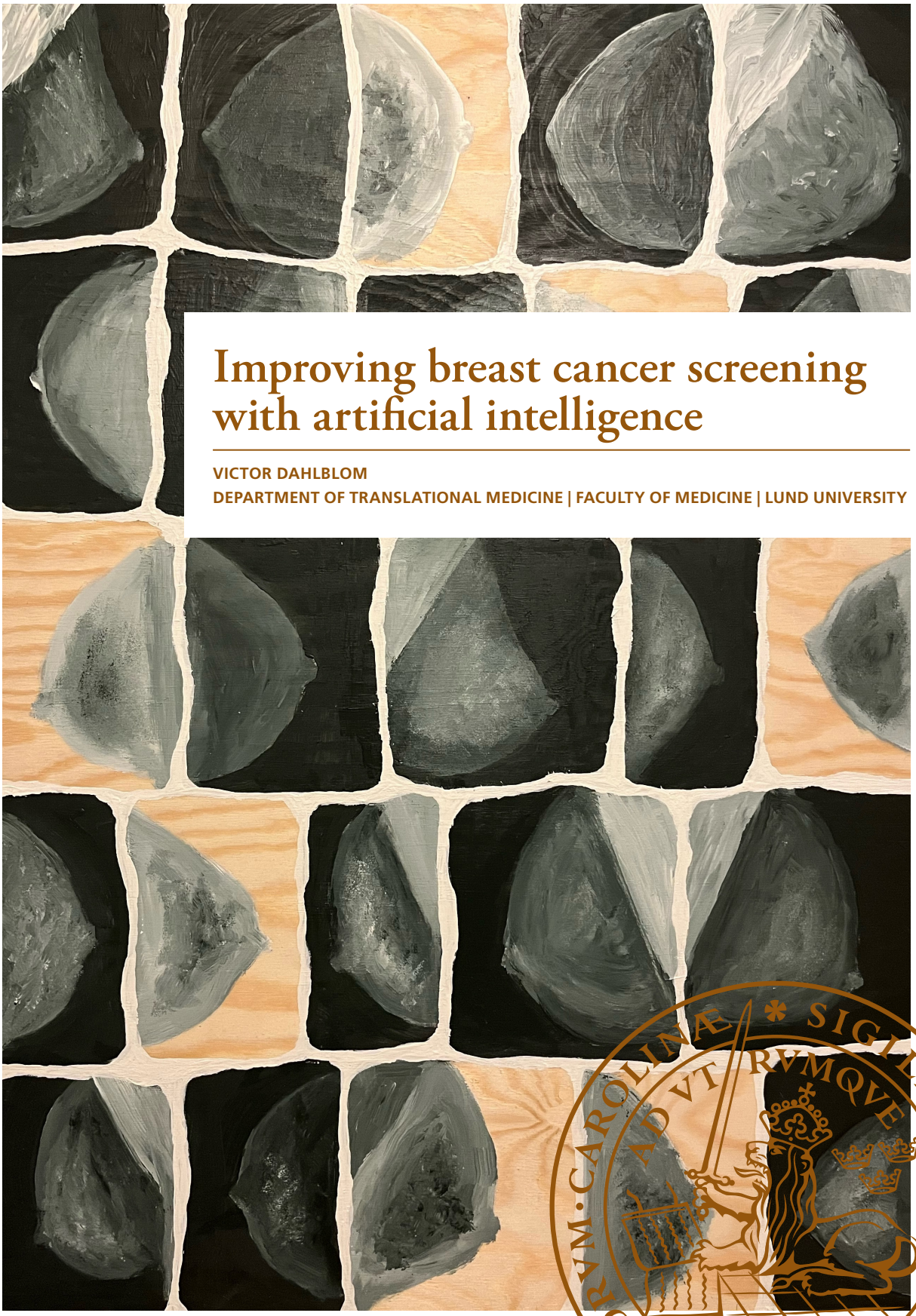
Improving breast cancer screening with artificial intelligence

Dahlblom, Victor

2024

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



Improving breast cancer screening with artificial intelligence

VICTOR DAHLBLOM
DEPARTMENT OF TRANSLATIONAL MEDICINE | FACULTY OF MEDICINE | LUND UNIVERSITY

Improving breast cancer screening with artificial intelligence

Improving breast cancer screening with artificial intelligence

Victor Dahlblom

LUND
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty of Medicine, Lund University, Sweden.
To be publicly defended on the 5th of April 2024 at 9.00 at the
Department of Radiology and Physiology, Room 2005/2007,
Skåne University Hospital, Malmö

Faculty opponent
Professor Matthias Dietzel
Universitätsklinikum Erlangen, Germany

Organisation: LUND UNIVERSITY

Document name: Doctoral Dissertation **Date of issue:** 2024-04-05

Author: Victor Dahlblom **Sponsoring organisation:**

Title and subtitle: Improving breast cancer screening with artificial intelligence

Abstract

Introduction: The current standard method for breast cancer screening is digital mammography (DM). Digital breast tomosynthesis (DBT) can detect more cancers but is more resource-demanding, not the least due to a more time-consuming reading, which hinders the implementation in screening. Artificial intelligence (AI) might open possibilities to overcome this, but different potential ways of using AI need to be tested using representative screening data. To facilitate the testing and further development of AI, it is necessary to collect and organise more data in a research-friendly form.

Aim: To create a breast imaging research database and explore different ways of using AI to improve breast cancer screening.

Methods: All DM and DBT examinations performed in Malmö, Sweden during 2004–2020 were collected and combined with other relevant information in a research database. A subset consisting of 14 848 women had been examined with paired DM and DBT as part of the Malmö Breast Tomosynthesis Screening Trial (MBTST). This cohort was used to test different ways of using an AI cancer-detection system, which scores examinations based on cancer risk. It was studied whether the AI system could be used on DM to exclude normal cases from human reading, detect additional cancers on DM that radiologists only detected on DBT, or add DBT in selected high-gain cases. Further, it was studied how the AI system can be utilised to reduce the workload of DBT screening.

Results: A research database was created that contained 449 000 examinations from 103 000 women, performed during a time span of 17 years. This includes 9 250 cancers in 7 371 women. It was found that the tested AI system can be used on DM to exclude 19% of examinations from human reading without missing any cancers and that AI can detect 44% of DBT-only detected cancers using only DM. Further, adding DBT for the 10% of the women with the highest AI risk score can detect 25% more cancers than DM screening. For DBT screening, the AI system can reduce the reading workload to the level of DM screening, either by replacing the second reader in a double reader setup or by discarding half of examinations from reading, thus focusing double reading on the half with the highest risk.

Discussion: The results indicate that AI can be used to improve the performance and efficiency of breast cancer screening in several ways, including making it possible to use DBT in screening without demanding more resources. The research database can facilitate larger retrospective studies on these and other subjects. However, before clinical implementation, prospective studies would also be necessary, where e.g. the interaction between radiologists and AI can be investigated.

Key words: breast cancer, artificial intelligence, screening, mammography, breast tomosynthesis

Classification system and/or index terms (if any) Supplementary bibliographical information

Language: English **ISSN and key title**: 1652-8220

ISBN: 978-91-8021-529-9

Recipient's notes **Number of pages**: 107

Price Security classification

Signature Date 2024-02-22

Improving breast cancer screening with artificial intelligence

Victor Dahlblom

LUND
UNIVERSITY

Faculty of Medicine, Department of Translational Medicine
Diagnostic Radiology, Malmö

ISBN 978-91-8021-529-9
ISSN 1652-8220
Lund University, Faculty of Medicine Doctoral Dissertation Series 2024:36

Printed in Sweden by Media-Tryck, Lund University
Lund 2024

To my family

Table of Contents

Abstract

Introduction: The current standard method for breast cancer screening is digital mammography (DM). Digital breast tomosynthesis (DBT) can detect more cancers but is more resource-demanding, not the least due to a more time-consuming reading, which hinders the implementation in screening. Artificial intelligence (AI) might open possibilities to overcome this, but different potential ways of using AI need to be tested using representative screening data. To facilitate the testing and further development of AI, it is necessary to collect and organise more data in a research-friendly form.

Aim: To create a breast imaging research database and explore different ways of using AI to improve breast cancer screening.

Methods: All DM and DBT examinations performed in Malmö, Sweden during 2004–2020 were collected and combined with other relevant information in a research database. A subset consisting of 14 848 women had been examined with paired DM and DBT as part of the Malmö Breast Tomosynthesis Screening Trial (MBTST). This cohort was used to test different ways of using an AI cancer-detection system, which scores examinations based on cancer risk. It was studied whether the AI system could be used on DM to exclude normal cases from human reading, detect additional cancers on DM that radiologists only detected on DBT, or add DBT in selected high-gain cases. Further, it was studied how the AI system can be utilised to reduce the workload of DBT screening.

Results: A research database was created that contained 449 000 examinations from 103 000 women, performed during a time span of 17 years. This includes 9 250 cancers in 7 371 women. It was found that the tested AI system can be used on DM to exclude 19% of examinations from human reading without missing any cancers and that AI can detect 44% of DBT-only detected cancers using only DM. Further, adding DBT for the 10% of the women with the highest AI risk score can detect 25% more cancers than DM screening. For DBT screening, the AI system can reduce the reading workload to the level of DM screening, either by replacing the second reader in a double reader setup or by discarding half of examinations from reading, thus focusing double reading on the half with the highest risk.

Discussion: The results indicate that AI can be used to improve the performance and efficiency of breast cancer screening in several ways, including making it possible to use DBT in screening without demanding more resources. The research database can facilitate larger retrospective studies on these and other subjects. However, before clinical implementation, prospective studies would also be necessary, where e.g. the interaction between radiologists and AI can be investigated.

Populärvetenskaplig sammanfattning

Screening för bröstcancer sker idag med digital mammografi (DM) i Sverige och många andra länder. Digital brösttomosyntes (DBT) kan upptäcka fler cancerfall och kan ses som en sorts 3D-mammografi. DBT visar bröstet i flera tunna snitt istället för bara en sammansatt 2D-bild, vilket minskar problemen med att vissa cancrar döljs av andra vävnadsstrukturer. Ett hinder för att införa DBT i screening är att med traditionell granskning tar det längre tid att granska DBT än DM. Detta är problematiskt eftersom det redan i dagsläget är brist på erfarna bröstradiologer. Under de senaste åren har det skett en revolutionerande utveckling inom artificiell intelligens (AI), vilket öppnar nya möjligheter inom bildgranskning och kan vara en möjlighet att effektivisera screeninggranskningen. Detta skulle kunna frigöra resurser för införande av DBT eller andra förbättringar. Det finns flera olika AI-system som kan användas som stödverktyg för radiologen i granskningen, men också exempelvis för att sortera undersökningarna efter cancerrisk så att screeningen kan anpassas efter varje kvinnas cancerrisk.

I delarbete 1 byggde vi en forskningsdatabas för bröstcancer med fokus på radiologisk bilddiagnostik. Där samlade vi alla DM- och DBT-undersökningar som utförts i Malmö sedan digitaliseringen 2004 till 2020 tillsammans med tillhörande granskningsresultat, radiologutlåtanden och uppgifter om cancer från olika register. För ändamålet har vi byggt en plattform där data från olika källor kan sökas och länkas samman. Databasen innehåller nästan 450 000 undersökningar från 103 000 kvinnor.

I delarbete 2 undersökte vi om ett AI-system kan användas för att utesluta normalfall i DM-screening. Vi analyserade DM-undersökningar från knappt 10 000 kvinnor med ett AI-system och undersökte ifall lågriskfall skulle kunna uteslutas från radiologgranskning. Resultaten visade att nästan 19 % av undersökningarna med lägst risk skulle kunna uteslutas utan några cancerfall missas.

I delarbete 3 studerade vi om ett AI-system kan användas för att hitta ytterligare cancerfall på DM, som enbart kan hittas på DBT vid traditionell radiologgranskning. Studien visade att AI-systemet genom att analysera DM-undersökningar kunde hitta 44 % av cancerfallen som utan AI enbart hittades på DBT.

I delarbete 4 testade vi om AI skulle kunna användas för att individualisera screeningen genom att lägga till DBT i högriskfall. Resultaten visade att genom att undersöka 10 %

av kvinnorna med DBT kan 25 % fler cancerfall hittas, vilket motsvarar mer än hälften av de cancerfall som enbart hittades med DBT.

I delarbete 5 undersökte vi hur AI skulle kunna användas för att effektivisera granskningen av DBT i screening. Vi testade om AI kan användas för att antingen ersätta den ena granskaren, eller för att sortera bort hälften av undersökningarna med lägst risk från granskning. Resultaten visade att båda sätten att använda AI fungerar ungefär lika bra. Oavsett metod upptäcktes 95 % av cancerfallen som hittades med dubbelgranskad DBT-screening, men med halverad total granskningstid. Detta skulle kunna göra det möjligt att införa DBT i screening med samma tidsåtgång som DM.

Resultaten från de olika studierna visar att AI skulle kunna användas för att förbättra screeningen på flera olika sätt. Fler och större studier behövs för att säkerställa att det fungerar, och där kan forskningsdatabasen vara en bra grund. Även om AI verkar fungera bra i studier på gamla undersökningar, så behöver det också undersökas under mer verkliga förhållanden, där det exempelvis går att se hur radiologerna samverkar med AI.

List of papers

This thesis encompasses the five published papers listed below. The full papers are included at the end of the printed thesis as appendices.

Apart from these published journal papers, I have participated in writing a number of conference abstracts that have been published as proceedings, which are listed separately. Further, I have written two conference abstracts that have not yet been published as journal papers, and these are also listed.

Papers included in thesis

Paper 1

Malmö Breast ImaginG database: objectives and development

Victor Dahlblom, Magnus Dustler, Anetta Bolejko, Predrag R. Bakic, Henrik Granberg, Kristin Johnson, Daniel Förnvik, Kristina Lång, Anders Tingberg and Sophia Zackrisson

Journal of Medical Imaging 2023; Vol. 10 Issue 6: 061402: 1–18.

Paper 2

Identifying normal mammograms in a large screening population using artificial intelligence

Kristina Lång, Magnus Dustler, **Victor Dahlblom**, Anna Åkesson, Ingvar Andersson and Sophia Zackrisson

European Radiology 2021; 31: 1687–92.

Paper 3

Artificial intelligence detection of missed cancers at digital mammography that were detected at digital breast tomosynthesis

Victor Dahlblom, Ingvar Andersson, Kristina Lång, Anders Tingberg, Sophia Zackrisson and Magnus Dustler

Radiology: Artificial Intelligence 2021; 3(6): e200299: 1–10.

Paper 4

Personalized breast cancer screening with selective addition of digital breast tomosynthesis through artificial intelligence

Victor Dahlblom, Anders Tingberg, Sophia Zackrisson and Magnus Dustler

Journal of Medical Imaging 2023; Vol. 10 Issue S2: S22408: 1–17.

Paper 5

Breast cancer screening with digital breast tomosynthesis: comparison of different reading strategies implementing artificial intelligence

Victor Dahlblom, Magnus Dustler, Anders Tingberg and Sophia Zackrisson

European Radiology 2022; 33: 3754–65.

Conference abstracts published as proceedings

Personalised breast cancer screening with selective addition of digital breast tomosynthesis through artificial intelligence

Victor Dahlblom, Anders Tingberg, Sophia Zackrisson and Magnus Dustler

15th International Workshop on Breast Imaging, IWBI 2020. Bosmans, H., Marshall, N. & Van Ongeval, C. (red.). SPIE, 115130C. (Proceedings of SPIE: The International Society for Optical Engineering; vol. 11513).

The effect of breast density on the performance of deep learning-based breast cancer detection methods for mammography

Magnus Dustler, **Victor Dahlblom**, Anders Tingberg and Sophia Zackrisson

15th International Workshop on Breast Imaging, IWBI 2020. Bosmans, H., Marshall, N. & Van Ongeval, C. (red.). SPIE, 1151324. (Proceedings of SPIE: The International Society for Optical Engineering; vol. 11513).

Correspondence between areas causing recall in breast cancer screening and artificial intelligence findings

Victor Dahlblom, Anders Tingberg, Sophia Zackrisson and Magnus Dustler

16th International Workshop on Breast Imaging, IWBI 2022. Bosmans, H., Marshall, N. & Van Ongeval, C. (red.). SPIE, 122860K. (Proceedings of SPIE: The International Society for Optical Engineering; vol. 12286).

Simultaneous digital breast tomosynthesis and mechanical imaging in women recalled from screening - A preliminary analysis

Rebecca Axelsson, **Victor Dahlblom**, Anders Tingberg, Sophia Zackrisson, Magnus Dustler and Predrag Bakic

16th International Workshop on Breast Imaging, IWBI 2022. Bosmans, H., Marshall, N. & Van Ongeval, C. (red.). SPIE, 1228607. (Proceedings of SPIE: The International Society for Optical Engineering; vol. 12286).

Tumor growth rate estimations in a breast cancer screening population

Hanna Tomic, Akane Ohashi, **Victor Dahlblom**, Anna Bjerkén, Daniel Förnvik, Magnus Dustler, Sophia Zackrisson, Anders Tingberg and Predrag Bakic

16th International Workshop on Breast Imaging, IWBI 2022. Bosmans, H., Marshall, N. & Van Ongeval, C. (red.). SPIE, 1228613. (Proceedings of SPIE: The International Society for Optical Engineering; vol. 12286).

Conference abstracts not yet published

Mammography screening with stand-alone artificial intelligence compared to single and double reading with and without consensus discussions

Victor Dahlblom, Magnus Dustler, Anders Tingberg and Sophia Zackrisson

European Congress of Radiology, ECR 2022.

Workload reduction of digital breast tomosynthesis screening using artificial intelligence and synthetic mammography

Victor Dahlblom, Magnus Dustler, Sophia Zackrisson and Anders Tingberg

European Congress of Radiology, ECR 2023.

Abbreviations

AI	Artificial Intelligence
AUC	Area under the curve (of ROC)
BI-RADS	Breast Imaging Reporting & Data System
CC	Craniocaudal (view)
CAD	Computer Aided Detection
CEM	Contrast Enhanced Mammography
CSAW	Cohort of Screen Aged Women
CT	Computerised Tomography
DBT	Digital Breast Tomosynthesis
DICOM	Digital Imaging and Communications in Medicine
DM	Digital Mammography
GDPR	General Data Protection Regulation
HER2	Human Epidermal growth factor Receptor 2
IC	Interval Cancer
IDC	Invasive Ductal Carcinoma
LM	Lateromedial (view)
M-BIG	Malmö Breast ImaginG (database)
MBTST	Malmö Breast Tomosynthesis Screening Trial
MLO	Mediolateral Oblique (view)
MRI	Magnetic Resonance Imaging
NKBC	National Quality registry for Breast Cancer
OMI-DB	OPTIMAM Mammography Image Database
PACS	Picture Archiving and Communication System
RIS	Radiology Information System
ROC	Receiver Operating Characteristic
SCAN-B	Sweden Cancerome Analysis Network-Breast
SM	Synthetic Mammography
TNM	Tumour, Node, Metastasis (staging)

Introduction

Globally, breast cancer is the most commonly diagnosed cancer, and it is also the cancer responsible for the largest number of cancer deaths among women.[1] In an attempt to combat this, many countries have implemented mammography screening for breast cancer with the aim of finding the cancers at an earlier stage when the prognosis is better. Unfortunately, a substantial number of breast cancers are still undiagnosed until they start to cause symptoms.[2] At that time, the possibilities for successful treatment and cure are often worse than if the cancer would have been detected at an earlier stage. Various potential ways of improving the currently usually digital mammography (DM) based breast cancer screening have been suggested, where, in particular, digital breast tomosynthesis (DBT) – or 3D mammography – has emerged as a better and feasible alternative to standard mammography.[3] This method is, however, hindered by a longer reading time that adds a further burden on the, in many places already understaffed, screening programmes.

In the last few years, artificial intelligence (AI) and deep learning has undergone rapid development and have begun to be employed within numerous different domains. In healthcare, there are higher requirements for a solid verification of the function and performance than in some other fields. Breast cancer screening is one of the areas in healthcare which has gained a significant amount of interest among AI developers, both because screening contains monotonous tasks suitable for AI and favourable conditions for developing AI systems thanks to the large amounts of data that have been collected over decades of breast cancer screening. However, the development of AI systems is not enough, as it is necessary to thoroughly investigate how these perform in a clinical context before they can be introduced in the clinical workflow.

At the time of initiation of this thesis in 2019, only a few AI systems were commercially available which were aimed at assisting the radiologist with breast cancer detection in their readings. Some studies had tested AI systems working stand-alone on cancer-enriched retrospective mammography datasets,[4] but this does not give a realistic estimation of the performance in real-world screening. No published studies had yet tested an AI system on a screening dataset from a population not used in developing the AI system. Apart from assisting the radiologist, different ways of using AI in breast

cancer screening remained largely uncharted. Further, despite the fact that huge numbers of mammography examinations had been performed in screening programmes, very little screening data were available for research and development purposes, which is a prerequisite for successful development and testing of AI systems.

This thesis proposes several different potential ways of using AI for the purpose of improving breast cancer screening, based on my testing of an AI system on data from a screening study with paired DM and DBT examinations. The use of screening data can give a more representative view on how AI would work in clinical use compared with studies based on cancer-enriched collections. The paired DM and DBT data opens possibilities for studying how AI can help to efficiently utilise DBT in screening, and this is explored from several perspectives. Additionally, a breast imaging research database is presented, including almost 450 000 examinations, which constitutes a solid foundation for future research on AI in breast cancer imaging.

