From The Department of Physiology and Pharmacology
Karolinska Institutet, Stockholm, Sweden

DEEP LEARNING IN BREAST CANCER SCREENING

Karin Dembrower

Karolinska Institutet

Stockholm 2022

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2022

ISBN **978-91-8016-533-4**
Cover illustration: Mammogram, after artistic post-processing

Deep Learning in Breast Cancer Screening

THESIS FOR Doctoral Degree (Ph.D.)

By

Karin Dembrower

The thesis will be defended in public at Capio Sankt Görans Sjukhus, Hörsalen, 9 am, April 1

Principal Supervisor:
Peter Lindholm
Karolinska Institutet
Department of Physiology and Pharmacology

Co-supervisors:
Kevin Smith
Royal School of Technolgy / Sci Life Lab
Division of Computational Science and Technology

Martin Eklund
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Fredrik Strand
Karolinska Institutet
Department of Oncology and Pathology

Opponent:
Emily F. Conant
University of Pennsylvania
Department of Radiology

Examination Board:
Torkel Brismar
Karolinska Institute
Department of Clinical Sciences, Intervention and Technology

Antonios Valachis
Örebro University
Department of Clinical Sciences

Fredrik Wärnberg
Göteborg University
Department of Clinical Sciences

Respect and kindness

To Gustaf♥, Oscar♥, Lovisa♥ and Charlotta♥

POPULAR SCIENCE SUMMARY OF THE THESIS

Breast cancer is the most common cancer form among women and the second leading cause of death among women after lung cancer. Mortality rates decreased by up to 40% when national breast cancer screening programs were introduced in the 1980s and 1990s. Risk factors connected to breast cancer have been identified, with female sex and older age being the most common. Other risk factors are breast density, hereditary factors, number of child births, age at first childbirth, breast feeding habits and alcohol consumption. One of these, breast density, is derived from examination of mammographic images.

In Sweden, all women between 40 to 74 years are invited to breast cancer screening every 18 to 24 months. The screening examination consists of two standard views of each breast, and questions about clinical breast symptoms. All mammograms are examined by two breast radiologists. If either of the radiologists flags an examination because they find something suspicious, or if the woman reports worrying symptoms, her exam will be discussed at a special meeting called a consensus discussion. During this discussion, at least two breast radiologists discuss whether the woman should be declared healthy or recalled for further examinations.

In Sweden there is a lack of breast radiologists, and it is important that their time is used efficiently. It is also important to make the screening process as effective as possible. We need to reduce the proportion of interval cancers - breast cancers that are clinically detected between two screening time points - which are associated with an increased mortality and morbidity. This might be achieved by making the screening process more individualized with the aim of detecting tumors as early as possible while the cancer is still curable.

The introduction of deep learning, or artificial intelligence, for mammographic image analysis, might contribute to make the screening process more individualized, efficient and, in the end, further reduce morbidity and mortality. My research has focused on the construction of a large retrospective cohort for deep learning, then exploring the potential use of this technique for risk assessment, for independent analysis of mammograms, and finally, to calibrate a commercial artificial intelligence (AI) algorithm for use in a prospective clinical study at a Stockholm Breast Center.

In study I, we described the underlying Cohort of Screen Aged Women from which the study populations of the following three studies are derived. I also described how the cohort has been used so far and the future opportunities for research. As expected, our research group found that there is a huge interest world-wide for population-based datasets. Parts of our dataset have been used in other research projects globally.

In study II, we analyzed how a deep learning risk score, developed in collaboration with academic computer scientists at the Royal Institute of Technology in Stockholm, performed compared with standard breast density measurements for predicting future breast cancer risk. We concluded that compared to density, a deep neural network can more precise predict which women are at risk for future breast cancer and more precisely detect more aggressive forms of breast cancer.

In study III, a retrospective simulation study, we analyzed the potential cancer yield when triaging screening examinations into two work streams, depending on the AI score related to the likelihood of cancer signs in the images - a ‘no radiologist’ work stream and an ‘enhanced assessment’ work stream. We found that the AI score could potentially reduce radiologist workload and detect a large proportion of breast cancers earlier.

In study IV, we analyzed the consequences of alternative choices of the abnormality threshold for an independent reading AI algorithm. We demonstrated that the extent of change in sensitivity and false positives depend on these choices. The results were then used to develop the study protocol for a prospective clinical study, which I continue to be involved in as local investigator at the study hospital.

In studies I to IV we have demonstrated promising results, shedding light on the possible introduction of AI and deep learning algorithms in breast cancer screening.

ABSTRACT

Breast cancer is the most common cancer form among women worldwide and the incidence is rising. When mammography was introduced in the 1980s, mortality rates decreased by 30% to 40%. Today all women in Sweden between 40 to 74 years are invited to screening every 18 to 24 months. All women attending screening are examined with mammography, using two views, the mediolateral oblique (MLO) view and the craniocaudal (CC) view, producing four images in total. The screening process is the same for all women and based purely on age, and not on other risk factors for developing breast cancer.

Although the introduction of population-based breast cancer screening is a great success, there are still problems with interval cancer (IC) and large screen detected cancers (SDC), which are connected to an increased morbidity and mortality. To have a good prognosis, it is important to detect a breast cancer early while it has not spread to the lymph nodes, which usually means that the primary tumor is small. To improve this, we need to individualize the screening program, and be flexible on screening intervals and modalities depending on the individual breast cancer risk and mammographic sensitivity. In Sweden, at present, the only modality in the screening process is mammography, which is excellent for a majority of women but not for all.

The major lack of breast radiologists is another problem that is pressing and important to address. As their expertise is in such demand, it is important to use their time as efficiently as possible. This means that they should primarily spend time on difficult cases and less time on easily assessed mammograms and healthy women.

One challenge is to determine which women are at high risk of being diagnosed with aggressive breast cancer, to delineate the low-risk group, and to take care of these different groups of women appropriately. In studies II to IV we have analysed how we can address these challenges by using deep learning techniques.

In study I, we described the cohort from which the study populations for study II to IV were derived (as well as study populations in other publications from our research group). This cohort was called the Cohort of Screen Aged Women (CSAW) and contains all 499,807 women invited to breast cancer screening within the Stockholm County between 2008 to 2015. We also described the future potentials of the dataset, as well as the case control subset of annotated breast tumors and healthy mammograms. This study was presented orally at the annual meeting of the Radiological Society of North America in 2019.

In study II, we analysed how a deep learning risk score (DLrisk score) performs compared with breast density measurements for predicting future breast cancer risk. We found that the odds ratios (OR) and areas under the receiver operating characteristic curve (AUC) were higher for age-adjusted DLrisk score than for dense area and percentage density. The numbers for DLrisk score were: OR 1.56, AUC, 0.65; dense area: OR 1.31, AUC 0.60, percent density: OR 1.18, AUC, 0.57; with $P < .001$ for differences between all AUCs). Also, the false-negative rates, in terms of missed future cancer, was lower for the DLrisk score: 31%, 36%, and 39% respectively. This difference was most distinct for more aggressive cancers.

In study III, we analyzed the potential cancer yield when using a commercial deep learning software for triaging screening examinations into two work streams – a 'no radiologist' work stream and an 'enhanced assessment' work stream, depending on the

output score of the AI tumor detection algorithm. We found that the deep learning algorithm was able to independently declare 60% of all mammograms with the lowest scores as "healthy" without missing any cancer. In the enhanced assessment work stream when including the top 5% of women with the highest AI scores, the potential additional cancer detection rate was 53 (27%) of 200 subsequent IC, and 121 (35%) of 347 next-round screen-detected cancers.

In study IV, we analyzed different principles for choosing the threshold for the continuous abnormality score when introducing a deep learning algorithm for assessment of mammograms in a clinical prospective breast cancer screening study. The deep learning algorithm was supposed to act as a third independent reader making binary decisions in a double-reading environment (ScreenTrust CAD). We found that the choice of abnormality threshold will have important consequences. If the aim is to have the algorithm work at the same sensitivity as a single radiologist, a marked increase in abnormal assessments must be accepted (abnormal interpretation rate 12.6%). If the aim is to have the combined readers work at the same sensitivity as before, a lower sensitivity of AI compared to radiologists is the consequence (abnormal interpretation rate 7.0%). This study was presented as a poster at the annual meeting of the Radiological Society of North America in 2021.

In conclusion, we have addressed some challenges and possibilities by using deep learning techniques to make breast cancer screening programs more individual and efficient. Given the limitations of retrospective studies, there is a now a need for prospective clinical studies of deep learning in mammography screening.

LIST OF SCIENTIFIC PAPERS

I. **Karin Dembrower**, Peter Lindholm, Fredrik Strand
A multi-million Mammography Image Dataset and Population-Based Screening Cohort for the training and Evaluation of Deep Neural Networks – the Cohort of Screen-Aged Women (CSAW)
Journal of digital Imaging 2020, vol 33 sid 408-413

II. **Karin Dembrower,** Yue Liu, Hossein Azizpour, Martin Eklund, Kevin Smith, Peter Lindholm, Fredrik Strand.
Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction
Radiology. 2020:190872. vol 294:2 sid 265-272

III. **Karin Dembrower**, Erik Wåhlin, Yue Liu, Mattie Salim, Kevin Smith, Peter Lindholm, Martin Eklund, Fredrik Strand
Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study
The Lancet Digital Health, 2020, 2.9: sid 468-sid 474

IV. **Karin Dembrower**, Mattie Salim, Martin Eklund, Peter Lindholm, Fredrik Strand
Implications for downstream workload and sensitivity based on calibrating an AI CAD algorithm by standalone-reader or combined-reader sensitivity matching
Manuscript

CONTENTS

LIST OF ABBREVIATIONS

AI	artificial intelligence
AUC	area under the curve
BCSC	Breast Cancer Surveillance Consortium
BI-RADS	Breast Imaging Reporting and Data System
BRCA1	breast cancer gene 1
BRCA2	breast cancer gene 2
BSE	breast self-examination
CC	craniocaudal
CIS	cancer in situ
CISH	chromogenic in situ hybridization
CSAW	Cohort of Screen Aged Women
DBT	digital breast tomosynthesis
DCIS	ductal cancer in situ
DLrisk	deep learning risk score
DM	digital mammography
Dnr	diarienummer
ERB	ethical review board
Er	estrogen
FISH	fluorescence in situ hybridization
FSH	follicle stimulating hormone
HER2	human epidermal growth factor receptor 2
HRT	hormone replace therapy
IC	interval cancer
IDC	invasive ductal carcinoma
KTH	Kungliga Tekniska Högskolan
LCIS	lobular cancer in situ
LH	luteinizing hormone
MD	mammographic density
ML	mediolateral view
MLO	mediolateral oblique view

MRI	magnetic resonance imaging
NST	no special type
OR	odds ratio
PR	progesterone
SD	standard deviation
SDC	screen detected cancer
SLN	sentinel lymph node
SLNB	sentinel lymph node biopsy
TDLU	terminal duct lobular unit
TNM	the tumor node metastasis classification of malignant tumors
95%CI	95% confidence interval

1 INTRODUCTION

Population-based breast cancer screening programs have been very successful. The mortality rates were reduced with up to 40% when nationwide breast cancer screening were introduced in the 90's (1-3). Despite the success with national breast cancer screening programs, there is room for improvement, e.g., by decreasing the number of women who are diagnosed with late-stage breast cancer, and, by addressing the shortage of breast radiologists in many countries, including Sweden.

In many developed countries, the only breast cancer risk factor that is used for inviting women is the age, and all who are invited is then offered the same "one size fits all" imaging method - mammography. Mammography is excellent for the majority of women but not for all. Some women invited to screening would benefit from being examined by other, more sensitive, modalities than mammography, for example magnetic resonance imaging with a considerably higher sensitivity (4).

To identify which women are likely to benefit from a modified screening process is challenging. Many breast cancer risk prediction models have been introduced, such as the Gail model (5) and the Tyrer-Cuzick model (6). These were primarily developed to assess life-time risk and not the relatively short-term horizon of two to three years applicable to the screening situation. Further, these risk prediction models do not generally take image-based factors into account; only the latest version of the Tyrer-Cuzick model takes mammographic density into account (7).

By introducing deep networks in the screening process, the information in the mammograms which is not consistently appreciated by the human eye might be used for cancer detection and risk estimation - if the networks are properly trained and validated. Since it is impossible or difficult to understand what the networks base their result on, proper validation and testing is paramount.

The results shown in my studies give hope that it may soon be time for deep networks to improve women's health by even better early detection of breast cancer, translating into less aggressive treatment being necessary and less lives being shortened by cancer.

2 LITERATURE REVIEW

2.1 THE BREAST

The breast is a glandular organ that develops from the milk line situated along the anterior part of the body wall from the groin to the axilla. The breast is eventually formed at the pectoral region. The breast consists of stroma, adipose tissue and glandular tissue which is connected by a loose framework of fibrous tissue (Cooper´s ligaments). The glandular tissue comprises the potentially milk producing lobules and the ducts eventually leading to the nipple. The nipple contains around 10 openings - each connected to a lactiferous sinus that receives a lobar collecting duct.

The lobule and its connecting duct are called the terminal duct lobular unit (TDLU), which is the likely starting point of the most common breast cancer form, the paradoxically named ductal carcinoma. The inner luminal layer of the duct is composed of epithelial cells and an outer layer of myoepithelial cells. An outer basal membrane encloses these layers (8). The breast undergoes all developmental stages if a woman experiences pregnancy and childbirth, and reaches its full function during lactation (9).

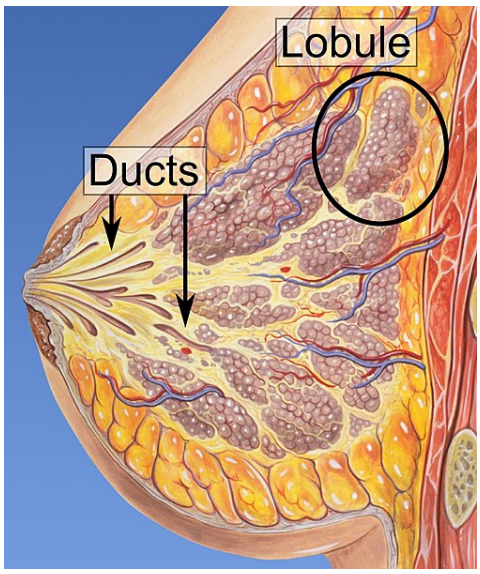

Figure 1. The breast with lobules and ducts.
(https://commons.wikimedia.org/wiki/File:Lobules_and_ducts_of_the_breast.jpg)

2.1.1 Biology, development, changes over time

In the fetus and in infants there is no relation between sex, age, and the stage of development of the breasts. At birth the infant has breast structures like adults, with well-defined lobules and terminal lobular duct units – sometimes with milk proteins. This means that both sexes have the TDLUs described above. A few months after birth, the glands involute in a similar pattern as the postmenopausal breast because of lack of breast stimulating hormones. Involution means that the glandular tissue decreases. During childhood, the breast grows in proportion to other tissues in the body. For women, the pubertal development of the breast commences before menarche, and changes drastically when average blood hormones such as estrogen (ER), prolactin, luteinizing hormone (LH), follicle stimulating hormone (FSH) and growth hormone levels rise. This process is gradually controlled by the hypothalamus, which in turn acts on the anterior pituitary gland, which increases the levels of FSH and LH. FSH stimulates the ovarian follicles to produce ER. Later in the menstrual cycle the ovaries also produce progesterone (PR).

During pregnancy, all parenchymal components of the breast change because of elevated levels of hormones. Similarly, but with an opposing effect, decreased levels of hormones lead to involution of the breast tissue in the postmenopausal period. The lobules shrink and the stromal tissue is replaced by fat. Menopause is initiated by atresia of ovarian follicles leading to a decrease of hormone levels. The menopause is a regressive phenomenon, and it occurs as a consequence of the atresia of around 400 000 follicles that were present in the fetus at the age of 5 months. Breast tissue in nulliparous (childless) women is less differentiated than that of parous women. Earlier differentiation stages are more vulnerable to carcinogenic damages than for more differentiated stages (10-12).

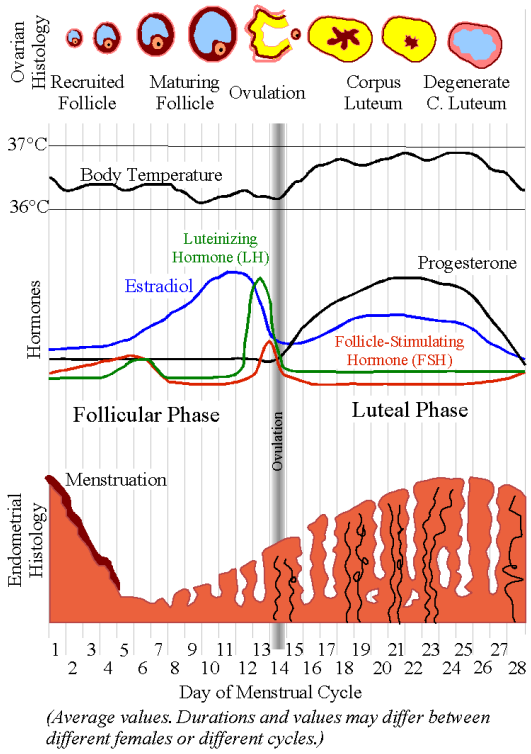

Figure 2. Hormones affecting the breasts and the uterine mucosa. (https://commons.wikimedia.org/wiki/File:MenstrualCycle2.png)

2.2 BREAST CANCER

2.2.1 Epidemiology

Breast cancer is the most common cancer form among women, and the second leading cause of deaths among women after lung cancer. Breast cancer counts for 23% of all cancers with an estimate of more than 2 million new cases worldwide yearly. It is now the most common cancer for women in both developed and in developing regions. The incidence rates vary greatly, with numbers ranging from high (more than 80 per 100,000 women) in developed regions to low (less than 40 per 100,000 women) in developing regions. In North America, the 5-year survival rate approached 90% between 2010 to 2014. The corresponding number for Western Europe was 85% or higher. Breast cancer survival is lower in Eastern Europe and Africa (13).

The overall lifetime risk for breast cancer diagnosis is 12.8% (1out of 8) and the lifetime risk for death for breast cancer is 2.6% (1 out of 39). The incidence of breast cancer is increasing, and most of the historic increase reflects changes related to fewer child births

and delayed childbearing. During the late 80s and 90s, the incidence rates of invasive breast cancer and DCIS increased rapidly because of the introduction of mammography screening programs, with increased attendance from 29% in year 1987, to 70% in year 2000. In contrast, there was a decrease (nearly 13%) for the invasive breast cancer rate between 1999 and 2004, mainly because studies were published concluding that hormone replace therapy (HRT) was linked to breast cancer and heart disease. Since 2004, the incidence of invasive breast cancer has risen by about 0.3% each year. Since the end of the 1980's, mortality rates in breast cancer have decreased with up to 40% to the present date. This can be explained by both improvements in treatment and by early detection with mammography screening programs.

