Automatic volumetric breast density in mammography screening using digital mammography and digital breast tomosynthesis

Bjørn Helge Østerås

Doctoral Thesis

Faculty of Medicine

University of Oslo

Department of diagnostic physics

Oslo University Hospital

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo

ISBN 978-82-8377-706-2

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

This thesis is dedicated to the many thousand women volunteering for mammography screening trials

Acknowledgement

This work was performed from 2013 to 2019 at the Department for Diagnostic Physics in collaboration with The Breast Imaging Center at Oslo University Hospital, Oslo, Norway.

I wish to express my gratitude to the many people who have contributed and made this work possible.

First, I would like to thank my main supervisor Anne Catrine Martinsen for always being positive to new ideas and projects. Your energy and positive attitude have inspired me, and made this thesis possible. I would like to thank you for your essential contribution of knowledge and scientific guidance, which has been invaluable, in this project and in the daily work as a medical physicist.

Secondly, I would like to thank my Co-supervisor Per Skaane for allowing me to take part in the Oslo Tomosynthesis Screening Trial. It has been a privilege to learn from a supervisor with such extensive experience in running well designed clinical trials. Your input is always valuable, and you have helped me focus on clinically relevant findings, rather than getting lost in technical details.

I am also very grateful to Randi Gullien for always being enthusiastic, and for your help in practical matters throughout the project. Without you, this project may never have landed.

I would like to thank my co-authors: Helene, Khalida, Ellen and Unni for their many hours of breast density assessment. I would also like to thank Ragnhild for valuable help in with statistics.

I would also like to thank Hologic, especially Loren and Ashwini for sharing their experience and providing valuable input regarding Quantra™.

I am grateful to my friends and colleagues at the Department of Diagnostic Physics for a good social environment and many interesting discussions. I am privileged to work among a group of such talented and interesting people. A special thanks to my leader Hilde, for always supporting me, and for allowing me the time to finish the PhD thesis.

I would like to thank The Breast Imaging Center at Oslo University hospital for always making me feel welcome.

I am grateful to the women who volunteered for the Oslo Tomosynthesis Screening Trial. It takes courage to enroll in trials testing new equipment.

I would like to thank my parents Karen Marie and Olav for always helping and being there for me.

I am grateful to my daughters, Martine and Emilie, for showing me what really is important, and for helping me take my mind of breast imaging.

And last, but not least, I would like to express my deepest gratitude to my wife, Carina. You have shown tremendous support throughout this work. Thank you for all your love and patience.

Table of contents

Abbreviations

2D	2-Dimensional
3D	3-Dimensional
ACR	American College of Radiology
AEC	Automatic Exposure Control
Ag	Silver
AGD	Average Glandular Dose
Al	Aluminum
AUC	Area Under the Curve
BI-RADS	Breast Imaging- Reporting and Data System
CAD	Computer Aided Detection
CBE	Clinical Breast Examination
CC	Craniocaudal
CI	Confidence Interval
CNR	Contrast to Noise Ratio
CR	Computed radiography
CT	Computed Tomography
DBT	Digital Breast Tomosynthesis
DICOM	Digital Imaging and Communications in Medicine
DM	Digital Mammography
DMIST	Digital Mammographic Imaging Screening Trial
Eq.	Equation
FDA	Food and Drug Administration
GE	General Electric

HHUS	Hand-Held Ultrasound
HVL	Half Value Layer
IAEA	International Atomic Energy Agency
KERMA	Kinetic Energy Released in MAtter
keV	kiloelectron Volts
kVp	kiloVolt peak
mA	milliAmpére (tube current)
mAs	milliAmpére-seconds (product of tube current and exposure time)
MLO	Mediolateral Oblique
Mo	Molybdenum
MRI	Magnetic Resonance Imaging
OTST	Oslo Tomosynthesis Screening Trial
RCT	Randomized Controlled Trial
Rh	Rhodium
ROC	Receiver Operating Characteristics
QD	Quantized Density
SCC	Six Category Classification
SFM	Screen-Film Mammography
SMF	Standard Mammographic Form
SNR	Signal to Noise Ratio
STORM	Screening with Tomosynthesis OR standard Mammography
TFT	Thin-Film Transistor
TMIST	Tomosynthesis Mammographic Imaging Screening Trial
U.S.	United States

U.K.	United Kingdom
VDG	Volpara Density Grade
W	Tungsten

List of papers

Paper I:

Classification of fatty and dense breast parenchyma: comparison of automatic volumetric density measurement and radiologists' classification and their inter-observer variation

Bjørn Helge Østerås, Anne Catrine T. Martinsen, Siri Helene B. Brandal, Khalida Nasreen Chaudhry, Ellen Eben, Unni Haakenaasen, Ragnhild Sørum Falk, Per Skaane

Acta Radiologica, 2016;57:1178-1185

Paper II:

BI-RADS density classification from areometric and volumetric automatic breast density measurements

Bjørn Helge Østerås, Anne Catrine T. Martinsen, Siri Helene B. Brandal, Khalida Nasreen Chaudhry, Ellen Eben, Unni Haakenaasen, Ragnhild Sørum Falk, Per Skaane

Academic Radiology, 2016;23:468-478

Paper III:

Average glandular dose in paired digital mammography and digital breast tomosynthesis acquisitions in a population based screening program: effects of measuring breast density, air kerma and beam quality

Bjørn Helge Østerås, Per Skaane, Randi Gullien, Anne Catrine T. Martinsen

Physics in Medicine and Biology, 2018;63(3):035006 (14 pp)

Paper IV

Digital mammography versus breast tomosynthesis impact of breast density on diagnostic performance in population-based screening

Bjørn Helge Østerås, Anne Catrine T. Martinsen, Randi Gullien, Per Skaane

Radiology, 2019;293(1):60-68

1 Introduction

A major challenge in mammography screening is women with dense breasts, as breast cancer can be masked by glandular tissue (1,2). This challenge still exists in modern digital mammography (DM) (3,4). To address this, breast density legislation has gradually been implemented in the United States (U.S.) from 2008 to 2019 (5). Under this law, all women with dense breasts (about half of all women) are informed of their breast density and encouraged to discuss supplemental screening using ultrasound or magnetic resonance imaging with their physician (6). Austria has implemented supplemental ultrasound for all women with dense breasts (7), and screening programs around the world are debating whether to implement breast density assessment into their screening programs.

Currently, radiologists evaluate mammographic breast density according to a scale called BI-RADS density (Breast Imaging- Reporting and Data System). The woman's breast density is classified into one of the following categories: almost entirely fatty, scattered fibroglandular, heterogeneously dense and extremely dense breasts. It has been shown that this assessment is highly subjective. Therefore, there is a high chance that two different radiologists will classify breasts differently (8). Recently, software has become available that calculates the breast density automatically and reproducibly. If implemented, this can lead to objective and reproducible breast density assessment in mammography screening.

Over the last decade a new way of performing mammography screening has been introduced, digital breast tomosynthesis (DBT) (9). Instead of performing one x-ray image of the breast, a series of projections are acquired at slightly different angles. These images are reconstructed into an image stack in which a plane of breast anatomy is in focus in each image. This helps the radiologist get an impression of the 3-dimensional distribution of the breast tissue. Potentially this can help the radiologist "see through" the dense tissue, thereby improving the results for women with dense breasts who undergo mammography screening.

There are two aims of this thesis: The primary aim was to compare density classification using BI-RADS- and automatically calculated breast density. Secondary, to compare digital mammography and digital breast tomosynthesis with respect to diagnostic accuracy and radiation dose for women of different breast density categories.

2 Background

2.1 The breast and breast cancer

2.1.1 Breast cancer

Breast cancer is the most common type of cancer among women. In 2018, world age-standardized incidence rate was 46.3 and 87. 5 per 100,000 person-years, with corresponding mortality of 11.0 and 13.0 worldwide and in Norway, respectively (10–12). Breast cancer has among highest mortality rate for women, only superseded by lung cancer in Norway and the U.S. (13,14). In the U.S. it is estimated that breast cancer caused 783,000 years of life lost, with 19 years lost on average per cancer death (14). The 5 year relative survival for different breast cancer stages are about: 100 % for stage 0 and I disease, 93 % for stage II, 72 % for stage III and 22 % for stage IV (15). Therefore, it is critical to detect and treat the cancer before it evolves into a metastatic disease (stage IV).

