarounds. Several other approaches were tried like other DL network architectures, multi-view training, a larger number of matched controls, loss weighing, and training for an extensive amount of epochs. However, none of the approaches improved stability.

### 6.7.3 Learning Density-independent Textural Features

Textural features, as described in Paper II, does not only refer to heterogeneity but also to a large extent breast density features. In early iterations of the texture model, the density bias was eliminated by matching cases and controls, not only on age, but also on PMD. Essentially, we removed the density signal from training to learn density-independent textural features. As expected, the AUC for identifying ICs and LTCs decreased and PMD and texture were much less correlated. This effect can be observed in Figure 17. When combining density-independent texture scores with PMD in a multivariate logistic regression model, the risk assessment ability was equal to that of density-dependent texture scores. However, the texture training stability became even worse. We discontinued this study as it was not possible to determine whether the combination of density-independent texture and PMD was truly equal to the density-dependent texture or simply due to random variation in the training.

![Figure 17: Scatter plot of texture risk scores vs. percentage mammographic density (PMD) (a) before training for density-independent textural features and (b) after training for density-independent textural features. Before, texture and PMD were monotonically correlated such that increasing PMD was associated with increasing texture as well. After matching cases with control based on PMD, texture was less correlated with PMD and increasing PMD was not necessarily associated with high texture.](image)

### 6.8 Future work

There are several future directions for this project that the candidate would like to explore that addresses the above-mentioned limitations and failed attempts to improve the developed risk models...
further.

The study on learning device-invariant representations did unfortunately not lead to reliable risk assessment. However, the proposed framework does provide a method for directly optimizing for a fourth loss term which is compatible with learning risk-predictive features. This should be pursued in future work.

In terms of texture model training stability, we mostly solved the problem with domain-specific workarounds and reducing the amount of noise in the training data. However, there are still many alternative methodical approaches worth pursuing described in published literature on noisy labels [147].

The study on learning density-independent texture is still quite relevant. If we can successfully separate texture into several risk factors, which is then subsequently combined in a simple interpretable model, e.g. multivariate logistic regression, we can with more certainty explain to clinicians and patients how the risk assessment was derived and which individual risk factors contributed most to the risk assessment.

It might be feasible to learn density-independent textural features methodically instead of using the data-specific solution where the PMD bias was eliminated between cancers and control which led to poor training stability. We might even repurpose the framework for device-invariance described in Section 6.7.1 to learn a latent feature space that does not encode density-related information. This topic should be investigated in the future to develop more fine-grained risk profiles.

Breast cancer risk is known to be partly inherited and partly dependent on external factors. It would be interesting to study whether certain genetic traits (genotypes) causes certain mammographic appearances (phenotypes), and conversely, if mammographic appearance could be mapped to likely genetic alterations. Ideally, mammographic features and genetic information should be combined into an aggregate feature space with meaningful mapping and be predictive of risk.

At the time of writing, the BOADICEA model, which includes all known clinical risk factors, polygenic risk score, and pathogenic variants estimates risk at the same level of the state-of-the-art mammography-based risk models. This raises another important question: Could genetic information, clinical risk factors, and mammography-based risk be combined to estimate risk even more reliably and accurately? If this is the case, then either risk-related mammographic appearance is not fully explained by known phenotypes/clinical risk factors or we are simply unable to extract that information yet. If it is not the case, then it would indicate that we might be nearing the upper limit of mammography-based risk. The state-of-the-art mammography-based risk models seems to converge around AUCs of 0.70 to 0.73 in terms of prediction of cancers five years from screening [98].

Health registry data for breast cancer risk has mostly been ignored in literature for breast cancer risk as it is likely less predictive and more difficult to process and analyse than mammograms or self-reported clinical risk factors. However, one interesting direction of future work would be to re-purpose the general framework developed for relapse patient identification for breast cancer risk.
The simplistic encoding scheme and the easily implemented random forest model allows for other endpoints such as future breast cancer. The framework might be combined with the mammography-based models to make predictions more reliable and create a richer risk profile.

Considering the overall aim of this PhD project and thesis, prospective studies and clinical implementation of mammography-based risk models are the ultimate goals. However, before this can happen, a number of limitations must be addressed.

Firstly, the risk models should be validated in women screened elsewhere than Europe in cohorts with larger demographic diversity, on different scanners, and with different screening intervals. Only with retrospective analyses and sufficient amount scientific evidence that risk models provide reliable risk estimates, a prospective study can be initiated. Another obstacle is that many US clinics use tomosynthesis for screening. To support such clinics, our risk models should support synthetically synthesized 2D FFDMs generated from the tomosynthesis slices. With enough training data, a texture model might even be trained directly on tomosynthesis images.
7 Concluding Remarks

In this thesis, personalized risk-stratified screening was proposed as a possible solution to the inherent challenges associated with large scale screening. Clinicians and radiologists need reliable tools to support them in safely and efficiently implementing personalized screening protocols.