In Sweden, the median age for being diagnosed with breast cancer is 64 years, and less than 5% of patients are under 40 years. Yearly, there are around 8,000-9,000 diagnoses of breast cancer, and every day around 20 women are diagnosed. A few (~ 40 to 60) men are diagnosed with breast cancer yearly and the prognosis is the same as for women. In 2018, 1,400 women died from breast cancer. The relative five year survival is around 90% and the relative 10 year survival is around 80 % (14, 15).

2.2.2 Biology and tumor characteristics

Breast cancer is a heterogeneous disease with several pathologic features and biological behaviors. Different breast cancer subtypes have varying clinical and histopathological features, outcomes, and they respond to different therapies.

Breast tumors are classified according to the location of origin. Of the histopathological types of breast cancer, around 70% of all breast cancers are of the ductal type. The second most common is lobular breast cancer which accounts for around 15%. The lobular cancer tends to be multifocal and bilateral. Other histological subtypes are medullary breast cancer (5%), tubular breast cancer (5%) cribriform breast cancer (2%), mucinous breast cancer (2%) and micro-papillary breast cancer (1–2%) (16).

Breast cancer survival varies by stage of the disease at diagnosis. Stage is one of the most important predictors for breast cancer prognosis (17). The different stages 1 to 3, describe the size and spread to the lymph nodes in different ways. Stage 4 indicates that the tumor has spread to other organs. The overall survival rate for diagnosed patients from 2009 to 2015 was 98% for stage 1 patients, 92% for stage 2 patients, 75% for stage 3 patients and 27% for stage 4 patients (18).

The tumor node metastasis classification of malignant tumors (TNM) classification is a structured tool developed by the American Committee on Cancer and the International Unit for Cancer Control. The system is applicable for all carcinomas with a histologic confirmation and describes the stages of the cancer. The system is defined by three letters:

T corresponds to the extent of the tumor and the relationship to surrounding tissue. In case of multifocal tumor burden the highest T value is used for the system.

N corresponds to eventual lymph node metastasis, and for breast cancer there are three levels (I–III). N0 refers to no spreading to the lymph nodes. N1 refers to spreading 1-3 axillary lymph nodes. N2 Refers to spreading to 4-9 axillary lymph nodes and N3 refers to spreading to > 9 axillary lymph nodes as well as lymph nodes infra- and supraclavicular and/or parasternal lymph nodes.

M corresponds to the extent of metastasis to other regions than lymph nodes (19).

The Elston grade (or Nottingham grade) describes the degree of differentiation in the tumors and is divided into three groups, where grade 1 is the most differentiated group and grade 3 is the least differentiated group (20).

The classical immunohistochemical (IHC) markers include the ER receptor, the PR receptor and the human epidermal growth factor receptor 2 (HER2). These receptors are known to mediate cell growth signaling. Breast tumors are divided into different subgroups according to these markers. In general, ER- and PR- tumors have a poorer prognosis than ER+ or PR+ tumors. (21, 22) It is suggested that ER+ and PR+ tumors are associated with exposure to ER and PR, while ER- and PR- tumors are independent of hormone exposure. Patients with hormone sensitive tumors have a longer disease-free life and a better prognosis.

The ER receptor is overexpressed in around 70% of all cancer cells; the hormone 17-oestraddiol activates the receptor which then leads to tumor growth and inhibition of apoptosis of the tumor cells (23). It is important to discriminate whether breast cancer is ER positive or not, as a targeted adjuvant therapy called tamoxifen is available, although 40% of ER positive tumors are resistant to this treatment (22, 24). Tamoxifen was introduced in the 1980's and was the first anti-ER therapy. For non-resistant tumors it effectively blocks ER stimulation by binding to the ER ligand (25).

The PR receptor is a positive prognostic factor in the presence of ER and its presence is associated with a favorable response to endocrine therapy and chemotherapy (26). ER and PR positive breast cancers have around a 70% chance of responding to any endocrine therapy; breast cancers that are only ER positive respond in 20–40% of cases and those that are only PR positive respond in 40–45%. Both ER and PR negative breast cancers respond to endocrine therapy in less than 10% of patients (27). However, there is a current debate regarding PR as a predictor and its clinical impact (28).

The HER2 receptor is normally related to cell proliferation and division, and if it is amplified in breast cancer cells it is a predictor for more advanced disease, increased risk for relapse and decreased patient survival (29). In around 15–30% of breast cancers HER2 is amplified, and if there is uncertainty regarding HER2 amplification the specimen should undergo confirmation testing with fluorescence in situ hybridization (FISH) or chromogenic in situ hybridization (CISH) (30, 31). There is currently one targeted treatment for HER2+ tumors, the antibody trastuzumab, which decreases tumor growth and acts as a sensitizer for chemotherapy (25).

A very important and widely used biological marker is the protein (antibody) Ki67 which indicates the proliferation activity in the tumor. The proliferation index is considered low when there are 14% or less stained nuclei, and considered positive or high when there are more than 14% stained nuclei. A high proportion of Ki67 is associated with lower overall survival and more often tumor recurrence (32, 33). Ki67 is also used to predict the neoadjuvant response, or the outcome from adjuvant chemotherapy (34). Posttreatment Ki67 levels can give prognostic information for patients with hormone positive tumors and for the risk of disease relapse (35).

Based on gene expression analysis, molecular subtypes of breast cancer have been defined: luminal A, luminal B, HER 2 enriched, and triple negative breast cancer (36). The following proxies based on receptor expression have been suggested:

- Luminal A proxy: ER+ and/or PR+ and HER2-, low grade, low proliferation.
- Luminal B proxy: ER+, low PR, high grade and/or high proliferation and HER 2-.
- HER2 enriched proxy: HER2+ and hormone receptor + or -.
- Triple negative breast cancer proxy: ER-, PR-, HER2-

The different subtypes are associated with different prognoses, where patients with luminal A have the best prognosis and patients with triple negative breast cancers have the worst prognosis. Based on the subtypes, patients have different treatment options. For patients with luminal A, B and HER2 positivity there are options for targeted treatments, while patients with triple negative breast cancer only have chemotherapy as an option (37).

Breast cancer can be either invasive, in situ cancer, or mixed. Historically the major subtypes of in situ cancers are ductal cancer in situ (DCIS) and lobular cancer in situ (LCIS). In 2020 in Sweden, 10.9% of all diagnosed breast cancers were non-invasive, with the majority of cancer in situ (CIS) being ductal (83%). The definition of CIS is that abnormal cells replace the epithelium while the basal membrane is intact. When the basal membrane is invaded, the cancer becomes invasive. DCIS counts as a precursor for invasive cancer and LCIS only acts as a marker for a higher risk of breast cancer diagnosis but the more aggressive pleomorphic LCIS is connected to invasive lobular cancer. DCIS often appears as microcalcifications and LCIS is often incidentally detected (38).

When the basal membrane is invaded, the cancer becomes invasive. Invasive breast cancer is a heterogeneous group of cancers; the largest group was previously known as invasive ductal carcinoma (IDC), but with the use of the new definitions, is now referred to as invasive carcinoma of no special type (NST). Other specific invasive breast cancers are invasive lobular, invasive medullary, invasive mucinous, invasive papillary, and metaplastic breast cancer (39).

It should be noted that an alternative classification system which I find very interesting and might provide a better correlation between imaging biomarkers, large 3D histologic format and prognosis has been suggested by Professor Tabár (40). He suggests that it is most important to take the histological site of origin into account for treatment planning and prognosis (41). For smaller cancers (up to 14 mm) the mammographic features are said to be tightly linked to the histological origin. Acinar adenocarcinomas of the breast, originating from the TDLU, have an excellent prognosis when they are small (up to 14 mm), and are often seen as spiculated or round masses in the mammogram. On the other hand, ductal adenocarcinomas of the breast have a poorer prognosis, and may appear as architectural distortions or microcalcifications arranged in duct-like patterns. Tabár argues that the current nomenclature of DCIS is a misnomer when the mammographic appearance are microcalcifications arranged in a duct-like pattern since those often represent invasive duct-forming cancers and are associated with a poor prognosis. This might also explain why they show contrast-enhancement on MRI even though they are supposedly “in situ”.

Symptoms of inflammatory breast cancer differ from non-inflammatory breast cancer, including lumps with red and swollen skin, sometimes with fluid running from the skin. Around 2% of all breast cancer diagnoses are inflammatory breast cancer and the histopathologic features are distinctive, with tumor cell emboli in the skin of the breast. Some data imply that inflammatory breast cancer is a special type of cancer form, while others suggest that it correlates to NST grade III (42).

Immunohistochemically markers (IHC) together with tumor size, tumor grade, histologic type, and nodal involvement, is used for prognosis and treatment decisions.

2.2.3 Risk factors

There are many risk factors for developing breast cancer. The most important besides female sex is age, with an incidence highly related to increasing age.

Mammographic density (MD) is the amount, or proportion, of pixels in the mammogram corresponding to radiodense breast tissue. Dense breast tissue appears bright, and non-dense appears dark (adipose tissue). Women with over 75% MD have a 4 to 6 times higher risk of developing breast cancer compared to women with a small proportion of dense breast tissue. Thus, high density is considered a strong risk factor for breast cancer. Another important consideration is that the dense tissue might mask tumors in the image (43, 44). MD is associated with both lifestyle and reproductive factors, and it has been hypothesized that MD might partially act as an intermediate marker of breast cancer risk (45). One study by Kerlikovske et al suggested that women with high breast density combined with the Breast Cancer Surveillance Consortium (BCSC) 5-year risk for breast cancer can identify women at high risk for interval cancer (IC), and thus inform them on supplemental breast cancer screening (46).

Another common risk factor for breast cancer is family history. Around one quarter of all breast cancer cases are related to family history. If a woman has a first degree relative with breast cancer, the risk is 1.75-fold higher to develop breast cancer than having no diagnosed relatives. The risk is 2.5-fold higher if a woman has two or more first degree relatives diagnosed with breast cancer (47).

Some of the hereditary cases are the results of mutations in the high- and medium penetrance genes including breast cancer gene 1 (BRCA1), breast cancer gene 2 (BRCA2), TP53, PTEN, STK11, CDH1 (48). The BRCA1 and BRCA2 genes are also associated with higher risk for ovary, prostate and pancreatic cancers. The presence of a BRCA1 or BRCA2 mutation can be predicted if a first degree relative is diagnosed with breast or ovarian cancer at a young age, the presence of bilateral breast tumors, as well as an increased number of affected relatives (25). The lifetime risk of developing breast cancer for BRCA1 and BRCA2 carriers varies between 45% to 87%, with a lifetime risk of between 15% to 45% for developing ovarian cancer (49). Carriers of BRCA1 often present with more aggressive cancers, such as triple negative breast cancer, while carriers of BRCA2 are more likely to present with ductal tumors such as DCIS or invasive ductal carcinoma (50). In Sweden, identified carriers of BRCA1 and BRCA2 mutation are offered yearly breast imaging including MRI from the age of 25. They are also offered prophylactic mastectomy and salpingo-oophorectomy after reproduction.

It is well known that reproductive factors have an impact on breast cancer risk. Childbirth and parity are associated with a decrease in developing luminal breast cancer, while higher age at first childbirth is associated with an increased risk. Breast feeding is associated with a reduced risk of developing both luminal and triple negative breast cancer (51).

ER levels play an important role in the risk of developing breast cancer, both endogenous and exogenous exposure. Endogenous ER is usually produced by the ovaries and, especially after menopause, by adipose tissue. The main source of exogenous ER are oral contraceptives and HRT. High ER levels in postmenopausal women are associated with an increased risk of developing breast cancer. The risk for developing breast cancer was decreased for women who stopped intake of oral contraceptives more than ten years ago while the risk for developing breast cancer was decreased two years after finishing treatment with HRT (52).

There are also lifestyle factors associated with breast cancer. Alcohol consumption is positively associated with ER+ and PR- breast cancer, and the association is even stronger for postmenopausal women. Alcohol can elevate the level of ER related hormones (53). There are conflicting results regarding the association of dietary fat with breast cancer, with some researchers suggesting that saturated fat that is more associated with breast cancer. Phytoestrogens and meat cooked at high temperatures have also been connected to an increased risk of developing breast cancer (54).

2.2.4 Breast cancer treatment

Oncologists, radiologists, surgeons, and pathologists are involved in the diagnostics and the treatment of breast cancer. Patients who are diagnosed with an operable tumor are treated with surgery and often with different combinations of systemic treatment and radiation therapy.

Of the surgical methods used, the most common are mastectomy and breast conserving surgery (BCS).

Mastectomy can be either total or simple, skin-sparing and nipple/areolar-sparing. The local recurrence rates vary with up to 7% for skin-sparing and with up to 5% for nipple/areolar-sparing mastectomy (55-57). The site of 80% of recurrences is the chest wall (58).

BCS is the most recommended surgical method, involving removal of the tumor and a rim of surrounding healthy tissue. BCS is most successful for DCIS and T1-T2 tumors if the woman can undergo radiation. For women with high risk of local recurrence, BCS is not recommended (59). Randomized studies show that BCS followed by radiotherapy has an equivalent survival rate to mastectomy for stage I to II invasive breast cancers (60). Tumor-free margins are important for patients who undergo BCS. For invasive breast cancers there should be ‘no tumor on ink’ and for DCIS the margins should be at least 2 mm (61). Re-excision occurs in around 20% but according to one study there were residual tumor cells in only 50% of the specimens (62). Many studies indicate that BCS gives the patient a better quality of life and similar satisfaction levels compared to mastectomy with immediate reconstruction (63). If more tissue than expected needs to be removed during BCS, there are several oncoplastic methods to fill the tissue-defect (64).

The first lymph node to drain the lymphatics from the breast is called the sentinel lymph node (SLN). Patients with an early stage invasive breast cancer and a clinically and radiologically negative axilla are recommended a sentinel lymph node biopsy (SLNB) (65). For around 90% of all patients the sentinel node can be found and the false negative rate is low, at around 5%to 10%, and the risk for a local axillary recurrence is less than 1% after a negative SLN (66). If a patient has three or more lymph node metastases, axillary lymph node dissection (ALND) is usually performed, which is associated with morbidity such as altered sensation, pain and lymphedema in the upper limb (67). Many studies have resulted in a trend towards less axillary surgery: the large SENOMIC and SENOMAC studies were designed to examine the usefulness of ALND vs SLNB (68).

A study published in Cancer 1995 showed that for 20% of mastectomy specimens, there were additional tumor foci within 2 cm of the index tumor (69). This is one reason for the introduction of radiotherapy; to remove unknown remaining tumor foci despite margins being free. Radiotherapy can be delivered to the whole breast, to a part of the breast, to the chest wall or to lymph nodes. After BCS the whole breast is treated (70). Adjuvant radiotherapy decreases the local recurrence rate by 50% and increases breast cancer specific survival rate (70). In a meta-analysis of 17 randomized trials, the local recurrence rate

decreased from 35% to 19.3% and breast cancer related deaths decreased from 25.2% to 21.4% when adding radiotherapy to breast conserving therapy (71). There is no benefit with radiotherapy for patients with low-risk tumors and no metastases. However, radiotherapy is beneficial for women undergoing BCS with unfavorable risk factors (72).

Patients with high or intermediate risk breast cancer should be treated with chemotherapy. Patients with small tumors (1-5 mm) and negative lymph nodes do not generally benefit from chemotherapy (73). Patients with triple-negative breast cancer, breast cancer negative for ER and progesterone, and positive for HER2 benefit more from chemotherapy than hormone positive tumors (74). Neoadjuvant chemotherapy is recommended for inoperable tumors to make them operable, for locally advanced breast tumors to allow BCS, and for the evaluation of drug sensitivity during treatment (75-77).

There are different endocrine treatments with varying mechanisms, including prevention of ER production or by blocking the action of ER. The patients' hormonal status is important for choosing the right treatment. Tamoxifen is a drug that blocks the binding of ER to the receptor. Goserelin is another therapy that blocks the ovarian production of ER by inhibiting the pituitary gland to produce hormones that stimulate the ovaries. To inhibit the conversion of androgens to ER, treatment with aromatase-inhibitors such as anastrozole, exemestane and letrozole is the option (78).

Today, the recommendation for endocrine treatment is five years, although there are studies reporting that 10 years of treatment reduces the risk of tumor recurrence further (79). If a woman experiences adverse side effects of an endocrine treatment, there are options to mix aromatase inhibitors with tamoxifen within certain intervals (80). Women treated with endocrine therapy over a long period often need additional treatment with zoledronic acid to strengthen the skeleton and to avoid pathological fractures.

Treatment recommendation for postmenopausal women is aromatase-inhibitor for five years and if there are lymph node metastases another five years with tamoxifen is recommended. For premenopausal women tamoxifen for five years is recommended and if the lymph nodes are affected tamoxifen for ten years is recommended and for younger women, additional treatment with goserelin is recommended (78).

The monoclonal antibody trastuzumab is available for targeted therapy for patients with HER2 overexpressing tumors. Trastuzumab is mediating cytotoxicity, cell cycle arrest and some level of apoptosis (81). Trastuzumab together with chemotherapy is synergistic and decrease the recurrence rate (82). Possible cardiotoxicity and treatment resistance are disadvantages with trastuzumab treatment (83, 84).

2.3 BREAST CANCER SCREENING

2.3.1 Imaging and sensitivity

Internationally, breast lesions in radiology are mainly described according to the BI-RADS (Breast Imaging Reporting and Data System) system. The BI-RADS system was developed in the United States of America and can be used for mammography, ultrasound, magnetic resonance imaging (MRI), and for density assessments (85). In Sweden, the BI-RADS system is generally not used, although some institutions do use the system for MRI assessments. The Swedish scoring system for mammography and ultrasound is partly similar to the BI-RADS system. The Swedish system also codes breast lesions from 1 to 5, where 1 is healthy and 5 is a clear cancer. The main difference is the expanded category 3, which contains a higher proportion of cancer in the Swedish system compared to the BI-

RADS system (where it should be below 2%). In the Swedish system, lesions of category 3 are always subject to biopsy while in the BI-RADS system, lesions of category 3 may instead be subject to radiological follow-up after six months. Another difference is that category 4 contains subgroups according to the BI-RADS system but not according to the Swedish system (86).