2.1.2 Normal breast anatomy

A sagittal cross section of the normal anatomy of the female breast with annotations is shown in Figure 1. A normal female breast consists of lobules (milk producing glands) that group together into 15 – 20 lobes in a spoke-like pattern (16). Ducts connect the milk-secreting lobular units to the nipple. Smaller ducts near the lobular units converge into larger collecting ducts that open into the lactiferous sinuses at the base of the nipple. These ducts are lined with epithelial cells and basal cells. Surrounding the glandular tissue is dense fibrous stroma mixed with adipose tissue. These glandular and fibrous tissues is often labelled fibroglandular tissues.

2.1.3 Variations of anatomy

The relative amount of fatty to fibroglandular tissues can vary greatly among women dependent on hereditary traits (17). Throughout the woman's life, the ratio of fatty- and fibroglandular tissue changes according to age, menopausal status, age at first birth and number of children among others (18), as the epithelium can atrophy and is replaced by fatty tissue (19). Other factors such as body habitus, use of hormones post menopause and alcohol consumption may also affect the relative amount of fibroglandular and fatty tissues (18).

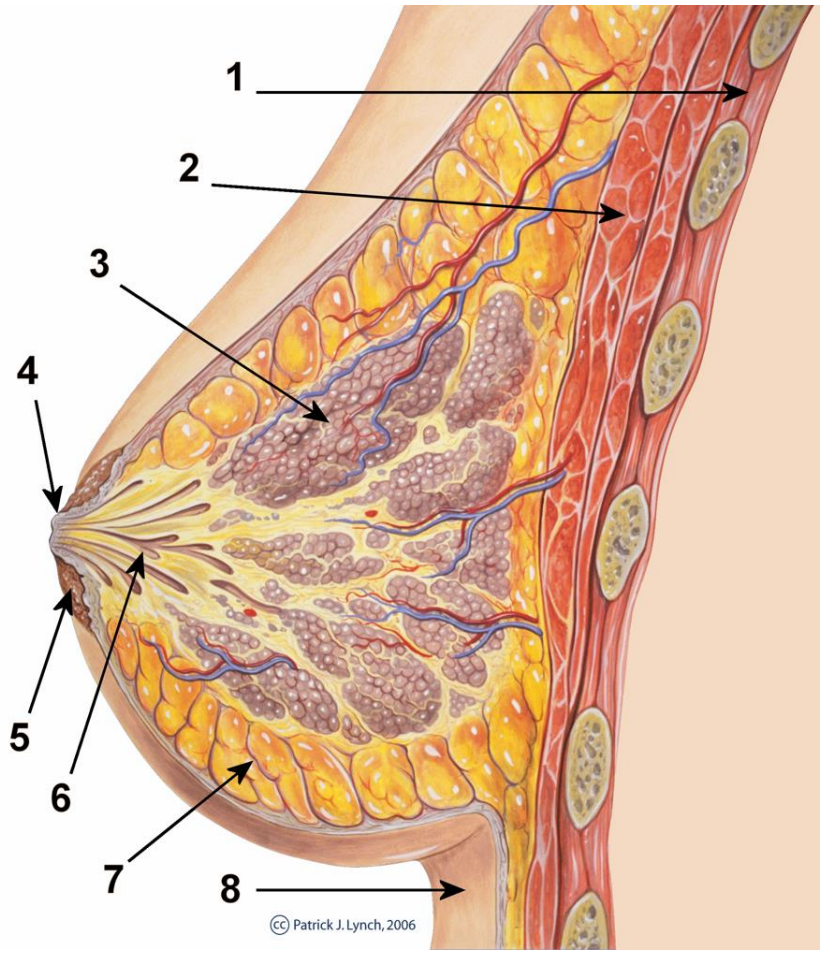

Figure 1: An annotated illustration of the sagittal cross section of a breast with normal anatomy. The organs shown is: 1 – The chest wall. 2 – The pectoralis muscle. 3 – Lobules. 4 – The nipple surface. 5 – Areola. 6 – Lactiferous ducts. 7 – Fatty tissue. 8 – The skin. Source: Patrick J. Lynch (Medical Illustrator) (*https://commons.wikimedia.org/wiki/File:Breast_anatomy_normal_scheme.png*)

2.2 Breast examination

2.2.1 Clinical breast examination

Clinical breast examination (CBE) consists of a visual inspection and palpation of both breast's tissue and axillae (20). A limitation of CBE is low sensitivity and large variation in performance depending on the experience and training of the physician (20,21). Younger women have firmer and more nodular breasts, reducing both sensitivity and specificity of CBE compared to older women (20). This firmness is not a good predictor of the relative amount of fibroglandular- and fatty tissue in the breast, which must therefore be determined radiographically, rather than through palpation (20,22).

2.2.2 Mammographic imaging

Mammographic imaging is the most common examination for early detection of breast cancer, due to its ability to detect cancers at an early stage at low-cost and low-radiation dose (23). Figure 2 shows (Figure 2a) the components of a mammography unit, (Figure 2b) a modern mammography unit with an operator console, and (Figure 2c) an illustration of the positioning of the patient for imaging.

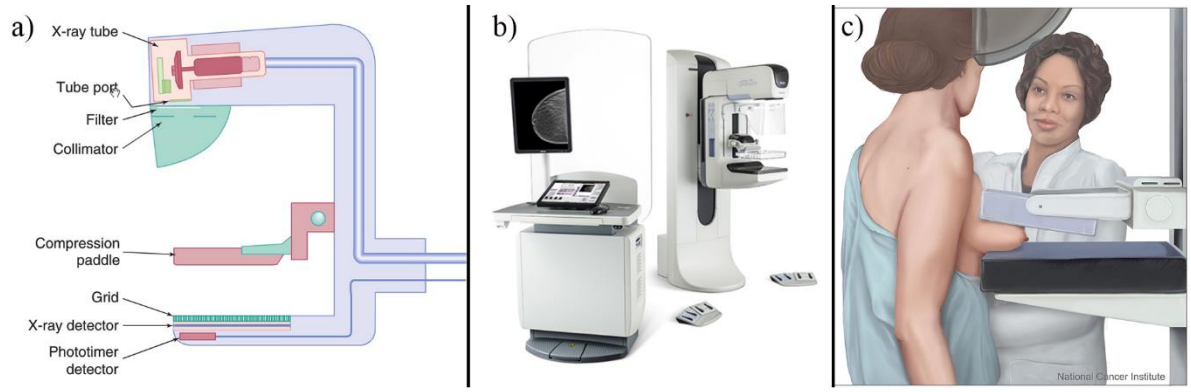

Figure 2: a) Shows an illustration of the components of a modern mammography unit. b) Shows the mammography unit with the operator console. c) shows an illustration of the mammography imaging procedure with the breast under compression. Sources: a) Lippincott Williams & Wilkins (with permission), The Essential Physics of Medical Imaging, Third edition. b) Hologic (Hologic Inc. Bedford, MA, U.S.) (with permission). c) Alan Hoofring, National Cancer Institute, https://visualsonline.cancer.gov/details.cfm?imageid=4361.

During the acquisition, the woman's breast is positioned on a flat surface containing the x-ray detection system. The breast is then compressed firmly and irradiated using an x-ray tube controlled by an x-ray generator. Mammography uses a lower energy (kiloVolt peak, kVp) compared to other radiological modalities. The beam is filtered to remove low energy radiation, which mostly contribute to absorbed breast dose. Filter and/or anode materials, such as rhodium (Rh) or molybdenum (Mo) that emit characteristic radiation at a suitable energy, is often used to optimize the x-ray beam energy. The energy of the beam is characterized by its half value layer (HVL), which is the thickness of aluminum (Al) required to reduce beam intensity to half (measured in mm Al). The amount of radiation emitted is determined by the product of the electric current in the anode of the x-ray tube (tube current or mA) and the exposure time (s), the mAs. Usually the mAs is controlled by an automatic exposure control

(AEC) system, where a sensor monitors the radiation dose to the detector and stops the radiation at the appropriate detector exposure (23).