Background

Breast cancer

Breast cancer is the most commonly diagnosed type of cancer worldwide, with 2.26 million new cases diagnosed in 2020 and accounting for 11.7% of all cancers but 24.5% of all cancers in women.[1] The lifetime risk of being diagnosed with breast cancer for women is 5.9% globally but varies considerably, and in e.g. Sweden, the risk is 12.2%.[5] Breast cancer also caused 685 000 deaths, making it the deadliest cancer among women worldwide. In Sweden on average 7 900 women have been diagnosed with breast cancer per year during the last 10 years, with 1 400 deaths yearly.[6] The incidence has been successively increasing over the last decades, in part due to changes in the population with longer lifespan and shifts in lifestyle, including having fewer children and at a higher age, but also due to higher awareness of breast cancer and introduction of breast cancer screening. During the same time span, the number of deaths from breast cancer has decreased, which in part is thanks to improved treatment methods but also the introduction and advancements of breast cancer screening, where about 60% of all breast cancer cases are diagnosed,[7] has an important impact.

The relative survival rate of the most common type of invasive breast cancer (invasive ductal carcinoma, IDC) is 90% after 5 years and 73% after 20 years.[8] Thus, despite relatively good short-term survival, the long-term survival remains relatively poor. Metastases from breast cancer can appear more than 15 years after seemingly successful treatment of the primary tumour.[9] The most common locations for metastases from breast cancer include the bones, lungs, brain and liver.[10]

Risk factors

There are several risk factors for breast cancer, with female sex being the most prominent. The risk also increases with age. Family history of breast cancer is an important risk factor; many breast cancer-associated genes have been identified, including the BRCA1 and BRCA2-genes, albeit more than half of the cases of hereditary breast cancer appear in women without any identified specific gene variants.[11]

Previous breast cancer and previous biopsy with benign findings are also risk factors.[12] Obesity is a risk factor for breast cancer in post-menopausal women.[13] Younger age at menarche and older age at menopause are also risk factors.[14] High breast density is associated with a higher risk of breast cancer.[15] Exposure to radiation therapy involving the breasts is also a risk factor, in particular at an early age.[16] Use of hormone replacement therapy is associated with an increased risk.[17] Alcohol intake is also a risk factor.[18] Higher parity and longer breast feeding are protective factors.[19]

Characterising breast cancer

Breast cancer is a heterogeneous disease, and there is a wide variation in characteristics, treatment options and prognosis. The complete biological background is not fully understood, but in order to understand a specific case as closely as possible, it is valuable to combine different properties.

As with most other cancer types, the size and localisation are usually essential in the description of breast cancer and are important in e.g. the planning of surgery and radiation treatment. The location of a tumour can be defined with radiological methods, such as by ultrasound examination or by combining information from at least two mammography projections. The size of a breast cancer can be measured with different methods, which can give some variations in results.[20] The visible extent of the lesion might differ between imaging modalities; namely, mammography, ultrasound and magnetic resonance imaging (MRI) can result in different measurements. After surgery, the size can also be measured on the pathological specimen. Here, the borders of the tumour can be defined with higher precision, but in the process of preparing the specimen for microscopy, the shape might be altered, and thus the size might differ from the size that the tumour had in vivo.

Regional lymph nodes are often the first site of metastases, and assessment of lymph nodes draining the area of the cancer is thus important. Among screening-detected breast cancers, lymph node metastases are found in about 23%–24% of the cases.[21,22] In some cases, enlarged lymph nodes, e.g. in the axilla, are palpable or visible with ultrasound or mammography. In these cases, the lymph nodes are biopsied and analysed cytologically. Unless the presence of lymph node metastases is already confirmed when the tumour is operated on, a sentinel node biopsy procedure is usually performed during tumour surgery. This means that the first one or few lymph nodes draining the cancer area are extirpated and analysed. The sentinel nodes are usually identified by the injection of a radioactive tracer and blue dye.

Apart from localisation and size, the radiological characterisation of breast cancer is, to a large extent, limited to radiographical appearance. Tumours can be described using different terms, including circumscribed mass, spiculated mass, calcifications and architectural tissue distortion. In some cases, enlarged lymph nodes are the only visible sign. While, in some cases, these properties may contribute information on prognosis and have some correlation with different histological and molecular types of breast cancer,[23–25] most of the characterisation relies on pathological examinations of samples from the suspected cancer lesion.

One categorisation of breast cancer is into different histological types based on the appearance of pathological specimen in microscopy. The most common histological type is invasive breast carcinoma of no special type, also called IDC, which accounts for about 65%–75% of all breast cancers.[26–28] The second most common invasive type is invasive lobular carcinoma, which accounts for about 10%–15% of all breast cancers.[26–28] The most common in situ form, i.e. a cancer that has characteristics of malignant cells but has not invaded the basal membrane, is ductal carcinoma in situ, which constitutes about 10%–20% of all breast cancers.[26–29] The proportion depends on the characteristics of the breast cancer screening programme as ductal carcinoma in situ is usually diagnosed after identifying calcifications on screening mammography, and, thus, has a higher apparent incidence in populations subject to screening. Other less common types of breast cancer account for about 5%.

The histological and nuclear grades describe the degree of differentiation and proliferative activity. The stage describes the extent and spread of the cancer and is defined according to the tumour, node, metastasis (TNM) staging system, where T depends on the size and spread of the primary tumour, N on the presence and extent of metastatic spread to regional lymph nodes and M on the presence and size of distant metastases.[30]

Breast cancer can also be classified by using molecular and genetic properties, which provides additional information about the prognosis and therapeutical possibilities.[31] Those definitions require analyses that until recently were too complex, slow and expensive for use in clinical routine, but with technical developments, these are now increasingly available for clinical use.[32] Surrogate measures based on immuno-histochemical staining for receptors of oestrogen, progesterone and human epidermal growth factor receptor 2 (HER2),[33] were originally introduced as a cheaper and faster alternative, but still play an important clinical role.[32] In the surrogate classification, invasive breast cancers are divided into Luminal A-like breast cancer accounting for 40%–50%, Luminal B-like breast cancer accounting for 20%–30% (often further divided into HER2- and HER2+), HER2-overexpressing breast cancer accounting for 15%–20% and triple negative breast cancer accounting for 10%–20%.[34] The surrogate

classification has been shown to have a prognostic value, for instance in predicting risk of lymph node metastases, recurrence patterns and disease-free survival.[35]

The histological and molecular subtypes are not clearly connected and provide complementary information; that is, a cancer with a particular histological type does not have to be of a certain molecular type.

The previously mentioned overall risk factors for breast cancer are dominated by the risk factors for the most common molecular subtype, Luminal A-like breast cancer.[36] Many of the risk factors are common among the subtypes but have different importance. However, higher parity increases the risk for triple negative breast cancer, while it is a protective factor for Luminal A-like breast cancer.[36]

Treatment

Surgery is, in most cases, the cornerstone in the treatment of breast cancer. If it is possible to attain a surgically radical treatment by using breast-conserving surgery, that approach is usually taken, as long as a cosmetically and functionally good result can be achieved.[37,38] Otherwise, or in cases with large tumours, multifocality, very high genetic risk or personal preference, mastectomy is performed.[37,38] As mentioned earlier, a sentinel node biopsy procedure is often performed during surgery. If any of the sentinel nodes contain metastases, complete axillary dissection is performed unless the woman has planned post-operative radiation therapy.[37]

Radiation therapy is always performed after breast-conserving surgery as well as after mastectomy with tumours >5 cm or lymph node metastases.[38] Systemic treatment includes hormonal drugs, antibody-based drugs, and chemotherapy. For ER-positive cancers, oestrogen effect inhibition, usually including Tamoxifen, is recommended.[38] Chemotherapy is recommended except in low-risk cancers.[37,38] Treatment regimens including specific antibodies are recommended for HER2-positive cancers. The purpose of systemic treatment is to reduce the risk of recurrence, and it can be given both after surgery, i.e. adjuvant therapy, and before surgery, i.e. neoadjuvant therapy, where the latter additionally has the purpose to downsize the tumour to enable a less extensive surgery.[37] In cases when the breast cancer is not curable, usually due to the presence of distant metastases, palliative treatment is normally offered with chemotherapy or radiation therapy aiming to prolong the lifetime and reduce symptoms.[38]

Breast imaging and diagnostics

Mammography

A mammography examination is performed by letting X-rays pass through the breast and collecting the transmitted rays on the other side of the breast, historically with an X-ray film, but in recent years with a digital detector. The resulting image shows the differences in the density of the structures inside the breast. The first experiments using X-rays for imaging the breasts had started already in 1913, just a few years after the discovery of X-rays in 1895.[39,40] While many other applications of plain X-rays usually focus on structures with high-contrast differences (e.g. bone or gas–fluid interfaces), the superficial placement of the breasts enables detailed imaging of the low-contrast attenuation differences in the soft tissue. A broader use of mammography began after the introduction of specialised mammography equipment in the 1960s.[39,40]

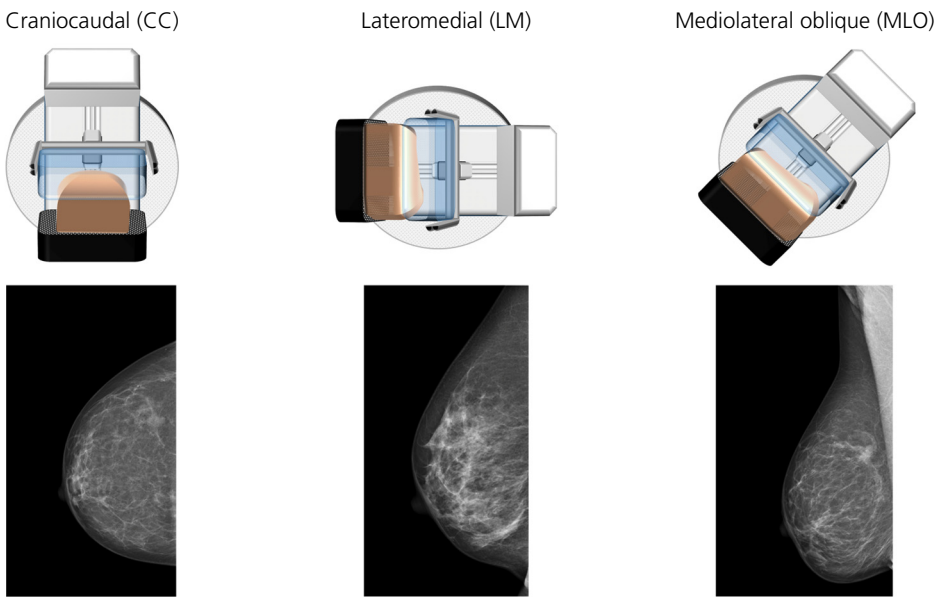


Figure 1. Common mammographic view projections.
Schematic illustration of the positioning of the breast and mammography machine in the upper line and examples of the corresponding resulting mammography images below. Portions of the illustration have been edited based on an original by Magnus Dustler.

Mammography technology has successively evolved, with a notable leap in the 2000s with the transition to digital imaging, where the X-ray film was replaced by a digital detector.[41] Even after the advent of other imaging techniques, mammography has remained the base method in breast imaging with short examination times, high resolution and good reproduction of calcifications as its specific strengths.

Common practice involves a few standardised views where the breast is imaged from different directions, as illustrated in Figure 1. The craniocaudal (CC) and lateromedial (LM) views depict the breast in two perpendicular planes, which makes it possible to determine the position of a specific structure in the breast, provided that the structure can be identified in both views. The mediolateral oblique (MLO) view is taken at an angle of about 30°–40° compared with a craniocaudal line, has a better inclusion of breast tissue close to the chest wall and parts of the axilla, and can improve cancer detection.[42] A breast cancer screening examination usually includes CC and MLO views, while an LM view is often performed when further investigating a suspected cancer. There are also several more specific supplementary views that can be used at assessment of suspicious findings, e.g. magnification views to better visualise small calcifications and focal spot compression views to better visualise masses and distortions. Usually, the breast is compressed, both in order to spread the tissues to a wider area, giving less superimposition of tissues and reduce the breast thickness in order to decrease scattered radiation and lower the required radiation dose necessary to achieve desirable image quality. The compression also has a fixating function and helps to avoid motion blur.

Mammography has limitations and is not ideal for imaging all cancers. In particular, in dense breasts, the abundant glandular tissue might obscure small cancer lesions. In some cases, it might be impossible to distinguish between a malignant and benign lesion, for example a circumscribed cancer or a simple benign cyst. Some kinds of lesions can also be subtle or invisible on mammography, and thus, other modalities are necessary for detection.

Digital breast tomosynthesis

DBT is basically an extension of mammography where the X-ray tube moves around the breast and images are taken from numerous different angles, as illustrated in Figure 2. From these so-called projection images, a layered stack of images can be reconstructed similarly as a computerised tomography (CT) 3D volume, albeit only one projection can be reconstructed with desirable results due to the limited data capture where not all angles are sampled. However, the layered stack can open possibilities to

separate structures at different depths, thereby reducing the issues with dense breast tissue obscuring small cancer lesions, as illustrated in Figure 3.

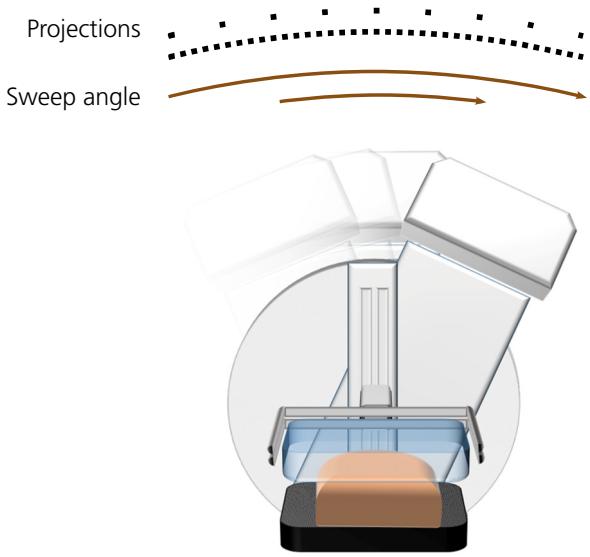


Figure 2. Illustration of DBT machine.
When conducting the DBT examination, the machine moves over the breast, and the sweep angle varies from 15° to 50°, depending on DBT machine vendor. During this movement, a number of projection images are acquired, and the number and distribution of projections also depend on the vendor. Portions of the illustration are edited based on an original by Magnus Dustler.

The sweep angle varies between manufacturers of DBT systems, with Hologic on the lower end at 15° and Siemens on the higher end at 50°.[43] The other manufacturers are in between, e.g. GE has 25°, and Fujifilm 15° or 40° depending on mode. The number of projections also varies from 9 (GE) to 25 (Siemens). As the total radiation dose has to be limited, a higher number of projections means a higher noise level in each projection image. The raw images are reconstructed to readable images, where each image usually corresponds to a tissue slice of a specific thickness, e.g. 1 mm. Thus, the sweep angle and number of projections are not directly visible to the reader, but affect the characteristics of the images. The depth resolution is better with a wider sweep angle, but as not all angles are sampled, the depth resolution will always be limited compared with CT.[43] A narrow angle may instead visualise microcalcifications better due to less geometrical bluring.[44]

Several studies have shown that DBT screening has a higher sensitivity than the current standard of DM screening.[45–50] Still, there is hesitation about moving away from DM to DBT.[51] This is in some places motivated by DBT being more resource-intensive,[52,53]

as DBT usually takes longer to read than DM. Continuing DM can also be valuable in facilitating comparison with previous examinations performed with DM. In many places with early introduction of DBT in routine screening, mainly in the USA, DBT was thus at least initially combined with DM.[51] Unfortunately, combining DM and DBT leads to substantially higher radiation doses. A solution is to use synthetic mammography (SM), which is a method to generate an image that resembles a DM by aggregating the data from a DBT into a single image, which has now largely replaced a combination of DBT and DM.[54,55] The complexity of the task varies with the sweep angle of the DBT system, as a wider angle is more prone to give artefacts.

In centres where DBT is available, it often has an important role in the assessment of suspicious findings that have been recalled from screening as well as diagnostic examinations when the woman is referred due to symptoms. In these situations, DBT has, to a large extent, replaced supplementary DM views.

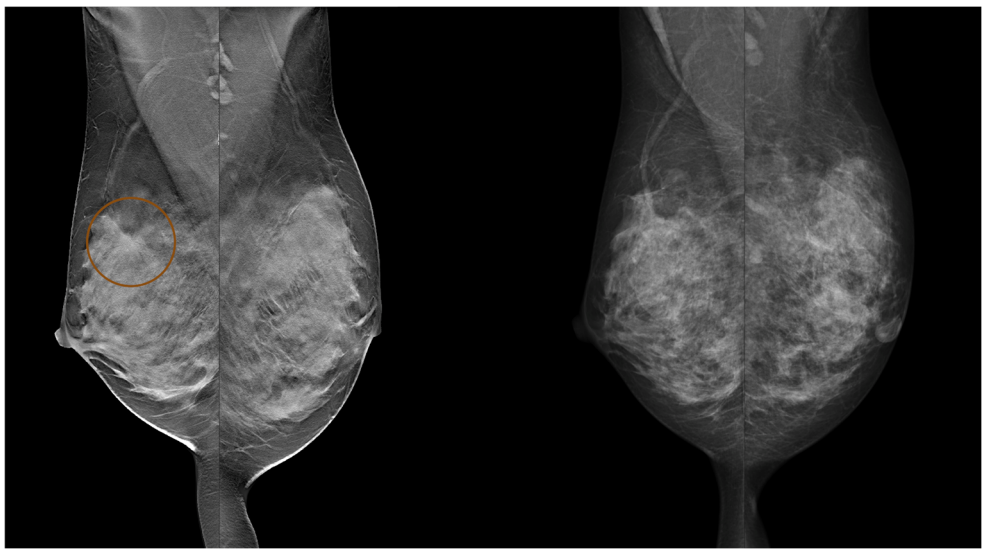


Figure 3. Example of a cancer visible on DBT but obscured by overlaying tissue on DM.
DBT on the left and DM on the right. The cancer (circled) is a 12 mm large invasive ductal carcinoma (IDC) with a spiculated mass appearance.

Other imaging methods

While DM is the most important method in breast cancer screening and plays an important role in investigating cases with cancer suspicion, there are several additional methods that are important in selected groups or cases.

Ultrasound is an important method for investigating suspected cancers and complements DM (and DBT). In Sweden, ultrasound is also the first method of choice for women under the age of 30 years and those who are pregnant or lactating.[38] The technology is based on sending high frequency sound waves into the tissues and registering the differences in echoes caused by stiffness variations in the tissues. As ultrasound depends heavily on the operator and is hard to standardise, its use as a primary screening method is limited, albeit it is important for defining the character of specific findings during assessment.[40] Further, ultrasound is the preferred method to guide biopsies and preoperative markings of structures with known or suspected malignancy.

Breast MRI with gadolinium contrast agents has a high sensitivity for detecting breast cancer but has a number of drawbacks, including long examination time, high costs and many false positive findings.[40,56] Also, administration of gadolinium contrast agents leads to deposition of gadolinium in the brain, which is a concern in particular with repeated examinations, albeit the biological significance is uncertain.[57] MRI mainly plays a role in screening of women with a known high risk of breast cancer and in characterisation of complex cancer cases, including preoperative staging and following the effect of chemotherapy.[58]

Contrast enhanced mammography (CEM) is an extension of DM where the contrast enhancement of breast cancers is depicted, similar to breast MRI, but instead using dual-energy X-ray and an iodine contrast medium. The availability and cost profile of CEM may be more favourable than for breast MRI.[59] The sensitivity for breast cancer is slightly lower than with MRI (91% compared with 97%), while the specificity of CEM is higher than that of MRI (74% compared with 69%).[59]

There are also a number of more experimental or niched breast imaging methods, which might gain more importance in the future, including breast CT,[60] phase contrast X-ray,[61] mechanical imaging,[62] molecular breast imaging[56] and optical breast imaging.[63]

Biopsy

Radiological imaging methods can detect and localise cancer suspicious areas, but the final diagnosis relies on microscopy and other pathological methods for analysing tissue samples from the suspected lesion. Usually, such samples are achieved through biopsy of the lesion, either freehand for palpable lesions or guided by imaging. Ultrasound guided biopsies are most common, but also DM, DBT and MRI can be used to guide biopsies, depending on the visibility of the lesion on different modalities and local traditions and availability.

Medium-sized core needle biopsy is the most common and allows a review of the histological appearance of the lesion. Fine needle aspirations can be used in selected cases, for instance when the localisation is unfavourable for larger biopsies; however, the gain of information is more limited as the analyses are restricted to cytological properties, i.e. the individual cells can be studied but not the full tissue. If larger tissue samples are needed than those obtained with medium-sized needle biopsy, vacuum-assisted core needle biopsy can be used.

Breast cancer screening

Background of breast cancer screening

As breast cancer is a common type of cancer, which also strikes many younger women, it has a large impact on both personal and social aspects as well as to society and the economy. In order to cure as many women as possible – or if cure is impossible, at least to prolong and improve the life quality of the remaining life – early diagnosis is necessary. Early diagnosis can also reduce the need for excessively extensive treatments with many side effects. Thus, breast cancer screening programmes with mammography have been implemented in many countries. The aim is to diagnose breast cancer before it reaches a size where it is palpable and begin to give symptoms.

The first steps to breast cancer screening were taken already in the 1960s with the start of the first clinical trial, which showed that screening with mammography could lead to a reduction in breast cancer mortality.[64] In the following decades, further randomised clinical trials confirmed the results in other populations and added more solid evidence. The effect of screening is most clearly proven in the 60–69 age group , where the effect is the largest, but a smaller but significant effect has been seen in the 50–59 age group.[65,66] There have been several studies focusing on younger women (40–49), but due to the small effect in this age group where breast cancer in less common, it is harder to reach significance. For women 70–74 years of age, where breast cancer is more common, the evidence is instead weak due to a relatively small total number of study participants in this age group, as few studies have included women over the age of 70 years of age and only to a limited extent. The shorter remaining life expectancy at a higher age also limits the potential gain from screening.