Mammography is the most common modality for breast imaging in screening programs. The sensitivity for mammograms varies between 48% to 98% depending on the structure and distribution of glandular tissue, fibrous tissue and fat in the breast. Mammograms from dense breasts, i.e., breasts with a lot of fibrous and glandular tissue confer a lower sensitivity than mammograms of more fatty breasts. The sensitivity also increases with higher age when women usually get more fatty breasts (87).

MD can be visually divided into different groups and classifications, including the BI-RADS classification, the Tabár classification (88), and the Wolfe classification (89). Internationally, the most commonly used classification system is the BI-RADS system where density is divided into four categories A to D and D represents the most dense tissue (85). In prior versions of BI-RADS it was a visual assessment of the quantity of density, but in the most recent version qualitative aspects are included. A qualitative difference between category C and B, is that category C should be chosen if there is a chance that density "may obscure small masses". This means that if there is a large blob of density in a small part of the breast it could still be category C, even if the total amount of density in the entire breast is not that high.

Mammography involves three different views: the craniocaudal view (CC), the mediolateral oblique view (MLO) and the mediolateral view (ML). CC and ML are perpendicular views, and the MLO is an oblique view which is oriented along the pectoral muscle towards the axilla and includes more glandular tissue than the ML and CC views. Mammograms of two women of the same age can look very different in terms of volume and the patterns of dense and non-dense tissue. The appearance of breast cancer in mammograms varies greatly, and includes microcalcifications, distortions, asymmetry, spiculated and non-spiculated masses (see figure 3).

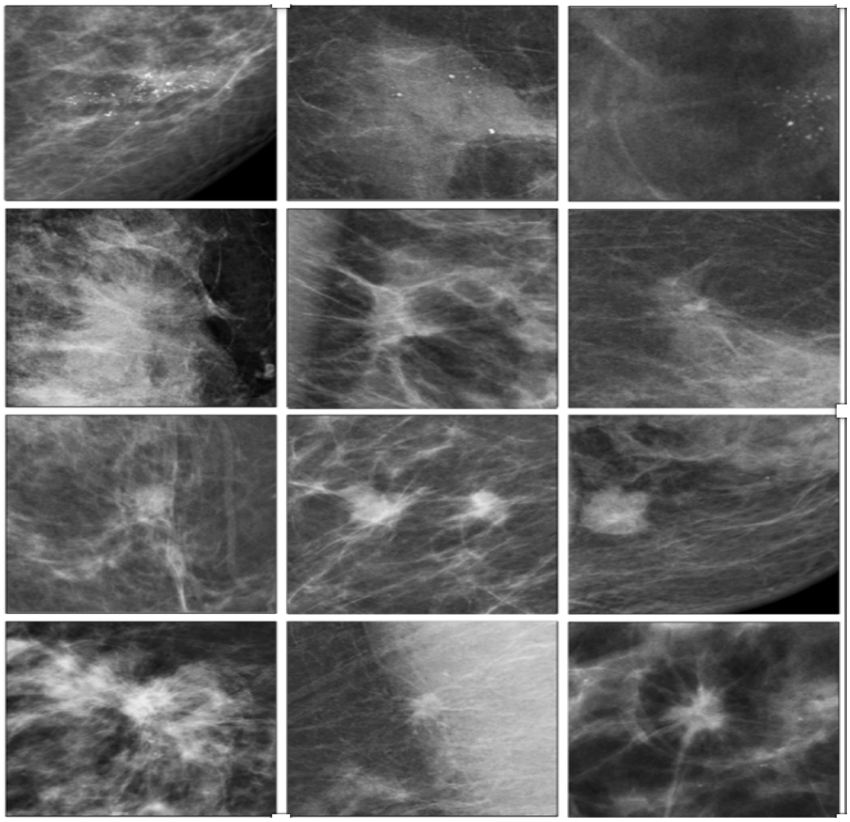

Figure 3. Mammographic appearances of breast cancer (courtesy of Fredrik Strand)

The sensitivity for ultrasound varies depending on how experienced the examiner is and how the breast tissue is comprised. A study by Berg et al from 2008 concluded that for women with elevated risk for breast cancer, the addition of ultrasound or MRI to mammography yielded a higher proportion of breast cancer diagnoses although the false positive rate increased (90). Ultrasound examination is well tolerated by women, and it is radiation-free. It has limited value as a single modality due to less sensitivity for the visualization of microcalcifications and a lower reproducibility than other techniques (91).

The combination of mammography and ultrasound can increase accuracy by up to 7.4% and the negative predictive value is greater than 98% when combining mammography and ultrasound when there is no palpable mass (92).

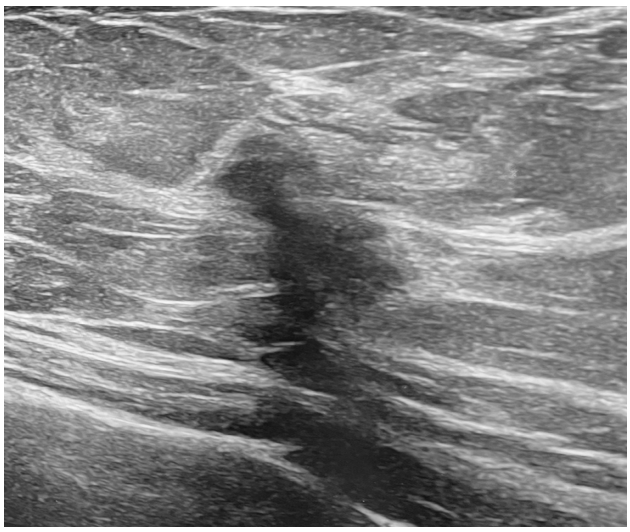

Figure 4. Image of a tumor from an ultrasound examination. (Karin Dembrower)

Examination with MRI has the highest sensitivity for finding malignant lesions in asymptomatic high risk women (71%-100%) compared to mammography (13-59%) and ultrasound (13-65%) (93). The MRI-findings are often classified according to the BI-RADS system. The randomized clinical trial performed by van Gils et al, the DENSE trial, implied that among women with extremely dense breasts invited to screening and examined with MRI, the proportion of IC increased with 80% compared to women who were examined only with mammography. In the second round of MRI examination in the same study, the proportion of false positive cases were strongly reduced (94).

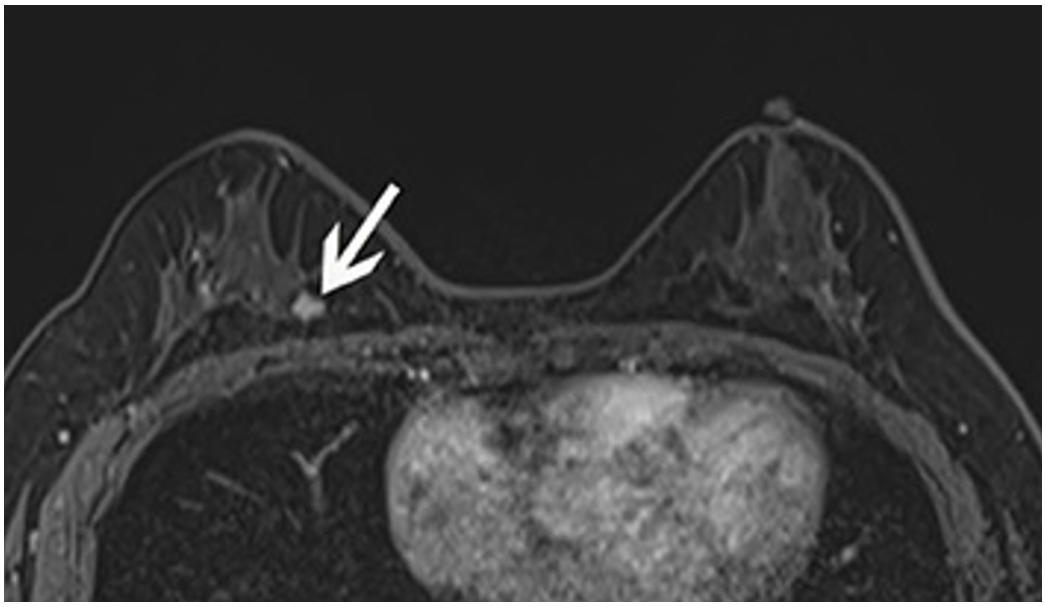

Figure 5. Example of a breast MRI examination with a contrast enhanced suspicious lesion close to the chest wall medially in the right breast. (https://commons.wikimedia.org/wiki/File:Breast_dce-mri.jpg)

Digital breast tomosynthesis (DBT) is a modality that involves multiple projections along an arc at small angular differences, then reconstructed into a stack of images. Depending on the manufacturer, the total arc along which the images are ensembled on varies between 15 to 60 degrees (95). In the “Malmö Breast Tomosynthesis Screening Trial” by Lång et al, the cancer detection rate increased with one view tomosynthesis (8,9%) compared to digital mammography (DM) (6.3%) (96). In the study by Conant et al “Five Consecutive Years of Screening with Digital Breast Tomosynthesis: Outcomes by Screening Year and Round”, the use of long term tomosynthesis demonstrated a higher detection of poor-prognosis cancers compared to DM (97). Another study by Conant et al comparing breast tomosynthesis with mammography demonstrated a higher proportion of smaller node negative breast cancers as well as a lower recall rate for DBT (98).

Since screening was introduced nationwide in 1989 in Sweden, mortality rates have decreased by up to 30 to 40% (99, 100). However, around 30% of all breast cancers from women attending screening programs are IC, detected clinically between screening intervals and some tumors are large (more than 2 cm) when detected at screening (101). These cases might be considered as failures, and show room for improvement of the screening programs.

Despite the general success of the screening programs, there are ongoing discussions regarding their harms and benefits. One detriment is the recall of healthy women for radiological work-up, which may impact their mental well-being due to worry and anxiety. There are also claims that treating cancer in situ is overdiagnosis (102, 103).

2.3.2 The current screening process in Sweden

Breast cancer is more compliant to treatment when detected early and therefore many countries have introduced screening programs (104). Swedish population-based national screening programs were introduced during the 1980s, and today all women aged 40 to 74 years are called for screening every 18 to 24 months. The attendance in the Swedish screening program is around 70-80% (105). The screening examination consists of two views, CC and MLO, of each breast, producing four images in total. In addition, nursing staff will ask questions regarding breast symptoms, hormonal medication and prior breast history.

All mammograms are assessed by two independent breast radiologists. If either flags for potential cancer in the images, or if the patient notes that she has serious symptoms, the mammograms will be discussed at a special meeting called consensus discussion. During the consensus discussion at least two breast radiologists finally discuss whether the woman should be declared as healthy or recalled for further work-up (106). The work-up is individualized, depending on the symptom or the suspicious finding in the image. The mammographic examination is often extended when the woman is recalled. Usually, additional imaging is needed such as magnification images, tomosynthesis, ultrasound or even MRI examination.

In Sweden the recall rate, i.e., the proportion of women who are recalled after attending screening, is around 2 to 3%, while the tumor detection rate worldwide and in Sweden is around 0.6 to 0.8 % per screening interval (107, 108). The recall rate is higher for women attending the first screening round. If the recall rate is too low there will be an increased number of false-negative women, increasing the risk of cancer cases, while if it is too high there will be an increased number of false positive women, unnecessarily worrying healthy women.

Between 15% to 35% of all cancers are missed in the screening programs because the cancer is not visible or the radiologist was not able to perceive the cancer in the mammogram. A majority of these cancers are later diagnosed symptomatically as IC (109). IC is associated with a higher morbidity and mortality (101).

Figure 6. The current screening process in Sweden

Worldwide, most countries recommend biennial screening between the ages of 50 to 74. Some countries, including Sweden, start screening at 40 years because of the higher incidence of breast cancer in those countries. In some countries women of 40 to 49 years and over 74 years are welcome, but they do not receive an invitation letter. Different methods are used although mammography by far is the most common method. Breast self-examination (BSE), DBT, ultrasound, MRI and identification of certain oncogene mutations are also methods used in the screening process. In some countries ultrasound is recommended together with mammography for women with dense breasts. MRI is not recommended as a primary screening modality in any country (110).

2.4 ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND DEEP LEARNING

Artificial intelligence (AI) is defined as any technique that mimics human decision-making. Machine learning is a subset of AI that enables machines to improve with experience and adaptation. The term 'machine learning' also includes logistic regression models fitted to empirical data. In this thesis, for convenience, I will use the term AI to refer to the newer types of AI, specifically deep learning.

Deep learning is a subset of machine learning techniques that has gained popularity during recent years. They are based on deep neural networks that allow more complex processing of input data. There are many different architectures of deep learning, but all are based on data nodes arranged in layers, going from the input data to output data and each layer can process data from earlier layers and affect the later layers. Each node contains a numeric value, and the connections between nodes are defined by mathematical formulas. When feeding deep learning models with large datasets, the discrepancy between the output of the network and the ground truth, is used to create an adjustment of the connections between the nodes through a method called 'backpropagation'. Backpropagation defines how the network weights, or coefficients, should be adjusted, based on minimizing the overall classification error of the model.

Training data with known outcomes is used to train the networks, validation data is used to adjust the training, and finally test data is used to test the network predictions. This kind of training is called supervised training. For the validity of the models, it is important that training data do not overlap with test data, neither by individual observations nor by including the same patients (111).

Deep learning methods have dramatically improved computer-based speech recognition (112), visual object recognition (113), language translation and object detection (114). It is sometimes stated that deep neural networks might find underlying relationships in a set of data in a way that mimics how the human brain operates.

The deep neural network in study II was developed with collaborating researchers and engineers from Kungliga Tekniska Högskolan (KTH). The network architecture was an Inception ResNet-v2 network (115). The input data were mammographic images, age at image acquisition, and image acquisition parameters such as exposure, tube current, breast thickness and compression force. The output of the network was a risk score, in which a higher number denoted women with a higher risk of breast cancer within five years.
The deep neural network in study III and IV was a commercial cancer detection algorithm trained on 170,230 images, 36,468 diagnosed women and 133,762 healthy women. The mammograms were both screening and clinical mammograms and came from South Korea, USA and the UK. The training images were acquired on equipment from GE, Hologic and Siemens. The output was a generated prediction score 0-1 for malignancy in the image, where 1 represented the highest level of suspicion.

2.4.1 Deep learning and tumor detection

Over the past 20 years, Computer Aided Detection (CAD) programs have been developed to assist the radiologists in analyzing screening mammograms. Traditional CAD programs usually mark a suspicious region in the mammogram and the radiologist will assess the suspicious area. The technique was spread quickly and in 2008 in the USA in the Medicare population, 74% of all screening mammograms were assessed by CAD programs (116, 117). However, it was never a success in Europe. There are controversial results for using CAD techniques. Initially, when CAD was introduced several studies indicated promising results with a higher sensitivity and increased cancer yield when adding CAD to the

analysis (118, 119). However, during the last ten years many studies indicate that the performance of CAD programs did not improve the performance of radiologists in the everyday practice in the USA (117, 120).

Figure 7. CAD program marking a tumor in a mammogram (Karin Dembrower)

During the last ten years, deep neural networks have been developed and in contrast to traditional CAD programs they do not normally involve handcrafted features (121, 122). In 2016, an international challenge (the DREAM challenge) was organized to analyze if artificial algorithms could outperform radiologists' performances. There were 126 teams participating in the challenge, assessing 144,231 screening mammograms. No single algorithm outperformed the human performance but a combination of algorithm and radiologist assessments improved the overall accuracy (123).

In the retrospective study by Salim et al (2020), three commercial tumor detection algorithms performance was evaluated on a single dataset (124). The study demonstrated that there were large differences in performance between the three algorithms they evaluated. They also showed that combining the first reader with the best one of the three AI algorithms identified more cancer cases than combining the first and the second reader. Other retrospective studies have indicated that deep learning systems are better than experienced radiologists and fewer cancers might be missed by fatigue or subjective diagnosis. Some have suggested that radiologists will be totally replaced by AI, whilst others believe that will not happen since our breast cancer patients need complex assessments and interventions that can only be performed by humans (125-127).

The results of the algorithmic assessments are presented in different ways such as continuous scales between 0 to 1, 0 to 10 or 0 to 100 where 0 demonstrates the lowest risk of having a tumor (128, 129). The sensitivity of different AI-algorithms differ between manufacturers and between datasets depending on many factors, e.g., the architecture and training of the algorithm as well as the size and quality of the dataset (130).

At my hospital (Capio Sankt Görans Hospital in Stockholm) we are conducting a prospective clinical AI study (ScreenTrust CAD, NCT04778670). We use a tumor detection algorithm as a third independent reader for our screening assessments. My impression is that the CAD system is very good at finding suspicious microcalcifications but tends to flag too many false positive findings because it is not able to compare with prior images. I think

that the systems might become even better when there is an ability to compare the actual images with priors. In the ScreenTrust CAD study, we will primarily analyze whether the AI algorithm plus one radiologist is non-inferior to two radiologists. In addition, it is possible to explore various reader set-ups such as AI as a single, double or third reader.

2.4.2 Deep learning as an independent reader

It is well known that there is a huge lack of breast radiologists. It would be advantageous if their time could be more focused on women who are at high risk of, or already diagnosed with, breast cancer, and less on assessing healthy women (131).

The introduction of AI in medical imaging might provide means to improve the efficacy of mammography screening by reducing the need of human readers. There are some studies indicating that AI-algorithms perform above or on par with an average radiologist (122, 132). There are a few retrospective studies published indicating that there is a span of low-risk mammographic examinations that could be assessed by an AI algorithm independently without missing any cancers and thereby save radiologist time for more important work. The numbers of the part of mammograms that could undergo independent reading by an AI-algorithm vary in different studies, by around 19% to 60% (133-135). As for radiologists, AI-systems can also miss cancer. One study demonstrated that three out of seven AI-missed cancers were small, low-grade invasive tubular breast cancers (133). Some studies have demonstrated that AI-systems increase the sensitivity for calcifications and could be more sensitive to invasive cancers (136, 137).

2.4.3 Deep learning and breast cancer risk

Breast cancer risk may be calculated based on many risk factors such as age, breast density, family history, hormone exposure and other risk factors (138). Examples of traditional models for assessing breast cancer risk are the Gail model and the Tyrer-Cuzick risk model. These models are based on questionnaires, taking into account clinical and demographic data and risk factors such as family history, hormone replacement therapy, parity, age at first birth, heredity etc (5, 6).

Breast density can be assessed visually or by automated procedures. Examples of automated systems are LIBRA and Volpara. Density can be described in different ways, such as a category, percent density and dense area (139, 140). Only the latest version of the Tyrer-Cuzick risk model takes breast density into account (7).