In a mammography screening examination two projections of each breast are acquired; craniocaudal (CC) and mediolateral oblique (MLO) views. This gives the radiologist two almost perpendicular views of each breast to interpret. In clinical mammography other types of acquisitions targeting specific suspicious areas, such as spot compression and fine focus magnification mammography is often performed.

The image is formed as the detector records energy deposition from the x-rays after being attenuated by the breast tissue. Fibroglandular tissue and cancer are displayed as bright areas and fat as dark grey. This makes detection of non-calcified cancers in breasts with a lot of fibroglandular tissues challenging in mammography.

2.2.3 Screen-film mammography

The first commercially available dedicated mammography units were introduced in 1969 (CGR Senographe), based on work by Charles-Marie Gros (24,25). The following dedicated screen-film units facilitated big improvement of image quality and radiation dose over previous general radiography units using industrial film (24). In screen-film mammography (SFM), the x-rays were absorbed by a screen containing phosphor, which emitted light absorbed by a film (26). The dose response had a maximum gradient within a limited dose range. Therefore, it was important to properly expose the film to optimize image contrast (26). Still, SFM had issues with providing optimal contrast to both the dense parts of the breast and the skin line (23). SFM was interpreted by the radiologists viewing the film in front of a light box, often aided by a magnifying glass. Thus, the film doubled as a detection system, display device and storage media.

SFM was used extensively until the first digital mammography unit (General Electric (GE) Senographe 2000D) was Food and Drug administration (FDA) approved in 2000 (24). Adaptation of digital mammography was a gradual process from the first clinical trials in the early 2000s (27–30) to 2012 (with 98 % coverage in the U.S.) (31). Still, it is worth noting that much of the literature concerning the efficacy of mammography and the issue of dense breasts is based on SFM (32).

2.2.4 Xeromammography and computed radiography

Xeromammography became available in the early 1970's, where an aluminum plate covered in amorphous selenium (a-Se) served as the x-ray absorber. The aluminum plate was charged

with positive ions. When the selenium was exposed to x-rays it would generate electron-hole pairs which would discharge the plate according to the absorbed dose. Dusting the plate with charged aerosol would produce an image of the breast (26). This system was discontinued and is no longer commercially available (24).

Computed radiography (CR) systems use phosphor to detect x-rays. Electrons in the phosphor crystals are freed from the crystal matric and stored in “traps”. The phosphor plate is then scanned using a laser beam, releasing the trapped electrons. These electrons emit light which can be detected and collected using a photomultiplier tube. This system was implemented using removable cassettes. CR systems have worse image quality and higher radiation dose than modern DM systems (33,34).

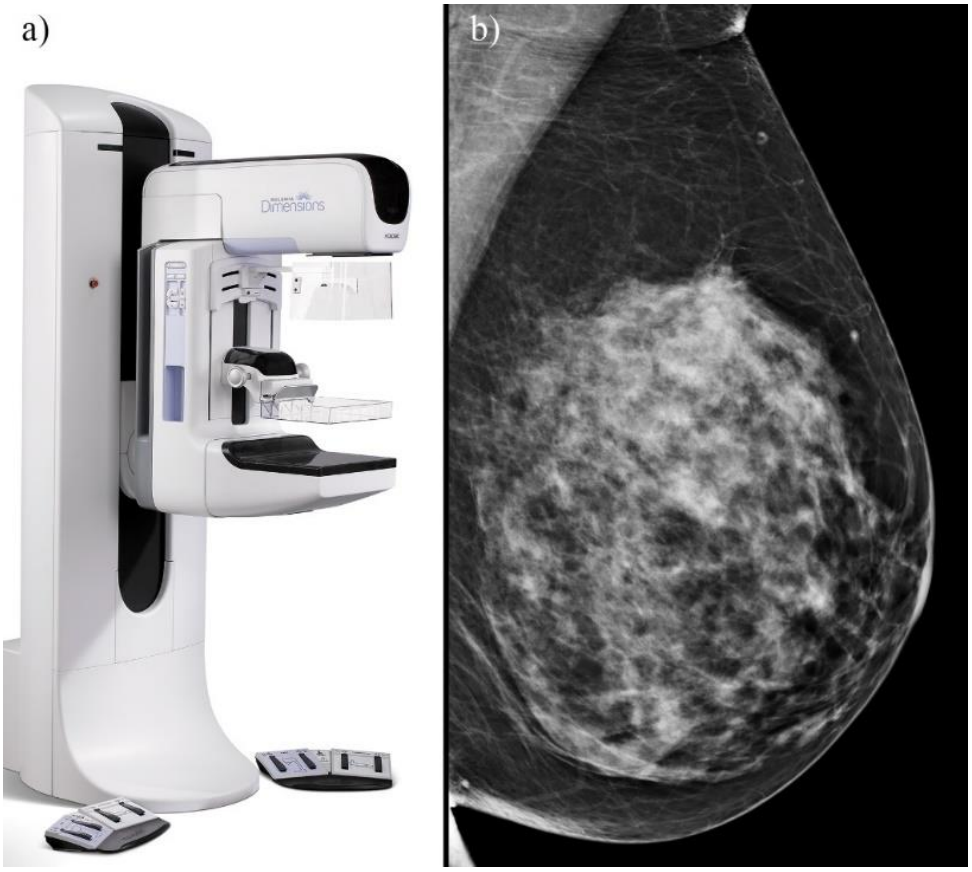

Figure 3: a) shows a modern mammography unit (Hologic Selenia Dimensions). b) shows a DM image of an MLO projection of a left breast. Source: a) Hologic Inc. (with permission).

2.2.5 Digital mammography

Today, digital mammography (DM) is the gold standard modality for breast cancer screening, which all new modalities will be evaluated against. The DM detectors consists of a thin-film transistor (TFT) array of amorphous silicon which collects electric charge. Early DM

equipment (such as GE Senographe 2000D) used a layer of cesium iodide to convert energy deposited by x-rays into light, which a photodiode converts back to charge for collection by the TFT (indirect detection). Modern DM (Figure 3a) instead uses a layer of amorphous Selenium (a-Se) to produce charge from the energy deposition by x-rays, which is collected by the TFT (direct detection). The latter yields the best spatial resolution (23). One major advantage of DM is the linear dose response of the detector, which facilitates optimal image contrast of the whole breast (23) (Figure 3b). In modern DM, the detector itself is used as input for the AEC, allowing the AEC to account for specific areas of high attenuation of x-rays.

Another major advantage of DM is that the image display and storage is decoupled from x-ray detection. This facilitates both improvements in image interpretation through postprocessing of images and convenient storage and retrieval of images. The images are stored in a Digital Imaging and Communications in Medicine (DICOM) format which attaches metadata describing the examination.

The initial trials comparing DM to SFM showed an improvement in cancer detection rate using DM, especially for women with dense breast parenchyma (29,30,35–37).

2.2.6 Digital breast tomosynthesis

The concept DBT was published in 1997 (9). DBT uses equipment similar to DM, but allows for the tube to be moved in an arc spanning ± 7.5^{o} to ± 50^{o} (depending on vendor) over the breast while acquiring low dose projections (Figure 4a). These projections are reconstructed to yield a stack of images where in every image, a single plane of the breast anatomy is in focus while the rest is blurred (pseudo 3D (3-Dimensional)) (Figure 4b). The images are typically reconstructed with 1 mm interval between focus planes with a slice sensitivity profile of about 2.6 mm full width half max (for the unit used in this thesis). As DBT acquires an incomplete set of raw data and a limited number of projections compared to CT, DBT cannot reconstruct true 3-dimensional volumes and produces out of plane image artifacts (38).

The pseudo 3D images produced by DBT can potentially reduce the issue with superposition of fibroglandular tissue over abnormalities in the breast. Potentially improving cancer detection and reducing the number of false positives due to pseudo lesions resulting from overlapping fibroglandular structures.

A challenge in implementation of DBT is that images in consecutive screening rounds often are compared to look for longitudinal changes in the breast. As a DBT images are

difficult to compare to previous DM acquisitions, an image comparable to DM must be included. As performing DM and DBT for all views leads to unacceptable radiation exposure of the women, vendors have invented a method of synthesizing DM images from DBT acquisitions (39) (Figure 4c).