Population-based screening programmes, i.e. all women in the population within the selected age span are systematically invited to screening, were introduced in some countries and regions already in the 1970s, with more following in the 1980s.[67,68]

However, the broad introduction came during the 1990s and 2000s. Many of the early breast cancer screening programmes have changed over time, e.g. by adjusting age spans or screening frequency.

The attention paid to breast cancer screening has caused many women to take part in opportunistic screening, i.e. actively booking mammography appointments despite having no symptoms. These women can either be well-informed women living in areas without organised screening or being outside of the age span targeted by the local screening programme or women choosing to have examinations outside of the screening programme either as a substitute or as a complement.[69,70]

Current status of breast cancer screening

Breast cancer programmes in some form exist in many countries all over the world, but population coverage, attendance, screening methods, intervals and age limits can vary. Most commonly, the targeted women are 50–70 years of age, but many programmes include younger women down to 40 years of age, and older women up to 74 years of age.[67,71,72]

Different screening programmes use various screening intervals, where a screening interval of two years is the most common. A few screening programmes use three-year intervals, e.g. the UK,[73] while annual screening may be used in some places.[72,74] Some screening programmes, including those in parts of Sweden, have more frequent screening among younger women due to higher average breast density and more fast-growing tumours.[38]

The cornerstone of breast cancer screening has long been mammography, but in recent years, DBT has emerged as a better alternative and is widely used in the USA, sometimes in companion with DM.[75,76] However, this leads to an increased radiation dose, and in some places, DM has instead been replaced by DBT with SM.[77] A generally higher breast density in some populations makes the sensitivity of DM insufficient, as is the case in South Korea, where DM is often combined with ultrasound in screening.[78]

Double reading, i.e. each screening examination is read by two radiologists in order to increase sensitivity, is commonly practiced in e.g. Europe,[79] Australia,[80] and Japan[81]. Single reading is commonly practiced in, for instance, the USA[82]. In order to increase performance, computer-aided detection (CAD) systems are often used in the USA, although the gain of this has been disputed.[83,84] The arrival of new systems – based on modern AI technology, which will be described more thoroughly later – with better performance has increased the interest in such systems also in Europe. This has, for instance, been proposed to be used for replacing one of the readers.[85,86]

There are differences in the culture, legislation and structure of screening programmes between countries. Thus, the recall rates for DM screening are generally in the range of 2.6% to 4.9% in Europe, while the USA has substantially higher recall rates in the range of 7.5% to 17.5%.[3]

Screening examinations are often scored according to level of cancer suspicion using the American Breast Imaging Reporting & Data System (BI-RADS) scale.[87] The examinations are scored from 0 to 6: (0) incomplete examination, (1) negative, (2) benign, (3) probably benign, (4) suspicious, (5) highly suggestive of malignancy and (6) known biopsy-proven malignancy, where 1–5 are most relevant in screening. Many screening programmes, including the Swedish programme, use similar, usually five grade classification systems, but with slight differences.[38,88,89]

Breast cancer screening in Malmö

In Sweden, there are some regional variations in the screening programmes. The screening programme in Malmö currently invites women 40–54 years of age to screening with 18-month intervals and women 55–74 years of age to screening with 24-month intervals. The screening examination is performed with two-view DM, and the examination is read by two non-blinded readers. The reading workflow is illustrated in Figure 4. The readers score each breast from 1 to 5: (1) normal, (2) benign findings, (3) nonspecific findings where malignance cannot be excluded, (4) findings suspicious for malignancy and (5) malignant findings. Usually, examinations with a score of 3 or higher are recalled for further examinations. Each of the readers has the option to put the examination up for a consensus discussion, instead of deciding on recall or not. At recall, additional imaging is performed, such as additional DM views, DBT or ultrasound.

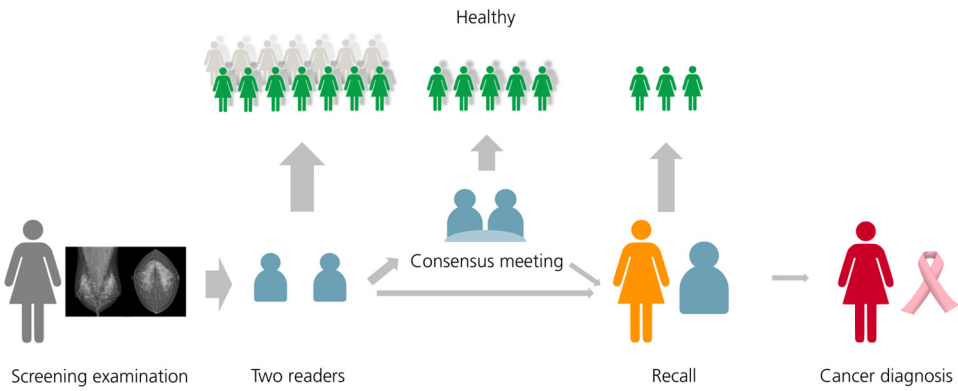


Figure 4. The path through the breast cancer screening workflow.
After conducting a screening examination, the examination is read by two different readers. Most examinations are identified as normal, but if something is unclear, the examination can be discussed on a consensus meeting. If something looks suspicious, the woman is recalled for further examinations, often including ultrasound and sometimes biopsy. After recall, many women are found to be healthy, but a few are diagnosed with breast cancer.

Limitations and issues with screening

Interval cancers

Despite the substantial resources and engagement required to conduct breast cancer screening with mammography – from the health care system, but not least by all healthy women taking time off from their everyday life to go to repeated appointments – many cancers are still not detected by breast cancer screening. Cancers that are diagnosed due to symptoms after a normal screening examination, but before the next scheduled screening, are considered interval cancers. This excludes cancers in women not following the screening programme in the time before the cancer diagnosis. Many cancers are also diagnosed among women outside the screening age. The interval cancer rate can be used as a measure of the performance of the screening, both in clinical trials and in clinical routine. The interval cancer rate varies depending on several factors, including underlying breast cancer incidence and the properties of the local screening programme, for example screening interval.[2] For biennial screening, the interval cancer rates usually range between 8.4 and 21.1 per 10 000 screenings.[2] The term interval cancer encompasses both cancers that to some extent were visible at previous screening but were missed or misinterpreted and fast-growing cancers that appeared after the screening (true interval cancers). Breast cancer can be considered to be a more definitely relevant end point than interval cancers, but the required long follow-up time makes it

an impractical measure; also, the mortality is affected by more factors outside of the screening programme, such as the efficiency of the provided breast cancer treatment.

False-positive recalls

Another issue is that many women are false positively recalled for further examinations despite being healthy. The false-positive recall rate is closely related to the overall recall rate, as this varies far more between screening programmes than the much lower cancer incidence. A false-positive recall can cause anxiety and other psychological consequences specific to breast cancer, albeit the effects on general well-being are limited.[90,91] A false-positive recall is associated with a higher risk of future interval cancer or screening-detected cancer as well as a new false-positive recall.[92] The attendance at the following screening appointment might be affected by a false-positive recall, but can both be increased (USA)[93] and decreased (Europe).[94,95]

Overdiagnosis

A few women are diagnosed with breast cancer at screening and go through surgery and other treatments, despite that their cancer would have been so slow-growing that it would never have caused any symptoms during their lifetime. Such women are subject to overdiagnosis and overtreatment. As it is hard to measure the extent of overdiagnosis, estimates in the wide range of 0%–76% have been reported when including several types of studies.[96] The best available source of data is generally be considered to be from randomised clinical trials comparing mammography screening and no screening, where the control group was not offered screening after the end of the trial.[97] This provides the opportunity to compare the cancer incidence over a longer time span, which allows to separate overdiagnosed cancers from cancers where the screening led to an earlier diagnosis of cancers that would otherwise give symptoms in a few years. A previous review identified three studies fulfilling these requirements, including one performed in Malmö and two in Canada, where the reported results yielded values of overdiagnosis of 10%–13%, defined as excess cancers compared with all breast cancers diagnosed in women not invited to screening[97–100]. All three studies are quite old, and the results might not correctly represent the current status, e.g. due to developments in technology and treatment, but it would be impossible to perform a similar study today when breast cancer screening is included in standard of care.

AI and deep learning

History of automation in breast cancer screening

The first attempts to use computers to automatise the reading of mammography images began in the 1970s.[101] As this was long before the digitalisation of the mammography equipment, the process started with scanning of the images in order to make them accessible to the computer. The limited computational power heavily restricted the possible resolution and colour depth. Only small segments could be analysed at the same time, and the methods relied on handcrafted rules, trying to imitate some of the aspects that human readers would assess. Apart from the then cumbersome process of using computer analysis, the usability was limited by large areas falsely marked as cancer suspicious.

The advent of more powerful computers and the transition to digital imaging opened possibilities to implement more powerful and user-friendly computer systems aimed at assisting radiologists in reading mammography examinations. Systems for CAD, based on traditional image analysis and handcrafted rules, were first introduced in the late 1990s and gained increasing usage mainly in the USA during the following decade.[102] The use of CAD can increase the cancer detection rate when applied in screening programmes where single reading is practiced but has less value when double reading is practiced.[103] CAD systems have been criticised for marking many normal areas as potentially malignant, leading to increased recall rates in screening.

Technological developments, including deep learning, have opened new opportunities for using computers to interpret breast cancer screening images.[103]

Overview of AI and deep learning

The term artificial intelligence was introduced in the 1950s and is a broad and unclearly defined term comprising numerous different technologies aimed to simulate human intelligent behaviours.[104] Over the years, there have on several occasions been great expectations that AI would revolutionise medicine, but unfortunately these have not yet been fulfilled.[104]

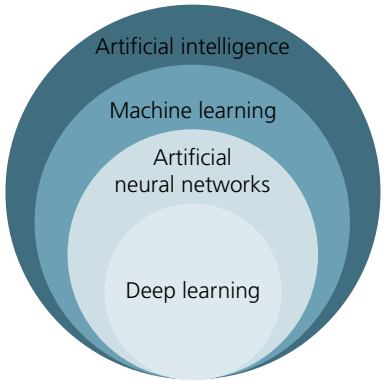


Figure 5. Illustration of the relationship between different methods in artifical intelligence.

Machine learning is a subfield of artificial intelligence, as illustrated in Figure 5, where the machine itself identifies patterns in the training data and generates its own rules for solving the specified task.[105] Machine learning covers both relatively simple methods as regressions and complex systems, such as artificial neural networks, which are loosely inspired by the properties and interplay of nerve cells. In artificial neural networks, multiple nodes are connected in several layers, as illustrated in Figure 6, where the first layers often encompass feature extraction, i.e. identifying different features in the input data, and the subsequent layers often combine these features in order to produce an output, for instance a classification of the input data. The layers in a model often have different sizes. The number of nodes of the input and output layers is determined by the number of dimensions in the input and output data, respectively. The layers between the input and output layers are hidden layers that are internal for the model, and their content cannot be readily examined. If the model contains multiple hidden layers, it is considered a deep neural network, which is the basis for deep learning systems.

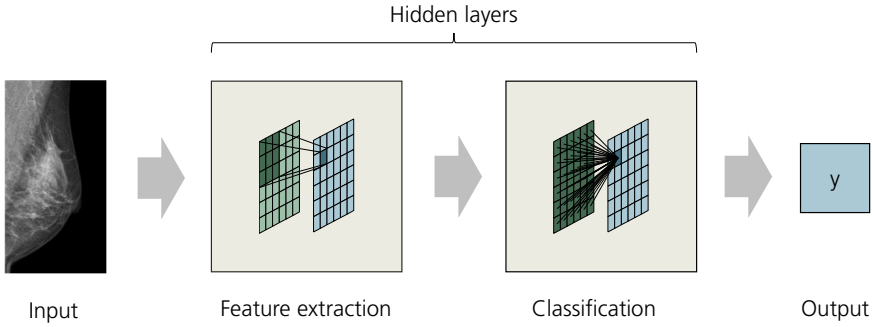


Figure 6. Simplified illustration of a neural network for an image classification task.
The input data, here an image, are fed into feature extraction layers, where different features in the input data are identified. The output of the feature extraction layers is fed into classification layers, where the different identified features are combined in order to produce an output, which in this case is a classification. The size of the output layer is in classification tasks usually small, for example a single binary value indicating sick or healthy. Both feature extraction steps and classification steps usually encompass multiple layers with a mix of different sizes and connection types (see Figure 7).

The layers in a model can be connected in various ways, as illustrated in Figure 7. If two layers are fully connected, every single node in one layer is connected to all the nodes in the other layer. Other ways of connecting the layers are convolutional layers, which combine information from adjacent nodes, e.g. to identify edges in images, and pooling layers, which reduce the number of layer nodes.

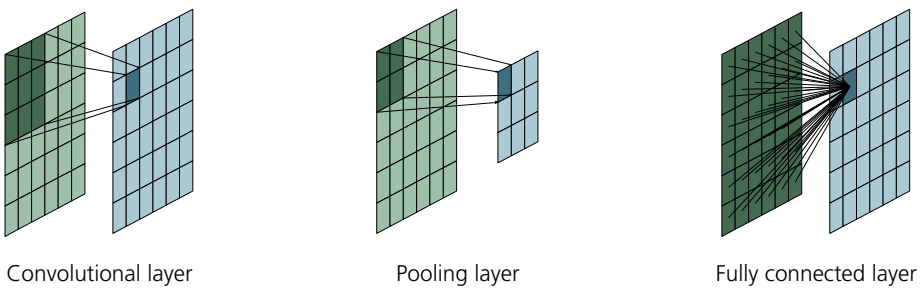


Figure 7. Examples of different ways of connecting layers.
Convolutional layers get input from a limited number of adjacent nodes and combine this information, e.g. to identify different features in the input data. Pooling layers scale down the number of nodes by combining the information in adjacent nodes. In fully connected layers, each node receives information from all the nodes in the previous layer.

An individual node of an artificial network is illustrated in Figure 8. Each node contains specific weights for the importance that is given to the input from each of the nodes in the feeding layer.[106] The input data from each input node are multiplied with the specific weight. Then the sum is calculated, and a node-specific offset is added. The value is fed into an activation function that generates the output of the node. Usually, all nodes in a layer are similar and, for instance, have the same activation function, but the weights and offsets are unique, resulting in an individual output from each node.

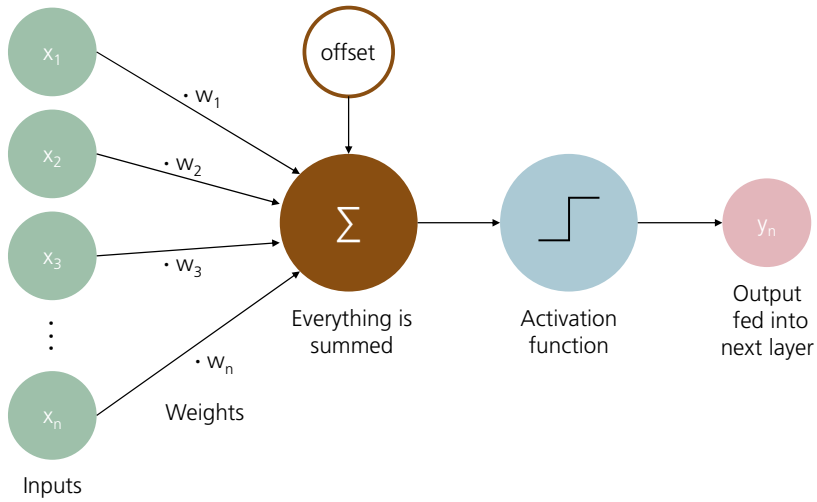


Figure 8. Illustration of an individual node in an artificial neural network.
Inputs are received from several nodes (x) in the input layer, which are multiplied by a specific weight for each input (w). The products are summed, and a node-specific offset is added. This is fed into an activation function, producing the output of the node.

The methods for training an artificial neural network vary depending on the task and the type of data. The typical task in mammography screening is the classification of images where the diagnosis of the training data is known. Thus, a supervised learning method can be used, that is, the desired result is used for training the model. This is illustrated in Figure 9. The training images are presented to the model as input, and the output is compared with the desired result by one or more performance metrics. The weights and offsets are adjusted in order to optimise the performance metrics. The process is repeated multiple times, and the model is gradually improved.

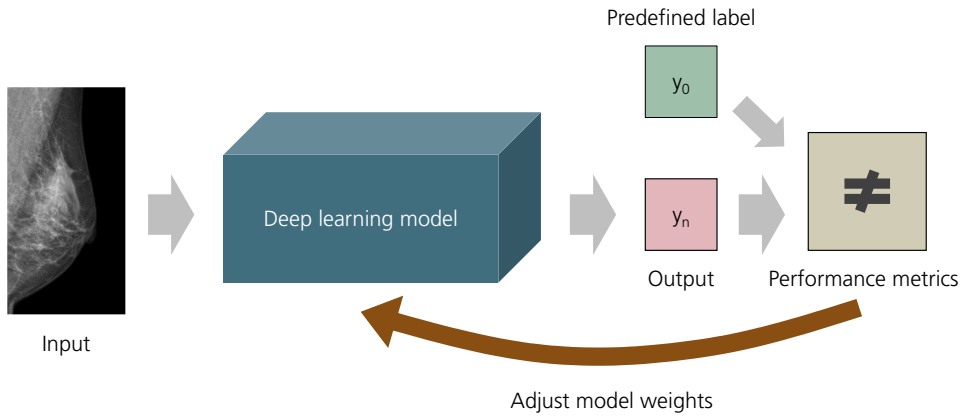


Figure 9. The training process of an image classification model using supervised learning. An example of an input image is presented to the deep learning model, and an output is produced. The output is compared with a predefined label, and one or more performance metrics is calculated, which is used to do a slight adjustment of the model weights. This process is repeated for each image in the training data and usually several times in order to optimise the performance metrics.

AI is a field that has gained huge attention in recent years, to a large extent owing to developments in deep learning. While the technology itself has been known for a long time, the computational power and methods necessary to train practically useful models first emerged in the 2010s.[107,108] The models depend on huge numbers of float number calculations in parallel, which are most efficiently calculated on graphic processors. The development and availability of this necessary hardware at reasonable prices is largely owed to enthusiast gamers.

AI in breast cancer screening

In the medical field, breast cancer screening is among the applications that are pioneering the introduction of AI and deep learning. The requirements of lots of training data are fulfilled by screening programmes, where large numbers of examinations have been collected, the ground truth – i.e. cancer diagnoses – is known, and the data are often relatively well structured.

A number of ways of using AI systems analysing DM or DBT in breast cancer screening have been proposed. The most mature niche is to try to improve sensitivity by assisting a human reader, where several commercial products are available, including ScreenPoint Transpara,[109] iCAD ProfoundAI,[110] Lunit Insight MMG,[78] Therapixel MammoScreen,[111] Vara[112] and Kheiron Mia.[113] The usage in the reading situation

resembles previous CAD systems based on classical image analysis but with improved performance.

While some commercial products for analysing DM were already available before the initiation of this thesis in 2019, the products have evolved in recent years. Only a limited number of scientific studies involving AI in breast cancer screening were published before 2019, with the vast majority being retrospective studies based on relatively small cancer-enriched datasets.[4] Also, many of the studies focused on the development of methods for cancer detection, rather than on clinical evaluation. The AI systems were usually used as standalone readers, which is a reasonable start, as it is much easier to perform studies where the AI system analyses retrospective data standalone, and a sufficient retrospective performance should be asserted before starting any prospective studies or clinical use.

Several retrospective studies of AI systems as standalone readers based on large screening datasets have been published since 2019.[78,114–116] However, the clinical role of standalone AI systems is currently a bit unclear, as AI-only reading of examinations requires a broad acceptance of AI both among women attending screening and healthcare staff. There might also be legal challenges, as the legislation regarding responsibility in healthcare often assumes that humans are making the clinical decisions, while it might be unclear who is responsible for mistakes made by an AI system. However, it has been suggested that AI can be used to prioritise which cases should be read first or triage between different reading strategies, e.g. single- or double-reading.[117,118] Several of the clinically available systems now also have versions aimed at such use.

Breast density is an important risk factor for breast cancer and has received public attention, particularly in the USA,[119] and there are several AI systems for measuring breast density.[120–122] Cancer risk models with a more comprehensive approach have also been suggested in order to predict individual cancer risk in a longer time perspective. Such models can be mainly image-based[123] or more complex and also include factors such as lifestyle and heritance.[124]

Big data

The term big data refers to collections of data that are so large that they cannot be processed by traditional methods of data processing.[125] As the digital era has now existed for about two decades, with an ever-increasing level of digital data production, many data collections have grown to a size at which they can be considered big data. Big data is particularly valuable in the training of deep learning systems, where performance

heavily depends on the amount and variation of the training data. In order to assert an intended performance, a comprehensive validation of medical AI systems is necessary before clinical implementation, which calls for the use of big data collections also in validation. In the healthcare sector, AI developers and data owners are often separated in different organisations, with academia and commercial companies on one side and healthcare providers on one side, necessitating sharing of data between organisations.

Mammography databases for research

Although years of mammography screening had led to large image collections in multiple hospitals, only a few relatively small mammography databases optimised for research purposes were available at the initiation of the project in 2018. These databases were cancer-enriched and included both digitised film mammography examinations, e.g. DDSM (2 620 cases, USA)[126] and MIAS (322 images, UK)[127], as well as DM examinations, e.g. INBreast database (410 images, Portugal)[128] and OPTIMAM Mammography Image Database (OMI-DB, 2 623+ cases, UK)[129]. Additionally, a number of internal datasets have been used for training or evaluating specific AI systems.[4]

The limited number of available databases called for the development of new research databases. Further, while cancer-enriched case collections can be valuable in some phases of training AI systems, the unrealistic proportion between cancer and normal cases might affect the performance when applying the system to real screening data. In particular, this is important when evaluating AI systems for use in screening, where it is crucial to assess the specificity in the screening situation. Thus, it is important to have databases with a large proportion of normal cases, and this should be taken into account in the creation of new databases.