In addition to the above-mentioned factors, many more, mathematically defined, image features have shown association with breast cancer risk (141). However, these and other human-specified features may not be able to catch all risk-relevant information in the images. By using a deep neural network, more risk-relevant information might be captured.

There are a few studies using deep neural networks for risk prediction. By using breast cancer risk models based on deep learning, it has been demonstrated that high risk women are more accurately selected. Risk scores based on deep neural networks have the strongest association with breast cancer and seem to be largely independent in relation to density measurements. It appears that deep neural networks might utilize more information from the mammograms than the density-based models (142-146).

3 RESEARCH AIMS

The overall aim of this thesis is to analyze how deep learning can be incorporated in the screening process. I have analyzed how deep learning can contribute to reduce radiologist workload without missing cancers, to perform short-term risk stratification by analyzing mammograms of supposedly healthy women as well as demonstrate different methods for setting the operating point for AI algorithms. One prerequisite for all studies was to have a proper and robust dataset which was described in manuscript I.

To improve our understanding of how deep learning can affect the screening process in different ways, the specific aims of my four studies were:

3.1 STUDY I

Aim: To develop a high-quality platform for training and testing of AI networks for screening mammograms

We knew that we needed a robust image dataset for analyzing AI performance. The dataset described in this study contains millions of mammograms from the Stockholm County breast centers using different mammography equipment manufacturers. Together with a well-established screening program with high attendance and linkage with nearly complete medical registers, the dataset provides an excellent platform for training and evaluating AI algorithms. Within this dataset we have created a smaller case-control subset for more efficient analyses of AI algorithms by reducing the abundant number of healthy women.

3.2 STUDY II

Aim: To evaluate and compare a deep learning risk score with standardized mammographic density for short term breast cancer risk prediction.

Our hypothesis was that the robust deep neural networks might extract more information in the mammograms than the traditional density-based models were able to. Our network was trained on one set of the images in the dataset described in study I, and then tested on another set of images. The images for the study population di not overlap with the training-set or the test-set.

3.3 STUDY III

Aim: To examine two roles for a commercially available AI cancer detector: as a single pre-reader to dismiss a proportion of normal mammograms; and as a final post-reader after a negative examination to identify women at highest risk of undetected cancer.

Our hypothesis was that a substantial proportion of the population with the lowest AI scores could be safely ruled out without missing cancers that would otherwise be screen detected by a radiologist. We hypothesized that many women with the highest AI scores after a negative examination would later show up with IC cancers or next-round screen-detected cancer, potentially detectable by another modality such as ultrasound or MRI earlier.

3.4 STUDY IV

Aim: To explore two different principles for choosing a sensitivity-based AI abnormality threshold

Our hypothesis was to set the abnormality threshold at a clinically meaningful and sustainable level by maintaining double-reading sensitivity of AI in combination with a radiologist rather than focusing on the independent sensitivity of AI compared to radiologists. We explored these two different principles for choosing the abnormality threshold to shed light on this issue, and to prepare for a prospective clinical study.

4 MATERIALS AND METHODS

4.1 UNDERLYING STUDY POPULATION - CSAW

The study populations for study II, III and IV were derived from the Cohort of Screen-aged Women, (CSAW), described in detail in study I. In short, CSAW contains all women invited to the national screening program within the Stockholm County area between 2008 to 2015. The purpose of this database is training and validating AI algorithms. We have also created a smaller case-control data subset within CSAW to more efficiently enable training and validation through random-sampling, rather than complete inclusion, of a large number of healthy women. All women were initially identified through the Regional Cancer Center Stockholm-Gotland from which we received data on radiologist assessments and clinical cancer data. Their images were extracted from the radiology databases of Karolinska University Hospital and Stockholm County joint image service.

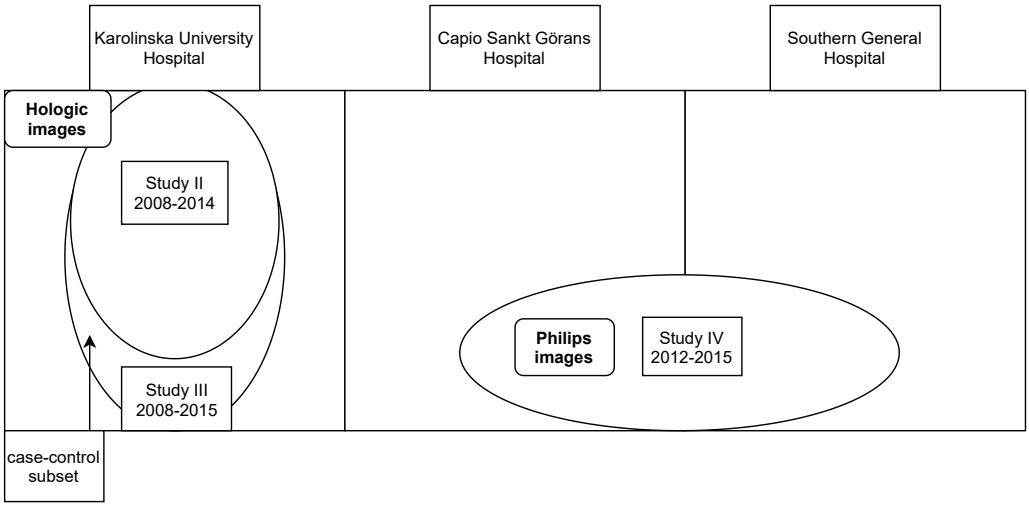

Figure 8. CSAW (study I). The distribution of the study populations within CSAW in the different studies II to IV.

4.2 REGISTER DATA

Population-based registers have a very long tradition in Sweden thanks to the personal number system, which was introduced by the Government in 1947. The personal number is assigned at birth and can only be changed under very rare circumstances. For Studies I to IV, participants were initially identified through the following registers:

- The Screening Register at the Regional Cancer Centre Stockholm-Gotland which contains data on attendance status, radiologist decisions and recall decisions.

Then, the personal numbers received were further linked to extract cancer data to the following register:

- The Breast Cancer Quality Register – a register that contains data on tumor receptor status, histological data, surgical margins, et cetera. This register in turn receives data from:
    - The Swedish Cancer Register which contains information about type of cancer, date of diagnosis, TNM stage, histological type. In 1978, 98.5% of all breast cancer diagnoses were reported to this register, which means there is a very small amount of missing data (147).

Finally, the personal numbers of all women with breast cancer were linked with:

- Karolinska University Hospital PACS (radiology image database), for the images pertaining to the Karolinska uptake area

- Stockholm county BFT (radiology service for all departments in Stockholm), for the images pertaining to the other breast centers of Stockholm (mainly Capio Sankt Görans Sjukhus and Södersjukhuset).

4.3 DENSITY MEASUREMENTS

The density-based measurements were calculated by the publicly available LIBRA software (version 1.0.4 University of Pennsylvania, Philadelphia, Pa) (148). In short, LIBRA provides a continuous measure of percentage density and dense area based on automated quantitative analysis of processed mammographic images. For study IV the density measurements were calculated by the software of the algorithm. The algorithm divides breast density into four categories 1 to 4.

4.4 EPIDEMIOLOGICAL STUDY DESIGN

Study I:
This study is a descriptive study of the cohort CSAW and its features and the areas of interest, as it has been described above.

Studies II– IV:
Study II is a case-control study containing 278 women diagnosed with breast cancer and 2005 randomly selected healthy controls without breast cancer through the end of follow-up in December 2015. All women were examined at the Karolinska University Hospital screening facility (on Hologic equipment). The study dataset is small since we needed a larger part for the prior development of the deep learning risk prediction algorithm.

Study III is a case-control study containing 547 diagnosed women and 6,817 randomly selected healthy controls. All women were examined at the Karolinska University Hospital screening facility (on Hologic equipment). This was an evaluation of an external AI algorithm, and therefore we could use the entire source population of women diagnosed with breast cancer with the main exclusion being women who did not fulfil the criteria to have visited two consecutive screening examinations.

Study IV was a case-control study containing 1,684 diagnosed women and 5,024 healthy controls. In contrast to studies II and III, we focused solely on images acquired on Philips equipment since the prospective clinical study is only on Philips equipment. These images were originally extracted from Capio Sankt Görans Sjukhus and Södersjukhuset.

In general, performing a case-control study is a practically efficient study design when the outcome is relatively rare, time to outcome is long, and the collection of exposure information is easy to assess. Given that around 0.6–0.8% of women receive a breast cancer diagnosis during a two-year period, the inclusion of all healthy women would in most cases constitute an inefficient study design. The starting point of a case-control study is the collection of individuals who are diagnosed with the outcome of interest. Then, individuals without the outcome, but at risk are collected. If the individuals without the outcome are sampled randomly, the results should be representative of the source population.

4.5 STATISTICAL CALCULATIONS

Odds ratio (Studies II-III)

Odds Ratio (OR) is often used to describe risk measurements in medical case-control studies. The OR can be defined as the ratio of exposed to non-exposed individuals with the outcome of interest, divided by the corresponding ratio among individuals without the outcome. The OR demonstrates how under- or overrepresented the exposure is among those who obtained the outcome.

Student's t-test (Studies I–IV)

Student's t-test or t-test are terms for statistical hypothesis testing and a process for rejecting or not rejecting a specific hypothesis, usually called the null hypothesis. With this test you can calculate if there is likely to be a difference between two samples from a normally distributed population. The possible rejection can be described by the so-called *p*-value. The *p*-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct (i.e., there is no true difference).

Student's t-test was used in Studies I–IV to compare normally distributed measures between groups. In Study I we compared measures between different age-groups diagnosed with breast cancer and in Study II we analyzed different predictors (follow-up time, age at mammography, dense area, percentage density) with the Student's t-test. In Study III we analyzed different predictors (AI score and percent density) with Student's t-test. In study IV the predictor (AI CAD score) was dichotomized to be similar to a radiologists dichotomized assessment - suspicious versus healthy. We tested for differences in subgroups for different methods of choosing the abnormal interpretation rate.

Characteristic	All Women (n = 2283)	Women with Cancer (n = 278)	Women without Cancer (n = 2005)	P Value	Missing Data
Patient characteristics*					
Follow-up time (y)	3.6 ± 2.2	4.1 ± 1.1	3.6 ± 2.3	<.001	0
Age at mammography (y)	54.6 ± 9.2	55.7 ± 9.1	54.6 ± 9.2	<.001	0
Dense area (cm^2)	34.4 ± 14.9	38.2 ± 15.3	34.2 ± 14.8	<.001	0
Percentage density	24.1 ± 13.0	25.6 ± 12.6	24.0 ± 13.0	<.001	0

Table 1. Comparing different predictors with Student's t-test (study II).

Logistic Regression (Studies II and III)

Logistic regression models are often used when the outcome is binary, i.e. the outcome can have two different values such as breast cancer or healthy. This is commonly used in medical research where regression models examine several potential predictors for patients having vs not having a particular disease or condition. The result is often presented as the estimated OR or as the area under the receiver-operating characteristics curve (AUC) rather than the actual calculated model coefficients. AUC provides a measure of the overall accuracy of a binary classification model.

Logistic regression modelling was used in Studies II and III with breast cancer or not as the outcome and mammographic percent density, mammographic dense area and DLrisk score as the predictors (Study II) and MD and AI score as the predictors (Study III). The results were presented as ORs with 95% confidence interval (CI) and AUCs.

Variable	OR	*P* Value	AUC	*P* Value
Without age adjustment				
DL risk score	1.55 (1.48, 1.63)	<.001	0.65 (0.63, 0.66)	<.001
Dense area	1.27 (1.20, 1.33)	<.001	0.58 (0.57, 0.60)	Reference
Percentage density	1.13 (1.06, 1.19)	<.001	0.54 (0.52, 0.56)	<.001
Composite*	…	…	0.66 (0.64, 0.67)	<.001
With age adjustment				
DL risk score	1.56 (1.48, 1.64)	<.001	0.65 (0.63, 0.66)	<.001
Dense area	1.31 (1.24, 1.38)	<.001	0.60 (0.58, 0.61)	Reference
Percentage density	1.18 (1.11, 1.25)	<.001	0.57 (0.55, 0.58)	<.001
Composite*	…	…	0.66 (0.64, 0.67)	<.001

Table 2. Deep learning risk score and mammographic measures associated with future breast cancer (study II).

Up-sampling and bootstrapping

We used the traditional cumulative sampling of healthy women in the study populations in the different studies. However, to calculate realistic performance measures in an enriched dataset with too many diagnosed women compared to the number of healthy women, we applied upsampling of healthy women. The two main approaches to resample a dataset, to obtain a desired proportion of observations between classes, are to delete examples from the over-represented class, called undersampling, or to duplicate examples from the under-represented class, called upsampling. Random upsampling has been used for a long time, and has been shown to be robust (149). It is important to note that it is not appropriate to perform statistical tests on the, after up-sampling, artificially enlarged study population. Bootstrapping may be applied, which involves sampling with replacement from the upsampled dataset to obtain the same sample size as the original dataset, permitting estimation of summary statistics with confidence intervals for measures involving both diagnosed and healthy women – such as the abnormal interpretation rate. Without bootstrapping, differences could be tested on the original, smaller, study population within diagnosed women (e.g., for sensitivity) or within healthy women (e.g., for specificity).

Standard Deviation

Standard deviation (SD) describes how the measures of the amount of a set of values varies from the mean. A low SD indicates that the values tend to be close to the mean (could be called as the expected value), while a high SD indicates that the values are spread out over a wider range.

5 RESULTS

5.1 STUDY I

This study is a description of the study population from which the study populations in the three following studies (and many more) are derived.

In total, 499,807 women were included in the CSAW cohort, based on 1,688,216 invitations between 2008 and 2015. In this cohort, 8,463 women were diagnosed with their first incident cancer (2,119 women were excluded due to a prior history of breast cancer outside the screening range). The average age was 53.2 years (SD 10.1) overall, and 57.8 (SD 9.3) for women diagnosed with breast cancer. The attendance during these years was 70%. Most of the women had three to four screening rounds during the study period. In total there were 4,703 SDC and 1,938 IC. The proportion of IC was 29%. The most common invasive cancer was ductal (67%), and the second most common was lobular (11%).

The total number of images in the CSAW dataset is more than 2 million, including all breast cancer cases, all healthy cases from the Karolinska University Hospital, and a random sample of healthy cases from the other breast centers in Stockholm.

Table 2. Mammography screening examinations.

	Invitation to screening		Completed examinations	
	n	%	n	%
Karolinska University Hospital	278,996	17%	198,820	17%
Sankt Göran Hospital	668,366	40%	454,341	38%
Southern General Hospital	482,883	29%	340,866	29%
Danderyd Hospital	257,717	15%	188,527	16%
Other	254	<1%	179	<1%
Total	1,688,216	100%	1,182,733	100%

Note: Each screening examination contains four images, two of each breast.

Table 3. Number of invited women and completed examinations within CSAW.

For the separately described case-control subset, only women from the Karolinska University Hospital were included. All images from women diagnosed with their first breast cancer (n=1 303) were included as well as 10,000 randomly selected healthy controls. Pixel-level annotations were made by the author in 1,891 mammograms from 898 women (Figure 9, below).

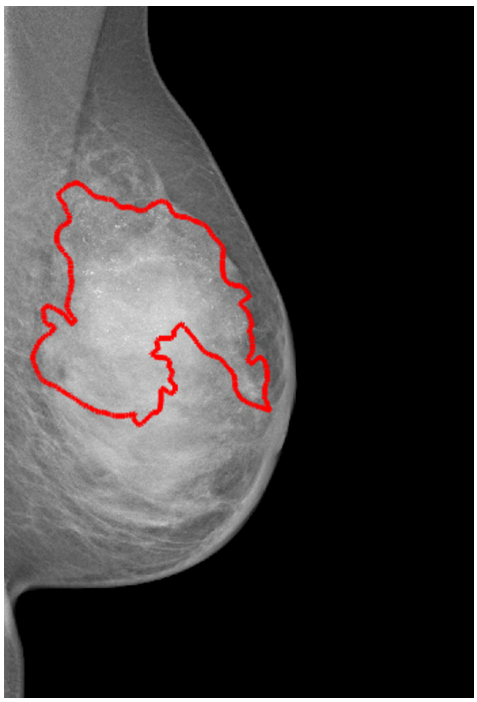

Figure 9. Example of a pixel-level annotation in a mammogram from CSAW.

5.2 STUDY II

The purpose of this study was to develop a deep learning risk (DLrisk) score associated with future breast cancer and compare it with density-based models. The deep learning network development was based on cases diagnosed from 2008 to 2012 and tested on cases between 2013 to 2014, along with healthy controls. The DLrisk score, dense area and percentage density were calculated for the earliest available digital mammographic examination for each woman.

In total, 2,283 women were included in the study population of which 278 women were diagnosed with breast cancer. The mean follow-up time was 4.1 (SD 1.1) years for women who received a diagnosis of breast cancer and 3.6 (SD 2.3) years for healthy women. The mean age at mammography, mean dense area and the mean percentage density were higher among women who were later diagnosed with breast cancer (55.7 years vs 54.6 years, 38.2 cm^2 vs 34.2 cm^2, and 25.6% vs 24%.

We calculated ORs and AUCs for the associations between each predictor and breast cancer during the follow-up period. The standardized ORs for DLrisk score, dense area and percent density were 1.55 (95% CI 1.48, 1.63), 1.27 (95% CI 1.20, 1.33) and 1.13 (95% CI 1.06, 1.19) respectively. The corresponding ORs after age-adjustment were 1.56 (95% CI 1.48, 1.64), 1.31 (95% CI 1.24, 1.38) and 1.18 (95% CI 1.11, 1.25) respectively.

In a multivariate model with DLrisk score, dense area and age as predictors, DLrisk (OR 1.52; 95% CI 1.42, 1.59) and dense area (OR 1.15; 95% CI 1.09, 1.22) were both associated with the outcome. In a multivariate model with DLrisk score, percent density and age as predictors, only DLrisk (OR 1.55; 95% CI 1.47, 1.64) was associated with the outcome However we had insufficient evidence to detect an association between percent density and future breast cancer (OR 1.02; 95% CI 0.96, 1.08). The AUCs for the DLrisk score, age and dense area, age and all three measurements were 0.65 (95% CI 0.63.0.66), 0.60 (95% CI 0.58, 0.61) and 0.66 (95% CI 0.64, 0.67).