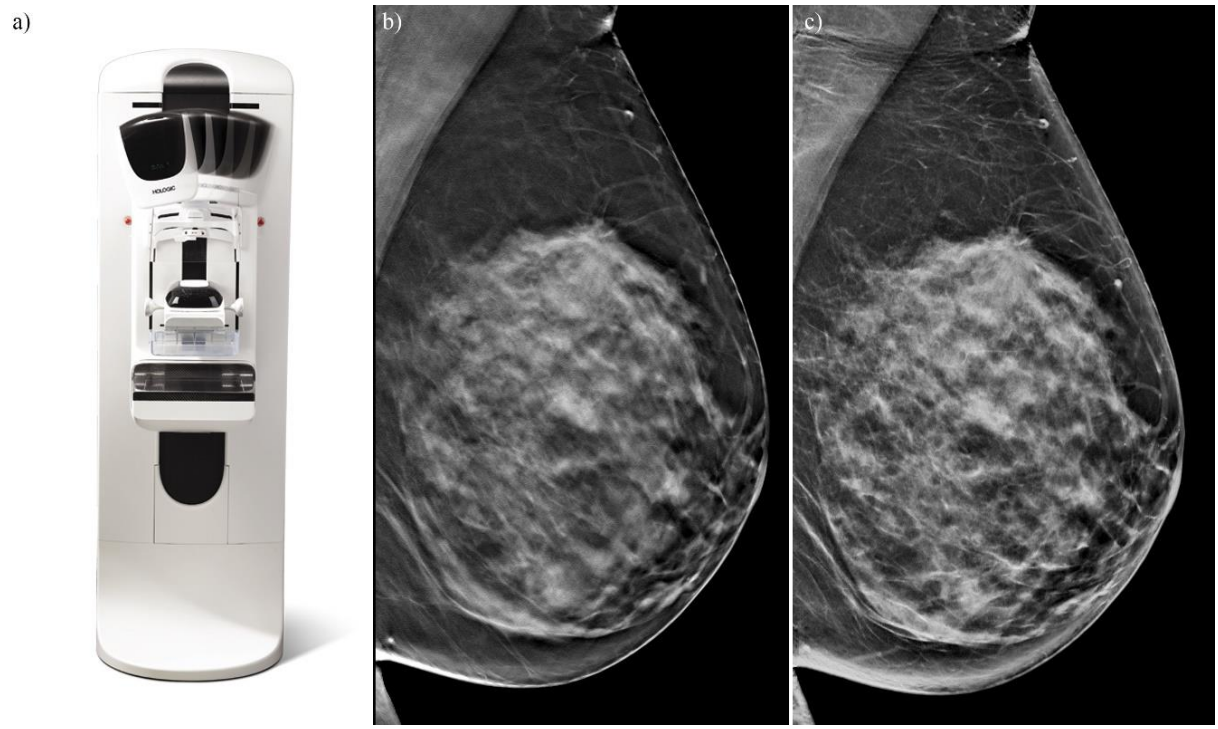

Figure 4: a) shows an illustration of the same mammography unit shown in Figure 3a performing a DBT examination with ± 7.5^o arc. b) Shows a DBT image where one slice of the anatomy is in focus, while the rest is blurred. c) Shows the corresponding synthetic DM image of the DBT acquisition. This is the same view and breast shown in Figure 3b. Source: a) Hologic Inc. (with permission).

In DBT, women have to stay under compression for longer than in DM, although some experiment with using less compression force (40). There is currently a lot of research comparing DBT to DM for breast cancer screening. Prospective studies show improvements in cancer detection and/or recall rate (40–44). So far there is insufficient evidence for implementation of DBT in population-based breast cancer screening (45,46).

Interpretation of mammographic images in screening

The Norwegian breast cancer screening program (BreastScreen Norway) uses an ordinal five point scale to assess the probability of malignancy based on screening mammograms (Table 1) (47). A score of 2 or more is considered positive.

Table 1: BreastScreen Norway categories and their associated assessment.

BreastScreen Norway Category	Assessment
1	Normal/Definitely benign
2	Probably benign
3	Indeterminate
4	Probably malignant
5	Malignant

2.2.7 Contrast enhanced mammography

The leaky blood vessels associated with angiogenesis in breast cancers can be imaged using dual-energy contrast enhanced mammography. Iodinated contrast media is administered prior to the examination. Then two acquisitions are made, one at 23-32 kVp and one at 45-49 kVp with an additional copper filter. The low energy image serves as a conventional DM examination, while the high energy image is uninterpretable (48). The attenuation of iodine increases dramatically at 33.2 keV (kiloelectron Volt) as above this energy photon has sufficient energy to eject the k-shell electrons through the photoelectric effect. Thus, subtracting the two images crates an iodine image which highlights cancers in a comparable manner as contrast enhanced MRI (48). It has been reported that contrast enhanced mammography increases the sensitivity in women with high risk of breast cancer and dense breasts (49).

2.2.8 Ultrasound

Ultrasound is an acoustic imaging technique. These images allow radiologists to detect some lesions that are not visible using mammography, especially in women with breasts containing extensive fibroglandular tissue (50–52). Hand-held ultrasound (HHUS) is the most common technique and is performed by a physician or a technician. It is limited by requiring well trained operators, a small field of view and that the lesion must be seen during the examination (thus HHUS is operator dependent) (50). The examination is also time consuming (53).

An alternative to HHUS is automated breast ultrasound, where a technologist positions a paddle on the breast. The transducer then moves across the paddle, scanning the breast automatically. A total of three views are acquired for each breast (front, inner and outer) for

an acquisition time of about 1 min per view (52). The images produced are standardized and allows for interpretation on a workstation using independent double reading. Which increases feasibility of implementation in population-based screening (53).

2.2.9 Magnetic resonance imaging

Magnetic resonance imaging (MRI) is an imaging technique images utilizing the magnetic properties of the hydrogen nucleus, which is abundant in the body. Using various examination protocols MRI can acquire images highlighting soft tissue, fat and/or liquids, and tissue characteristics such as proton density or diffusion. Thus, MRI can provide images of excellent soft tissue contrast. Usually gadolinium contrast media is administered, facilitating imaging of the vasculature and the leaking blood vessels usually present in breast cancer. MRI is a more sensitive technique compared to mammography and is not limited by presence of fibroglandular tissue. But as contrast media is administered the tolerance for using MRI in routine screening is low. Many women experience claustrophobia in the narrow tube if the scanner. The MRI scan is also time consuming. To increase MRI tolerance a 3-minute fast MRI protocol (abbreviated MRI), which still uses contrast media, has been suggested for screening (54).

2.3 Radiation physics and image quality

2.3.1 Photon interactions in tissue

The attenuation of x-rays in the mammographic energy range (10 – 50 keV) is primarily due to two processes: Photoelectric effect and Compton scatter. The probability of each interaction for a specific material is described by the materials' linear attenuation coefficient μ (cm^{-1}) or more conveniently (normalized to density) the mass attenuation coefficient $\mu \cdot \rho^{-1}$ in units ($cm^2 \cdot g^{-1}$).

Photoelectric effect

When a photon interacts through the photoelectric effect, all its energy is transferred to an electron, which is ejected from the atom. The probability of this interaction increases strongly with decreasing energy and increasing atomic number ($\sim Z^3 \cdot E^{-3}$) (23). Which is why the photoelectric effect contributes strongly to image contrast. Especially between tissues with substantially different atomic number, such as calcifications and soft tissue. As no photon is emitted in photoelectric effect, there is no scattered photon that can reach the detector and degrade image quality.

Compton scattering

Compton scatter is an interaction between the photon and an electron which is ejected from the atom. The leftover energy is released as a lower energy photon. The probability of Compton scatter is largely constant with energy and proportional to the number of electrons per gram. Therefore, it is nearly independent of Z (with the exception of hydrogen) (23). For higher energy x-ray spectra, a greater proportion of photons interact through Compton scatter, reducing image contrast compared to lower energies. Additionally, the photon released can reach the imaging detector and further reduce image contrast.