Data regulations and ethics

Ethical considerations are always important in research using human data, albeit no effort by the individual is necessary for the use of data that has already been collected as part of the clinical routine. However, it is practically impossible to find each individual in order to obtain informed consent. The legislation in the ethical field might vary between different countries, but at least in Sweden, it is possible to get the requirement for informed consent waived by the ethical review board in cases when the risks for the research person can be considered negligible and there is a potential scientific gain that can help people in the future. Particularly in large database projects, where the amount of data is large and often collected from different sources, privacy and data security are of paramount importance. The rules and knowledge about

handling personal data have seen major developments in recent years due to the introduction of the General Data Protection Regulation (GDPR). Although the previous Swedish laws on the subject had many similarities, the increasing concern about handling personal data led to the development of knowledge, nomenclature and structures – in the society as well as specific organisations – and many previously unclear aspects were clarified.

Aims

The overall aim of my research is to improve the breast cancer screening with AI in order to identify more cancers earlier. Specifically, the aims are as follows:

- To create a platform to make future research on breast cancer imaging more efficient.
- To investigate whether AI can improve the performance and resource efficiency of DM screening.
- To investigate how AI can make it possible to introduce DBT in screening without excessive additional demand for reader time or other resources.

Methods

Study populations

The studies in this thesis are all retrospective and based on breast images collected in Malmö. Paper 1 introduces the Malmö Breast ImaginG (M-BIG) population, which includes breast examinations with DM and DBT from 2004 through 2020. Setting up a large research database is a complex and time-consuming process, and although the database project was already initiated before I started my thesis work, the database has not been ready for use until the very end of my work. Thus, most of the papers in my thesis, i.e. Papers 2–5, are based on data collected during the Malmö Breast Tomosynthesis Screening Trial (MBTST)[47], which now encompass a portion of the M-BIG population.

Malmö Big ImaginG database – M-BIG

The project of building the M-BIG database was started in order to create a solid foundation for future studies where a large collection of organised and accessible data is necessary. This includes AI in breast cancer screening as well as cancer risk models and long-term follow-up of the breast cancer screening programme. The M-BIG database encompasses all DM and DBT examinations, screening or clinical, performed in Malmö from 2004 through 2020.

The M-BIG database contains the cohort from the MBTST, which is a clinical trial that evaluated the use of DBT in breast cancer screening compared to DM and was run in Malmö from 2010–2015. This cohort is further described in the next paragraph. The M-BIG also contains the MBTST control group. In these previous studies, much of the data collected during 2010–2015 were collected and organised in a structured form. However, for all other examinations, the M-BIG database relies on data collected from different registries. This includes the regional cancer registry and the National Quality Registry for Breast Cancer (NKBC). While the quality of the registries is usually good, some data might be missing, particularly for the oldest cases. Further, some data are not available from any registries, e.g. data on adherence to the screening

programme, which is needed for differentiation between interval cancers and cancers diagnosed outside of the screening programme.

The development of the M-BIG database is the focus of Paper 1, and more details on the population are given in the article.

Malmö Breast Tomosynthesis Screening Trial – MBTST

The MBTST was performed during 2010–2015. At the launch of the study, DBT was a relatively new method and had not been studied as a method for breast cancer screening. A few other studies also started about the same time, all using either a combination of DM and DBT or two-view DBT, which both led to a higher radiation dose compared with DM. In order to keep the radiation dose equivalent to the current standard two-view DM, the MBTST was designed to investigate whether one-view wide-angle DBT (MLO) can be used for improving the sensitivity of breast cancer screening without losing specificity. The rationale was that the depth view of wide-angle DBT is sufficient for examining the breast in only one view; thus, the second view can be left out.

A random selection of one-third of the women in the breast cancer screening programme in Malmö were invited to take part in the MBTST. Of the invited women, 68% gave informed consent to participate, resulting in a total of 14 848 women in the study.[47] Each woman was examined during the same visit with both two-view DM and one-view wide-angle (50°) using a Siemens Mammomat Inspiration. The DM and DBT examinations were separately double read and scored according to Swedish routine on a scale of 1–5, and the readers could flag the examinations for discussion. A slight difference from the usual routine was that all examinations were discussed at a consensus meeting before recall. The consensus meeting was collective for both the DM and DBT reading arms. Recalled cases were marked as recalled on either DBT, DM or both depending on whether the initial reading decision called for discussion on a consensus meeting or not. The reading times were not recorded in the MBTST, as the reading workflow included a number of additional steps and thus would not be representative of routine reading. Most of the examinations were manually classified for breast density following 4th edition of BI-RADS density classification, i.e. four categories based on percent fibroglandular tissue (<25%, 25%–50%, 50%–75%, and >75%).

For all the cancer cases, different characteristics were registered in a structured form. This included histological type (invasive ductal cancer, invasive lobular cancer, tubular cancer, ductal carcinoma in situ, and other invasive cancer), presence of lymph node

metastases, histological grade (invasive cancers) and nuclear grade (in situ cancers). This information was collected from pathology reports. Data were also recorded on radiological tumour size and radiological appearance (spiculated mass, circumscribed mass, microcalcifications, architectural distortion and enlarged lymph nodes).

The MBTST cohort has been the basis for a series of previously reported studies, initially mainly focused on the sensitivity and false positives of DBT screening[47,130–132], but also including aspects related to breast density[133,134] and cancer characteristics[135]. During the work on this thesis, further studies have been performed based on the cohort, in particular related to interval cancers[136], false-positive recalls compared with a control group screened with DM only[137], and cancers detected during the following screening rounds (not yet published). This means that a comprehensive amount of data about the cohort has been collected, including the number and characteristics of screening-detected cancers and interval cancers. Further, the image data for DBT examinations and the majority of DM examinations were stored in a research-accessible form.

Papers 2 through 5 all study different aspects of using AI based on the MBTST cohort.

Database tools

There are several different ways to implement a research database. In the beginning of the database project, we evaluated several different solutions. The main purpose of the research database was to make access to images easier and more efficient than when the images are stored in a clinical picture archiving and communication system (PACS). Further, the requirements set by ethical approval made it necessary to have a pseudonymisation framework.

A separate research PACS could have been a possibility, either by using a commercial PACS or a free open-source solution. Due to the differences in requirements and needs between our purposes and a general PACS system, any PACS system would include unnecessary functions and lack important functions. Commercial products would likely be expensive, including both licensing costs and consultant costs for any adaptions, and it might also be hard to integrate into a closed product. Some of these limitations could be overcome by using an open-source system. However, extending an existing system with necessary functions always has barriers, limitations and risks.

The implemented solution followed an approach in which the basic parts were implemented first and then more functions could be added gradually. The first step was a file storage solution where all the examinations were simply stored as Digital imaging and

communications in medicine (DICOM) files. After identification of relevant examinations, these were exported from the clinical PACS to the file storage server. A directory was created for each person containing all the files. In order to make it possible to identify and access specific examinations and also make the files searchable in more complex ways, all files were scanned with a MATLAB script and indexed in a registry. This was saved to a PostgreSQL database together with selected metadata from the DICOM headers. In the next step, clinical data were imported to the PostgreSQL database from different sources, including medical records and registries. This made it possible to integrate data from all different sources to select and extract specific pieces of information or examinations. A pseudonymisation pipeline was implemented as part of the framework for loading images and clinical data into the PostgreSQL database, where all identifiers were replaced by randomly generated pseudonymised identifiers. The improved knowledge and structures around data regulations following the introduction of the GDPR was valuable in creating the database and probably led to a more solid structure with better protection of personal data.

Evaluation of ways of using AI in screening

In all studies in which the use of an AI system is evaluated, the selection of the AI system poses an integral part of the study. The results may or may not be similar to what would have been achieved if another AI system were to be used, depending on the similarities of the AI systems for that specific purpose.

ScreenPoint Transpara

For this thesis, the commercial product ScreenPoint Transpara was used in the studies described in Papers 2–5. The system is one of the most established products in the field and is probably also the most scientifically studied system in independent studies.[109,138–141] Since there is a continuous development of AI systems with frequent releases of updated versions, two different versions were used in in the papers, with Transpara 1.4 used in Paper 2, while Transpara 1.7 was used in Papers 3–5.

The system, which is illustrated in Figure 10, uses a combination of machine learning technology and hand-crafted rules to identify and grade suspicious areas in the images. Soft tissue lesions and calcifications are analysed separately. Each suspicious area is given a score from 1 to 100, where 100 indicates the highest risk of cancer. Areas with scores above a predefined threshold, at about 35 for calcifications and 55 for soft tissue lesions, are recorded as findings, while those below the threshold are discarded. All the

findings can be seen in the user interface if clicked directly, but those with high scores are also presented as CAD marks and can be shown by clicking a button in the user interface. The thresholds vary slightly between different versions of Transpara.

Figure 10. Screenshot of ScreenPoint Transpara Viewer.
In this example, the AI system has identified an area with a suspected malignancy (indicated with circles), which has gotten high finding scores (91 and 95) and led to a high examination score (10). This corresponds to a DM screening-detected 20 mm large invasive ductal carcinoma (IDC).

All findings in all images of an examination are combined into a composite score for the full examination using a proprietary calculation algorithm. This examination score ranges from 1 to 10, where 10 indicates the highest risk of cancer. The score is calibrated to place approximately 10% of each score in a screening material. However, this may vary depending on the characteristics of the examinations, such as number of images in each examination and type of mammography equipment.

The examination score is usually presented as an integer value, which is derived from a decimal value through rounding upwards. The decimal score with several decimals is available for export from the system, and this was used in all the studies.

Data management

Even if the size of the MBTST dataset used in Papers 2–5 may seem small compared with the M-BIG dataset, it still consists of a lot of data, and some tools are necessary to handle this data efficiently. At the time when the projects of Papers 2–5 were performed, the infrastructure of the database that was implemented as part of the M-BIG

project described in Paper 1 was not available. Thus, these projects were performed by using data collected in the MBTST project at the time available in a separate infrastructure. The data were handled with project-specific scripts and data structures in MATLAB. This included orchestrating automated processing of the data, e.g. sending all examinations to Transpara, gathering the results and to verify that the results were valid and complete. There was also a substantial amount of data handling even before the AI analyses could be started, for example the DBT examinations at the research-accessible image storage were stored as raw data and had to be reconstructed to a suitable format (i.e. 'for presentation') before they could be processed by Transpara.

Statistics

Diagnostic performance

When studying the performance of a diagnostic test, sensitivity and specificity are both basic measures. Sensitivity describes the ability of the test to detect the target condition, such as what percentage of all sick individuals are detected by the test.[142] On the other hand, specificity describes the ability to detect the absence of the target condition, i.e. what percentage of healthy individuals are correctly identified as healthy by the test. The positive predictive value is the proportion of positive diagnostic tests that are true positives. Conversely, the negative predictive value is the proportion of negative diagnostic tests that are true negatives. All the values rely on comparing the diagnostic test with the truth. However, the truth is often not easily defined, and commonly it is necessary to settle on another better diagnostic test – if available – or simply see what happens during a long follow-up time.

Descriptive statistics (Papers 1–5)

Most of the data in the papers included in this thesis were binary, i.e. each woman either had cancer or was healthy, which meant that the data had a binomial distribution. For the purpose of calculating confidence intervals, this could have been handled in a number of different ways. In Paper 2, the binominal distribution was approximated by the normal distribution, and Wald intervals were used. In Papers 3–5, the confidence intervals were instead calculated as Clopper–Pearson intervals, which use the binominal distribution. When the number of cases is reasonably large, a normal approximation can be used, and the differences should be limited. In some cases, Wald intervals can include negative values, which can never appear in a binomial distribution,

and thus Clopper–Pearson intervals might be a better alternative. Clopper–Pearson intervals, on the other hand, might be excessively conservative, and there might be better alternatives to calculate confidence intervals from the binomial distribution; however, Clopper–Pearson intervals is the most widely used method. All confidence intervals were calculated with the commonly accepted 95% significance level.

Statistical tests

Receiver operating characteristic (ROC) analysis

Receiver operating characteristic (ROC) analysis is a commonly used method for analysing the ability of a diagnostic test to discriminate between two states, usually having a specific condition or not. While originally developed for measuring the performance of radar operators during the Second World War, it is now an important tool in studies of diagnostic tests in different fields, including medicine.[143] The ROC analysis is a graphical plot of the relation between the true positive rate and the false positive rate for all possible reader operating points. In many cases, the area under the curve (AUC) is used as a simple way of quantifying and comparing different tests. However, for many applications, the shape of a specific segment of the curve is of the most importance, and thus, only comparing the AUC is not enough and a visual comparison is preferable. ROC-analyses were included in Papers 3 and 5.

McNemar's test

When comparing two different diagnostic methods performed on the same individuals, i.e. paired data, McNemar's test is one way to test for differences in a specific measure. McNemar's test tests the null hypothesis that there is no real difference between the methods. The calculated value follows the χ^2-distribution and can also be transformed into a p-value. McNemar's test was used in Papers 3 and 5.

Kruskal–Wallis test/one-way analysis of variance (ANOVA) on ranks

The Kruskal–Wallis test is a method to assess if two or more samples come from the same distribution, or if one sample stochastically dominates another sample, i.e. the values in one sample are always higher than in another sample. The test is non-parametric and can be applied to values where normal distribution cannot be assumed. The Kruskal–Wallis test was used in Paper 3 to test for differences in AI scores between breast cancers with different characteristics. The AI scores sometimes have a stepwise and skewed distribution, and thus, normal distribution cannot be assumed.

Summary of papers

Research database

Creation of a database for radiological breast imaging research (Paper 1)

The focus of Paper 1 was the creation of a database for breast cancer research with a focus on radiological imaging where all DM and DBT examinations that have been acquired in Malmö from 2004 to 2020 were collected. The data collection also included reading results for screening examinations and free text radiological reports for clinical examinations as well as information about cancer and treatments from different registries, for instance the regional cancer registry and the NKBC.

A purpose-built platform was created where data from different sources can be linked together for searches and data extraction. The platform also contains tools for the pseudonymisation of personal data. A total of 449 000 examinations from 103 000 women were included in the database, and it contains consecutive screening examinations with a follow-up of up to 17 years. Screening mammography examinations dominate the database by number, with over 343 000 examinations from 84 800 women. Diagnostic mammography (64 500 examinations) and recall from screening (9 800 examinations) follow as the second and third most common examinations. Regarding screening examinations, it is most common to have only one examination (16 800 women), but there are about 9 000–11 000 women in each of the categories 2–7 examinations. The median number of screening examinations is 4, and thus, most women have multiple screening examinations, which is valuable for studying changes between several screening occasions. Almost 20 000 examinations were performed with DBT, where most originate from the MBTST, while the rest are diagnostic examinations.

A total of 9 250 breast cancers were diagnosed from 2004 through 2020 in 7 371 of the women included in the database. For 5 913 of these, relatively detailed information on cancer characteristics are available from the NKBC, while the rest only have basic cancer information, e.g. histological type, available from the regional cancer registry. This includes 1 485 cancers in 1 399 women diagnosed prior to the start of NKBC in

2008. There is also some inconsistency between the two registries in later years, with some cancers included in one of the registries but missing in the other.

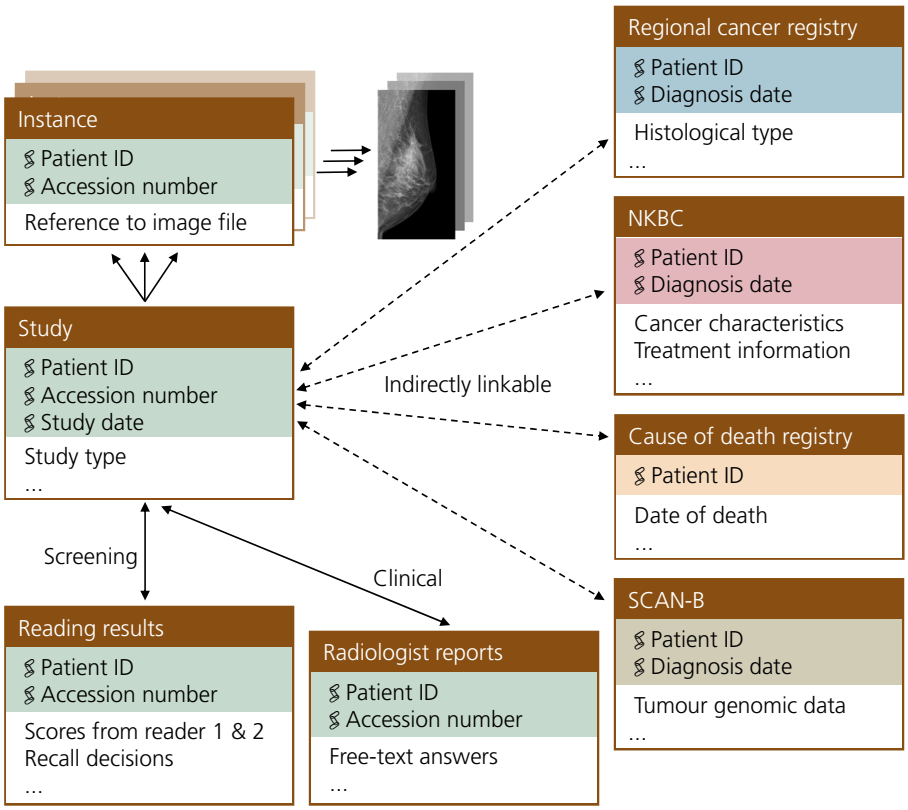


Figure 11. Linkability of data in tables imported to M-BIG from different sources. Radiological examinations can be directly linked to reading results and radiologist reports by using unique identifiers of the study, e.g. accession number. Each study usually encompasses several images, which can be identified through the instance table. The regional cancer registry, NKBC (National Quality Registry for Breast Cancer) and cause of death registry can be linked to each other and to radiology information through patient ID and dates. Data from SCAN-B (Sweden Cancerome Analysis Network-Breast) is not yet available, and the way to link to other data is thus still uncertain. Most women attending screening are by definition not patients, but here, the term 'Patient ID' is used for conformity with the DICOM names.

The data from different sources that have been imported into the PostgreSQL database as separate tables are linkable to different extents, as illustrated in Figure 11. For images and information from directly related sources, such as the radiology information system (RIS), the information can be directly linked to a specific examination by using e.g. accession number. Data from other sources are linkable through patient ID and dates,

i.e. a cancer diagnosis can be linked to an examination performed in the weeks before the diagnosis.

The database is still under development and new kinds of data are continuously being added. This includes e.g. data on cause of deaths from the national cause of death registry and genomic data for cancer cases included in the Sweden Cancerome Analysis Network-Breast (SCAN-B) database. Further, the process of further data curation is going on, with e.g. adding data on the mode of detection of cancers, i.e. screening-detected, interval cancer or cancer in non-screening participants.

Evaluation of ways of using AI in screening

Papers 2–5 of this thesis all describe studies where different potential ways of using AI in breast cancer screening were tested. All the studies were retrospective and based on the same cohort of women from the MBTST, where paired separately double-read DM and DBT examinations were available. Paper 2–4 used AI on the DM examinations, while Paper 5 used AI on the DBT examinations. A brief overview of the conceptual differences between the Papers 2–5 is presented in Table 1.

Table 1. Overview of thesis papers regarding DM and DBT for AI and radiologists, respectively.

Paper	AI on DM	AI on DBT	Radiologist DM	Radiologist DBT	Purpose
2	X*		X		Exclude normal cases
3	X		X	X	Detect additional cancers
4	X		X	Partly	Add DBT in high-gain cases
5		X	Comparison	X	Exclude normal cases

AI on DM and DBT describes whether the AI system was used to analyse DM or DBT examinations. Radiologist DM and DBT describes if the study used the radiologist reading results from DM and DBT arms, respectively. *Only about two-thirds of the dataset was used in Paper 2.

AI can identify normal DM screening examinations (Paper 2)

The aim of Paper 2 was to determine if some normal DM examinations could be safely removed from the human reading by AI. The study population encompassed a subset of 9 581 out of the 14 848 women in the MBTST cohort who were examined in the later part of the MBTST when DM examinations were stored in a form readily available for research. DM examinations from the included women were analysed with the AI system ScreenPoint Transpara 1.4, which gave an AI score ranging from 1–10 for each

examination. The DM examinations for the full MBTST cohort were not readily available when the analyses were run.

At publication, the paper was the first evaluation of an AI system on real screening material. The results showed that the DM examinations with AI risk scores of 1 or 2 can be considered to be normal and can be excluded from reading without missing any cancers, while also avoiding a few false-positive recalls. That means that 19.1% of the examinations did not have to be read by radiologists and 10 false-positive recalls could have been avoided.

AI can detect additional cancers on DM that would otherwise only be detected on DBT (Paper 3)

In Paper 3, we investigated whether AI analysing DM examinations can detect some of the cancers that radiologists missed on DM but detected on DBT. We analysed all the DM examinations from the MBTST cohort with ScreenPoint Transpara 1.7 and compared this with the results from radiologists' readings DM and DBT. The AI performance on the examination level, i.e. a score from 1–10, was analysed with ROC analyses with different definitions of ground truth: DM screening-detected cancers, DM or DBT screening-detected cancers, and all screening-detected cancers plus interval cancers. When comparing the ROC curves of AI with the operating point of radiologist DM double reading, the AI system could not reach the performance of human double reading on DM for any of the ground truth definitions.