False negative predictions were defined as patient cases with a DLrisk prediction score below the median, who nevertheless were diagnosed with breast cancer later on. The false negative rate for DLrisk score, age-adjusted dense area and age-adjusted percent density were 31% (95% CI 29%, 34%), 36% (95% CI 33%, 39%; *P*=0.006 compared with the DLrisk score) and 39% (95% CI 37%, 42%; *P*<0.001 compared with the DLrisk score).

Table 2: Deep Learning Risk Score and Mammographic Measures Associated with Future Breast Cancer

Variable	OR	*P* Value	AUC	*P* Value
Without age adjustment				
DL risk score	1.55 (1.48, 1.63)	<.001	0.65 (0.63, 0.66)	<.001
Dense area	1.27 (1.20, 1.33)	<.001	0.58 (0.57, 0.60)	Reference
Percentage density	1.13 (1.06, 1.19)	<.001	0.54 (0.52, 0.56)	<.001
Composite*	…	…	0.66 (0.64, 0.67)	<.001
With age adjustment				
DL risk score	1.56 (1.48, 1.64)	<.001	0.65 (0.63, 0.66)	<.001
Dense area	1.31 (1.24, 1.38)	<.001	0.60 (0.58, 0.61)	Reference
Percentage density	1.18 (1.11, 1.25)	<.001	0.57 (0.55, 0.58)	<.001
Composite*	…	…	0.66 (0.64, 0.67)	<.001

Note.—Numbers in parentheses are the 95% confidence interval. AUC = ten-fold cross-validated area under the receiver operating characteristics curve, DL = deep learning, OR = per–standard deviation odds ratio.

* Composite score is based on a logistic regression model combining all three measures.

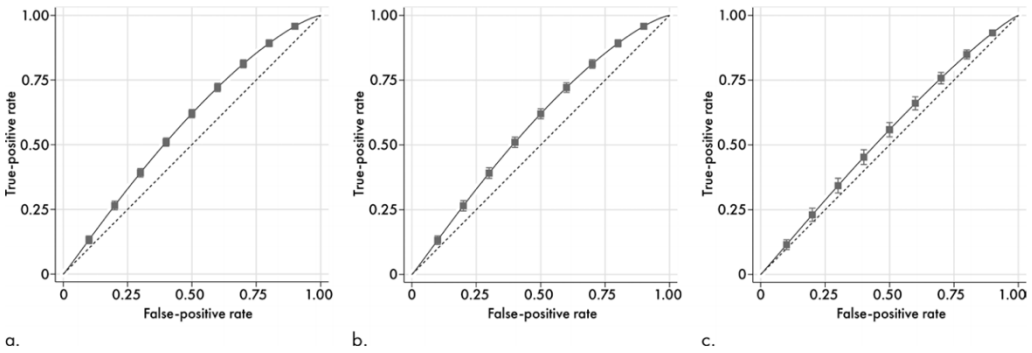

Table 4. ORs and receiver operating characteristic curves for the prediction of future breast cancer based on mammographic image evaluation according to three alternative predictors: a. deep learning risk score AUC 0.65 (95% CI 0.63, 0.66) b. dense area AUC 0.58 (95% CI 0.52, 0.56) c. percentage density AUC 0.54 (95% CI 0.52, 0.56)

Table 3: False-Negative Predictions

Variable	FN Rate for DL Risk Score (%)	Age-adjusted Dense Area		Age-adjusted Percentage Density	
		FN Rate (%)	*P* Value	FN Rate (%)	*P* Value
Overall (*n* = 278)	31 (29, 34)	36 (33, 39)	.006	39 (37, 42)	<.001
Lymph node status					
Negative (*n* = 193)	31 (28, 34)	35 (32, 38)	.06	36 (33, 39)	.002
Positive (*n* = 71)	31 (26, 37)	42 (36, 47)	.02	48 (42, 54)	.01
Tumor size, invasive					
≤15 mm (*n* = 162)	32 (28, 35)	37 (33, 41)	.01	40 (37, 44)	<.001
>15 mm (*n* = 80)	28 (23, 33)	35 (30, 41)	.01	38 (33, 43)	<.001
Molecular subtype					
Luminal A (*n* = 158)	32 (28, 36)	40 (36, 44)	.002	42 (38, 45)	<.001
Other (*n* = 88)	28 (24, 33)	33 (28, 37)	.12	36 (31, 41)	.002
Tumor grade					
Low (*n* = 61)	29 (23, 35)	29 (24, 35)	.91	36 (30, 42)	.04
Medium or high (*n* = 205)	31 (28, 34)	38 (35, 41)	<.001	41 (37, 44)	<.001

Note.—Data are for women with a prediction below the median within each model (the three columns) who had a later diagnosis of breast cancer. Numbers in parentheses are the 95% confidence interval (based on the sandwich variance estimator). Differences were tested with the McNemar method. DL = deep learning, FN = false-negative.

Table 5. Women with a prediction score below median and diagnosed with breast cancer – false negative predictions

Figure 10. Examples of mammograms with concordance between DLrisk score and outcome of breast cancer (true predictions). These mammograms had low DLrisk scores and the women did not receive a diagnosis of breast cancer.

Figure 11. Examples of mammograms with concordance between DLrisk score and outcome of breast cancer (true predictions). These mammograms had high DLrisk scores and the women received a diagnosis of breast cancer.

Figure 12. Examples of mammograms with discordance between DLrisk score and outcome of breast cancer (false predictions). These mammograms had low DLrisk scores and the women received a diagnosis of breast cancer.

Figure 13. Examples of mammograms with discordance between DLrisk score and outcome of breast cancer (false predictions). These mammograms had high DLrisk scores and the women did not get a diagnosis of breast cancer.

5.3 STUDY III

In this study we examined the potential change in cancer detection when using an AI cancer-detection software to triage (pre-read) certain screening examinations into a 'no radiologist' work stream and after regular assessment triaging mammograms with a high suspicion of cancer into an 'enhanced assessment-work' stream.

All women diagnosed with breast cancer who attended two consecutive screening rounds were included. A total of 7,364 women were included in the study sample, of which 547 women were diagnosed with cancer and 6,817 women were randomly selected as healthy controls. A total of 347 cancers were detected at screening and 200 were detected clinically as IC. Healthy women from the case-control source population were up-sampled to mimic a realistic frequency of 0.7% cancer per interval, and the simulated screening population contained 75,534 women.

Within the no radiologist work stream, the AI score did not miss any cancer that would otherwise have been screen detected for 60% of the lowest AI scores. For the 70%, 80% and 90% lowest AI scores, there were 1, 9, and 14 cancers missed by the AI score respectively.

	n	Proportion (95% CI)
Lowest 10%	0	0 (NA)
Lowest 20%	0	0 (NA)
Lowest 30%	0	0 (NA)
Lowest 40%	0	0 (NA)
Lowest 50%	0	0 (NA)
Lowest 60%	0	0 (NA)
Lowest 70%	1	0·3% (0·0–4·3)
Lowest 80%	9	2·6% (1·1–5·4)
Lowest 90%	14	4·0% (2·1–6·9)
All	347	100·0% (NA)

Table 6. Number of screen-detected cancers that would be missed in the no radiologist work stream depending on the proportion of the population lowest scores included.

In the enhanced assessment work stream we calculated that among the top 1% of AI scores for women with negative mammograms after double reading, there were 24 IC and 48 next round SDC. The corresponding numbers for the top 5% were 53 IC and 121 next round SDC. The ORs for predicting IC for AI score and density was 2.01 (95%CI 1.98, 2.18); AUC 0.74) and 1.59 (95%CI 1.50, 1.68; AUC 0.67) respectively. The OR was markedly higher in the ipsilateral breast than in the contralateral breast, while the OR was similar in both breasts for MD.

	Interval cancer (n=200)	Next-round screen-detected cancer (n=347)	Cancer of both categories (n=547)	Additional cancer detection rate*
Highest 1% (n=633)	24 (12%)	48 (14%)	72 (13%)	114/1000
Highest 2% (n=1445)	32 (16%)	71 (21%)	103 (19%)	71/1000
Highest 5% (n=5073)	53 (27%)	121 (35%)	174 (32%)	34/1000
Highest 10% (n=8746)	73 (37%)	155 (45%)	228 (42%)	26/1000
Highest 15% (n=12 571)	86 (43%)	183 (53%)	269 (49%)	21/1000
Highest 20% (n=16 181)	100 (50%)	204 (59%)	304 (56%)	19/1000
All (n=75 534)	200 (100%)	347 (100%)	547 (100%)	7/1000

Table 7. Potential detection of IC and next round SDC in the enhanced assessment work stream depending on the proportion of the population highest scores included (after negative double-reading)

	Interval cancer (n=200)	Next-round screen-detected cancer (n=347)	Cancer of both categories (n=547)	Additional cancer detection rate*
Highest 1% (n=633)	24 (12%)	48 (14%)	72 (13%)	114/1000
Highest 2% (n=1445)	32 (16%)	71 (21%)	103 (19%)	71/1000
Highest 5% (n=5073)	53 (27%)	121 (35%)	174 (32%)	34/1000
Highest 10% (n=8746)	73 (37%)	155 (45%)	228 (42%)	26/1000
Highest 15% (n=12 571)	86 (43%)	183 (53%)	269 (49%)	21/1000
Highest 20% (n=16 181)	100 (50%)	204 (59%)	304 (56%)	19/1000
All (n=75 534)	200 (100%)	347 (100%)	547 (100%)	7/1000

Data are n (%) or n/n. *The ratio was calculated with the total number of women in the population selected as the denominator.

Table 2: **Potential detection of interval and next-round screen-detected cancer in the enhanced assessment work stream depending on the proportion of the population highest scores (after negative double-reading) included**

Table 8. Number of cancers potentially detected pre-emptively by the enhanced assessment work stream minus the screen-detected cancers missed in the no radiologist work stream.

5.4 STUDY IV

In this study, we analyzed two different principles as to how the abnormality threshold can be set when using an AI algorithm in a prospective screening setting. For the first principle, we aimed at setting the AI operating point at the same sensitivity as the second human reader. For principle two, we aimed at setting the AI operating point at the combined sensitivity of reader one and two. In our workgroup, less than 10% of all examinations are subject to consensus discussion. Therefore, setting the threshold too low in order to definitively include all breast cancer cases would create an impossible workload; a more realistic aim is to set the threshold in relation to the sensitivity or specificity of the radiologists of today.

In this study, we aimed to maintain sensitivity when replacing a radiologist with AI, which we thought would serve to maintain the public trust in screening better than focusing on specificity. We included all women receiving a diagnosis of breast cancer at screening or within 23 months of screening and a random selection of healthy women. In total, there were 1,684 women diagnosed with breast cancer and 5,024 healthy women. To mimic the proportion of cancer (0.7%) in a true screening population, we upsampled (duplicated) the observations of healthy women, and the total number of the study population became 235,428.

The overall sensitivity for reader 1, 2 and 1+2 was 69.7%, 75.7% and 78.6% respectively. The proportion of abnormal assessments for reader 1, 2 and 1+2 were 4.4%, 4.6% and 6.1% respectively. AI alone had a sensitivity of 75.6% by principle 1 and 65.9% by principle 2. Reader 1 and AI together had a sensitivity of 82.4% and a proportion of abnormal assessment of 12.6% by principle 1. AI and reader 1 together with the combined sensitivity of reader 1 and 2 had a sensitivity of 78.6% and a proportion of abnormal assessment of 7.0% by principle 2.

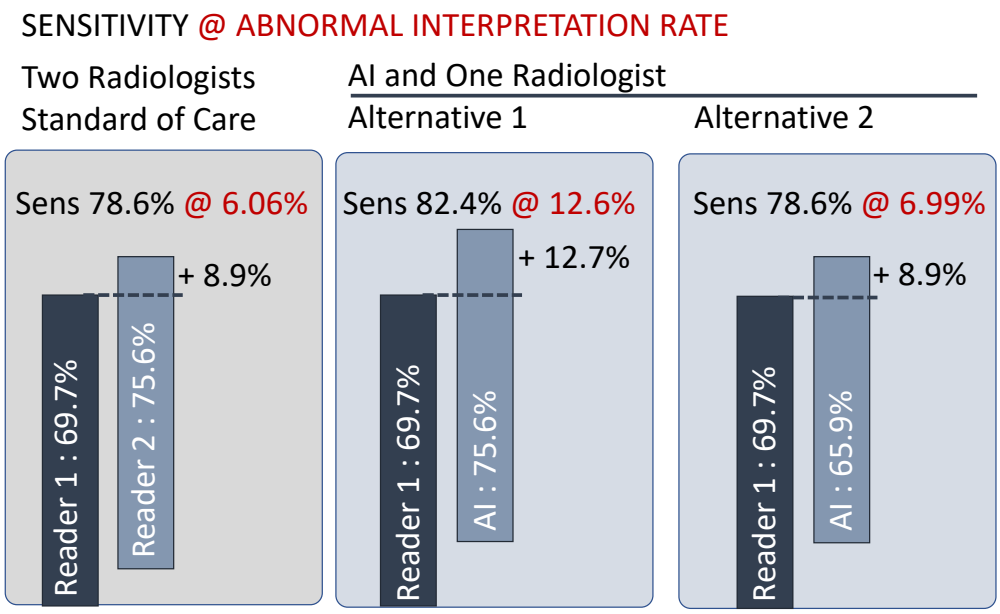

Figure 14. This figure illustrates the percent of women with cancer that are flagged as suspicious by each reader. The vertical offset between the light blue and dark blue bars illustrates the discordant assessments were only one of the readers correctly flagged the case as suspicious. We show the observed measures for two radiologists and the estimated measures for AI and one radiologist, where the operating point was chosen based on standalone-reader matching (Alternative 1) and on combined-reader matching (Alternative 2).

Table 3. Operating Characteristics in Double-Reading

	Reader 1 + Reader 2	Reader 1 + AI@Thr1		Reader 1 + AI@Thr2		
	% (95% CI)	% (95% CI)	*p*- value*	% (95% CI)	*p*- value*	*p*-value**
Specificity (%)	94.4 (93.8 to 95.1)	87.9 (87.0 to 88.8)	p < 0.001	93.5 (92.8 to 94.2)	p = 0.049	*p* < 0.001
AIR* (%)	6.06 (5.49 to 6.63)	12.6 (11.8 to 13.4)	p < 0.001	6.99 (6.38 to 7.59)	p < 0.001	p < 0.001
Sensitivity (%)	78.6 (76.6 to 80.5)	82.4 (80.6 to 84.2)	p = 0.005	78.6 (76.6 to 80.5)	p = 1.0	*p* = 0.005
Density Category 1	84.0 (77.5 to 90.6)	89.9 (84.5 to 95.3)	p = 0.178	86.6 (80.4 to 92.7)	p = 0.583	*p* = 0.421
Density Category 2	90.1 (87.3 to 92.9)	90.6 (87.8 to 93.3)	p = 0.170	87.2 (84.1 to 90.3)	p = 0.170	*p* = 0.110
Density Category 3	76.5 (73.8 to 79.2)	80.8 (78.3 to 83.3)	p = 0.022	77.6 (74.9 to 80.2)	p = 0.586	*p* = 0.080
Density Category 4	55.4 (47.9 to 63.0)	64.5 (57.2 to 71.7)	p = 0.093	55.4 (47.9 to 63.0)	p = 1.0	*p* = 0.093

AI@Thr 1= Threshold set to achieve same sensitivity of AI as replaced Reader 2
AI@Thr 2= Threshold set to achieve same sensitivity of Reader 1 + AI as Reader 1 + Reader 2
*) p-value for the comparison with Reader 1 + Reader 2
**) p-value for the comparison with Reader 1 + AI@Thr1

Table 9. Operating characteristics in double reading (sensitivity)

The highest increase in sensitivity was for women with dense breast, category 4. The sensitivity increased from 55.4% to 64.5% for the standalone-reader matching.

6 DISCUSSION

6.1 STUDY I

We observed that older age was a risk factor for diagnosis of breast cancer compared to healthy controls, which is in line with prior studies (150). We also found that nearly 30% of all breast cancers were clinically detected in the interval between two screenings, which is also in line with previous studies (101). We can be confident that the diagnoses are correct, as 99% of the breast cancers were biopsy-verified and underreporting to the cancer registry is very low (1.1% to 1.6%) (151). Several research areas can be addressed using this dataset, with some of those being tackled at present listed below:

- Developing risk prediction networks (study II)
- Developing tumor detection networks (ongoing prospective study, ScreenTrust MRI, Karolinska University Hospital)
- Developing sensitivity assessment networks
- Evaluating and validating third-party networks (implemented)
- Interactive education and continuous training (implemented in a wide context among residents nationwide in Sweden) In November 2020 we held a course for residents in Stockholm. The participants used I-pads showing selected cases from CSAW to learn about breast pathology. The teaching was more compressed than going through a screening population with mainly healthy women. The course was well-received, and the participants were very happy to be able to identfy different tumors after short training.

The strengths of the CSAW dataset are that all women within a specific geographic uptake area are included, without exclusions, a large number of diagnosed women, clinical data and image acquisition parameters are available, as well as the free-hand pixel-level annotation dataset, which can make precise comparisons by location. Other available mammography datasets are available but they are a lot smaller than CSAW, with content in the range of hundreds or thousands of images and small numbers of cases. CSAW contains millions of images and thousands of cases making it very robust (152).

A possible limitation of the dataset is that for a task that requires a huge amount of training data it might be too small. However, Study II demonstrated that from a small amount of data from CSAW we were able to develop an algorithm that performed better than breast density measurements for breast cancer risk prediction.

6.2 STUDY II

It is of great importance to be able to stratify at what risk a woman is for developing breast cancer and for determining whether she would benefit from enhanced screening. Previous models have taken important factors such as age, parity and heredity into account. When MD was considered, these risk models improved markedly.

In this study, we found that the DLrisk score could more accurately help to predict which women were at risk for future breast cancer compared to age-adjusted dense area (OR 1.56, AUC 0.65 and OR 1.31, AUC 0.60 respectively). The DLrisk score was an independent predictor for breast cancer in relation to density predictions and this is in line with previous deep learning studies regarding risk assessments for breast cancer (142, 144). The AUC of 0.65 implies that there is a 65% probability that the DLrisk score assigns a higher risk score to a woman who will get a diagnosis of breast cancer than a women who will remain healthy. This is better than the density measurement, AUC 0.60, but far from ideal. The breast cancer risk assessment study by Yala et al indicated an AUC of 0.70 for their hybrid

DL score (142). It is likely that AUCs for deep learning models have a great potential to improve, although it is not realistic that they might attain a perfect score of 1.0.