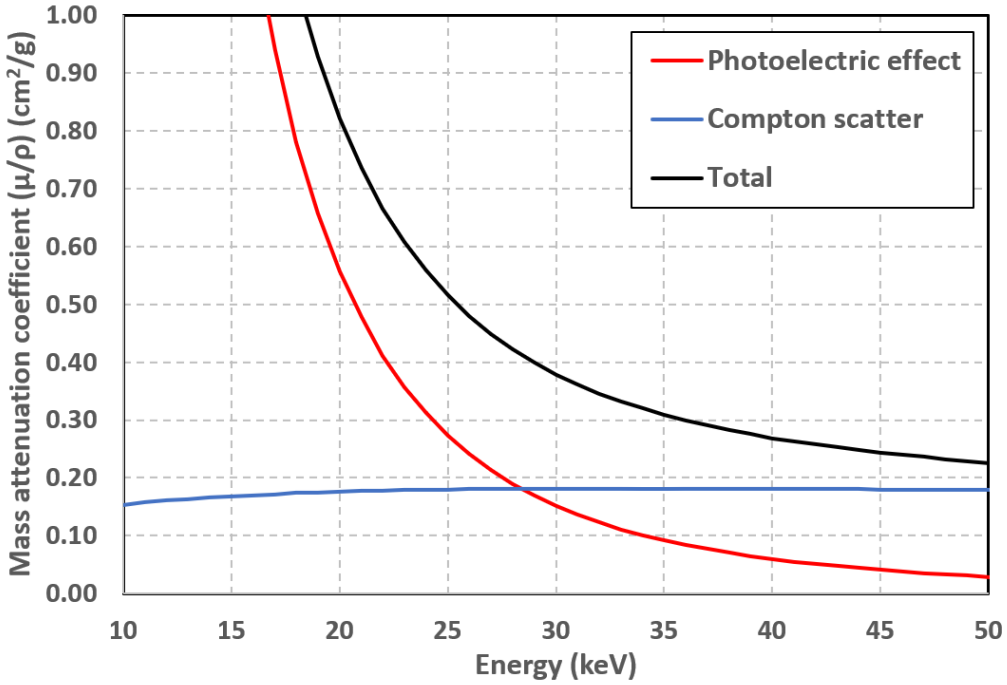

Figure 5: The mass attenuation coefficients for soft tissue (55) *(Z_{eff} = 7.64) for photo electric effect, Compton and total mass attenuation coefficient. Calculated using XMuDat* (56)*, using mass attenuation coefficient from Boone et al.* (57)*.*

Total attenuation

The total linear attenuation coefficient for a photon is the sum of the individual coefficients (eq. 1).

$$\mu_{Total} = \mu_{Photoelectric} + \mu_{Compton\ scatter} \quad (1)$$

Figure 5 shows the mass attenuation coefficient for photoelectric effect, Compton scatter and total attenuation coefficient. for soft tissue (55) calculated using XMuDat (56), using the attenuation data published by Boone et al. (57).

2.3.2 Radiation dose

A beam of photons deposits energy in a two-step process. First, the energy is transferred to charged particles (electrons) due to photoelectric effect and Compton scatter. Then, the charged particles interact with matter through ionization and excitation. This process is often described using the kinetic energy released in matter, or KERMA (K). KERMA is the kinetic energy transferred to charged particles by the photons per unit mass of the absorber and has unit of Gray (Gy) or S.I. units Joules per kilogram ($J \cdot kg^{-1}$). To calculate KERMA the energy fluence (Ψ), the amount of energy passing through a unit area at each energy E, is multiplied by the respective mass energy transfer coefficient ($\mu_{tr} \cdot p^{-1}$). The mass energy transfer coefficient is the mass attenuation coefficient multiplied with the fraction of energy transferred to charged particles. Thus, KERMA at energy E can be calculated as (eq. 2):

$$K = \Psi \left(\frac{\mu_{tr}}{\rho}\right)_E, [Gy\,] \tag{2}$$

Absorbed dose (D) is the energy imparted by all ionizing radiation per unit mass of the irradiated material (eq. 3).

$$D = \frac{E}{m}, [Gy] \tag{3}$$

To estimate absorbed dose, the mass energy absorption coefficient ($\mu_{en} \cdot \rho_0^{-1}$), which also accounts for radiative losses due to bremsstrahlung. Thus, the absorbed dose at energy E can be calculated as (eq. 4):

$$D = \Psi \left(\frac{\mu_{en}}{\rho}\right)_E, [Gy] \tag{4}$$

However, at energies and tissues relevant to diagnostic radiology the mass energy absorption- and mass energy transfer coefficients are close to numerically identical, due to the low amount of bremsstrahlung produced.

Measurement of radiation dose

Ionizing radiation is often measured using an ion chamber which measures electric charge collected from the ions produced in the air of the dosimeter, which is usually converted to air KERMA within the dosimeter system. As the effective atomic numbers in soft tissue and air is similar, dose measurements using ion chambers is a good approximation of dose to soft tissue in the energy range relevant for mammography.

2.3.3 Image quality

Contrast

For a breast irradiated using a homogeneous beam, there will be a difference in the emerging beam intensity at points A and B reflecting the tissue in the respective beam path. This difference is often normalized to intensity A and called the subject contrast C_s (eq. 5) (23).

$$C_s = \frac{(A-B)}{A} \tag{5}$$

Figure 6 (a) shows the differences in attenuation coefficients for fat-, glandular tissue and infiltrating ductal carcinoma. (b) shows the resulting subject contrast. Notice how the contrast is noticeably reduced as the beam energy increases as more photons undergo Compton scatter.

Subject contrast is modulated by the dose response of the detection system. For digital images, raw-data image is non-linearly postprocessed in order to condense the large dynamic range to a useful "for presentation" image. Additionally, the radiologist may adjust the grayscale (window width and window level). All these steps modulate the subject contrast originating in the raw-data image.

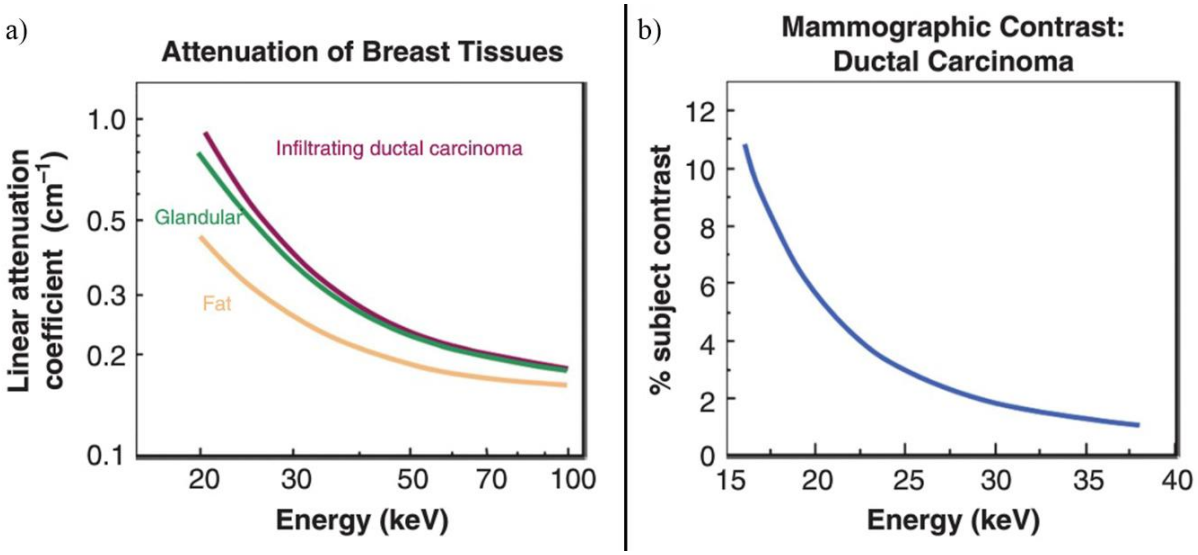

Figure 6: a) The attenuation coefficients of fat-, glandular tissue and infiltrating ductal carcinoma. b) Subject contrast of glandular and ductal carcinoma versus fat due to differences in linear attenuation coefficient. Source: a) Lippincott Williams & Wilkins (with permission), The Essential Physics of Medical Imaging, Third edition (23).