The locations of the AI findings were compared with the actual locations of the diagnosed cancer lesions. Of the cancers diagnosed on DBT but missed on DM, 44% had matching highly scored AI findings, which indicates that at least some of these cancers could potentially be detected by using an AI system when reading DM examinations. In the same manner, 9% of the interval cancers could potentially be detected. The AI cancer detection was evaluated in relation to different cancer characteristics, and the AI score distribution was tested with Kruskal–Wallis one-way ANOVA on ranks without finding any notable differences.

High-gain cases for DBT screening can be identified by AI on DM (Paper 4)

The possibility of enhancing a DM-based screening programme by adding DBT in AI-identified high-risk cases was explored in Paper 4. This was performed by analysing the DM examination of the MBTST cohort with ScreenPoint Transpara 1.7 and simulating how different score thresholds for adding a DBT would affect the number of detected cancers and false positives. We found that by using a threshold value where

10% of the women would be examined with DBT, 25% more cancers would be detected, which is 59% of the cancer cases that were detected exclusively on DBT. The number of false-positive recalls would be increased by 22%. This means that more than half of the effect of DBT screening could be obtained by only screening 10% of the women with DBT, but there are some challenges with the proposed workflow, not the least of the logistical nature, because the scheduling of busy screening clinics would be complicated if larger variations in the examination time are introduced.

AI can speed up reading of DBT screening to make it workload equivalent with DM (Paper 5)

The aim of Paper 5 was to evaluate different ways of using AI to reduce the workload of reading DBT screening examinations. Three different workflows were tested: excluding normal cases from reading, replacing the second reader and replacing both readers. ScreenPoint Transpara 1.7 was used to analyse the DBT examinations, and score thresholds were set to levels to exclude half of the examinations from reading for the excluding normal cases approach or to keep the number of consensus discussions at the same level as with DBT double reading for the replacing the second reader and both readers approaches. The results showed that by excluding half of the examinations from reading while double reading the others, 95% of the cancer cases that were detected with DBT double reading would be detected, with retained recall frequency. If instead replacing the second reader, 95% of the cancer cases that were detected with DBT double reading would still be detected, but with a 53% increase in recalls. The approach of replacing both readers with AI could detect slightly more cancers than DM double reading with somewhat fewer recalls.

Overarching summary of results

Papers 2–5 all include AI analyses performed with ScreenPoint Transpara of DM or DBT examinations from the MBTST cohort. While there are differences in Transpara versions and whether DM or DBT examinations were analysed, the papers contain similar plots presenting the AI scores of all examinations. A compilation of these results is presented in Figure 12 in order to facilitate comparison. Since Paper 2 only included a sub-cohort of the MBTST, these results are not directly comparable; thus, the results for the full MBTST cohort using the same version of Transpara are also provided. Similarly, the scores of all screening cancer cases and DM screening-detected cancer cases are presented in Figure 13.

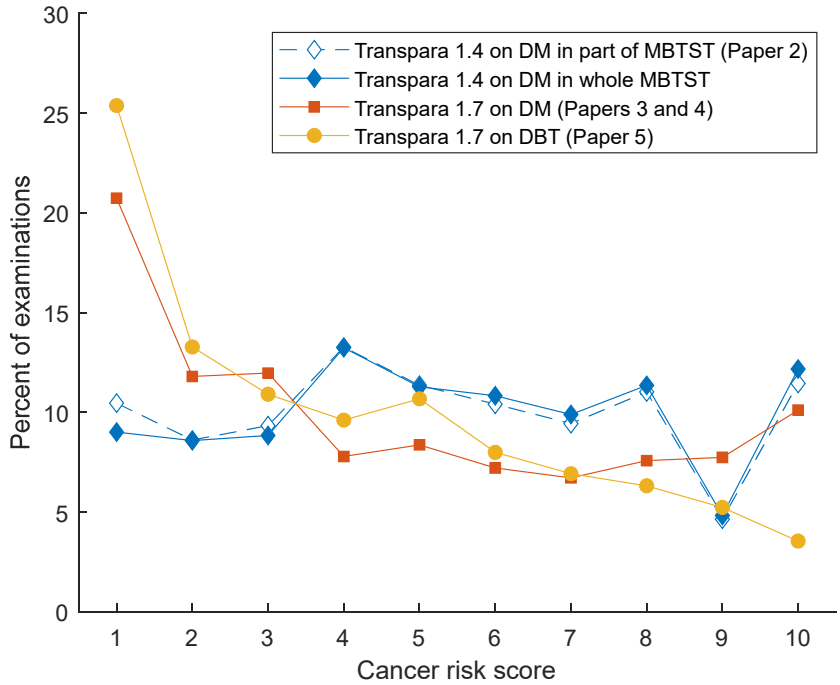


Figure 12. Distribution of cancer risk scores among all examinations in the different papers. The results from Paper 2 are not completely comparable with the other papers as a sub-cohort was used; thus, the results for the full population with the same version (Transpara 1.4) are also included for comparison.

In Figure 12, it can be seen that there is substantial difference in score distribution between the different versions of ScreenPoint Transpara, where there is a general shift towards lower scores in the later version. A further shift is seen for DBT compared to DM. Among the cancers (Figure 13), there are only small differences between the versions, but on DBT, the number of cancers with score 10 is a bit lower than on DM, while the number of cancers scored 8 and 9 increased slightly. The general redistribution towards lower scores in the newer version comes at the price of a few cancer cases among the 1 and 2 scoring categories, where there were no cancers in Paper 2.

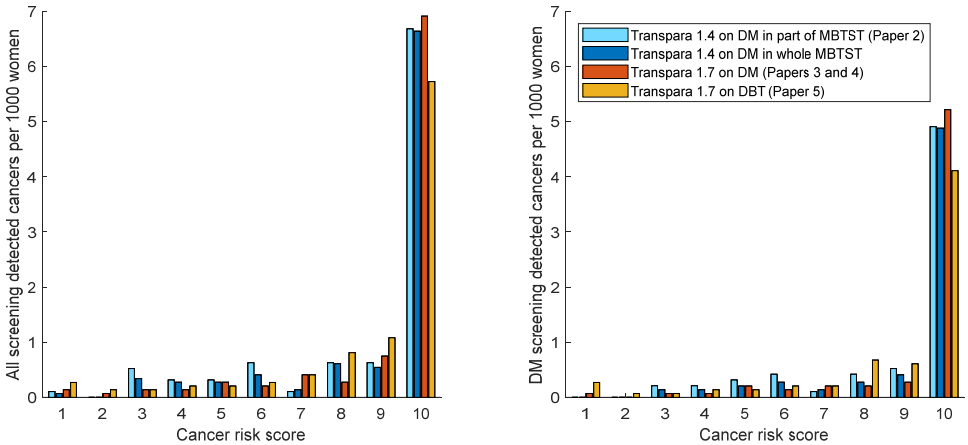


Figure 13. Distribution of cancer risk scores among cancers in the different papers
The score distribution among all screening-detected cancers, i.e. detected on either DM or DBT (left), and among the DM detected cancers only (right), are presented separately. The results from Paper 2 are not completely comparable with those of the other papers as a sub-cohort was used; thus, the results for the full population with the same version (Transpara 1.4) are also included for comparison.

In Figure 14, the ROC curves for Transpara 1.7 on DM (Paper 3) and DBT (Paper 5), respectively, are presented together. The figure also includes operating points for radiologist single and double reading with and without consensus for DM and DBT. The AUC for AI on DBT is higher than for DM, but the curves are relatively close to each other and are crossed a few times. While the AI system has results largely on par with the human readers on DM, only surpassed by double reading with consensus, the human readers are clearly superior on DBT compared to AI on DBT.

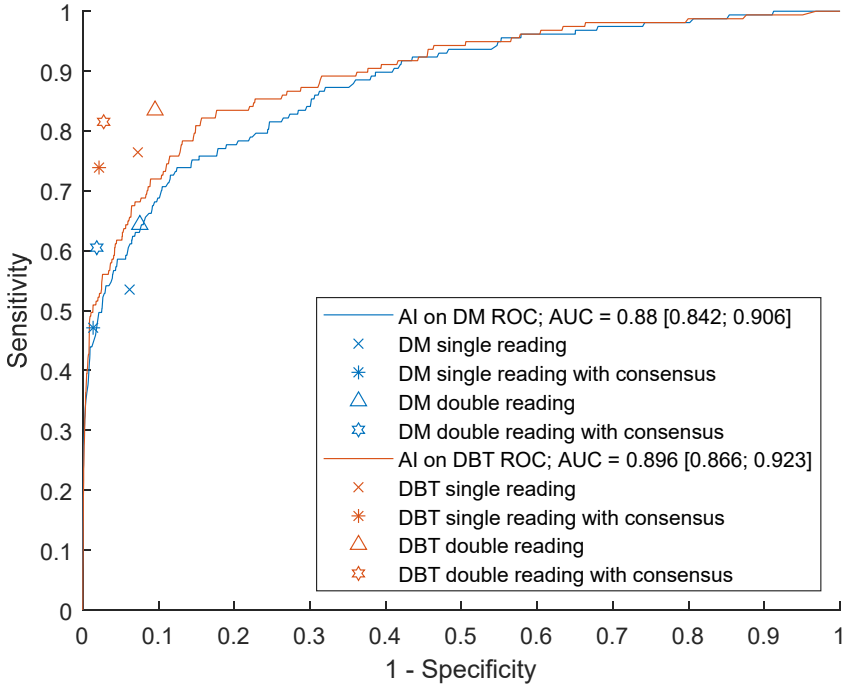


Figure 14. ROC for Transpara 1.7 for DM and DBT.
Ground truth defined by all screening-detected cancers and interval cancers during follow-up until the next screening round. Corresponding operating points for radiologist single and double reading with and without consensus are provided for reference.

Discussion

This thesis includes two types of papers, where establishment and implementation of a research database is described in Paper 1, while Papers 2–5 all evaluate different ways of using AI in breast cancer screening. The two categories of papers are first discussed separately, followed by a discussion of more overarching aspects.

Breast imaging databases for research (Paper 1)

Paper 1 describes the creation and basic characteristics of the mammography research database M-BIG, which in a research-accessible form collects 449 000 DM and DBT examinations from 103 000 women performed during a span of 17 years. At the initiation of the project, only a few research databases existed, most of which were small and with several limitations. Many previously published case collections and databases contained exclusively or mainly cancer cases, which poses limitations both in the development and evaluation of AI systems.

Table 2. Mammography research databases and studies using large internal datasets

Name	Examinations	Women	Cancers	Centres	Countries	Vendors	Years	Usage	Type	Level of annotations
Dedicated databases										
M-BIG	449 000	103 000	9 250	1	Sweden	Siemens, GE, Fuji	2004–2020	Database	All	Examination
OMI-DB[144]	373 000	170 000	8 586	3	UK	Hologic, GE, Siemens	2010–2020	Database	Screening	Polygonal
CSAW[145]	1 180 000	499 800	10 582	4	Sweden	Hologic	2008–2015	Database	Screening	Pixel
VAI-B[a][146]	105 000+	44 000+	8 000+	3+	Sweden	N/A	2008–2021	Database	Case-control	Examination
Specific studies										
Lunit[b][78]	170 230	Unknown	36 468	5	South Korea, USA, UK	GE, Hologic, Siemens	2000–2018	Training / validation	Cancer-enriched	N/A
Kheiron[113]	280 594	180 542	2 783	7	UK, Hungary	Hologic, Siemens, GE, IMS Giotto	2009–2019	Validation	Screening	N/A
Vara[112]	738 107	322 381	14 858	8	Germany	Siemens, Hologic, Fuji	2007–2020	Training / validation	Screening	N/A
Lauritzen et al. / ScreenPoint[116]	114 421	114 421	1 118[c]	5	Denmark	Siemens	2014–2015	Validation	Screening	Examination
Larsen et al.[147]	122 969	48 877	752	4	Norway	Siemens	2009–2018	Validation	Screening	Examination
Marinowich et al.[148]	108 970	108 970	995	N/A	Australia	Siemens	2015–2016	Validation	Screening	Examination
Hickman et al.[149]	78 849	78 849	1 326	2	UK	GE, Philips	2015–2018	Validation	Screening	Examination

a) Some overlap with M-BIG and CSAW.
b) Potential overlap of 6 543 examinations with OMI-DB.
c) +1 473 long-term cancers.
d) +688 subsequent screening-detected cancers.

Other databases and case collections

The attention paid to AI and deep learning in the field of breast imaging has led to a rising interest in creating large image databases in order to facilitate the development and evaluation of such systems. Thus, a number of new databases and large case collections emerged during the project time, and some of the largest ones are presented in Table 2.

The OMI-DB has been substantially expanded and now includes 373 000 examinations from 170 000 women collected from several screening centres in the UK.[144] A database has been created containing about 1.2 million examinations from 500 000 women aged 40–74 years of age in the Stockholm region (Cohort of Screen-Aged Women, CSAW).[145] Another project which is called VAI-B (validation platform for AI in breast radiology) and is led from the Karolinska Institute in Stockholm, collects mammography data from different Swedish regions. It focuses on women in screening ages with the goal of creating a platform where several AI systems can be evaluated on locally collected data in order to identify suitable systems prior to procurement and to verify expected performance of version updates before clinical installation.[146] As Malmö is one of the participants in the VAI-B project, there is a partial overlap between the VAI-B and M-BIG databases.

Apart from the databases that have been described in dedicated publications, there are several internal, more purpose-built case collections that have been used to train or internally validate different AI systems (Table 2).[78,112,113] These are usually less well-described and are in some cases at least partially cancer-enriched. Further, as these case collections were often created on commercial grounds, they are unlikely to be available for external collaborators, which is usually the case for dedicated databases, at least to some extent. Further, a number of studies independently evaluating AI systems have used large datasets, which could potentially be used for other purposes in the future, e.g. evaluating competing AI systems (Table 2).[116,147–149]

The large number of normal cases is demanding, both with respect to storage and computational power. Thus, some databases, such as the VAI-B, have chosen to include cancer-enriched datasets, where all available cancer cases are included, but normal cases are restricted to a random subset using case-control approaches. This can be rational if the cancers are the main focus and resource limitations do not allow for including all the available examinations. On the other hand, when evaluating AI systems for use in screening, the specificity is often at least as important as the sensitivity, and thus, it is important to also thoroughly test the AI system on normal cases. Full screening datasets give the opportunity to retrospectively validate AI systems in datasets with a cancer

frequency that is realistic for screening, and using weighting factors in calculations and analyses can be avoided.

Strengths and limitations with the M-BIG database

The M-BIG database opens possibilities for numerous retrospective studies, including large-scale image-based studies using AI, but also studies aggregating clinical data. The database can be used as a basis for creating reader studies targeting specific questions or relatively rare conditions, thanks to the large amount of data. The M-BIG database spans a longer time period than most other databases, which is a potential advantage in doing longitudinal studies, for example, comparing AI scores between consecutive examinations. As the database contains all mammography examinations, and is not restricted to just screening examinations and screening-detected cancers, it is also possible to study cancers that are not detected in screening, e.g. in women above screening ages or women not taking part in screening.

In contrast to most of the other databases, the M-BIG database also contains DBT examinations and opens up possibilities for studying DBT, though the study possibilities related to DBT are more limited than DM, as the number of screening DBT examinations is much smaller (i.e. 14 848 acquired in the MBTST). However, it is possible to do studies using a DM examination prior to the DBT screening, e.g. investigating if AI on previous DM can be used to select high-gain cases for DBT. This would be similar to what was described in Paper 4, but radiation dose and logistical issues could be reduced.

The M-BIG database benefits from the well-developed Swedish registries and, thus, contains information about all breast cancer diagnoses in the included cohort, even if the woman has moved to another part of the country. As the database contains all examinations from the only breast radiology clinic in Malmö, the population should be well represented, spanning all socioeconomical groups. Further, the population of Malmö is diverse, with many ethnicities represented. Another unique feature of the M-BIG database, which is currently in progress and will be included, is the linkage to data from the SCAN-B project, including tumour genomic data and gene expression analyses from most cancer breast cases diagnosed from 2010 onwards.

There are also a number of limitations with the M-BIG database compared with the other databases (Table 2). Although the M-BIG is among the larger databases, the number of examinations, included women and cancers are still larger in some of the other databases. As the data inclusion was restricted to one clinic, the generalisability might be less than some of the databases with a broader inclusion. In contrast to some

of the other databases, the M-BIG currently does not contain any annotations of the locations of cancer lesions, and annotations are restricted to the examination level. The M-BIG database is currently missing information regarding the mode of cancer detection for non-screening-detected cancers, i.e. interval cancer or cancer in a woman not attending screening, but the work of adding this is in progress.

With its size and unique characteristics, the M-BIG database is an important complement to other available databases. While the database in its current form can be a valuable resource to pursue research projects in an efficient way, at the time of this writing, some additional data curation and annotations are still necessary in order to unleash its full value, for instance, annotating locations of cancers and identifying interval cancers and other modes of detection.

AI to enhance breast cancer screening (Papers 2–5)

In this thesis, several ways of using AI to improve breast cancer screening have been presented: speeding up reading of DM by excluding normal cases (Paper 2), increasing sensitivity on DM (Paper 3), increasing sensitivity through selective addition of DBT (Paper 4) and using AI to speed up reading of full DBT screening (Paper 5).

Speeding up reading of DM with AI (Paper 2)

In Paper 2, we used an AI system to analyse DM examinations from screening, focusing on excluding normal cases from manual reading. At the time of publication, it was among the first studies evaluating an AI system on a real screening dataset. A few studies reporting novel AI systems had included evaluations on screening datasets that in some cases partly had been used in the development of the AI model.[78,115,150] These studies focused on a general comparison between AI and radiologist performance and are thus not completely comparable to Paper 2, which had a more specific focus related to using AI in reading workflow and did not include an ROC analysis.

A number of recently published studies have focused on strategies for making the reading workflow more efficient. As there are uncountable potential workflows, comparing these studies tend to be complex due to differences in design. Some studies have simulated workflows where the AI system is used to completely exclude normal cases from reading while the rest are double read, similarly to Paper 2; this is, however, usually combined with sending the cases with highest level of AI suspicion either to consensus discussion or directly to recall.[116,139,147,151] Compared with Paper 2, these

studies classified a larger proportion of the examinations as normal and discarded them from radiologist reading, in most cases about 60%–70%, but have in some cases also accepted a small proportion (1.5%–2.3%) of missed cancers.[116,139,147,151] If using a threshold excluding 53% of the cases from reading in Paper 2, 10.3% of the screening-detected cancers would be missed which is more than in the other studies. The differences might to some extent be due to use of more modern AI systems as these studies were performed a few years later. However, it is always hard to guarantee that no cancers at all will be missed, regardless of the threshold. The result in Paper 2 showing that 19% of the examinations can be excluded from normal reading without missing any cancers at all is probably mostly a matter of random variation and does not appear in Paper 3, where a more complete version of the same dataset was analysed with a newer version of the same AI system. A more aggressive approach regarding workload reduction is to exclude normal cases from reading while single reading the high risk cases, but this led to over 10% missed screening-detected cancers also in recently published studies.[112,147] However, this might be an alternative for DBT-based breast cancer screening, as the gain in sensitivity from moving from DM to DBT would still result in a higher sensitivity compared with double-read DM.

Replacing the second reader with AI

A related but more conservative approach than excluding normal cases from reading – not the least in terms of psychological and legal aspects – is to replace one of the readers with AI, which has been studied in several retrospective studies summarised in Table 3.[113,147–149,152,153] There are some variations in proposed workflows, where the AI system can send high-risk cases either to a second reader or consensus meeting. The table also includes two prospective studies, each with a design representing one of the two approaches.[85,86]

In studies where AI sends high-risk cases to double reading, the proportion of examinations sent to double reading varied between 13%–65%. This means that the potential reduction in reading workload has a wide range. To some extent, this is related to the study design and performance of specific AI systems, but it is also heavily dependent on how many missed cancers can be acceptable. If focusing only on cancers detected on DM with double reading, as mentioned earlier, the AI system can by design not reach a higher sensitivity than the two human readers. In studies where 13%–24% of the cases are sent to double reading, the proportion of missed screening-detected cancers varied from 0%–10%.[113,148,153] However, detecting some of the interval cancers can at least partly compensate for the loss of screening-detected cancers.

Table 3. Studies of workflows in which the second reader is replaced by AI

	Name	Double reading rate	Consensus meeting rate	Recall rate without AI	Recall rate with AI	Missed screening-detected cancers
AI triages between single reading or double reading	Larsen et al.[147] (scenario 8)	30%	7.2%	3.2%	2.9%	0.8%
	Balta et al.[153]	23.6%	11.1%	5.5%	4.8%	0.0%
	Ng et al.[113a] (workflow C)	13.0%	2.0%	6.3%	5.8%	N/A (2.1%)[b]
	Marinovich et al.[148]	20.2%	N/A	3.4%	3.1%	9.6%
	Hickman et al.[149c] (scenario A)	44%–65%	2.6%	3.5%	3.4%	0%–0.1%[d]
	*Lång et al.[85]**	*13.7%*	*4.0%*	*2.0%*	*2.2%*	*N/A (20% more detected)*[e]
AI sends to consensus	Larsen[147] et al. (scenario 2)		14%	3.2%	2.8%	3.1%
	Ng et al.[113] (workflow B)		13%	6.3%	5.8%	N/A (2.1%)[b]
	Leibig et al.[112] (pathway B)		N/A	N/A	N/A	2.6%
	*Dembrower et al.[86]**		*8.95%*	*2.93%*	*2.80%*	*0 (4% more detected)*

* Prospective studies (in italics).
a) Results from two different populations (UK and Hungary) are separately presented in the study, and the overall results have been calculated and included in the table.
b) Results for screening-detected cancers are not presented in the paper. The number in the table refers to the difference between double reading and single reader+AI, which includes interval cancers.
c) Comparison of three different AI systems where the range of results is presented in the table.
d) Missed screening-detected cancers defined as <1% and double reading rate adjusted accordingly.
e) By study design, the number of missed screening-detected cancers cannot be defined, but 20% more cancers were detected in the AI group than in the control group receiving traditional double reading without AI.