The fact that the correlation between DLrisk score and density-based measurements was quite weak shows that the DLrisk score is not a density estimator, and thus it can extract more type of information than just density from the images. Other studies imply that image-based models are superior to the traditional Tyrer-Cuzick model (142), (153). Why do the DL models perform better than the density-based models? Visually assessed MD is limited by inter-reader variability. Quantitative density, using a single density-measurement, is unlikely to capture all relevant risk information from an image that is useful for predicting breast cancer. It is likely that the AI algorithm and MD might capture complementary information and more precise delineate women at higher risk of developing breast cancer. The false negative rate was lower for the DLrisk score than for age-adjusted dense area overall, and especially for more density measurements regarding more aggressive cancers, for example lymph node positive disease (31% and 42% respectively). For these women it is important to improve early detection, i.e., before the breast cancer has spread to lymph nodes, to improve the prognosis of the disease. This reasoning is in line with a previous study by Ding et al analyzing density and future breast cancer types, which indicated that the association between the two parameters was stronger for less aggressive subtypes than those that were more aggressive (154).

It is important to consider where to set the cutoff to define false versus true predictions. In our study we chose the median as the threshold. A prospective study must take many parameters into account when deciding the cutoff point, for example the ability to further examine positive women, cost/benefit ratio, the disadvantages of causing healthy women mental anxiety during screening, and so on.

This study is based on a screening population that did not have any cases of breast cancer excluded, which is quite rare and contributes to the strength of the study. Thanks to the Swedish personal number system, the cancer registers are almost complete (97.7% for non-myeloma and non-leukemia cancers) (147). We used a temporal approach for validation of this model that might not always be an advantageous method and could make generalization to other settings and manufacturers difficult. We excluded mammograms within 12 months of diagnosis, and this meant that the algorithm could have been influenced by subtle tumor signs that it was able to catch more than 12 months before diagnosis. Another limitation is that our dataset, especially for training, could have some variety in the numbers and types of tumors, which might not be generalizable to other settings.

6.3 STUDY III

We demonstrated that a commercial AI breast cancer detection algorithm could be used both as a single reader to assess easily read mammograms without any radiologist involvement, and to select women for enhanced screening after negative double-reading by radiologists. We demonstrated that the AI breast cancer detection algorithm would not miss any screen-detected breast cancer among women with the lowest 60% scores. This is remarkable, as other studies have observed more modest results with a 19% cancer-free detection rate (135). If the algorithm were to solely assess 90% of all low score mammograms, we would miss only 4% of cancers that would otherwise be screen-detected; this is a relatively small number compared to all IC (28%) for all women invited to screening in a biennial basis (101). Given the lack of breast cancer radiologists, it would be most valuable to use their competence for women at a higher risk of developing breast cancer than examining healthy women. By using the AI algorithm to assess mainly healthy

women, we believe that we can reduce the workload substantially for radiologists without missing too many cancers. The AI algorithm has a great potential as an independent rule-out reader.

We also demonstrated that within the enhanced assessment work stream, the AI breast cancer algorithm could find a potential additional cancer detection rate of 71 cancers per 1000 examinations among women with the 2% highest AI scores. This is a remarkable performance of the AI algorithm, and it is important to point out that the algorithm has not been trained on any image from our institution. If we implement the enhanced assessment work stream we could promote earlier detection of screen-detected cancers and thus, a reduction of IC at the first screening as well as a reduction of SDC during the first round. Going forward – a shift towards smaller SDC would be expected.

Women are generally positive towards using computer programs to assess mammograms and to triage for MRI screening (155). However, in this context we think that those women ending up with a clinically detected cancer (IC) form an important target group for a discussion between policy makers and politicians when it comes to changing screening programs. In the USA, many states have decided that women should be informed whether they have an increased breast density and thus a high risk factor for IC (156, 157). These women can then discuss with their health-care provider whether supplemental screening is necessary to get a reliable answer from the screening examination (158). If we examined 20% of all women with the highest AI scores and placed them in an enhanced screening program, the additional IC detection rate would be as high as 6.2 per 1000 examinations. This would be an enviable performance and superior to another American study by Kerlikovske et al, which demonstrated an additional IC detection rate of 1.4 per 1000 woman when combining breast density with a traditional breast cancer risk model (46). However, that study differed from ours in two ways: breast cancer screening is mostly annual in the USA and IC rate is “as low as” 13% according to the BCSC (159).

When trying to determine how the AI functions so well, we speculate that the AI algorithm finds subtle tumor signs in the image that our eyes are not able to capture due to density masking any tumor signs present. Upon AI assessment, the image is marked at the location of the suspicious area and enables targeted ultrasound for women at high risk. MRI has the highest sensitivity for finding malignant lesions in the breast, but it is time- and cost consuming, and a targeted ultrasound examination is a good, safe and cheap alternative for a screening population.

In July 2021, and article in the British Medical Journal discussed the use of AI for image analysis in breast cancer. This was a systematic review of 12 studies published between 2010 and 2021, and included our Study III as one of the twelve. The authors claimed that almost all studies were connected to a lot of biases, it was not yet clear whether the use of AI in screening programs is beneficial, and that prospective studies are needed to further investigate this area. Certainly more prospective studies are needed, but I don´t agree with the discussion concerning bias. The authors noted a bia when choosing randomly selected controls, but in my point of view the random selection process removes bias. Some studies described screening processes under ‘laboratory conditions’ and this might be associated with bias. They also wrote that the applicability to European or UK breast cancer programs is low, meaning that the lack of a British study population made the studies not applicable to UK. However, in our study we used a true population-based screening cohort. Finally, the 12 studies under review were different in many ways, making comparison between them difficult (130).

The AI algorithm used in this study has never been exposed to images from our department, and is commercially available which is very advantageous. A weakness is that we had to use a case-controlled study design to improve computing efficiency. Another limitation is that women needed to have a prior mammogram not more than 30 months before diagnosis to be included, which affected the proportion of IC from 20% to 37%. We were not informed about the location of the tumor, and we could not confirm the tumor location via the AI algorithm findings. In addition, all women were from Sweden and the results might differ in another population.

6.4 STUDY IV

We found that when setting the abnormality threshold for AI to match the radiologist sensitivity, the abnormal assessment rate was almost twice as high compared to matching the combined readers sensitivity. This is important to keep in mind in the light of a severe lack of breast radiologists and a tradition of very low recall rates in Sweden. One explanation for the increased number of abnormal assessments when replacing one of the two double-reading radiologists with AI is that AI does not compare current images with prior images, which is central for radiologists. In prior retrospective studies, the method of choosing the AI abnormality threshold has been either to match the AI threshold to sensitivity, or specificity of human readers (123, 130, 146, 160-162). The abnormal assessment rate was 6.1% in our simulated screening study and that is in line with the Swedish breast cancer environment where the abnormal interpretation rates vary between 5% to 7% (105). It is not possible or correct to have a twice as high abnormal interpretation rate, which would be the case if the sensitivity was supposed to match that of the radiologist. Our study has taken a completely new approach by matching the combined sensitivity of AI and reader 1 to the combined sensitivity of reader 1 and 2.

Our group has published a study analyzing the performance of algorithms in another dataset (160). The double reader sensitivity in that study was 85.0% compared to 78.6% in our study. Differences between these studies are that the follow up period was 23 months in our study, compared to 12 months in the other, the mammograms were made only using Philips equipment in our study while the other used Hologic equipment, and the images originated from different breast centers in Stockholm in our study and only from one breast center in the other study. When matching the threshold on radiologists' sensitivity in our study, the overall sensitivity resulted in 82.4% compared to an overall sensitivity of 88.6% in the other study.

Why does the result differ? One reason could be that the AI algorithm in our study is mainly trained on GE and Hologic images and only a small set of Philips images. This could imply that the AI algorithm is better adapted to interpret Hologic. The highest increase in sensitivity was for women with most dense breasts (category 4), (increase from 55.4% to 64.5%) when choosing the standalone reader approach. This means that for women with dense breasts it is favourable to choose the standalone reader approach rather than the combined-reader approach.

This study is based on a large true screening population with a high attendance rate, and all women between 40 to 74 years were invited without exclusions. The screening registers and breast cancer registers are almost complete. A limitation is that images are derived only from one manufacturer, Philips, and the algorithm is trained mainly on mammograms from GE and Hologic. The study population was upsampled for healthy women because the retrospective study population was enriched. All radiologists were from Sweden, whith a tradition of very low recall rates.

7 CONCLUSIONS

AI holds great promise to improve breast cancer screening. This thesis shows the great potential of AI for both risk assessment and tumor detection. During the last few years, several retrospective studies have been published indicating that AI can perform on par with or even above radiologist performance. However, the settings are artificial, not least because all women recalled historically were recalled by the human radiologist. Also, to what extent radiologists may appreciate the findings of AI cannot be accurately studied retrospectively. Therefore, how the recall decisions of AI would have played out cannot be known with certainty. For this and other reasons, prospective studies using AI in large screening cohorts are needed. Prospective studies will tell us how well AI works in real clinical settings with no exclusions and in a 'non-laboratory' environment.

When it comes to understanding how well AI may perform, in Study II, we showed that AI was better than density as a risk predictor and in Studies III and IV, that AI performed well in detecting existing cancer in the images. However, it is not easy to shed light on which information in the images is most important or to obtain some sort of reasoning around how AI came to its conclusion. AI algorithms are to a large extent black boxes. Even though some techniques, such as saliency maps, can be used to visualize where in the image information is most important, this is not enough to create a human understanding of what AI is doing. Perhaps other techniques will become available, or we might have to accept that AI are largely black boxes – then we must make sure that those black boxes are accurate and robust in retrospective and prospective studies alike, and that we understand the limitations of their use. We are only in the infancy of development of AI for breast imaging.

8 ETHICAL CONSIDERATIONS

The ethical Review Board (ERB) waived the requirement for informed consent for our studies. As a researcher, it is important to fully protect study participants against the risk of physical injury and violation of integrity, as well as examining women using the right modality in the best possible work stream as early as possible when there might be a suspicion of breast cancer. This reasoning is derived from the utilitarianism, a version of consequentialism, which means that the consequences of any action are the only standards of right and wrong. Women might say, “Look at this – the AI algorithm assessed my images as high-risk images, and I got the diagnosis of breast cancer. Why don´t you use this in the everyday work?” or “The DLrisk score assessed my images as having high risk of developing cancer – what do I do now? Can I get an MRI examination?” Due to the retrospective nature of the current studies, these questions did not arise. However, for prospective studies, they may very well arise and they are not easy to address. One possibility is to consider the AI as a medical device for which the manufacturer is responsible given that the clinicians have used the device in accordance with the approved use cases.

Colleagues and patients are interested in these questions quite often and I currently have not reached a conclusion. I think that AI mistakes may be comparable to false positive and negative screening decisions made by radiologists, which to some extent is unavoidable given the very subtle image findings in some cases. However, it should not be taken for granted that the public perceives mistakes made by AI compared to mistakes made by humans in the same way.

Another ethical consideration is that even if in theory, AI algorithms may be deployed across geographies and even out differences in diagnostic quality, it all depends on cost. If the AI companies price their algorithms too high, this evening out of inequalities will not happen, and the inequality may become even more pronounced between AI-equipped well-staffed departments and AI-lacking short-staffed departments.

Finally, another consideration is the desire of management to decrease cost. Is it a good use of AI to replace radiologists without re-deploying the radiologists to further improve diagnostic work-up or screening of more difficult cases requiring their expertise? I hope that we focus on using AI to improve the health and well-being of our patients instead of using it as a mere cost-cutting method.

It is crucial that algorithms demonstrate both accuracy and robustness and are also evaluated according to ethical, legal, and social criteria. Excitement about deep learning algorithms could encourage a rush to implementation and to counteract that I think scientists and radiologists need to make policy-makers and the public aware of the need for controlled studies and continued surveillance of AI algorithms.

Ethical approval for the studies was granted by the ERB in Stockholm with diarienummer (dnr) : 2016/2600-31.

9 POINTS OF PERSPECTIVE

There is a need to improve the breast cancer screening process to detect breast cancer earlier and to reduce mortality and morbidity among breast cancer patients. Until now the breast cancer screening process has been the same for all women in a screening population, despite more and more studies, of AI and of density, showing that it is possible to risk-stratify women for different screening regiments, using varying time intervals between screening rounds and with varying modalities. Mammography is a robust and proven modality for breast cancer screening, especially for women with less dense breasts, but there are other modalities that are more sensitive and that might suit some women better. The challenge is to identify which women need this kind of enhanced screening and for which women screening with only mammography is sufficient. Also, the time intervals between screening appointments might be reconsidered. To reduce the proportion of IC some women would benefit from more frequent screening. The further development of mammography with tomosynthesis and contrast-enhanced mammography as well as shorter MRI protocols (163) might make the incorporation of individualized screening easier to handle.

There are many research groups worldwide studying risk-stratified breast cancer screening in different models with incorporated AI techniques. Many promising studies have been performed in retrospective datasets, but due to inherent biases in retrospective studies, such as exclusion of study participants, random up-sampling of healthy controls, there is a need for prospective clinical studies before wider clinical implementation.

The development towards individual screening has started, and I firmly believe that AI will be an important contribution to identify women at different risks for developing breast cancer as well as to improve the efficacy for radiologists and to make the screening process more sophisticated.

I am also convinced that the algorithms will improve even more and be able to make even more precise assessments, as well as having the ability to compare actual images with priors and thus reduce the proportion of false positives which is a challenge today. There might be algorithms trained to detect more aggressive breast cancers, indolent breast cancers as well as benign lesions.

Further studies are needed to demonstrate whether AI is a robust and reliable tool that will work alongside physicians. It is also important to define a legal-ethical framework to make patients, physicians and researchers as comfortable as possible when using AI in the daily practice. We need to clarify who is responsible if AI fails in different situations and how the AI providers are responsible visavi the users of the AI systems.

SAMMANFATTNING PÅ SVENSKA (SWEDISH ABSTRACT)

Bröstcancer är den vanligaste cancerformen bland kvinnor och incidensen ökar. När populationsbaserad screeningmammografi introducerades på 1980- och 1990-talen i Sverige sjönk mortaliteten med upp till 30-40%.

Idag inbjuds alla kvinnor mellan 40-74 års ålder till screeningundersökning vartannat år (i vissa regioner varje 18 månader). Vid mammografiundersökningen tar man i regel två bilder på varje bröst i två olika projektioner, mediolateral oblique projektion och kraniokaudal projektion. Screeningprocessen är lika för alla kvinnor i Sverige.

Även om det var framgångsrikt att införa nationell bröstcancerscreening så kvarstår ändå utmaningar, till exempel förekomsten av en stor andel intervallcancrar och stora screeningupptäckta cancrar som är kopplade till en ökad mortalitet och morbiditet. Idag är den svenska screeningmodellen endast baserad på ålder och inte på andra riskfaktorer. Den enda modaliteten som används är mammografi. Det finns ett behov att förändra den svenska screeningprocessen för att ytterligare minska mortalitet och morbiditet i bröstcancer. Det är viktigt att en kvinnas bröstcancer upptäcks tidigt när den är liten och inte har spridit sig till lymfkörtlar för att ha den bästa prognosen. För att uppnå det så tror jag att screeningprocessen behöver individualiseras och bli mer flexibel avseende längden på screeningintervall och avseende vilka modaliteter som är lämpligast beroende på den individuella risken att utveckla bröstcancer och den mammografiska sensitiviteten.

Ett annat dilemma är den stora bristen på bröstradiologer i Sverige och vikten av att använda deras kompetens så effektivt som möjligt. Mest prioriterat torde bröstradiologers kompetens utnyttjas för svårbedömda fall och inte för friska kvinnor.

Hur ser lösningarna ut på dessa frågor? En utmaning är att kartlägga vilka kvinnor som har hög risk att utveckla bröstcancer. Det är också viktigt att identifiera de mammografiundersökningar som har hög respektive låg känslighet för att uppvisa tumörtecken. I studie II till IV har vi analyserat hur man med hjälp av deep learning skulle kunna adressera dessa utmaningar.

I studie I beskrevs kohorten CSAW. Ur denna kohort kommer studiepopulationerna för studie II till IV (även studiepopulationer från andra publikationer från vår forskargrupp). I CSAW ingår alla kvinnor som inbjudits till screeningundersökninng mellan 2008 och 2015 inom Region Stockholm. Vi beskrev kohorten och hur den har använts. Vi beskrev också framtida möjligheter att använda kohorten samt den separata fall-kontroll databasen med annoterade tumörer och friska kontroller. Denna studie presenterades vid RSNA 2019.

I studie II jämförde vi en AI algoritm, DLrisk med brösttäthet avseende risken att drabbas av framtida bröstcancer. Vi kom fram till att odds OR och AUC var högre för åldersjusterad DLrisk än för dense area och percentage density: 1.56; AUC, 0.65, 1.31; AUC, 0.60, och 1.18 AUC, 0.57 ($P < .001$ for AUCs). Andelen falskt negativa var även lägre för DLrisk än för dense area och percentage density; 31%, 36% och 39%. Skillnaden var störst för mer aggressiva cancrar.

I studie III analyserade vi mammografibilderna i två olika arbetsflöden. En AI algoritm bedömde förekomsten av tumörtecken i mammografibilderna. Varje mammografiundersökning fick en poäng mellan 0 till 1 där 1 representerade högst sannolikhet för tumörtecken i bilden. I det ena arbetsflödet bedömdes mammografibilderna av endast en AI-algoritm och ingen radiolog. I detta arbetsflöde kunde AI-algoritmen

bedöma 60% av mammografibilderna korrekt utan att missa någon cancer. I det andra arbetsflödet undersöktes kvinnorna med en negativ screening och de 1% respektive 5% högsta poängen avseende risk för tumörtecken i bilden med “en perfekt radiologisk undersökning”. I detta flöde kunde man hitta 24 (12%) respektive 53 (27%) intervallcancrar (av 200 senare diagnosticerade intervallcancrar) och 48 (14%) respektive 121 (35%) av 347 senare diagnosticerade screeningupptäckta cancrar.