Image noise

Image noise deteriorates the image quality, obscuring details representing the clinically interesting features. Typically, noise is considered to have four components (23):

- Electronic noise, which originates in electronic components of the imaging chain. This noise is independent of radiation dose. Therefore, it becomes more important at low doses.
- Structured noise is signal variations which are constant in time. They typically represent image artifacts, which usually are corrected by calibration.
- Obscuring patterns generated by the anatomy of the patient is called anatomical noise. In mammography there is substantial anatomical noise generated by the glandular tissue which can obscure tumors embedded in, or covered by this tissue.
- Quantum noise is due to the statistical nature of x-rays. This noise is reduced by increasing the radiation dose. A doubling the dose will reduce noise by a factor of $\sqrt{2}$.

The magnitude of noise is often assessed using the standard deviation of the pixel values in an image of a homogeneous object. The visibility of a lesion is tied to the image signal of the lesion compared to the background noise (58), which is often measured using contrast to noise ratio (CNR) or signal to noise ratio (SNR). Objects with lower contrast and smaller objects require less noise in order to be reproducibly detected.

The abovementioned measures have limitations as they ignore the graininess of the noise. The magnitude and frequency dependence of the noise can be assessed using the noise power spectrum. CNR and SNR also ignores the reduction in contrast transfer of the system due to limitations in the spatial resolution, which can be assessed using the modulation transfer function. These can be combined into a measure of SNR as function of spatial frequencies called detective quantum efficiency (23).

2.4 Breast density measurement

2.4.1 Background

The breast consists mainly of two types of tissues; fibroglandular tissue and fat. As fibroglandular tissue attenuates x-rays more than fat it appears white on a mammogram. Therefore, breasts with a lot of fibroglandular tissue are labeled as dense. This density is different and not correlated with the firmness of the breast (22). In 1976 Wolfe discovered

breast density as a risk factor for breast cancer (59) (further explained in chapter 2.5.2). In 1977 Egan and Mosteller found that mammography is less sensitive in dense breasts (1), due to masking of cancer by fibroglandular tissue (2) (further explained in chapter 2.5.3). Since these discoveries breast density has been studied extensively, both in relation to risk and masking. The task of measuring breast density has been performed using various methods. The following chapter gives a short overview of the most commonly used measurement methods.

2.4.2 Subjective breast density classification

Wolfe classification

In 1976 Wolfe was the first to classify breasts into four categories based on the pattern of fibroglandular tissues, with the purpose of providing an index of increased risk of developing breast cancer (Table 2) (59,60).

Table 2: The Wolfe classifications with description of the categories (60).

Wolfe classification	Description
N1	Lowest risk. Parenchyma composed primarily of fat with at most small amounts of "dysplasia". No ducts visible.
P1	Low risk. Parenchyma chiefly fat with prominent ducts in anterior portion up to one-fourth of volume of breast. Also, may be a thin band of ducts extending into a quadrant.
P2	High risk. Severe involvement with prominent duct pattern occupying more than one-fourth of the breast.
DY	Highest risk. Severe involvement with "dysplasia". Often obscures an underlying duct pattern.

Here the N1 and P1 patterns indicate breasts with low risk and P2 and DY indicate breasts with high risk of breast cancer. Women classified with DY breasts was found to have 37 times the risk of breast cancer compared to N1 (60), but these results could not be replicated, perhaps due to issues with inter-observer variability in pattern assessment (61).

Tabár classification

In 1997 Tabár made a classification based on histologic-mammographic correlations and the relative proportions of nodular densities, linear densities, homogeneous fibrous tissue and

radiolucent fat tissue (61,62) (Table 3). Like Wolfe, Tabár's goal was to identify women at increased risk of breast cancer. But the agreement with respect to high and low risk using Wolfe and Tabár classifications was poor (62).

Patten-based breast classification are not widely used today as they are subjective, lack reproducibility and are difficult to assess (61).

Table 3: The Tabár classification with a short description of categories (61,62).

Tabár classification	Short description
I	Balanced proportion of all components of breast tissue with a slight predominance of fibrous tissue.
II	Predominance of fat tissue (fat breast).
III	Predominance of fat tissue with retroareolar residual fibrous tissue.
IV	Predominantly nodular densities.
V	Predominantly fibrous tissue (dense breast).

Boyd classification

Encouraged by the results of Wolfe, Boyd studied breast cancer risk using the percent area of the breast image covered by dense breast tissue. This classification (called SCC) had six categories: A: = 0 %, B: > 0 - < 10 %, C: 10 – 25 %, D: 25 – 50 %, E: 50 – 75 % and D: ≥ 75 %, which were assessed visually by radiologists (63). This classification resulted in consistent breast cancer risk relationships. But even though the scale was quantitative, there was still issues with subjectivity, as the threshold for considering a patch of breast as dense is not clearly defined (61).

BI-RADS density

In 1993 the ACR (American College of Radiology) included reporting of the breast density in BI-RADS 3rd edition (Table 4) (64). After a woman attends breast cancer screening, radiologists are more interested in whether breast density could mask a lesion, rather than the effect of breast density on breast cancer risk, therefore BI-RADS density mainly focus on masking (65). Additionally, BI-RADS density focus on the volume of dense tissue, not the area. This has been criticized as radiologists only have planar images which lack sufficient information for such assessment (65).

Table 4: BI-RADS density 3rd ed. (64).

BI-RADS density 3rd ed.	Description
1	Entirely fat
2	Scattered fibroglandular
3	Heterogeneously dense
4	Extremely dense

In 2003, ACR updated BI-RADS to 4th edition. They elaborated on the description of the categories and added area-based percentages breast density (66), which is confusing as BI-RADS focus on volumetric assessment (61). The BI-RADS 4th edition categories are shown in Table 5, with examples of cases in Figure 7.

Table 5: BI-RADS density 4th ed. (66).

BI-RADS density 4th ed.	Description
I	The breast is almost entirely fat (<25 % glandular)
II	There are scattered densities (approximately 25 - 50 % glandular)
III	The breast tissue is heterogeneously dense, which could obscure detection of small masses (approximately 51 - 75 % glandular)
IV	The breast tissue is extremely dense. This may lower the sensitivity of mammography (> 75 % glandular)

In 2013 ACR updated BI-RADS to the 5th edition. The addition of the area percentages did not change the evaluation of density on average from the 3rd to 4th edition, so they were removed in the 5th edition. The wording was also changed slightly to emphasize the risk of masking and he labels were changed to letters to avoid confusion with the BI-RADS assessment for probability of malignancy (Table 6). In the 5th edition a few breasts with extremely dense tissue in a small part of the breast can be given a higher density compared to the 4th edition.