A recently published randomised clinical trial by Lång et al. compared standard of care double reading, with AI triaging between single reading (86.3%) and double reading (13.7%), where all readers had access to the AI results, which is a more realistic way of clinical use, but is impossible to simulate in retrospective data.[85] The design did not allow for direct comparison where the number of missed screening-detected cancers could be calculated, but as 20% more cancers were diagnosed in the AI arm, this seems to be at least as good as double reading. A more definite measure for comparison between the workflows would be the interval cancer rate, which will not be available until the follow-up time has passed for all the participants in the trial.[154]

A few retrospective studies also included workflows where the AI system sent high-risk cases directly to consensus meetings instead of a second reader.[112,113,147] Obviously, this would further reduce the reading workload, but instead, the number of the usually more resource-intense consensus discussions would be increased, while more cancers would be missed. A recently published prospective study by Dembrower et al. followed this approach, i.e. the examinations were read by a single reader complemented by an

AI system working standalone and sending cases with an AI score above a predefined threshold to consensus meeting. This study actually showed a higher cancer detection when using single reading + AI than in traditional double reading.[86] The AI results were not used in the initial readings, but in contrast to the retrospective studies, the consensus meetings could benefit from the AI results, and this was apparently valuable.

Some retrospective studies have also included the possibility for the AI system to directly recall cases with the highest risk. When including interval cancers, where the AI system has the chance to detect additional cancers that were not detected on double-read DM screening, some studies reported that the total sensitivity could instead potentially be increased.[147,148]

Increase sensitivity on DM with AI (Paper 3)

In Paper 3, the main focus was to investigate if part of the higher sensitivity with DBT screening can be achieved by using AI on DM. This can only be studied in a dataset with paired DM and DBT examinations. A Spanish study had some similarities, as using a paired DM and DBT and the same version of Transpara, but differed by using two-view narrow-angle DBT and a higher AI score threshold for considering a cancer as detected.[155] The study showed that the AI system could detect 45.5% of the cancers, which human readers only detected on DBT, could have been detected on DM with AI. This is very similar to the corresponding value of 44% in Paper 3. The study focused on evaluating AI as a standalone sole reader, which could potentially be implemented in the screening workflow but would also lead to a substantial number of missed cancer cases that would be detected with double reading. The approach in Paper 3 is of a more explorative nature, and no concrete way of implementation in clinical workflow is investigated. The results should instead be seen as an estimation of the highest achievable sensitivity when using the AI system in DM screening, at least regarding cancers that are actually present and potentially detectable, in contrast to cancers detected in the next round, where at least a portion of them might have emerged since the index screening. In a prospective clinical setting, the potential detection of the additional DBT-only detected cancers comes with a risk of increasing the false-positive recalls. Thus, it would be interesting to compare this with prospective studies comparing DM+AI with DBT.

Another recently published Spanish study compared reading of DM and DBT with concurrent use of an AI system for both DM and DBT examinations with historical data from the same population prior to introduction of AI support.[156] DM and DBT are used in parallel in screening depending on available resources. This design can in many respects be considered as a prospective study, although some other factors might

have changed since the collection of the control group. Sensitivity cannot be calculated at this time, as not enough time has passed for interval cancers to be diagnosed. The cancer detection rate for DM+AI was higher than for DBT without AI (8.1‰ compared with 5.8‰), but at the cost of a higher recall rate. Such a strong effect from AI seems unlikely according to our results in Paper 3. However, the data in the study differed substantially from our study in that the difference in cancer detection rate between DM and DBT was negligible, while the MBTST showed a 34% higher cancer detection rate with DBT. A previous study comparing DM and DBT without AI in the same population showed a 17% increase in cancer detection with DBT compared with DM – an effect that seems to have diminished, while the recall rate for DM has doubled.[48,156] One speculation is that in light of the shown superiority of DBT, the local readers may have lost confidence in DM and are now likelier to recall DM in order to compensate for this inequality. Another part of the difference could be related to the fact that the study is based on examinations collected with narrow-angle DBT (Hologic), while MBTST used wide-angle DBT (Siemens), but differences in the screening programme and underlying population should also be considered.

Table 4: Comparison with previous studies of AI on DM

Name	Year of publication	AI system	Country	Type of data	AUC	AI results compared with radiologist		
						Single reading	Double reading	Consensus
McKinney et al.[150]	2020	Google Health / DeepMind	UK	Screening	0.889	Better	Worse	Worse
McKinney et al.[150]	2020	Google Health / DeepMind	USA	Cancer-enriched	0.811	Better	N/A	N/A
Salim et al.[114]	2020	Undisclosed (Lunit)	Sweden	Cancer-enriched	0.956	Better	N/A	N/A
Salim et al.[114]	2020	Undisclosed	Sweden	Cancer-enriched	0.922	Worse	N/A	N/A
Salim et al.[114]	2020	Undisclosed	Sweden	Cancer-enriched	0.92	Worse	N/A	N/A
Schaffter et al.[115]	2020	Undisclosed (Therapixel)	USA	Screening	0.858	Worse	Worse	Worse
Schaffter et al.[115]	2020	Undisclosed (Therapixel)	Sweden	Screening	0.903	Worse	Worse	Worse
Kim et al.[78a]	2020	Lunit	South Korea/UK/USA	Cancer-enriched	0.959			
Romero-Martín et al.[155b]	2021	Transpara 1.7.0	Spain	Screening	0.93	Better	Better	N/A
Lotter et al.[157c]	2021	DeepHealth	UK/USA/China	Screening / diagnostic	0.927–0.971			
Leibig et al.[112d]	2022	Vara	Germany	Screening	0.944	Worse	Worse	Worse
Lauritzen et al.[116e]	2022	Transpara 1.7.0	Denmark	Screening	0.97			
Paper 3	2022	Transpara 1.7.0	Sweden	Screening	0.925[f]	Better[g]	Equal[g]	Worse[g]
Marinovich et al.[148]	2023	DeepHealth Saige-Q 2.0	Australia	Screening	0.83	Worse	Worse	N/A
Hickman et al.[149]	2023	Undisclosed	UK	Screening	0.87			
Hickman et al.[149]	2023	Undisclosed	UK	Screening	0.89			
Hickman et al.[149]	2023	Undisclosed	UK	Screening	0.90			
de Vries et al.[158]	2023	Kheiron Mia 2.0	UK	Screening	0.95			

a) Trained on other portions of same datasets.
b) DBT screening-detected cancers and interval cancers included in ground truth.
c) Trained on other portions of same datasets. AUC is reported separately for different datasets; thus, the range of results is presented in the table.
d) Only screening-detected cancers included in the dataset. Cases without normal follow-up were excluded.
e) Dataset included interval cancers and long-term cancers, but results refer to screening-detected cancers.
f) Ground truth based on DM screening data only.
g) Cancers detected on DM or DBT and interval cancers included in ground truth, see Figure 14.

Standalone performance of AI on DM

Paper 3 also includes a general comparison with reader results, including ROC analysis, which can be compared with other studies where an ROC analysis has been included. An overview of such studies and corresponding AUC values are provided in Table 4. As the composition of the datasets differs, most importantly between cancer-enriched and full screening datasets, the AUC values are not completely comparable. However, the AUC of 0.925 in our study is somewhere in the middle among the other studies. The value is slightly lower than in two other studies using the same version of the same AI system on Danish and Spanish data.[116,155] When comparing AUC values, it should be taken into account that non-screening data might give misleading results, and even in cases when the data originates from screening, the datasets might be biased in different ways; for example, some studies excluded all women without a normal follow-up, while others included all available data. The performance measures of the AI systems can also be affected by differences in the screening programmes, including screening interval, workflow, vendors of mammography equipment and use of other modalities in the screening as well as differences in the underlying populations.

Several studies have retrospectively compared the performance of AI systems standalone with historical radiologist reading results from clinical screening workflow, and these showed an AI performance that is as least as good as single reader screening results.[114,150,155] However, radiologist double reading with consensus has usually been superior to the AI system performance, which is in accordance with the results in Paper 3.[114,148,150] There is a substantial variation particularly in specificity but, to some extent, also in the general performance of radiologists between different screening centres, which limits the value of comparing AI systems with radiologist reading.

In some studies, the performance of the AI system was compared with radiologists in reading studies, where the AI system clearly outperformed the radiologists.[78,150] The performance of the radiologists in these reading studies was clearly inferior to results usually achieved in clinical reading, which might to some degree be due to using highly cancer-enriched datasets. Further, it may in part be related to limited reader experience, but this can be expected even with experienced readers due to the laboratory effect, which cause the readers to perform significantly worse in retrospective reading studies than in prospective clinical reading.[159]

Personalisation of screening by selective addition of DBT (Paper 4)

Paper 4 investigates the potential of personalising screening by adding DBT for women based on an AI risk assessment and shows that a large part of the sensitivity gain with DBT screening could be achieved by only adding DBT to the 10% of women with the highest risk.

Several cancer risk models have been proposed, where different types of risk factors are included.[160,161] Some models focus on clinical and lifestyle factors, e.g. personal and family history of breast cancer, age, parity and age at first birth, while others also include genetic analyses or image-derived information, such as breast density. There are also risk models where more general image-based factors are included by using an AI system designed for cancer detection.[124] Collecting a large number of factors can be challenging, and risk models based mainly on images have also been developed.[123]

The proposed workflows in Paper 4 have similarities with using an image-based risk model to adapt the screening to the individual risk, but instead of predicting future risk, the AI system is used to assess the risk of cancer at the DM examination at the screening appointment so DBT can be added directly. While including other risk factors, such as family history of breast cancer, might have some additional gain, it would add more complexity to the screening process. Another possibility could be to analyse previous examinations and use DBT for women who had a high risk at the previous examination, which is more similar to other risk models. However, this means that any changes since the last examination would not be taken into account. This could potentially be solved by analysing both previous DM, and then the current DM in cases not triaged to DBT. While this approach may solve some logistical issues, it cannot be used when no previous examinations are accessible, which would be a limitation in fragmented screening programmes where previous images may be inaccessible if a woman attends screening at a different centre than the previous time.

No retrospective studies with approaches similar to those in Paper 4 have been found, and the concept should be tested in prospective studies before clinical implementation. Then, the assumptions that were necessary in the retrospective study can be verified, and the actual behaviour of readers can be studied. As mentioned previously, in particular, the effects on recall rate can be unpredictable when mixing DM and DBT screening, but this effect would probably be less if the allocation to DM or DBT follows a risk-based pattern rather than being random.

Speeding up reading of screening DBT with AI (Paper 5)

The possibilities of using AI to reduce the reading burden with DBT screening are investigated in Paper 5 by analysing the DBT examinations with an AI system and simulating different ways of using it in the reading workflow. Using AI to exclude low-risk cases from any reading and double reading the high-risk cases led to slightly better results than replacing the second reader with AI, but the latter approach is probably more feasible for psychological and legal reasons.

Compared with AI on DM, there are relatively few studies using AI on DBT. One retrospective study based on a similar cohort with paired DM and DBT examinations also investigated a workflow excluding low-risk cases from reading, but replacing the second reader was combined with excluding low-risk cases from all reading.[139] That study was more aggressive regarding workload reduction than that described in Paper 5 and discarded about 70% of the examinations as normal, but still the sensitivity was slightly higher. This might be due to differences in study design, including a simulated recall of the 2% of the examinations with the highest score, use of two-view narrow-angle DBT and a slightly higher overall recall rate.

In a study from the USA, an AI system was developed specifically for the task of identifying normal cases that can be removed from human reading.[162] In the internal evaluation, including data from sources other than the training data, the level of workload reduction was about 40%, which is a bit lower than that shown in Paper 5. The sensitivity was at the same level as single-reading, which was used as a reference since the study was performed in the USA.

AI for DBT in general

Paper 5 also includes more general results about AI on DBT with ROC analyses, which can be compared with other studies on the subject. The AUC values from some previous studies of AI on DBT are presented in Table 5. Our study has an AUC in the middle among all studies, but towards the lower end among studies based on screening data. This could be due to the use of one-view wide-angle DBT, as the studies with higher AUC were all based on two-view narrow-angle Hologic examinations. These are more similar to DM images, which constitute the majority of the training data for the AI systems. However, comparisons between studies are complicated due to differences in study design and type of datasets.

Table 5: Comparison with previous studies of AI on DBT

Name	Year of publication	AI system	DBT vendor	Country	Type of data	Number of cases	Number of cancers	AUC	Excluded from reading
Conant et al.[110]	2019	iCAD PowerLook Tomo Detection 2.0	Hologic	USA	Reader study / cancer-enriched	260	65	0.82[a]	N/A
Lotter et al.[157b]	2021	DeepHealth	Hologic	USA	Screening	11 687	78	0.959	N/A
van Winkel et al.[141]	2021	Transpara 1.6.0	Siemens	USA	Reader study / cancer-enriched	240	65	0.84	N/A
Pinto et al.[140]	2021	Transpara 1.6.0	Siemens[c]	Netherlands[d]	Reader study / cancer-enriched	190	75	0.90	N/A
Romero-Martín et al. / Raya-Povedano et al.[139,155]	2021	Transpara 1.7.0	Hologic	Spain	Screening + IC	15 987	113	0.94	70%
Shoshan et al.[162e]	2022	IBM in-house	Hologic	USA	Cancer-enriched	4 310	453	0.89	40%
Paper 5	2022	Transpara 1.7.0	Siemens[c]	Sweden	Screening + IC	14 772	157	0.896	50%

a) AUC not reported in publication but has been calculated graphically from the included ROC curve.
b) Trained on other portions of same datasets.
c) One-view DBT.
d) Refers to source of data. Readers from Norway.
e) Trained on other portions of same datasets. Results for validation dataset.

Concurrent use of AI for DBT has been studied in a number of reader studies where the sensitivity increased with AI, while false-positive recalls and reading time were retained or reduced (Table 5).[110,140,141] However, as previously mentioned, reader studies based on retrospective data might not represent the real behaviour of readers due to the laboratory effect.

A study in the USA aimed to assess the results of using AI in a clinical workflow by comparing the results from two different screening centres, one with and one without AI available in the reading situation.[163] The study reported a slightly higher cancer detection rate with the use of AI, but also a higher rate of false positives. However, the study has important limitations, where the use of two different centres serving different populations is probably the most important.

The previously mentioned Spanish study comparing reading of DM and DBT with and without concurrent use of an AI system also includes data on DBT, which can be compared with Paper 5.[156] Using the AI system on DBT was associated with a higher cancer detection rate but at the cost of a slight increase in recall rate. The increase in the cancer detection rate was slightly higher on DBT than on DM, but this difference was smaller than the difference between with and without AI support. The recall rate also increased slightly when using AI, but was still lower than for DM. Due to the differences in recall rate and the lower gain from DBT compared with DM, it is hard to determine how these results would transfer to a context similar to the MBTST. However, it is possible that introduction of a mixed DM and DBT screening, at least if the selection of modality is not controlled by the risk, could lead to a similar rise in the recall rate of DM as seen in the Spanish context.

Ethical considerations and trust in AI

Trust in AI

A prerequisite for introducing AI in breast cancer screening is that it is accepted and trusted by the women taking part in screening. A few studies have investigated the opinions about the use of AI in the context of breast cancer screening.[164–167] While being open to start using AI in breast cancer screening if this could improve the screening programme, the importance to thoroughly test and validate the AI systems both prior to and after introduction was stressed. A common opinion was also that human readers should always be involved in decisions and be responsible for the outcomes, while using

AI as the only reader was met with more scepticism. Equity was also an important aspect, and it has to be ensured that the AI systems perform well in all subgroups.

Privacy and use of training data

AI systems are usually trained using historical examinations from breast cancer screening programmes, which might have some ethical implications. Informed consent or explicit permission from each woman is usually not collected when using medical data for training AI systems. The ownership of medical data can be a bit diffuse, where the healthcare system and individual women both have some rights to the data.[168] Commonly, healthcare data are considered to not be personal data after anonymisation and are thus sharable with external partners. As an alternative, methods have been proposed where the data can be used to train AI models locally at the hospital.[169] That means that the personal information in itself never leaves the healthcare provider. In both cases, the trained model can eventually be sold to external instances by commercial companies. However, as all training data affect the model, the personal data can in some way be considered to be incorporated into the model that then becomes a product. There has recently been much publicity on generative AI systems, which have been trained on copyrighted data without permission, where data very similar to the training data can be exported.[170] However, in the case of classification models, such as cancer detection systems, where the model cannot export any data similar to the training data, the incorporation of training data is probably more of a philosophical matter.

Overarching discussion

Which way of using AI is the best?

The papers included in this thesis, together with numerous other articles published during the last few years, have explored different ways of using AI to enhance breast cancer screening. There seem to be several promising approaches, and the continuous developments of AI systems might lead to further improvements and unleash even more new possibilities. While most studies in the field have been retrospective, with all limitations this entails, the first results from a prospective randomised controlled trial have indicated that AI can actually be used to replace the second reader with a performance even better than expected from retrospective studies.[85] The other approaches have yet to be studied prospectively.

In order to optimise the use of resources and maximise performance, it would probably be ideal to implement a combination of several approaches, namely exclude low-risk cases from human reading and single read intermediate-risk cases with AI support, while high-risk cases are double read, or even are more thoroughly examined (e.g. with DBT, contrast enhanced mammography or MRI). However, a prerequisite for excluding cases from human reading is that all stakeholders trust the AI system for this use. This might not yet be the case, but with the rapid introduction of AI in different domains, we will likely all be more familiar with AI in just a few years, and this could lead to a larger acceptance and trust of AI in the medical field as well.

The new AI-enhanced workflows and reading processes of breast cancer screening can also open possibilities for introducing DBT in screening. While studies both by us and others have shown promising results from using AI on DBT, there might still be some development to be done before AI reaches the same level on DBT as on DM. As an example, Figure 14 shows that the AI system performs on par with radiologists on DM, only superseded by double reading with consensus, while the AI system clearly does not reach the level of radiologists on DBT. This is probably due to insufficient training data from DBT screening, where the training of the AI systems often, to a large extent, has to rely on mostly DM data with only a smaller portion of DBT data. With enough training data, logically, the potential for AI to outperform human readers on DBT should rather be greater than on DM, as an AI system can spend as much processing time on each slice of a DBT stack as on a full DM image, while a human reader has to use a more scrolling approach to retain a reasonable reading time.

Role of databases

Research databases with breast cancer images have several important roles. They are an important source for training data in the further development of AI systems, not the least to add more diversity of data as well as better coverage of uncommon cases and presentations. Databases are also important in evaluating the performance of AI in further retrospective studies, where the aggregation of large amounts of data can potentially make it possible to do specific analyses in subgroups and of rare types of breast cancer. Databases can also be used to test an AI system on retrospective data from a particular screening centre in order to verify the performance in the local context prior to procurement. As databases currently predominately contain DM examinations, the usability for DBT is more limited at the moment.

Overdiagnosis and overtreatment

An increased screening sensitivity results in more cancers diagnosed at an early stage, leading to a better prognosis. However, there will also be an increase in overdiagnosis, which will lead to overtreatment. It is hard to estimate the extent of overdiagnosis as it today would be unethical to keep a control group out of screening. A longer life expectancy also means that some of the cancers that were overdiagnosed in the now quite aged studies of overdiagnosis would today actually lead to symptomatic disease. Thus, we probably have to accept that the extent of overdiagnosis is uncertain. The risk of increasing overdiagnosis should not be taken as a reason to not improve the sensitivity of screening, as only part of the detected cancers represent overdiagnosis, while others will become symptomatic and should be treated as soon as possible. While overdiagnosis can have psychological consequences, overtreatment can in addition lead to serious somatic side effects.

In cases where the cancer has characteristics that indicate a low risk of progression, diagnosis may not necessarily have to lead to treatment. Instead, concepts such as active surveillance could be introduced, which are being investigated for ductal carcinoma in situ in some ongoing studies.[171,172] As cancers are usually removed after diagnosis, moving to an active surveillance approach might be challenging and discourage patients from entering such studies.[173,174] However, in one of the studies, removing the randomisation step and allowing patients to choose between active surveillance and conventional treatment accelerated the recruitment and, interestingly, led to most patients entering the active surveillance arm.[174]

Methodological considerations and overall limitations

An overarching limitation of all the studies in this thesis is that they were based on retrospective data. This means that it has been necessary to make some assumptions that may or may not be correct. Some aspects cannot at all be studied when applying AI to retrospective data and comparing it with historical radiologist readings, in particular the interaction between radiologists and AI.

Results on DM from the MBTST cohort are not fully comparable with other studies, as some of the cancers detected on DBT screening would likely not have been diagnosed until after the next or second-to-next screening rounds and, thus, would not have been included in the ground truth – or appeared as interval cancers – in studies without paired DBT. This might to same extent affect Papers 2–4. Using data collected during

a trial also introduces a selection bias, as the women refraining from taking part in the study might differ from those who take part.

All the studies were based on data from a single screening centre, serving an almost exclusively urban population. All examinations in Papers 2–5 are from Siemens equipment. The M-BIG database is more diverse on the mammography machine vendor, with three different vendors represented, although there is a temporal difference. Further, only one AI system was used, and using other systems might result in different outcomes. The development in the field has been very rapid with several updates of the AI system since the analyses were performed. Using a newer version of the AI system might give somewhat different results.

As the M-BIG database is a work in progress where the data have been further curated since publication of Paper 1, some numbers have changed slightly since the time of publication, for instance, the total number of examinations and women have been slightly reduced.