I studie IV analyserade vi retrospektivt hur man kan välja abnormalitetspoäng i en miljö där en AI algoritm ska agera som oberoende tredje granskare av screeningmammografier i en klinisk prospektiv studie enligt två alternativ. Vi kom fram till att om man vill att en AI algoritm ska ha samma sensitivitet som en annan granskare så får man acceptera att en stor mängd undersökningar kommer att läggas för konsensusdiskussion (alternativ 1). Om man vill att AI-algoritmen ska ha samma sensitivitet som den samlade sensitiviteten av två radiologer (alternativ 2) men ändå hitta lika mycket cancer som vid dubbelgranskning så får man acceptera en lägre sensitivitet av AI algoritmen vilket innebär att en mindre mängd fall läggs till konsensusdiskussion jämfört med alternativ 1. Sensitiviteten för radiolog 1, 2 och 1+2 var 69,66%, 75,69% respektive 78,56%. Andelen fall som lades till diskussion för radiolog 1, 2 och 1+2 var 4,45%, 4,56% respektive 6,06%. Granskare 1 och AI hade tillsammans en sensitivitet på 82,42% och lade 12,63% av fallen till diskussion enligt alternativ 1. AI tillsammans med den sammanlagda sensitiviteten av granskare 1 och 2 hade en sensitivitet på 78,56% och lade 6,99% av fallen till diskussion enligt alternativ 2. Denna studie presenterades som poster vid the annual meeting of the Radiological Society of North America 2021.

Sammantaget har vi försökt att adressera en del av utmaningarna med en reformerad, individualiserad screeningprocess med deep learning.

10 ACKNOWLEDGEMENTS

Peter Lindholm, my main supervisor. For being always supportive, positive and encouraging both in periods of flow and in periods of less motivation and for being one of those who started CSAW.

Kevin Smith, my co-supervisor. For always going that extra mile after reviewers feedback and for always being supportive and motivating. For making his team working splendid in the collaborations with our research group.

Fredrik Strand, my co-supervisor. For always listening about numerous discussions about methods and study populations and for encouraging me before the exams at the research school. For making me motivated to always make an extra effort and for giving me very honest feedback to improve the work.

Martin Eklund, my co-supervisor. For being an excellent statistical expert who guides me regarding statistical methods and for supporting me regarding discussions with reviewers.

Torkel Brismar, Antonio Valachis, Brigitte Wilczek for good and constructive feedback during my half-time session, sharing a lot of experience within the field of breast cancer care and research.

Erik Wåhlin, physicist. For the never-ending patience with data collection.

Evaldas Laurencikas, senior general/child/neuroradiologist and former colleague at Danderyds Sjukhus. For inspiring me to start researching and for supporting me when I wrote my first-ever manuscript in the field of neuroradiology.

Anders Byström, head of the department of radiology at Capio Sankt Görans Sjukhus. For always being positive when it comes to research questions, for letting me to go the research school and for always finding solutions regarding research collaborations and time issues.

Hossein Azizpour, assistant professor. For always listening thoroughly and for coming with smart ideas.

Yue Liu, fellow PhD student. For always contributing to the technical parts of the studies and for the patience with data curation.

Mattie Salim, colleague and fellow PhD student. For always taking your time and for all support during nervous presentations and for all memories!

Sophie Norenstedt, breast surgeon and mentor. For always listening and giving advice, and for your kindness and happiness.

Astrid Rocchi, senior breast radiologist at the department of breast radiology at Capio Sankt Görans Sjukhus. For always supporting me and for always solving issues with planning and for giving me time for research. For being an excellent and loyal colleague and a smashing line-dancer!

Marina Janicijevic, senior breast radiologist, close friend and colleague. For being the complete person and the complete breast radiologist! You are my role model in life. The world would be perfect with Marinas!

Brigitte Wilczek, senior (French) breast radiologist. For always being supportive, encouraging, humoristic and an excellent colleague and friend. For always sending us wonderful videos and for your caring generosity.

Johanna Swärd, one of my best friends and colleague at the breast radiology department, Capio Sankt Görans Sjukhus. For always being encouraging, positive, humoristic and sharing your life with me. For sharing the same gut-feeling as me and for being just you!

Ingrid Bråkenhielm, one of my best friends, study buddy at KI and colleague at the department of breast radiology, Capio Sankt Görans Sjukhus. For all our wonderful and crazy memories and for your never-ending loyalty and big heart.

Kjell Hågemo, senior breast radiologist at the department of breast radiology, Capio Sankt Görans Sjukhus. For your expertise and willingness to help and discuss breast cancer cases. And for sharing your travel tips, your cat stories and advices regarding horse-riding!

Karin Thorneman, senior breast radiologist at the department of breast radiology, Capio Sankt Görans Sjukhus. For always being supportive and for always listening. For your kindness and for sharing your artistical skills with us.

Maria Balarova, senior breast radiologist at the department of breast radiology, Capio Sankt Görans Sjukhus. For always being friendly and supportive and for all support during our intense summer periods.

Edith Herterich, senior breast radiologist at the department of breast radiology, Capio Sankt Görans Sjukhus. For always being positive and for your willingness to always discuss breast cancer cases. For sharing your life and thoughts and for being, humoristic, understanding and encouraging.

Anca Plotoaga, senior breast radiologist at the department of breast radiology, Capio Sankt Görans Sjukhus. For being that very sharp breast radiologist. For always being kind, positive and giving advice in life and at work and for giving me your perfect sense of humour!

All colleagues, doctors and nurses, at the Breast Radiology department at Capio Sankt Görans Sjukhus - you are my extended extended family! For always making me motivated to go to work and for all the small chats and joyful and important everyday moments we share.

Edward Azavedo, senior breast radiologist, pathologist and associate professor. For always being positive and supportive and for the good collaboration during research sessions and courses.

Jonathan Waldenström, IT-manager at the Radiology department Capio Sankt Görans Sjukhus. For always helping me instantly when it comes to questions regarding algorithms and IT-solutions.

Bröstcancerförbundet, for always being encouraging and interested in my work and in breast radiology.

All colleagues at Capio Sankt Görans Sjukhus Breast Center for great support and always encouraging me to do my very best in the daily job.

Laura Juskaite, close friend and my daughters´ best friend. For your big heart and loyalty. Without you, no life and no career.

Karin Sandstedt, my oldest best friend. For your never-ending patience, understanding, humor and support during tough periods and for all memories the latest 40 years!

Anna Gunnerbeck, one of my best friends and study buddy at KI. For your understanding and for your gut-feeling and always correct analyses. For your support during tough periods and for sharing all crazy and memorable memories!

Ebba Swedenhammar, close “new” friend and breast surgeon at Capio Sankt Görans Breast Center. For your wonderful sense of humour, for your mindset and for always listening and giving honest advices! And for our future Megaformer-classes!

Charlotta Flodström, Laila Hellkvist, Margit Anell, Lisa Lenerius, Susanna Andersin, Elisabeth Hallan, Louise de la Gardie, Anna Janse, Kristina Lind, Åsa Eckerbom, Kaisa Ahopelto, Anna Malmfors, Anna Niciolaysen, Åsa Winge, Sofie Bonde, Erika Bartholdson, close friends. For being such nice, interesting, funny, understanding friends and for all memories!

Inger, Leif, Lennart, Inger, Petra, Isa and Linnea, my husband’s parents, sister and her children. For being always loving, friendly and welcoming!

Mum and Dad. For always supporting and encouraging me to study and for always giving me the best opportunities in life. For motivating me to always do the little extra. And mum, for always listening and sharing your life experience with me.

Maria and David Dembrower, sister and brother. For being supportive and positive and for sharing all memories with me during life! And Maria, for all our memories after moving to Stockholm!

Teo, Ingrid, the perfect bonus children. For all wonderful memories!

Fredrik, my husband and the love in my life! For everything and for all the joy and fire! I love you.

Gustaf, Oscar, Lovisa, Charlotta, my never ending beloved children.

11 REFERENCES

1. Tabár L, Dean PB, Chen THH, Yen AMF, Chen SLS, Fann JCY, et al. The incidence of fatal breast cancer measures the increased effectiveness of therapy in women participating in mammography screening. Cancer. 2019;125(4):515-23.
2. Duffy SW, Tabár L, Yen AM-F, Dean PB, Smith RA, Jonsson H, et al. Beneficial effect of consecutive screening mammography examinations on mortality from breast cancer: a prospective study. Radiology. 2021;299(3):541-7.
3. Seely J, Alhassan T. Screening for breast cancer in 2018—what should we be doing today? Current Oncology. 2018;25(s1):115-24.
4. Warner E, Messersmith H, Causer P, Eisen A, Shumak R, Plewes D. Systematic review: using magnetic resonance imaging to screen women at high risk for breast cancer. Annals of internal medicine. 2008;148(9):671-9.
5. Gail MH. Personalized estimates of breast cancer risk in clinical practice and public health. Statistics in medicine. 2011;30(10):1090-104.
6. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Statistics in medicine. 2004;23(7):1111-30.
7. Brentnall AR, Cuzick J, Buist DS, Bowles EJA. Long-term accuracy of breast cancer risk assessment combining classic risk factors and breast density. JAMA oncology. 2018;4(9):e180174-e.
8. Shore AN, Rosen JM. Regulation of mammary epithelial cell homeostasis by lncRNAs. The international journal of biochemistry & cell biology. 2014;54:318-30.
9. Geddes DT. Inside the lactating breast: the latest anatomy research. Journal of midwifery & women's health. 2007;52(6):556-63.
10. Russo J, Russo IH. Development of the human breast. Maturitas. 2004;49(1):2-15.
11. Gusterson BA, Stein T, editors. Human breast development. Seminars in cell & developmental biology; 2012: Elsevier.
12. Harris JR, Lippman ME, Osborne CK, Morrow M. Diseases of the Breast: Lippincott Williams & Wilkins; 2012.
13. Zaidi Z, Dib HA. The worldwide female breast cancer incidence and survival, 2018. AACR; 2019.
14. Cancerfonden. 2020 [https://www.cancerfonden.se/om-cancer/cancersjukdomar/brostcancer]
15. Hartman J, Ehinger A, Kovacs A, Olofsson H, Colon-Cervantes E, Stemme S, et al. Kvalitetsbilaga för bröstpatologi (KVAST-bilaga): Tillhörande Nationellt vårdprogram för bröstcancer. 2020.
16. Ellis I, Galea M, Broughton N, Locker A, Blamey R, Elston C. Pathological prognostic factors in breast cancer. II. Histological type. Relationship with survival in a large study with long-term follow-up. Histopathology. 1992;20(6):479-89.
17. Saadatmand S, Bretveld R, Siesling S, Tilanus-Linthorst MM. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients. Bmj. 2015;351.
18. DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, et al. Breast cancer statistics, 2019. CA: a cancer journal for clinicians. 2019;69(6):438-51.
19. Wittekind C, Asamura H, Sobin LH. TNM atlas: John Wiley & Sons; 2014.
20. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology. 1991;19(5):403-10.

21. Dai X, Xiang L, Li T, Bai Z. Cancer hallmarks, biomarkers and breast cancer molecular subtypes. Journal of Cancer. 2016;7(10):1281.
22. Speirs V, Carder PJ, Lane S, Dodwell D, Lansdown MR, Hanby AM. Oestrogen receptor β: what it means for patients with breast cancer. The lancet oncology. 2004;5(3):174-81.
23. Saji S, Hirose M, Toi M. Clinical significance of estrogen receptor β in breast cancer. Cancer chemotherapy and pharmacology. 2005;56(1):21-6.
24. Scherbakov A, Krasil'nikov M, Kushlinskii N. Molecular mechanisms of hormone resistance of breast cancer. Bulletin of experimental biology and medicine. 2013;155(3):384-95.
25. Banin Hirata BK, Oda JMM, Losi Guembarovski R, Ariza CB, Oliveira CECd, Watanabe MAE. Molecular markers for breast cancer: prediction on tumor behavior. Disease markers. 2014;2014.
26. Barnes D, Hanby A. Oestrogen and progesterone receptors in breast cancer: past, present and future. Histopathology. 2001;38(3):271-4.
27. Osborne CK, Yochmowitz MG, Knight III WA, McGuire WL. The value of estrogen and progesterone receptors in the treatment of breast cancer. Cancer. 1980;46(S12):2884-8.
28. Group EBCTC. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. The Lancet. 2005;365(9472):1687-717.
29. Iqbal N, Iqbal N. Human epidermal growth factor receptor 2 (HER2) in cancers: overexpression and therapeutic implications. Molecular biology international. 2014;2014.
30. Burstein HJ. The distinctive nature of HER2-positive breast cancers. New England Journal of Medicine. 2005;353(16):1652-4.
31. Kiyose S, Igarashi H, Nagura K, Kamo T, Kawane K, Mori H, et al. Chromogenic in situ hybridization (CISH) to detect HER2 gene amplification in breast and gastric cancer: comparison with immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH). Pathology international. 2012;62(11):728-34.
32. Soliman NA, Yussif SM. Ki-67 as a prognostic marker according to breast cancer molecular subtype. Cancer biology & medicine. 2016;13(4):496.
33. De Azambuja E, Cardoso F, de Castro G, Colozza M, Mano MS, Durbecq V, et al. Ki-67 as prognostic marker in early breast cancer: a meta-analysis of published studies involving 12 155 patients. British journal of cancer. 2007;96(10):1504-13.
34. Yerushalmi R, Woods R, Ravdin PM, Hayes MM, Gelmon KA. Ki67 in breast cancer: prognostic and predictive potential. The lancet oncology. 2010;11(2):174-83.
35. Von Minckwitz G, Schmitt WD, Loibl S, Müller BM, Blohmer JU, Sinn BV, et al. Ki67 measured after neoadjuvant chemotherapy for primary breast cancer. Clinical Cancer Research. 2013;19(16):4521-31.
36. Prat A, Pineda E, Adamo B, Galván P, Fernández A, Gaba L, et al. Clinical implications of the intrinsic molecular subtypes of breast cancer. The Breast. 2015;24:S26-S35.
37. Hon JDC, Singh B, Sahin A, Du G, Wang J, Wang VY, et al. Breast cancer molecular subtypes: from TNBC to QNBC. American journal of cancer research. 2016;6(9):1864.
38. Ward EM, DeSantis CE, Lin CC, Kramer JL, Jemal A, Kohler B, et al. Cancer statistics: breast cancer in situ. CA: a cancer journal for clinicians. 2015;65(6):481-95.

39. Sinn H-P, Kreipe H. A brief overview of the WHO classification of breast tumors. Breast care. 2013;8(2):149-54.
40. Tabár L, Dean PB, Lee Tucker F, Yen AM-F, Chen SL-S, Jen GHH, et al. A new approach to breast cancer terminology based on the anatomic site of tumour origin: The importance of radiologic imaging biomarkers. European Journal of Radiology. 2022;149:110189.
41. Tabár L, Dean PB, Chen TH-H, Yen AM-F, Chiu SY-H, Tot T, et al. The impact of mammography screening on the diagnosis and management of early-phase breast cancer. Breast Cancer. 2014:31-78.
42. Robertson FM, Bondy M, Yang W, Yamauchi H, Wiggins S, Kamrudin S, et al. Inflammatory breast cancer: the disease, the biology, the treatment. CA: a cancer journal for clinicians. 2010;60(6):351-75.
43. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, et al. Mammographic density and the risk and detection of breast cancer. New England Journal of Medicine. 2007;356(3):227-36.
44. Pollán M, Ascunce N, Ederra M, Murillo A, Erdozáin N, Alés-Martínez JE, et al. Mammographic density and risk of breast cancer according to tumor characteristics and mode of detection: a Spanish population-based case-control study. Breast Cancer Research. 2013;15(1):R9.
45. Rice MS, Bertrand KA, VanderWeele TJ, Rosner BA, Liao X, Adami H-O, et al. Mammographic density and breast cancer risk: a mediation analysis. Breast Cancer Research. 2016;18(1):94.
46. Kerlikowske K, Zhu W, Tosteson AN, Sprague BL, Tice JA, Lehman CD, et al. Identifying women with dense breasts at high risk for interval cancer: a cohort study. Annals of internal medicine. 2015;162(10):673-81.
47. Sun Y-S, Zhao Z, Yang Z-N, Xu F, Lu H-J, Zhu Z-Y, et al. Risk factors and preventions of breast cancer. International journal of biological sciences. 2017;13(11):1387.
48. Campeau PM, Foulkes WD, Tischkowitz MD. Hereditary breast cancer: new genetic developments, new therapeutic avenues. Human genetics. 2008;124(1):31-42.
49. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, et al. Gene-expression profiles in hereditary breast cancer. New England Journal of Medicine. 2001;344(8):539-48.
50. Krammer J, Pinker-Domenig K, Robson ME, Gönen M, Bernard-Davila B, Morris EA, et al. Breast cancer detection and tumor characteristics in BRCA1 and BRCA2 mutation carriers. Breast cancer research and treatment. 2017;163(3):565-71.
51. Lambertini M, Santoro L, Del Mastro L, Nguyen B, Livraghi L, Ugolini D, et al. Reproductive behaviors and risk of developing breast cancer according to tumor subtype: a systematic review and meta-analysis of epidemiological studies. Cancer treatment reviews. 2016;49:65-76.
52. Bernstein L, Ross RK. Endogenous hormones and breast cancer risk. Epidemiologic reviews. 1993;15(1):48-65.
53. Jung S, Wang M, Anderson K, Baglietto L, Bergkvist L, Bernstein L, et al. Alcohol consumption and breast cancer risk by estrogen receptor status: in a pooled analysis of 20 studies. International journal of epidemiology. 2016;45(3):916-28.
54. Kotepui M. Diet and risk of breast cancer. Contemporary Oncology. 2016;20(1):13.