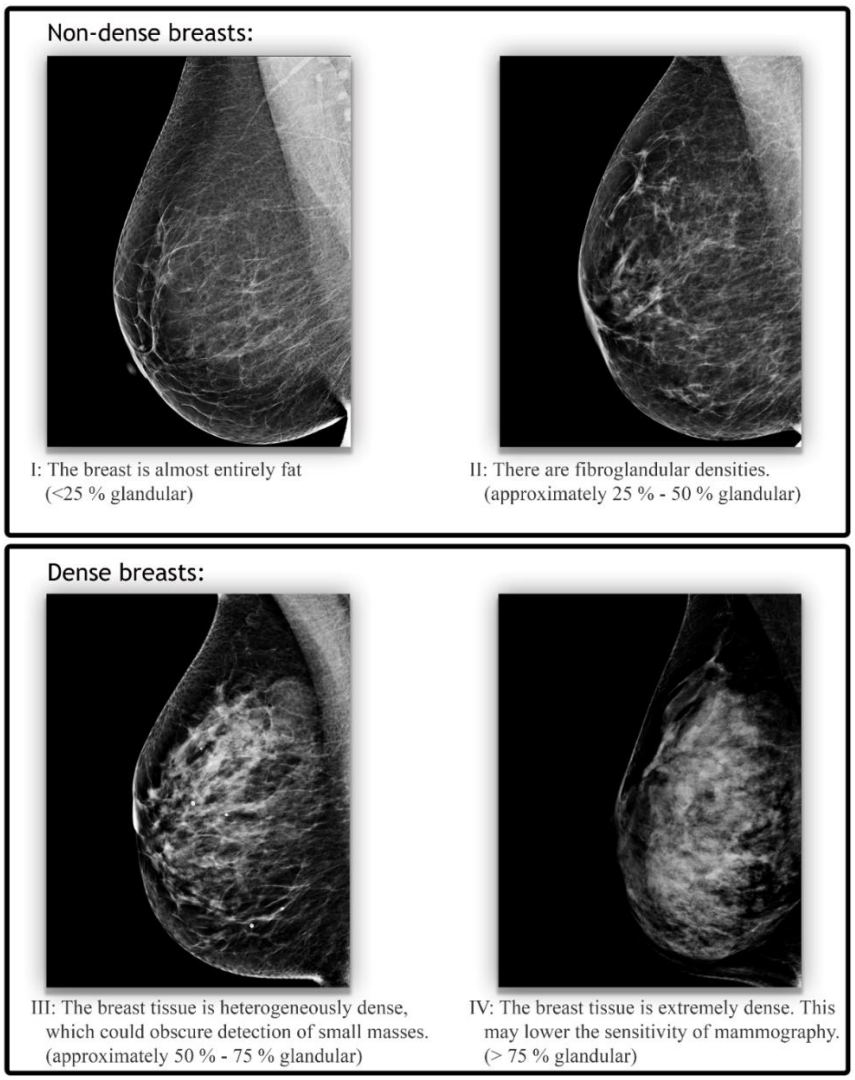

Figure 7: DM MLO views of breasts uniformly categorized in the respective BI-RADS density 4th edition categories in paper I and II.

Table 6: BI-RADS density 5th ed. (67).

BI-RADS density 5th ed.	Description
A	The breast is almost entirely fatty
B	There are scattered areas of fibroglandular density.
C	The breast tissue is heterogeneously dense, which could obscure small masses.
D	The breast tissue is extremely dense, which lowers the sensitivity of mammography.

BI-RADS density is currently the most used density assessment method in breast cancer screening and is included in the standard report. The distribution of the categories is about 10, 40, 40, 10 % from lowest to the highest in screening populations (67). Women in the two highest categories are considered dense, and the two lowest is considered non-dense. Therefore, about 50 % of women in breast cancer screening is considered to have dense breasts. Even though BI-RADS density has been a part of breast cancer screening for almost 30 years, there is still major issues with inter-observer variability (8).

2.4.3 Computer aided mammographic density assessment

Planimetric

As there was major subjectivity in the assessment using SCC classification, an effort was made to make a semi-automatic software called Cumulus (68). This software calculated the ratio of the breast area occupied by dense tissue to the total area of the breast, like SCC. But rather than using an ordinal scale, Cumulus used a continuous scale (68). To use this software the screen-film images had to be digitized, and the operator had to manually set the threshold for what the software considered dense and non-dense. Cumulus then calculated the fraction of the breast area that was dense tissue. As this assessment required manual input, it did not eliminate the variability in assessment (63). This method was also too labor intensive to be implemented in routine mammography screening, but for the next two decades it was considered the gold standard in breast density assessment for research (61). Other planimetric software has been written for research purposes, but none were as popular as Cumulus (61).

A modern commercial area-based density assessment software Densitas™ (Halifax, NS, Canada) uses an Artificial Intelligence driven method on processed DM images to assess area-based density and provides a BI-RADS density like score.

Volumetric

In order to accurately assess the physical amount of fibroglandular tissue, volumetric breast density assessment is necessary. Here, breast density is defined as the ratio of volume fibroglandular tissue to the total volume of the breast.

Volumetric breast density can be estimated using projection images from mammographic imaging. Initially a method called the standard mammographic form (SMF) was used (eq. 6) (61,69–71).

$$E^{imp}(\boldsymbol{x}) = \phi(V_t, \boldsymbol{x}) A_p t_s \int_0^{E_{max}} N_0^{rel}(V_t, \varepsilon) G(\varepsilon) D(\varepsilon) e^{-\mu_{lucite(\varepsilon)} h_{plate}} e^{-h\mu(\varepsilon)} d\varepsilon \quad (6)$$

$E^{imp}(\boldsymbol{x})$ is the energy imparted on the detector element **x** by primary photons. $\phi(V_t, \boldsymbol{x})$ is the photon flux at the tube voltage V_t. A_p and the $t_s$is the area of the pixel and exposure time, respectively. $N_0^{rel}(V_t, \varepsilon)$ is the relative number of photons at energy ε. $G(\varepsilon)$ describes transmission through the grid and $D(\varepsilon)$ describes the absorption ratio of the detector for primary photons. $\mu_{lucite}(\varepsilon)$ is the linear attenuation coefficient of lucite and h_{plate} is the height of the (lucite) compression paddle. The energy imparted is known from the raw data image. The amount, and energy of the radiation is known from the DICOM metadata, which allows for determination of the x-ray spectrum (72) and linear attenuation coefficients (73). The height of the breast (H) is known from the compression thickness, and the breast is modelled as two compartments: height of fibroglandular tissue (h_{int}) and height of fatty tissue (h_{fat}) (eq. 7 – 9).

$$h\mu(\varepsilon) = h_{int}\mu_{int}(\varepsilon) + h_{fat}\mu_{fat}(\varepsilon) \quad (7)$$

$$H = h_{int} + h_{fat} \quad (8)$$

$$h\mu(\varepsilon) = h_{int}\left(\mu_{int}(\varepsilon) - \mu_{fat}(\varepsilon)\right) + H\mu_{fat}(\varepsilon) \quad (9)$$

Eq.9 can be substituted into eq. 6, which can be solved for h_{int}. This type of method is often labeled as an absolute physics approach. As the equation ignores scattered radiation it must be corrected for (71). However, small uncertainties such as a 1 % error in compressed breast thickness could lead to large errors in the estimated breast density using this absolute physics model (61,74). Therefore, these methods require calibration of the image, either using a step wedge in each image (70,75) or using imaging parameters and the breast images for calibration (76,77). The imaging based calibrations are based on finding a pixel representing pure fat, which can be difficult to find in very dense breasts, resulting in potential underestimation on the breast density in these women (78,79). Prior to volumetric density estimation, the algorithm must segment the breast, removing the edges of the compression paddle (if present) and the pectoral muscle. Several implementations for volumetric breast density assessment exists and is explained below. As some of these are commercial software, they are likely to involve elements unknown to the authors. The following explanation serve as a rough guide to the philosophy behind the algorithms.