Conclusions

- A breast imaging research database has been created where images, reader results and free text radiology reports have been gathered together with cancer data from different registries. This platform provides more convenient access to data and will facilitate future research on breast cancer imaging.
- AI can improve the resource efficiency of DM screening by eliminating the need for manual reading of normal cases.
- AI has the potential to improve the performance of DM screening by detecting some of the cancers that would otherwise only be detected on DBT screening. Further, AI can be used on DM for selecting high-risk cases where addition of DBT would be beneficial.
- AI can make it possible to introduce DBT in screening with an unchanged reader workload by either excluding normal cases from reading or replacing the second reader. This would improve the sensitivity in breast cancer screening, probably without notable changes in recall rates.

Future perspectives

Breast cancer screening in the future

AI will likely play a successively increasing role in breast cancer screening. It is already used as support for readers in some places and has in this role been shown to be able to replace the second reader with retained sensitivity. This can mainly be motivated by saving resources, which obviously is attractive for healthcare providers, but it can be more questionable from the perspective of women taking part in screening, who would likely appreciate enhanced performance more. Although there have been some indications of minor improvements in both sensitivity and specificity by using AI in DM screening, it might be favourable to link the introduction of AI to other improvements of the screening programme that are made possible by using AI. It has been suggested to use AI for adding breast MRI in high-risk cases,[175] but the most immediate advancement is probably to introduce DBT in breast cancer screening, either in selected high-risk cases or for all participants. Adding DBT only in high-risk cases does not unleash the full potential of DBT screening, but this approach could be useful as a transition phase to full DBT screening. As this approach primarily relies on AI on DM, the relative immaturity of AI for DBT would largely be avoided.

A screening programme based on DBT and AI not only has several possible workflows, e.g. excluding normal cases or replacing the second reader, but there are also different ways of reading the DBT. If a human reader is required to gain public or legal acceptance, reading an SM might be sufficient in cases marked as low-risk by AI (suggested in my conference abstract from European Congress of Radiology, ECR 2023).

Role of breast imaging databases

Present evaluations of AI using retrospective data have shown that AI can work on screening data as a whole, but the performance in different subgroups of women and for specific types of breast cancer is still to be tested. As large amounts of data are

important in order to do subgroup analyses, breast imaging databases will be crucial in this process. Further, databases can be utilised to collect training data from underrepresented groups. Current breast cancer imaging databases predominantly contain DM, while DBT screening data are more limited. More DBT training data are necessary to unleash the full potential of AI on DBT. Clinical DBT can be used in training, but this comes with the risk of restricting the models to cancers detectable with DM if not enough DBT screening-detected cancers are included. Subgroup analyses of AI on DBT also require more DBT screening data. Also, as the characteristics of the images vary more between vendors for DBT than DM, it might be necessary with even more extensive evaluations of AI on DBT than on DM.

Acknowledgements

Jag vill börja med att tacka mina handledare som har väglett mig genom doktorandtiden:

Min huvudhandledare Sophia Zackrisson har varit både ett stort stöd och inspirationskälla. Trots en späckad kalender så har du alltid funnits till hands för frågor om både stort och smått oavsett var i världen du befunnit har dig – och dessutom svarat snabbt när det varit bråttom.

Anders Tingberg som alltid har funnits tillgänglig för frågor eller om jag bara behövt prata lite. Ditt lugn och din avslappnade inställning har varit en trygghet i det jäkt och prestationsfokus som ofta råder inom forskningsvärlden.

Magnus Dustler har varit ett ovärderligt stöd inom såväl praktiska saker som programmering och teknik som med statistik och att skriva och granska artiklar. Runt dig brukar det också alltid vara intressanta diskussioner om precis vad som helst.

Ni har alla en bred kompetens och kompletterar varandra på olika sätt. Dessutom har er akademiska erfarenhet och goda insyn i vad som ska göras under doktorandtiden varit en stor hjälp för att nå i mål.

Jag är tacksam för att jag har fått doktorera tillsammans med ett så härligt gäng doktorander och andra juniora forskare i LUCI. Gänget på plan 2 – Anna Bjerkén och Hanna Tomic, och tidigare Rebecca Axelsson och Gustav Hellgren – har bidragit med både stöd och en glad stämning i vardagen vilket har förgyllt min doktorandtid. Mina andra doktorandkollegor Kristin Johnsson (nu disputerad), Li Sturesdotter, Nadia Chaudhry, Jakob Olinder och Ann-Sofi Sjöqvist har också bidragit med praktiskt stöd och givande diskussioner såväl som med många fina minnen.

Jag vill också tacka de andra medlemmarna i LUCI: Daniel Förnvik, Predrag Bakic, Anetta Bolejko, Hannie Förnvik, Pontus Timberg, Maria Inasu och Akane Ohashi som på olika sätt har hjälpt mig genom att komma med värdefulla synpunkter eller har bidragit till intressanta diskussioner. Kajsa Trens som alltid varit hjälpsam med olika praktiska och administrativa saker. Därutöver vill jag tacka Kristina Lång och Ingvar Andersson som är medförfattare på flera av artiklarna i min avhandling.

Eftersom jag varit deltidsdoktorand har jobbet inte enbart varit forskning, utan det finns ju en klinisk sida också. Jag har under hela doktorandtiden samtidigt varit ST-läkare på Bild- och funktionsmedicin och där vill jag tacka:

Min kliniska handledare Ylva Gårdinger för att du har stöttat mig när det har behövts och dessutom har insyn i utmaningarna med att balansera klinik och forskning.

Min nuvarande chef Håkan Sjunnesson och mina tidigare chefer för att ni har gett mig möjlighet att forska.

Mina nuvarande och tidigare ST-kollegor och specialister på Röntgen som alla bidrar till att kliniken är en trivsam arbetsplats, trots alla de svårigheter som vi tidvis har drabbats av.

Jag är tacksam för de forskningsanslag som har gjort det möjligt för mig att få inte bara forska, utan dessutom få tillräckligt med tid för att kunna fördjupa mig ordentligt. Detta gäller de ST-ALF-medel som jag har fått för större delen av min doktorandtid, men också det AIDA-fellowship som gjorde det möjligt för mig att börja forska. Jag är också tacksam för att jag fått forskningstid från Bild- och funktionsmedicin under slutet av doktorandtiden som gjorde att jag kunde förbereda inför disputationen.

Min forskning hade inte gått att genomföra utan MBTST och M-BIG och jag vill därför tacka alla som har bidragit till att dessa projekt har kunnat genomföras – alltifrån anslagsgivare, granskare och teknisk personal till de kvinnor som har deltagit i antingen MBTST eller vanlig mammografiscreening. Tack också till ScreenPoint Medical som har låtit oss använda Transpara och har hjälpt mig med både teknisk support och värdefulla synpunkter.

Jag vill också tacka mina föräldrar Ola och Kerstin och min bror Oscar som alltid har stöttat, inspirerat och hjälpt mig på alla tänkbara sätt.

Till sist vill jag tacka min älskade fru Malin som alltid finns där för mig. Du har fått både lyssna på mina presentationer och utstå ett ständigt "ska bara"-ande alltifrån hemma i soffan till i transferbussen på Kreta. Under sista tiden har vår älskade dotter Elvira också hjälpt till på sitt sätt.

References

1 Ferlay J, Colombet M, Soerjomataram I, *et al.* Cancer statistics for the year 2020: An overview. *Intl Journal of Cancer* 2021; **149**: 778–89.

2 Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *npj Breast Cancer* 2017; **3**: 12.

3 Marinovich ML, Hunter KE, Macaskill P, Houssami N. Breast Cancer Screening Using Tomosynthesis or Mammography: A Meta-analysis of Cancer Detection and Recall. *JNCI: Journal of the National Cancer Institute* 2018; **110**: 942–9.

4 Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. *Expert Review of Medical Devices* 2019; **16**: 351–62.

5 Zheng R, Wang S, Zhang S, *et al.* Global, regional, and national lifetime probabilities of developing cancer in 2020. *Sci Bull (Beijing)* 2023; **68**: 2620–8.

6 Socialstyrelsen. Statistik om bröstcancer. 2023; published online Oct 26. https://www.socialstyrelsen.se/statistik-och-data/statistik/alla-statistikamnen/cancer/.

7 Åhlin E, editor. Cancer i siffror 2023. 2023. https://www.cancerfonden.se/om-cancer/statistik/cancer-i-siffror.

8 Chamalidou C, Fohlin H, Albertsson P, *et al.* Survival patterns of invasive lobular and invasive ductal breast cancer in a large population-based cohort with two decades of follow up. *The Breast* 2021; **59**: 294–300.

9 Pan H, Gray R, Braybrooke J, *et al.* 20-Year Risks of Breast-Cancer Recurrence after Stopping Endocrine Therapy at 5 Years. *N Engl J Med* 2017; **377**: 1836–46.

10 Berman AT, Thukral AD, Hwang W-T, Solin LJ, Vapiwala N. Incidence and Patterns of Distant Metastases for Patients With Early-Stage Breast Cancer After Breast Conservation Treatment. *Clinical Breast Cancer* 2013; **13**: 88–94.

11 Wendt C, Margolin S. Identifying breast cancer susceptibility genes – a review of the genetic background in familial breast cancer. *Acta Oncologica* 2019; **58**: 135–46.

12 Singletary SE. Rating the Risk Factors for Breast Cancer: *Annals of Surgery* 2003; **237**: 474–82.

13 Pujol P, Galtier-Dereure F, Bringer J. Obesity and breast cancer risk. *Human Reproduction* 1997; **12**: 116–25.

14 Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *The Lancet Oncology* 2012; **13**: 1141–51.

15 Bodewes FTH, van Asselt AA, Dorrius MD, Greuter MJW, de Bock GH. Mammographic breast density and the risk of breast cancer: A systematic review and meta-analysis. *Breast* 2022; **66**: 62–8.

16 Ronckers CM, Erdmann CA, Land CE. Radiation and breast cancer: a review of current evidence. *Breast Cancer Res* 2004; **7**: 21.

17 Vinogradova Y, Coupland C, Hippisley-Cox J. Use of hormone replacement therapy and risk of breast cancer: nested case-control studies using the QResearch and CPRD databases. *BMJ* 2020; : m3873.

18 Colditz GA. Cumulative Risk of Breast Cancer to Age 70 Years According to Risk Factor Status: Data from the Nurses' Health Study. *American Journal of Epidemiology* 2000; **152**: 950–64.

19 Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50 302 women with breast cancer and 96 973 women without the disease. *The Lancet* 2002; **360**: 187–95.

20 Cuesta Cuesta AB, Martín Ríos MD, Noguero Meseguer MR, *et al.* Precisión de la resonancia magnética, ecografía y mamografía en la medida del tamaño tumoral y su correlación con el tamaño histopatológico en el cáncer de mama primario. *Cirugía Española* 2019; **97**: 391–6.

21 Voogd AC, Coebergh J-WW, Driel OJRV, *et al.* The risk of nodal metastases in breast cancer patients with clinically negative lymph nodes: a population-based analysis. *Breast Cancer Res Treat* 2000; **62**: 63–9.

22 the SCREENREG Working Group, Bucchi L, Barchielli A, *et al.* Screen-detected vs clinical breast cancer: the advantage in the relative risk of lymph node metastases decreases with increasing tumour size. *Br J Cancer* 2005; **92**: 156–61.

23 Sartor H, Borgquist S, Hartman L, Olsson Å, Jawdat F, Zackrisson S. Do mammographic tumor features in breast cancer relate to breast density and invasiveness, tumor size, and axillary lymph node involvement? *Acta Radiol* 2015; **56**: 536–44.

24 Baré M, Torà N, Salas D, *et al.* Mammographic and clinical characteristics of different phenotypes of screen-detected and interval breast cancers in a nationwide screening program. *Breast Cancer Res Treat* 2015; **154**: 403–15.

25 Sturesdotter L, Sandsveden M, Johnson K, Larsson A-M, Zackrisson S, Sartor H. Mammographic tumour appearance is related to clinicopathological factors and surrogate molecular breast cancer subtype. *Sci Rep* 2020; **10**: 20814.

26 Suhrke P, Zahl P. Breast cancer incidence and menopausal hormone therapy in Norway from 2004 to 2009: a register-based cohort study. *Cancer Medicine* 2015; **4**: 1303–8.

27 Hofvind S, Sørum R, Thoresen S. Incidence and tumor characteristics of breast cancer diagnosed before and after implementation of a population-based screening-program. *Acta Oncologica* 2008; **47**: 225–31.

28 Eheman CR, Shaw KM, Ryerson AB, Miller JW, Ajani UA, White MC. The Changing Incidence of *In situ* and Invasive Ductal and Lobular Breast Carcinomas: United States, 1999-2004. *Cancer Epidemiology, Biomarkers & Prevention* 2009; **18**: 1763–9.

29 Ernster VL, Ballard-Barbash R, Barlow WE, *et al.* Detection of ductal carcinoma in situ in women undergoing screening mammography. *J Natl Cancer Inst* 2002; **94**: 1546–54.

30 Compton CC. AJCC cancer staging atlas: a companion to the seventh editions of the AJCC cancer staging manual and handbook, 2nd edition. New York: Springer, 2012.

31 Sørlie T, Perou CM, Tibshirani R, *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 2001; **98**: 10869–74.

32 Szymiczek A, Lone A, Akbari MR. Molecular intrinsic versus clinical subtyping in breast cancer: A comprehensive review. *Clinical Genetics* 2021; **99**: 613–37.

33 Goldhirsch A, Wood WC, Coates AS, Gelber RD, Thürlimann B, Senn H-J. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Annals of Oncology* 2011; **22**: 1736–47.

34 Nascimento RGD, Otoni KM. Histological and molecular classification of breast cancer: what do we know? *Mastology* 2020; **30**: e20200024.

35 Vasconcelos I, Hussainzada A, Berger S, *et al.* The St. Gallen surrogate classification for breast cancer subtypes successfully predicts tumor presenting features, nodal involvement, recurrence patterns and disease free survival. *The Breast* 2016; **29**: 181–5.

36 Barnard ME, Boeke CE, Tamimi RM. Established breast cancer risk factors and risk of intrinsic tumor subtypes. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 2015; **1856**: 73–85.

37 Burstein HJ, Curigliano G, Thürlimann B, *et al.* Customizing local and systemic therapies for women with early breast cancer: the St. Gallen International Consensus Guidelines for treatment of early breast cancer 2021. *Annals of Oncology* 2021; **32**: 1216–35.

38 Vårdprogramgruppen för bröstcancer. Nationellt vårdprogram - Bröstcancer Version 4.3. 2023; published online March 28. https://kunskapsbanken.cancercentrum.se/diagnoser/brostcancer/vardprogram/ (accessed Nov 28, 2023).

39 Van Steen A, Van Tiggelen R. Short history of mammography: a Belgian perspective. *Jbr Btr* 2007; **90**: 151.

40 Nicosia L, Gnocchi G, Gorini I, *et al.* History of Mammography: Analysis of Breast Imaging Diagnostic Achievements over the Last Century. *Healthcare* 2023; **11**: 1596.

41 Skaane P. Studies comparing screen-film mammography and full-field digital mammography in breast cancer screening: Updated review. *Acta Radiol* 2009; **50**: 3–14.

42 Andersson I, Hildell J, Muhlow A, Pettersson H. Number of projections in mammography: influence on detection of breast disease. *American Journal of Roentgenology* 1978; **130**: 349–51.

43 Tirada N, Li G, Dreizin D, *et al.* Digital Breast Tomosynthesis: Physics, Artifacts, and Quality Control Considerations. *RadioGraphics* 2019; **39**: 413–26.

44 Hadjipanteli A, Elangovan P, Looney PT, *et al.* Detection of microcalcification clusters by 2D-mammography and narrow and wide angle digital breast tomosynthesis. In: Kontos D, Flohr TG, Lo JY, eds. . San Diego, California, United States, 2016: 978306.

45 Ciatto S, Houssami N, Bernardi D, *et al.* Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study. *The Lancet Oncology* 2013; **14**: 583–9.

46 Skaane P, Bandos AI, Gullien R, *et al.* Prospective trial comparing full-field digital mammography (FFDM) versus combined FFDM and tomosynthesis in a population-based screening programme using independent double reading with arbitration. *Eur Radiol* 2013; **23**: 2061–71.

47 Zackrisson S, Lång K, Rosso A, *et al.* One-view breast tomosynthesis versus two-view mammography in the Malmö Breast Tomosynthesis Screening Trial (MBTST): a prospective, population-based, diagnostic accuracy study. *The Lancet Oncology* 2018; **19**: 1493–503.

48 Romero Martín S, Raya Povedano JL, Cara García M, Santos Romero AL, Pedrosa Garriguet M, Álvarez Benito M. Prospective study aiming to compare 2D mammography and tomosynthesis + synthesized mammography in terms of cancer detection and recall. From double reading of 2D mammography to single reading of tomosynthesis. *Eur Radiol* 2018; **28**: 2484–91.

49 Pattacini P, Nitrosi A, Giorgi Rossi P, *et al.* A Randomized Trial Comparing Breast Cancer Incidence and Interval Cancers after Tomosynthesis Plus Mammography versus Mammography Alone. *Radiology* 2022; **303**: 256–66.

50 Heindel W, Weigel S, Gerß J, *et al.* Digital breast tomosynthesis plus synthesised mammography versus digital screening mammography for the detection of invasive breast cancer (TOSYMA): a multicentre, open-label, randomised, controlled, superiority trial. *The Lancet Oncology* 2022; **23**: 601–11.

51 Monticciolo DL. Digital Breast Tomosynthesis: A Decade of Practice in Review. *Journal of the American College of Radiology* 2023; **20**: 127–33.

52 Socialstyrelsen. Screening för bröstcancer – Socialstyrelsens rekommendation – Slutversion 2023. 2023; published online May. https://www.socialstyrelsen.se/kunskapsstod-och-regler/regler-och-riktlinjer/nationella-screeningprogram/slutliga-rekommendationer/brostcancer/.

53 Socialstyrelsen. Bilaga – Översyn av rekommendation om screening för bröstcancer – Organisatoriska underlag 2023 – Ökad användning av DBT. 2023; published online May. https://www.socialstyrelsen.se/kunskapsstod-och-regler/regler-och-riktlinjer/nationella-screeningprogram/slutliga-rekommendationer/brostcancer/.

54 Chikarmane SA, Offit LR, Giess CS. Synthetic Mammography: Benefits, Drawbacks, and Pitfalls. *RadioGraphics* 2023; **43**: e230018.

55 Zuckerman SP, Sprague BL, Weaver DL, Herschorn SD, Conant EF. Survey Results Regarding Uptake and Impact of Synthetic Digital Mammography With Tomosynthesis in the Screening Setting. *Journal of the American College of Radiology* 2020; **17**: 31–7.

56 Fallenberg EM, Fuchsjäger M, editors. Screening & Beyond Medical imaging in the detection, diagnosis and management of breast diseases, 1. Auflage. [S.l.] @: European Society of Radiology (ESR), 2016.

57 Gulani V, Calamante F, Shellock FG, Kanal E, Reeder SB. Gadolinium deposition in the brain: summary of evidence and recommendations. *The Lancet Neurology* 2017; **16**: 564–70.

58 Mann RM, Cho N, Moy L. Breast MRI: State of the Art. *Radiology* 2019; **292**: 520–36.

59 Pötsch N, Vatteroni G, Clauser P, Helbich TH, Baltzer PAT. Contrast-enhanced Mammography versus Contrast-enhanced Breast MRI: A Systematic Review and Meta-Analysis. *Radiology* 2022; **305**: 94–103.

60 Berger N, Marcon M, Saltybaeva N, *et al.* Dedicated Breast Computed Tomography With a Photon-Counting Detector: Initial Results of Clinical In Vivo Imaging. *Invest Radiol* 2019; **54**: 409–18.

61 Taba ST, Gureyev TE, Alakhras M, Lewis S, Lockie D, Brennan PC. X-Ray Phase-Contrast Technology in Breast Imaging: Principles, Options, and Clinical Application. *American Journal of Roentgenology* 2018; **211**: 133–45.

62 Dustler M, Förnvik D, Timberg P, *et al.* Can mechanical imaging increase the specificity of mammography screening? *Eur Radiol* 2017; **27**: 3217–25.

63 Poplack SP, Park E-Y, Ferrara KW. Optical Breast Imaging: A Review of Physical Principles, Technologies, and Clinical Applications. *Journal of Breast Imaging* 2023; **5**: 520–37.

64 Shapiro S. Evidence on screening for breast cancer from a randomized trial. *Cancer* 1977; **39**: 2772–82.

65 Houssami N, Miglioretti D, editors. Breast cancer screening: an examination of scientific evidence. London ; San Diego, CA: Academic Press is an imprint of Elsevier, 2016.

66 Canelo-Aybar C, Ferreira DS, Ballesteros M, *et al.* Benefits and harms of breast cancer mammography screening for women at average risk of breast cancer: A systematic review for the European Commission Initiative on Breast Cancer. *J Med Screen* 2021; **28**: 389–404.

67 Altobelli E, Rapacchietta L, Angeletti P, Barbante L, Profeta F, Fagnano R. Breast Cancer Screening Programmes across the WHO European Region: Differences among Countries Based on National Income Level. *IJERPH* 2017; **14**: 452.

68 Klabunde CN, Ballard-Barbash R. Evaluating Population-Based Screening Mammography Programs Internationally. *Seminars in Breast Disease* 2007; **10**: 102–7.

69 Quintin C, Chatignoux E, Plaine J, Hamers FF, Rogel A. Coverage rate of opportunistic and organised breast cancer screening in France: Department-level estimation. *Cancer Epidemiology* 2022; **81**: 102270.

70 Boncz I, Sebestyén A, Pintér I, Battyéany I, Ember I. The effect of an organized, nationwide breast cancer screening programme on non-organized mammography activities. *J Med Screen* 2008; **15**: 14–7.