55. Meretoja T, Rasia S, Von Smitten K, Asko-Seljavaara S, Kuokkanen H, Jahkola T. Late results of skin-sparing mastectomy followed by immediate breast reconstruction. Journal of British Surgery. 2007;94(10):1220-5.
56. Lanitis S, Tekkis PP, Sgourakis G, Dimopoulos N, Al Mufti R, Hadjiminas DJ. Comparison of skin-sparing mastectomy versus non–skin-sparing mastectomy for breast cancer: a meta-analysis of observational studies. LWW; 2010.
57. Moo TA, Pinchinat T, Mays S, Landers A, Christos P, Alabdulkareem H, et al. Oncologic outcomes after nipple-sparing mastectomy. Annals of surgical oncology. 2016;23(10):3221-5.
58. Jagsi R, Abi Raad R, Goldberg S, Sullivan T, Michaelson J, Powell SN, et al. Locoregional recurrence rates and prognostic factors for failure in node-negative patients treated with mastectomy: implications for postmastectomy radiation. International Journal of Radiation Oncology* Biology* Physics. 2005;62(4):1035-9.
59. Moran MS, Schnitt SJ, Giuliano AE, Harris JR, Khan SA, Horton J, et al. Society of Surgical Oncology–American Society for Radiation Oncology consensus guideline on margins for breast-conserving surgery with whole-breast irradiation in stages I and II invasive breast cancer. International Journal of Radiation Oncology* Biology* Physics. 2014;88(3):553-64.
60. Litière S, Werutsky G, Fentiman IS, Rutgers E, Christiaens M-R, Van Limbergen E, et al. Breast conserving therapy versus mastectomy for stage I–II breast cancer: 20 year follow-up of the EORTC 10801 phase 3 randomised trial. The lancet oncology. 2012;13(4):412-9.
61. Untch M, Gerber B, Harbeck N, Jackisch C, Marschner N, Möbus V, et al. 13th st. Gallen international breast cancer conference 2013: primary therapy of early breast cancer evidence, controversies, consensus-opinion of a german team of experts (zurich 2013). Breast care. 2013;8(3):221-9.
62. Tamburelli F, Maggiorotto F, Marchiò C, Balmativola D, Magistris A, Kubatzki F, et al. Reoperation rate after breast conserving surgery as quality indicator in breast cancer treatment: A reappraisal. The Breast. 2020;53:181-8.
63. Cocquyt VF, Blondeel PN, Depypere HT, Van De Sijpe KA, Daems KK, Monstrey SJ, et al. Better cosmetic results and comparable quality of life after skin-sparing mastectomy and immediate autologous breast reconstruction compared to breast conservative treatment. British journal of plastic surgery. 2003;56(5):462-70.
64. Clough KB, Lewis JS, Couturaud B, Fitoussi A, Nos C, Falcou M-C. Oncoplastic techniques allow extensive resections for breast-conserving therapy of breast carcinomas. Annals of surgery. 2003;237(1):26.
65. Reintgen D, Giuliano R, Cox CE. Sentinel node biopsy in breast cancer: an overview. The breast journal. 2000;6(5):299-305.
66. Andersson Y, de Boniface J, Jönsson P, Ingvar C, Liljegren G, Bergkvist L, et al. Axillary recurrence rate 5 years after negative sentinel node biopsy for breast cancer. Journal of British Surgery. 2012;99(2):226-31.
67. Sackey H, Magnuson A, Sandelin K, Liljegren G, Bergkvist L, Fülep Z, et al. Arm lymphoedema after axillary surgery in women with invasive breast cancer. Journal of British Surgery. 2014;101(4):390-7.
68. de Boniface J, Frisell J, Andersson Y, Bergkvist L, Ahlgren J, Rydén L, et al. Survival and axillary recurrence following sentinel node-positive breast cancer without completion axillary lymph node dissection: the randomized controlled SENOMAC trial. Bmc Cancer. 2017;17(1):1-7.

69. Holland R, Veling SH, Mravunac M, Hendriks JH. Histologic multifocality of tis, T1–2 breast carcinomas implications for clinical trials of breast-conserving surgery. Cancer. 1985;56(5):979-90.
70. Fisher B, Anderson S, Bryant J, Margolese RG, Deutsch M, Fisher ER, et al. Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. New England Journal of Medicine. 2002;347(16):1233-41.
71. Darby S, McGale P, Correa C, Taylor C, Arriagada R, Clarke M, et al. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10,801 women in 17 randomised trials. Lancet (London, England). 2011;378(9804):1707-16.
72. Group EBCTC. Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10 801 women in 17 randomised trials. The Lancet. 2011;378(9804):1707-16.
73. Anampa J, Makower D, Sparano JA. Progress in adjuvant chemotherapy for breast cancer: an overview. BMC medicine. 2015;13(1):1-13.
74. Heys SD, Hutcheon AW, Sarkar TK, Ogston KN, Miller ID, Payne S, et al. Neoadjuvant docetaxel in breast cancer: 3-year survival results from the Aberdeen trial. Clinical breast cancer. 2002;3:S69-S74.
75. Giordano SH. Update on locally advanced breast cancer. The oncologist. 2003;8(6):521-30.
76. Wang M, Hou L, Chen M, Zhou Y, Liang Y, Wang S, et al. Neoadjuvant chemotherapy creates surgery opportunities for inoperable locally advanced breast cancer. Scientific reports. 2017;7(1):1-7.
77. Mieog J, Van der Hage J, Van De Velde C. Neoadjuvant chemotherapy for operable breast cancer. Journal of British Surgery. 2007;94(10):1189-200.
78. Lindman H. Bröstcancer, primärbehandling. Internetmedicin se Hämtad den från. 2019.
79. Davies C, Pan H, Godwin J, Gray R, Arriagada R, Raina V, et al. Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of oestrogen receptor-positive breast cancer: ATLAS, a randomised trial. The Lancet. 2013;381(9869):805-16.
80. Group EBCTC. Aromatase inhibitors versus tamoxifen in early breast cancer: patient-level meta-analysis of the randomised trials. The Lancet. 2015;386(10001):1341-52.
81. Asif HM, Sultana S, Ahmed S, Akhtar N, Tariq M. HER-2 positive breast cancer-a mini-review. Asian Pacific Journal of Cancer Prevention. 2016;17(4):1609-15.
82. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. New England journal of medicine. 2001;344(11):783-92.
83. Fiszman GL, Jasnis MA. Molecular mechanisms of trastuzumab resistance in HER2 overexpressing breast cancer. International journal of breast cancer. 2011;2011.
84. Sparano JA, editor Cardiac toxicity of trastuzumab (Herceptin): implications for the design of adjuvant trials. Seminars in oncology; 2001: Elsevier.
85. Eberl MM, Fox CH, Edge SB, Carter CA, Mahoney MC. BI-RADS classification for management of abnormal mammograms. The Journal of the American Board of Family Medicine. 2006;19(2):161-4.

86. Liberman L, Menell JH. Breast imaging reporting and data system (BI-RADS). Radiologic Clinics. 2002;40(3):409-30.
87. Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V. Effect of age, breast density, and family history on the sensitivity of first screening mammography. Jama. 1996;276(1):33-8.
88. Gram IT, Funkhouser E, Tabár L. The Tabar classification of mammographic parenchymal patterns. European journal of radiology. 1997;24(2):131-6.
89. Wolfe JN. Breast parenchymal patterns and their changes with age. Radiology. 1976;121(3):545-52.
90. Berg WA, Blume JD, Cormack JB, Mendelson EB, Lehrer D, Böhm-Vélez M, et al. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. Jama. 2008;299(18):2151-63.
91. Kolb TM, Lichy J, Newhouse JH. Occult cancer in women with dense breasts: detection with screening US--diagnostic yield and tumor characteristics. Radiology. 1998;207(1):191-9.
92. Zonderland HM, Coerkamp EG, Hermans J, van de Vijver MJ, van Voorthuisen AE. Diagnosis of breast cancer: contribution of US as an adjunct to mammography. Radiology. 1999;213(2):413-22.
93. Lehman CD, Smith RA. The role of MRI in breast cancer screening. Journal of the National Comprehensive Cancer Network. 2009;7(10):1109-15.
94. Veenhuizen SG, de Lange SV, Bakker MF, Pijnappel RM, Mann RM, Monninkhof EM, et al. Supplemental breast MRI for women with extremely dense breasts: results of the second screening round of the DENSE trial. Radiology. 2021;299(2):278-86.
95. Chong A, Weinstein SP, McDonald ES, Conant EF. Digital breast tomosynthesis: concepts and clinical practice. Radiology. 2019;292(1):1-14.
96. Lång K, Andersson I, Rosso A, Tingberg A, Timberg P, Zackrisson S. Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the Malmö Breast Tomosynthesis Screening Trial, a population-based study. European radiology. 2016;26(1):184-90.
97. Conant EF, Zuckerman SP, McDonald ES, Weinstein SP, Korhonen KE, Birnbaum JA, et al. Five consecutive years of screening with digital breast tomosynthesis: outcomes by screening year and round. Radiology. 2020;295(2):285-93.
98. Conant EF, Barlow WE, Herschorn SD, Weaver DL, Beaber EF, Tosteson AN, et al. Association of digital breast tomosynthesis vs digital mammography with cancer detection and recall rates by age and breast density. JAMA oncology. 2019;5(5):635-42.
99. Marmot MG, Altman D, Cameron D, Dewar J, Thompson S, Wilcox M. The benefits and harms of breast cancer screening: an independent review. British journal of cancer. 2013;108(11):2205-40.
100. Tabar L, Fagerberg G, Chen HH, Duffy SW, Smart CR, Gad A, et al. Efficacy of breast cancer screening by age. New results swedish two-county trial. Cancer. 1995;75(10):2507-17.
101. Törnberg S, Kemetli L, Ascunce N, Hofvind S, Anttila A, Seradour B, et al. A pooled analysis of interval cancer rates in six European countries. European journal of cancer prevention. 2010;19(2):87-93.
102. Van Luijt P, Heijnsdijk E, Fracheboud J, Overbeek L, Broeders M, Wesseling J, et al. The distribution of ductal carcinoma in situ (DCIS) grade in 4232 women and its impact on overdiagnosis in breast cancer screening. Breast Cancer Research. 2016;18(1):1-10.

103. Yen M-F, Tabar L, Vitak B, Smith R, Chen H-H, Duffy S. Quantifying the potential problem of overdiagnosis of ductal carcinoma in situ in breast cancer screening. European journal of cancer. 2003;39(12):1746-54.
104. Youlden DR, Cramb SM, Dunn NA, Muller JM, Pyke CM, Baade PD. The descriptive epidemiology of female breast cancer: an international comparison of screening, incidence, survival and mortality. Cancer epidemiology. 2012;36(3):237-48.
105. Lind H, Svane G, Kemetli L, Törnberg S. Breast Cancer Screening Program in Stockholm County, Sweden–Aspects of Organization and Quality Assurance. Breast Care. 2010;5(5):353-7.
106. Lidbrink EK, Törnberg SA, Azavedo EM, Frisell JO, Hjalmar M-L, Leifland KS, et al. The general mammography screening program in Stockholm: organisation and first-round results. Acta Oncologica. 1994;33(4):353-8.
107. Baines CJ, Miller A, Wall C, McFarlane D, Simor I, Jong R, et al. Sensitivity and specificity of first screen mammography in the Canadian National Breast Screening Study: a preliminary report from five centers. Radiology. 1986;160(2):295-8.
108. Houssami N, Macaskill P, Bernardi D, Caumo F, Pellegrini M, Brunelli S, et al. Breast screening using 2D-mammography or integrating digital breast tomosynthesis (3D-mammography) for single-reading or double-reading–evidence to guide future screening strategies. European Journal of Cancer. 2014;50(10):1799-807.
109. Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. NPJ Breast Cancer. 2017;3(1):1-13.
110. Shah TA, Guraya SS. Breast cancer screening programs: Review of merits, demerits, and recent recommendations practiced across the world. Journal of microscopy and ultrastructure. 2017;5(2):59-69.
111. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.
112. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A-r, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine. 2012;29(6):82-97.
113. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25:1097-105.
114. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. Journal of machine learning research. 2011;12(ARTICLE):2493− 537.
115. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al., editors. Tensorflow: A system for large-scale machine learning. 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16); 2016.
116. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. Scientific reports. 2018;8(1):1-7.
117. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA internal medicine. 2015;175(11):1828-37.
118. Brem RF, Baum J, Lechner M, Kaplan S, Souders S, Naul LG, et al. Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial. American Journal of Roentgenology. 2003;181(3):687-93.
119. Morton MJ, Whaley DH, Brandt KR, Amrami KK. Screening mammograms: interpretation with computer-aided detection—prospective evaluation. Radiology. 2006;239(2):375-83.

120. Fenton JJ, Abraham L, Taplin SH, Geller BM, Carney PA, D'Orsi C, et al. Effectiveness of computer-aided detection in community mammography practice. Journal of the National Cancer institute. 2011;103(15):1152-61.
121. He K, Zhang X, Ren S, Sun J, editors. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision; 2015.
122. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. JNCI: Journal of the National Cancer Institute. 2019;111(9):916-22.
123. Schaffter T, Buist DS, Lee CI, Nikulin Y, Ribli D, Guan Y, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA network open. 2020;3(3):e200265-e.
124. Salim M, Wåhlin E, Dembrower K, Azavedo E, Foukakis T, Liu Y, et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. JAMA oncology. 2020.
125. Dustler M. Evaluating AI in breast cancer screening: a complex task. The Lancet Digital Health. 2020;2(3):e106-e7.
126. Chockley K, Emanuel E. The end of radiology? Three threats to the future practice of radiology. Journal of the American College of Radiology. 2016;13(12):1415-20.
127. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. The New England journal of medicine. 2016;375(13):1216.
128. Stephens K. Recent Studies Examine Lunit AI in Breast Cancer Detection. AXIS Imaging News. 2020.
129. Sasaki M, Tozaki M, Rodríguez-Ruiz A, Yotsumoto D, Ichiki Y, Terawaki A, et al. Artificial intelligence for breast cancer detection in mammography: experience of use of the ScreenPoint Medical Transpara system in 310 Japanese women. Breast Cancer. 2020:1-10.
130. Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. bmj. 2021;374.
131. Gulland A. Staff shortages are putting UK breast cancer screening "at risk," survey finds. British Medical Journal Publishing Group; 2016.
132. Salim M, Dembrower K, Eklund M, Lindholm P, Strand F. Range of Radiologist Performance in a Population-based Screening Cohort of 1 Million Digital Mammography Examinations. Radiology. 2020:192212.
133. Lång K, Dustler M, Dahlblom V, Åkesson A, Andersson I, Zackrisson S. Identifying normal mammograms in a large screening population using artificial intelligence. European radiology. 2020:1-6.
134. Dembrower K, Wåhlin E, Liu Y, Salim M, Smith K, Lindholm P, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. The Lancet Digital Health. 2020;2(9):e468-e74.
135. Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A deep learning model to triage screening mammograms: a simulation study. Radiology. 2019;293(1):38-46.
136. Kim H-E, Kim HH, Han B-K, Kim KH, Han K, Nam H, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. The Lancet Digital Health. 2020;2(3):e138-e48.

137. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94.
138. Cummings SR, Tice JA, Bauer S, Browner WS, Cuzick J, Ziv E, et al. Prevention of breast cancer in postmenopausal women: approaches to estimating and reducing risk. Journal of the National Cancer Institute. 2009;101(6):384-98.
139. Highnam R, Brady M, Yaffe MJ, Karssemeijer N, Harvey J, editors. Robust breast composition measurement-Volpara TM. International workshop on digital mammography; 2010: Springer.
140. Keller BM, Chen J, Daye D, Conant EF, Kontos D. Preliminary evaluation of the publicly available Laboratory for Breast Radiodensity Assessment (LIBRA) software tool: comparison of fully automated area and volumetric density measures in a case–control study with digital mammography. Breast Cancer Research. 2015;17(1):1-17.
141. Tan M, Zheng B, Leader JK, Gur D. Association between changes in mammographic image features and risk for near-term breast cancer development. IEEE transactions on medical imaging. 2016;35(7):1719-28.
142. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology. 2019;292(1):60-6.
143. Kallenberg M, Petersen K, Nielsen M, Ng AY, Diao P, Igel C, et al. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. IEEE transactions on medical imaging. 2016;35(5):1322-31.
144. Ha R, Chang P, Karcich J, Mutasa S, Van Sant EP, Liu MZ, et al. Convolutional neural network based breast cancer risk stratification using a mammographic dataset. Academic radiology. 2019;26(4):544-9.
145. Zhu X, Wolfgruber TK, Leong L, Jensen M, Scott C, Winham S, et al. Deep Learning Predicts Interval and Screening-detected Cancer from Screening Mammograms: A Case-Case-Control Study in 6369 Women. Radiology.0(0):203758.
146. Lotter W, Diab AR, Haslam B, Kim JG, Grisot G, Wu E, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nature Medicine. 2021;27(2):244-9.
147. Mattsson B, Wallgren A. Completeness of the Swedish cancer register non-notified cancer cases recorded on death certificates in 1978. Acta Radiologica: Oncology. 1984;23(5):305-13.
148. Keller BM, Chen J, Daye D, Conant EF, Kontos D. Preliminary evaluation of the publicly available Laboratory for Breast Radiodensity Assessment (LIBRA) software tool: comparison of fully automated area and volumetric density measures in a case-control study with digital mammography. Breast cancer research : BCR. 2015;17(1465-542X (Electronic)):117.
149. Ling CX, Li C, editors. Data mining for direct marketing: Problems and solutions. Kdd; 1998.
150. McPherson K, Steel C, Dixon J. ABC of breast diseases: breast cancer—epidemiology, risk factors, and genetics. BMJ: British Medical Journal. 2000;321(7261):624.
151. Barlow L, Westergren K, Holmberg L, Talbäck M. The completeness of the Swedish Cancer Register–a sample survey for year 1998. Acta oncologica. 2009;48(1):27-33.

152. Heath M, Bowyer K, Kopans D, Kegelmeyer P, Moore R, Chang K, et al. Current status of the digital database for screening mammography. Digital mammography: Springer; 1998. p. 457-60.
153. Eriksson M, Czene K, Pawitan Y, Leifland K, Darabi H, Hall P. A clinical model for identifying the short-term risk of breast cancer. Breast Cancer Research. 2017;19(1):1-8.
154. Ding J, Warren R, Girling A, Thompson D, Easton D. Mammographic density, estrogen receptor status and other breast cancer tumor characteristics. The breast journal. 2010;16(3):279-89.
155. Jonmarker O, Strand F, Brandberg Y, Lindholm P. The future of breast cancer screening: what do participants in a breast cancer screening program think about automation using artificial intelligence? Acta radiologica open. 2019;8(12):2058460119880315.
156. Boyd NF, Lockwood GA, Byng JW, Tritchler DL, Yaffe MJ. Mammographic densities and breast cancer risk. Cancer Epidemiology and Prevention Biomarkers. 1998;7(12):1133-44.
157. Pollán M, Ascunce N, Ederra M, Murillo A, Erdozáin N, Alés-Martínez JE, et al. Mammographic density and risk of breast cancer according to tumor characteristics and mode of detection: a Spanish population-based case-control study. Breast Cancer Research. 2013;15(1):1-11.
158. Freer PE. Mammographic breast density: impact on breast cancer risk and implications for screening. Radiographics. 2015;35(2):302-15.
159. Lehman CD, Arao RF, Sprague BL, Lee JM, Buist DS, Kerlikowske K, et al. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. Radiology. 2017;283(1):49-58.
160. Salim M, Wåhlin E, Dembrower K, Azavedo E, Foukakis T, Liu Y, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA oncology. 2020;6(10):1581-8.
161. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94.
162. Rodríguez-Ruiz A, Krupinski E, Mordang J-J, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology. 2019;290(2):305-14.
163. Leithner D, Moy L, Morris EA, Marino MA, Helbich TH, Pinker K. Abbreviated MRI of the breast: does it provide value? Journal of Magnetic Resonance Imaging. 2019;49(7):e85-e100.