An AI-based screening protocol was designed and simulated. The study demonstrated that AI could safely reduce radiologists’ workload while preserving high screening quality. This study contributed to clinical implementation of AI as first reader in the Capital Region of Denmark breast cancer screening program. The candidate continues to monitor screening quality indicators at Gentofte Hospital.

A mammographic texture model was developed for long-term breast cancer risk assessment and was able to generalize across mammographic devices from different vendors. In combination with an AI system for lesion detection, a rich mammography-based risk profile was created which advanced the state-of-the-art for breast cancer risk assessment. Further studies are needed to examine whether mammography-based risk assessment is nearing the upper limit of performance by utilizing genetic data. The combination model should further be validated in other screening regimes and populations.

A machine learning framework for identifying patients using health register data was developed and applied to breast cancer relapse patients. Such a framework might contribute to a richer breast cancer risk profile. The encoding scheme should be extended to encode registry data into a more expressive feature space.

Overall, the candidate believe that this work have contributed to and advanced the research of personalized breast cancer screening by development of reliable risk models. With prior rigorous validation, prospective studies on risk-stratified screening should be designed and conducted in collaboration with clinicians.
References


8 Appendix

This appendix contains four abstracts. The abstracts themselves were not discussed as they were preliminary studies ultimately leading to the papers included in this thesis.
Reducing the radiologist's workload by detecting normal mammograms with an AI system

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

Purpose:
To investigate whether an AI system can detect normal mammographies in a breast cancer screening cohort.

Methods and materials:
This retrospective study analysed 18,020 doubly read studies from the Danish Capital Region breast cancer screening program, comprised of 143 screen-detected cancers and 447 non-cancer recalls (false-positives). Using the deep learning-based image analysis tool, Transpara v1.5, all studies were sorted into 10 categories based on findings from four views. A high category (10) indicated a high probability of malignancy, while a low category (1) indicated a very low probability of malignancy. Normal studies were identified as being in category 5 or less. This study examined the number of studies, and non-cancer recalls, that can possibly be avoided by detecting normal studies before radiologist reading.

Results:
Using category 5 or less as a threshold, 10,545 (58.52%) studies were classified as normal. Included were 5 screen-detected cancers (3.5%) and 106 non-cancer recalls (23.71%).

Category 1 and 2 comprised of 4,738 (26.29%) studies, 26 non-cancer recalls (5.82%), and 2 screen-detected cancers (1.36%).

Category 1 comprised of 2,627 (14.58%) studies, 12 non-cancer recalls (2.68%), and 0 screen-detected cancers.

Conclusion:
The results show that the AI system can successfully identify normal mammographies with very few missed screen-detected cancers. Furthermore, a substantial amount of false-positive studies were identified as normal. The results suggest that AI systems could potentially effectively and safely reduce the number of studies that radiologists would have to examine by a considerable amount, and several false-positives could be avoided.

Limitations:
Transpara identifies a few screen-detected cancers as normal. Having radiologists examine missed cases, future improvements could be made. The number of cancer cases was limited. Results should be confirmed on a larger study.
Measuring short and long-term breast cancer risk by combining mammographic texture models, an AI-based CAD system, and established risk factors

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

Purpose:
To investigate the combined effect of mammographic texture and the exam score of an AI-based CAD system in terms of quantifying short- and long-term breast cancer risk

Methods and materials:
This retrospective study comprised a cohort of 52637 double-read screens of four FFDMs from the Danish Capital Region breast cancer screening program, including 154 interval cancers (IC, diagnosed 6-24 months after screening) and 808 long-term cancers (LTC, diagnosed 2-5 years after screening). For each exam, three metrics were computed. Percent mammographic density (PMD) was measured using a deep-learning-based tool. Mammographic texture-based risk (TBR) was computed using the deep-learning architecture, ResNet34, trained to detect images of women with high probability of developing cancer in the future. An AI-CAD system, (Transpara, ScreenPoint Medical), analyzed all studies providing an exam score from 1 to 10, where a high score indicates a high probability of visible malignancy.

Using baseline screening mammograms, we examined risk segregation performance of the models for IC (short-term) and LTC using area under the ROC curve (AUC-ROC, 95%CI). Logistic regression was used to combine co-variates and validated using 5-fold cross-validation.

Results:
TBR yielded an AUC-ROC of 0.69 (0.64-0.73) for ICs and 0.66 (0.64-0.68) for LTCs. The AI-CAD yielded an AUC-ROCs of 0.67 (0.62-0.71) for ICs and 0.64 (0.62-0.66) for LTCs. Combining TBR, AI-CAD, PMD, and age yielded an AUC-ROCs of 0.72 (0.68-0.76) for ICs and 0.68 (0.66-0.69) for LTCs.