71 U.S. Preventive Services Task Force. Final Recommendation Statement - Breast Cancer: Screening. 2016; published online Jan 11. https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/breast-cancer-screening.

72 Lim YX, Lim ZL, Ho PJ, Li J. Breast Cancer in Asia: Incidence, Mortality, Early Detection, Mammography Programs, and Risk-Based Screening Initiatives. *Cancers* 2022; **14**: 4218.

73 Cancer Research UK. Breast screening. 2023; published online May 26. https://www.cancerresearchuk.org/about-cancer/breast-cancer/getting-diagnosed/screening-breast.

74 American Cancer Society. American Cancer Society Recommendations for the Early Detection of Breast Cancer. 2022; published online Jan 14. https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/american-cancer-society-recommendations-for-the-early-detection-of-breast-cancer.html (accessed Nov 29, 2023).

75 Gao Y, Babb JS, Toth HK, Moy L, Heller SL. Digital Breast Tomosynthesis Practice Patterns Following 2011 FDA Approval. *Academic Radiology* 2017; **24**: 947–53.

76 Richman IB, Hoag JR, Xu X, *et al.* Adoption of Digital Breast Tomosynthesis in Clinical Practice. *JAMA Intern Med* 2019; **179**: 1292.

77 Conant EF, Zuckerman SP, McDonald ES, *et al.* Five Consecutive Years of Screening with Digital Breast Tomosynthesis: Outcomes by Screening Year and Round. *Radiology* 2020; **295**: 285–93.

78 Kim H-E, Kim HH, Han B-K, *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**: e138–48.

79 Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition—summary document. *Annals of Oncology* 2008; **19**: 614–22.

80 BreastScreen Australia. National Accreditation Standards. 2022; published online March. https://www.health.gov.au/resources/publications/breastscreen-australia-national-accreditation-standards-nas?language=en.

81 Uematsu T. Rethinking screening mammography in Japan: next-generation breast cancer screening through breast awareness and supplemental ultrasonography. *Breast Cancer* 2023; published online Oct 12. DOI:10.1007/s12282-023-01506-w.

82 Brennan PC, Ganesan A, Eckstein MP, *et al.* Benefits of Independent Double Reading in Digital Mammography: A Theoretical Evaluation of All Possible Pairing Methodologies. *Acad Radiol* 2019; **26**: 717–23.

83 Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Internal Medicine* 2015; **175**: 1828.

84 Keen JD, Keen JM, Keen JE. Utilization of Computer-Aided Detection for Digital Screening Mammography in the United States, 2008 to 2016. *Journal of the American College of Radiology* 2018; **15**: 44–8.

85 Lång K, Josefsson V, Larsson A-M, *et al.* Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *The Lancet Oncology* 2023; **24**: 936–44.

86 Dembrower K, Crippa A, Colón E, Eklund M, Strand F. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. *The Lancet Digital Health* 2023; **5**: e703–11.

87 Sickles EA, D'Orsi CJ, Bassett LW, et al. ACR BI-RADS® Mammography. In: American College of Radiology, ed. ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. Reston, VA: American College of Radiology, 2013.

88 Maxwell AJ, Ridley NT, Rubin G, Wallis MG, Gilbert FJ, Michell MJ. The Royal College of Radiologists Breast Group breast imaging classification. *Clinical Radiology* 2009; **64**: 624–7.

89 National Breast Cancer Centre, Camperdown, NSW, Australia. Synoptic breast imaging report. 2007. https://www.canceraustralia.gov.au/sites/default/files/publications/big-2-synoptic-breast-imaging-report_504af02c46210.pdf.

90 Salz T, Richman AR, Brewer NT. Meta-analyses of the effect of false-positive mammograms on generic and specific psychosocial outcomes. *Psycho-Oncology* 2010; **19**: 1026–34.

91 Bolejko A, Hagell P, Wann-Hansson C, Zackrisson S. Prevalence, Long-term Development, and Predictors of Psychosocial Consequences of False-Positive Mammography among Women Attending Population-Based Screening. *Cancer Epidemiol Biomarkers Prev* 2015; **24**: 1388–97.

92 Román M, Hofvind S, Von Euler-Chelpin M, Castells X. Long-term risk of screen-detected and interval breast cancer after false-positive results at mammography screening: joint analysis of three national cohorts. *Br J Cancer* 2019; **120**: 269–75.

93 Taksler GB, Keating NL, Rothberg MB. Implications of false-positive results for future cancer screenings. *Cancer* 2018; **124**: 2390–8.

94 Román R, Sala M, De La Vega M, *et al.* Effect of false-positives and women's characteristics on long-term adherence to breast cancer screening. *Breast Cancer Res Treat* 2011; **130**: 543–52.

95 McCann J, Stockton D, Godward S. Impact of false-positive mammography on subsequent screening attendance and risk of cancer. *Breast Cancer Res* 2002; **4**: R11.

96 Houssami N. Overdiagnosis of breast cancer in population screening: does it make breast screening worthless? *Cancer Biology & Medicine* 2017; **14**: 1–8.

97 Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. *Lancet* 2012; **380**: 1778–86.

98 Zackrisson S, Andersson I, Janzon L, Manjer J, Garne JP. Rate of over-diagnosis of breast cancer 15 years after end of Malmö mammographic screening trial: follow-up study. *BMJ* 2006; **332**: 689–92.

99 Miller AB. Canadian National Breast Screening Study-2: 13-Year Results of a Randomized Trial in Women Aged 50-59 Years. *Journal of the National Cancer Institute* 2000; **92**: 1490–9.

100 Miller AB. The Canadian National Breast Screening Study-1: Breast Cancer Mortality after 11 to 16 Years of Follow-up: A Randomized Screening Trial of Mammography in Women Age 40 to 49 Years. *Ann Intern Med* 2002; **137**: 305.

101 Kimme C, O'Loughlin BJ, Sklansky J. Automatic Detection of Suspicious Abnormalities in Breast Radiographs. In: Data Structures, Computer Graphics, and Pattern Recognition. Elsevier, 1977: 427–47.

102 Rao VM, Levin DC, Parker L, Cavanaugh B, Frangos AJ, Sunshine JH. How Widely Is Computer-Aided Detection Used in Screening and Diagnostic Mammography? *Journal of the American College of Radiology* 2010; 7: 802–5.

103 Henriksen EL, Carlsen JF, Vejborg IM, Nielsen MB, Lauridsen CA. The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review. *Acta Radiol* 2019; **60**: 13–8.

104 Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. *Gastrointestinal Endoscopy* 2020; **92**: 807–12.

105 Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to Machine Learning, Neural Networks, and Deep Learning. *Transl Vis Sci Technol* 2020; **9**: 14.

106 Maier A, Syben C, Lasser T, Riess C. A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik* 2019; **29**: 86–101.

107 Raina R, Madhavan A, Ng AY. Large-scale deep unsupervised learning using graphics processors. In: Proceedings of the 26th Annual International Conference on Machine Learning. Montreal Quebec Canada: ACM, 2009: 873–80.

108 Ciresan DC, Meier U, Gambardella LM, Schmidhuber J. Deep Big Simple Neural Nets Excel on Handwritten Digit Recognition. 2010. DOI:10.48550/ARXIV.1003.0358.

109 Rodríguez-Ruiz A, Krupinski E, Mordang J-J, *et al.* Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. *Radiology* 2019; **290**: 305–14.

110 Conant EF, Toledano AY, Periaswamy S, *et al.* Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis. *Radiology: Artificial Intelligence* 2019; **1**: e180096.

111 Pacilè S, Lopez J, Chone P, Bertinotti T, Grouin JM, Fillard P. Improving Breast Cancer Detection Accuracy of Mammography with the Concurrent Use of an Artificial Intelligence Tool. *Radiology: Artificial Intelligence* 2020; **2**: e190208.

112 Leibig C, Brehmer M, Bunk S, Byng D, Pinker K, Umutlu L. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. *The Lancet Digital Health* 2022; **4**: e507–19.

113 Ng AY, Glocker B, Oberije C, *et al.* Artificial Intelligence as Supporting Reader in Breast Screening: A Novel Workflow to Preserve Quality and Reduce Workload. *Journal of Breast Imaging* 2023; **5**: 267–76.

114 Salim M, Wåhlin E, Dembrower K, *et al.* External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. *JAMA Oncol* 2020; published online Aug 27. DOI:10.1001/jamaoncol.2020.3321.

115 Schaffter T, Buist DSM, Lee CI, *et al.* Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Netw Open* 2020; **3**: e200265.

116 Lauritzen AD, Rodríguez-Ruiz A, Von Euler-Chelpin MC, *et al.* An Artificial Intelligence–based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload. *Radiology* 2022; **304**: 41–9.

117 Yi PH, Singh D, Harvey SC, Hager GD, Mullen LA. DeepCAT: Deep Computer-Aided Triage of Screening Mammography. *J Digit Imaging* 2021; **34**: 27–35.

118 Rodriguez-Ruiz A, Lång K, Gubern-Merida A, *et al.* Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur Radiol* 2019; **29**: 4825–32.

119 Raj SD, Fein-Zachary V, Slanetz PJ. Deciphering the Breast Density Inform Law Movement: Implications for Practice. *Seminars in Ultrasound, CT and MRI* 2018; **39**: 16–24.

120 Ciatto S, Bernardi D, Calabrese M, *et al.* A first evaluation of breast radiological density assessment by QUANTRA software as compared to visual classification. *The Breast* 2012; **21**: 503–6.

121 Destounis S, Johnston L, Highnam R, Arieno A, Morgan R, Chan A. Using Volumetric Breast Density to Quantify the Potential Masking Risk of Mammographic Density. *American Journal of Roentgenology* 2017; **208**: 222–7.

122 Gastounioti A, Kasi CD, Scott CG, *et al.* Evaluation of LIBRA Software for Fully Automated Mammographic Density Assessment in Breast Cancer Risk Prediction. *Radiology* 2020; **296**: 24–31.

123 Yala A, Mikhael PG, Strand F, *et al.* Multi-Institutional Validation of a Mammography-Based Breast Cancer Risk Model. *JCO* 2021; : JCO.21.01337.

124 Eriksson M, Czene K, Strand F, *et al.* Identification of Women at High Risk of Breast Cancer Who Need Supplemental Screening. *Radiology* 2020; **297**: 327–33.

125 De Mauro A, Greco M, Grimaldi M. A formal definition of Big Data based on its essential features. *Library Review* 2016; **65**: 122–35.

126 Heath M, Bowyer K, Kopans D, *et al.* Current status of the digital database for screening mammography. In: Digital Mammography: Nijmegen, 1998. Springer, 1998: 457–60.

127 Suckling J, Parker J, Dance D, *et al.* Mammographic image analysis society (mias) database v1. 21. 2015.

128 Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast. *Academic Radiology* 2012; **19**: 236–48.

129 Halling-Brown MD, Looney PT, Patel MN, Warren LM, Mackenzie A, Young KC. The oncology medical image database (OMI-DB). In: Law MY, Cook TS, eds. . San Diego, California, USA, 2014: 903906.

130 Rosso A, Lång K, Petersson IF, Zackrisson S. Factors affecting recall rate and false positive fraction in breast cancer screening with breast tomosynthesis - A statistical approach. *Breast* 2015; **24**: 680–6.

131 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**: 184–90.

132 Lång K, Nergården M, Andersson I, Rosso A, Zackrisson S. False positives in breast cancer screening with one-view breast tomosynthesis: An analysis of findings leading to recall, work-up and biopsy rates in the Malmö Breast Tomosynthesis Screening Trial. *European Radiology* 2016; **26**: 3899–907.

133 Sartor H, Lång K, Rosso A, Borgquist S, Zackrisson S, Timberg P. Measuring mammographic density: comparing a fully automated volumetric assessment versus European radiologists' qualitative classification. *Eur Radiol* 2016; **26**: 4354–60.

134 Förnvik D, Förnvik H, Fieselmann A, Lång K, Sartor H. Comparison between software volumetric breast density estimates in breast tomosynthesis and digital mammography images in a large public screening cohort. *Eur Radiol* 2019; **29**: 330–6.

135 Johnson K, Zackrisson S, Rosso A, *et al.* Tumor Characteristics and Molecular Subtypes in Breast Cancer Screening with Digital Breast Tomosynthesis: The Malmö Breast Tomosynthesis Screening Trial. *Radiology* 2019; **293**: 273–81.

136 Johnson K, Lång K, Ikeda DM, Åkesson A, Andersson I, Zackrisson S. Interval Breast Cancer Rates and Tumor Characteristics in the Prospective Population-based Malmö Breast Tomosynthesis Screening Trial. *Radiology* 2021; **299**: 559–67.

137 Johnson K, Olinder J, Rosso A, Andersson I, Lång K, Zackrisson S. False-positive recalls in the prospective Malmö Breast Tomosynthesis Screening Trial. *Eur Radiol* 2023; published online May 5. DOI:10.1007/s00330-023-09705-x.

138 Rodriguez-Ruiz A, Lång K, Gubern-Merida A, *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**: 916–22.

139 Raya-Povedano JL, Romero-Martín S, Elías-Cabot E, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. AI-based Strategies to Reduce Workload in Breast Cancer Screening with Mammography and Tomosynthesis: A Retrospective Evaluation. *Radiology* 2021; **300**: 57–65.

140 Pinto MC, Rodriguez-Ruiz A, Pedersen K, *et al.* Impact of Artificial Intelligence Decision Support Using Deep Learning on Breast Cancer Screening Interpretation with Single-View Wide-Angle Digital Breast Tomosynthesis. *Radiology* 2021; **300**: 529–36.

141 van Winkel SL, Rodríguez-Ruiz A, Appelman L, *et al.* Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study. *Eur Radiol* 2021; published online May 4. DOI:10.1007/s00330-021-07992-w.

142 Kirkwood BR, Sterne JAC. Essential medical statistics, 2. ed., [Nachdr.]. Oxford: Blackwell Science, 2009.

143 Junge MRJ, Dettori JR. ROC Solid: Receiver Operator Characteristic (ROC) Curves as a Foundation for Better Diagnostic Tests. *Global Spine J* 2018; **8**: 424–9.

144 Halling-Brown MD, Warren LM, Ward D, *et al.* OPTIMAM Mammography Image Database: A Large-Scale Resource of Mammography Images and Clinical Data. *Radiology: Artificial Intelligence* 2021; **3**: e200103.

145 Dembrower K, Lindholm P, Strand F. 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). *J Digit Imaging* 2020; **33**: 408–13.

146 Cossío F, Schurz H, Engström M, *et al.* VAI-B: a multicenter platform for the external validation of artificial intelligence algorithms in breast imaging. *J Med Imag* 2023; **10**. DOI:10.1117/1.JMI.10.6.061404.

147 Larsen M, Aglen CF, Hoff SR, Lund-Hanssen H, Hofvind S. Possible strategies for use of artificial intelligence in screen-reading of mammograms, based on retrospective data from 122,969 screening examinations. *Eur Radiol* 2022; **32**: 8238–46.

148 Marinovich ML, Wylie E, Lotter W, *et al.* Artificial intelligence (AI) for breast cancer screening: BreastScreen population-based cohort study of cancer detection. *EBioMedicine* 2023; **90**: 104498.

149 Hickman SE, Payne NR, Black RT, *et al.* Mammography Breast Cancer Screening Triage Using Deep Learning: A UK Retrospective Study. *Radiology* 2023; **309**: e231173.

150 McKinney SM, Sieniek M, Godbole V, *et al.* International evaluation of an AI system for breast cancer screening. *Nature* 2020; **577**: 89–94.

151 Dembrower K, Wåhlin E, Liu Y, *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**: e468–74.

152 Sharma N, Ng AY, James JJ, *et al.* Multi-vendor evaluation of artificial intelligence as an independent reader for double reading in breast cancer screening on 275,900 mammograms. *BMC Cancer* 2023; **23**: 460.

153 Balta C, Rodriguez-Ruiz A, Mieskes C, Karssemeijer N, Heywang-Köbrunner SH. Going from double to single reading for screening exams labeled as likely normal by AI: what is the impact? In: Van Ongeval C, Marshall N, Bosmans H, eds. 15th International Workshop on Breast Imaging (IWBI2020). Leuven, Belgium: SPIE, 2020: 66.

154 Lång K. Study protocol: Mammography Screening With Artificial Intelligence (MASAI). 2022; published online Dec 13. https://clinicaltrials.gov/study/NCT04838756.

155 Romero-Martín S, Elías-Cabot E, Raya-Povedano JL, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. Stand-Alone Use of Artificial Intelligence for Digital Mammography and Digital Breast Tomosynthesis Screening: A Retrospective Evaluation. *Radiology* 2022; **302**: 535–42.

156 Elías-Cabot E, Romero-Martín S, Raya-Povedano JL, Brehl A-K, Álvarez-Benito M. Impact of real-life use of artificial intelligence as support for human reading in a population-based breast cancer screening program with mammography and tomosynthesis. *Eur Radiol* 2023; published online Nov 17. DOI:10.1007/s00330-023-10426-4.

157 Lotter W, Diab AR, Haslam B, *et al.* Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat Med* 2021; **27**: 244–9.

158 de Vries CF, Colosimo SJ, Staff RT, *et al.* Impact of Different Mammography Systems on Artificial Intelligence Performance in Breast Cancer Screening. *Radiol Artif Intell* 2023; **5**: e220146.

159 Gur D, Bandos AI, Cohen CS, *et al.* The 'laboratory' effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 2008; **249**: 47–53.

160 Clift AK, Dodwell D, Lord S, *et al.* The current status of risk-stratified breast screening. *Br J Cancer* 2022; **126**: 533–50.

161 Harkness EF, Astley SM, Evans DG. Risk-based breast cancer screening strategies in women. *Best Practice & Research Clinical Obstetrics & Gynaecology* 2019; : S1521693419301695.

162 Shoshan Y, Bakalo R, Gilboa-Solomon F, *et al.* Artificial Intelligence for Reducing Workload in Breast Cancer Screening with Digital Breast Tomosynthesis. *Radiology* 2022; **303**: 69–77.

163 Letter H, Peratikos M, Toledano A, *et al.* Use of Artificial Intelligence for Digital Breast Tomosynthesis Screening: A Preliminary Real-world Experience. *Journal of Breast Imaging* 2023; **5**: 258–66.

164 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 Radiol Open* 2019; **8**: 2058460119880315.

165 Ongena YP, Yakar D, Haan M, Kwee TC. Artificial Intelligence in Screening Mammography: A Population Survey of Women's Preferences. *Journal of the American College of Radiology* 2021; **18**: 79–86.

166 Pesapane F, Rotili A, Valconi E, *et al.* Women's perceptions and attitudes to the use of AI in breast cancer screening: a survey in a cancer referral centre. *Br J Radiol* 2023; **96**: 20220569.

167 Carter SM, Carolan L, Saint James Aquino Y, *et al.* Australian women's judgements about using artificial intelligence to read mammograms in breast cancer screening. *DIGITAL HEALTH* 2023; **9**: 20552076231191057.

168 Rowell C, Sebro R. Who Will Get Paid for Artificial Intelligence in Medicine? *Radiology: Artificial Intelligence* 2022; **4**: e220054.

169 Sheller MJ, Edwards B, Reina GA, *et al.* Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep* 2020; **10**: 12598.

170 Samuelson P. Generative AI meets copyright. *Science* 2023; **381**: 158–61.

171 Francis A, Thomas J, Fallowfield L, *et al.* Addressing overtreatment of screen detected DCIS; the LORIS trial. *European Journal of Cancer* 2015; **51**: 2296–303.

172 Elshof LE, Tryfonidis K, Slaets L, *et al.* Feasibility of a prospective, randomised, open-label, international multicentre, phase III, non-inferiority trial to assess the safety of active surveillance for low risk ductal carcinoma in situ – The LORD study. *European Journal of Cancer* 2015; **51**: 1497–510.

173 Wheelwright S, Matthews L, Jenkins V, *et al.* Recruiting women with ductal carcinoma in situ to a randomised controlled trial: lessons from the LORIS study. *Trials* 2023; **24**: 670.

174 Schmitz RSJM, Engelhardt EG, Gerritsma MA, *et al.* Active surveillance versus treatment in low-risk DCIS: Women's preferences in the LORD-trial. *European Journal of Cancer* 2023; **192**: 113276.

175 Strand F. Study protocol: Using AI to Select Women for Supplemental MRI in Breast Cancer Screening (ScreenTrustMRI). 2023; published online Oct 5. https://clinicaltrials.gov/study/NCT04832594.

Errata

Paper 4

There are some numbers in the results paragraph of the abstract that is inconsistent with the main text. The paragraph should read (changes are underlined):

"If using a threshold of 9.0, 24 (25%) more cancers would be detected compared to using DM alone. Of the 41 cancers only detected on DBT, 59% would be detected, with only 1493 (10%) of the women examined with both DM and DBT. The detection rate for the added DBT would be 16/1000 women, whereas the false-positive recalls would be increased with 60 (22%)."

There is a similar error in the second sentence of the first paragraph of the discussion section, which should read (changes are underlined):

"We found that using a threshold of 9.0, 10% of the women would have DBT added, and with DM + DBT combination 25% more cancers would be detected, at a cost of 22% increase in false positives."

Printed by Media-Tryck, Lund 2024 NORDIC SWAN ECOLABEL 3041 0903

The overall aim of this thesis is to find ways of using artificial intelligence (AI) to improve breast cancer screening. It was investigated if an AI system analysing either mammography or breast tomosynthesis examinations could be used to remove normal examinations from reading, replace the second reader, detect additional cancers or individualise the screening of high-risk cases. Further, a breast imaging research database was created to facilitate future research.

Victor Dahlblom M.D. is a radiology resident at Skåne University Hospital in Malmö, Sweden.

Diagnostic Radiology, Malmö
Department of Translational Medicine

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2024:36
ISBN 978-91-8021-529-9
ISSN 1652-8220