Conclusion:
Results indicate that combining texture-based risk, AI-CAD exam score, PMD, and age improved risk segregation for both ICs and LTCs compared to texture and CAD alone suggesting that both systemic and localised findings can contribute to risk modelling.

Limitations:
The cross-validated performance should be validated independently in future studies.
Improving Mammographic Texture Risk Models with Long-term Cancers
Andreas D. Lauritzen, My Catarina von Euler-Chelpin, Elsebeth Lynge, Ilse Vejborg, Mads Nielsen, and Martin Lillholm

Purpose:
To investigate the impact on mammographic texture-based breast cancer risk segregation developed with and without knowledge of long-term follow-up cancers.

Methods and Materials:
This retrospective study comprised a cohort of 52637 double-read screens from the Danish Capital Region breast cancer screening program, including 418 screen-detected, 150 interval (within 24 months), and 813 long-term (2-5 years follow-up) cancers. Screens were assigned the maximum BI-RADS density of two radiologists. Percent mammographic density (PMD) was measured using a deep learning-based tool.

The deep learning architecture, ResNet18, has been trained with an age-matched subset of the cohort to detect images of women with high probability of developing cancer. Training images are full-field digital mammographies (MLO and CC). Training images were chosen as contra-lateral to the diagnosed cancer.

In this study, we examined the texture model's risk segregation using area under the receiver operating characteristic curve (AUC-ROC, 95% CI). Texture models are evaluated on the full cohort in a 5-fold cross-validation setup.

Logistic regression (LR) models were used to combine multiple co-variates. Age has been included as a co-variate in all models below.

Results:
BI-RADS and PMD resulted in AUC-ROCs of 0.59 (0.58–0.61) and 0.59 (0.57-0.60), respectively.

A texture model trained without long-term cancer follow-up status yielded an AUC-ROC of 0.63 (0.62-0.65). Training the same model with follow-up status included resulted in an AUC-ROC of 0.65 (0.63-0.66).

A combination model of PMD and BI-RADS produced an AUC-ROC of 0.60 (0.58-0.61). Combining texture and PMD produced an AUC-ROC of 0.65 (0.64-0.67). Combining texture and BI-RADS produced an AUC-ROC of 0.65 (0.64-0.66). Combining texture, PMD, and BI-RADS produced an AUC-ROC of 0.65 (0.64-0.67).

Conclusion:
Training a texture model using long-term cancers improved overall cancer risk segregation. The enhanced model provided mammographic texture-based risk-segregation which, in this study, cannot be improved by incorporating density using LR.
An AI-based Breast Cancer Reading Protocol Evaluated Across Density Categories in a Danish Screening Cohort
Andreas D. Lauritzen, Alejandro Rodriguez-Ruiz, My Catarina von Euler-Chelpin, Elsebeth Lynge, Ilse Vejborg, Mads Nielsen, Nico Karssemeijer, Martin Lillholm

Purpose:
To investigate how AI-supported reading during screening performs in a cohort study across breast density strata.

Methods and Materials:
This retrospective simulation study analyzed 53948 screens from the Danish Capital Region breast cancer screening program comprising 418 screen-detected and 150 interval cancers (IC).

Using an AI-based CAD system, Transpara v1.6, screens were scored from 0 to 10 where high scores indicates high probability of malignancy. Each screen was categorized into four breast density categories using BI-RADS.

Screens with a score less than 5 were only read by Transpara. Any screen with a score greater than 9.91 were flagged as recalls. For the remaining screens, standard double reading was used as outcome.

This study examined screening outcomes with the reading protocol above across different combinations of BI-RADS categories.

Results:
Standard screening of the cohort yielded sensitivity and specificity of 73.6% and 97.6%. 1291 women were false positive recalls (FPs).

AI-supported screening for all screens yielded sensitivity and specificity of 72.0% and 97.8%. 134 FPs were avoided. 39.7% of the cohort was read by radiologists. 3.8% (16) screen-detected cancers were not found but 7 ICs would have been recalled.

AI-supported screening on BI-RADS 1 and 2 yielded sensitivity and specificity of 72.9% and 97.7%. 42 FPs were avoided. 59.4% of the cohort was read by radiologists. 2.15% (9) screen-detected cancers were not found but 5 ICs would have been recalled.

AI-supported screening on BI-RADS 1 and 4 yielded sensitivity and specificity of 73.4% and 97.6%. 1 FP were avoided. 78.9% of the cohort was read by radiologists. 0.72% (3) screen-detected cancers were not found but 2 ICs would have been recalled.

Conclusion:
AI-supported reading applied to select density strata maintain overall screening quality in terms of FPs and cancer detection while reducing radiologists’ workload by 41% and 21% depending on strata.