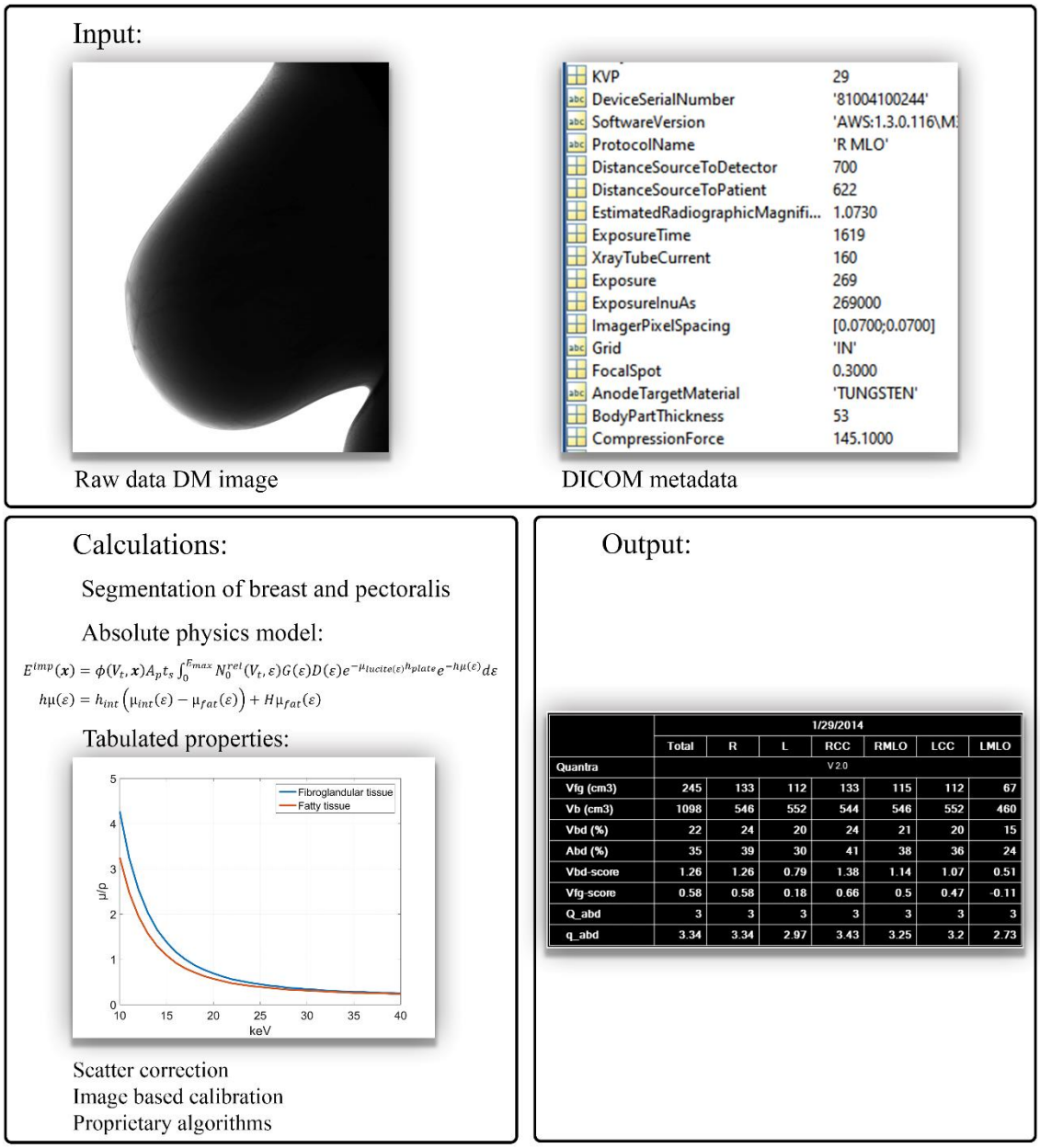

Figure 8: An overview of the input, calculation steps and output, of the Quantra™ version 2.0 algorithm. Source: (output) Elsevier, (83) (with permission).

Hologic Inc. (Bedford, MA, U.S.) was in 2008 the first vendor to have a fully automatic volumetric breast density assessment tool commercially available (80). An overview of the Quantra™ (Hologic Inc.) algorithm is shown in Figure 8. The software, is implemented using a method based on SMF, explained above. However, it includes a number of improvements such as better breast thickness estimation, using more information from the DICOM header and exploiting beneficial properties of DM images compared to SFM (78). Thus, Quantra™ uses an absolute physics approach with improved calibration compared to

the SMF (78). Since the release of their version 1.2 in 2008, version 2.0 (used in this thesis) implemented a correction for the skin. In addition to calculating volumetric density, Quantra™ version 2.0 maps density to BI-RADS 4th edition density as scored by radiologists in the Digital Mammographic Imaging Screening Trial (DMIST) trial (30,81). Thereby providing a BI-RADS density like ordinal score from 1 to 4 called Quantized density (QD). In version 2.0 the thresholds in volumetric density was approximately QD 1: < 5 %, QD 2: 5 – 13 %, QD 3: 13 – 26 % and QD 4 > 26 %. These thresholds are applied on a per image basis. To get the QD score on a per breast basis, the QD score (with decimal precision) is averaged, then rounded to yield a breast-based score. To get a per subject score the maximum QD score of the left and right breast is used (81). Quantra™ can also provide an area-based breast density by applying a threshold to the proportion of fibroglandular tissue to classify each pixel as dense or non-dense. Recently, Quantra™ version 2.2 has been released. This version is directed towards BI-RADS density 5th edition. BI-RADS density 5th edition can classify a woman as having dense breasts if a small portion of the breast is very dense, even if most of the breast contains fatty tissue. Therefore, Quantra™ version 2.2 uses a new algorithm, based on machine learning which includes pattern and texture analysis (82).

Volpara™ (Matakina, Wellington, New Zealand) is another software providing volumetric density assessment. Unlike Quantra™ they have implantation based on the relative physics approach (eq. 10) (77,79).

$$h_d(\boldsymbol{x}) = \frac{ln(P(\boldsymbol{x})/P_{Fat})}{\mu_{fat} - \mu_{dense}} \tag{10}$$

Where h_d is the thickness of dense tissue, $P(\boldsymbol{x})$ and P_{Fat} is the pixel value at location **x** and reference pixel representing fat only (84). In addition to volumetric density, Volpara™ also provides a BI-RADS density like score from 1 to 4 called Volpara Density Grade (VDG), based on thresholds in volumetric density. For version 1.5.0, VDG 1: 0,0 – 4.5 %, VDG 2: 4.5 – 7.5 %, VDG 3: 7.5 – 15.5 % and VDG 4 ≥ 15.5 % (85,86).

Cumulus V is a volumetric breast density algorithm developed at the University of Toronto. This algorithm uses a refinement of the step-wedge phantom calibration (75,87), which is a simplified version of the absolute physics approach explained previously. Although, used in research, the calibration requirement makes this approach infeasible to implement in population-based screening.

These volumetric approaches have been validated towards MRI which produces 3 dimensional datasets showing that both Quantra™ and Volpara™ can predict breast density accurately (77,88,89). The automatic volumetric algorithms have been compared in several studies. Despite being based on different approaches they well correlated and reproducible (84–86,90–92). With the Volpara™ algorithm resulting in higher density estimates compared to Quantra™, especially for very dense breasts. Additionally, the proportion of women given a BI-RADS density like score indicating dense breasts was higher using Volpara™ (version 1.5.0) compared to Quantra™ (version 2.0) (86).

2.5 Breast density and breast cancer screening

2.5.1 Screening for breast cancer

Rationale for screening

Breast cancer screening is based the assumption that breast cancer is a progressive disease for which early detection improves prognosis. As mammography facilitates early detection of breast cancer, one can start treatment of cancer while the prognosis still is favorable. This is critical as even with modern treatment the survival for most stage IV breast cancers remain below 20 % (93). But even with frequent screening, some tumors have growth rates which results in clinical detection between screening rounds, e.g. interval cancers.

Cohort and screening interval

In Norway, mammography screening is offered biannually for women between 50 and 69 years (94). In the United States (U.S.) and elsewhere in Europe screening is generally recommended annually or biannually from 45 - 50 years up to 75 years (14,95,96). Women emphasizing the benefit of screening more than the risk can opt to start screening at 40 in the U.S (14,95).

Mammography screening workflow

In BreastScreen Norway mammography screening is performed using double reading. Two radiologists read the images and if at least one reader scores the case positive, the case is discussed at an arbitration/consensus meeting. Here, a team of radiologists review the case and decide whether to recall the woman for additional assessment. In the U.S. mammography screening is generally performed using a single reader, and consequently without arbitration/consensus meeting. It is worth noting that in the U.S. the recall rate is much higher than in Europe at about 10 % (97), while in Norway it is about 3 % (94).

Benefits of mammography screening

Mammography is the only screening method shown to reduce mortality from breast cancer (98–100). The mortality reduction due to screening is closely followed by the reduction in node positive disease (stage II or worse) (94,101). Breast conserving surgery is preferred for patients with early stage disease, additionally tumor size and node status also influence whether adjuvant radiotherapy should be used (102). Therefore, early stage disease can be treated with less treatment morbidity compared to late stage disease.

Potential harms of mammography screening

Overdiagnosis is: "the diagnosis of a cancer as a result of screening that would not have been diagnosed in the patient's lifetime if breast cancer screening had not taken place" (103). These cancers can be slow growing, indolent or cancers that might regress. Unfortunately, it is currently not possible to determine which cancers will turn out fatal and which are overdiagnosed. Overdiagnosis remains a controversial issue as it is difficult to assess.

Another limitation of mammography is false positive screening results. As the incidence of cancer in the general population is low, most positive interpretations in mammographic screening are false. It is estimated that about 20 % of women in European screening programmes (10 biennial screens between 50 – 70 years) had a false positive result and was recalled for further assessment (104). For U.S. women starting annual screening at 50, there was a 61.3 % chance of a false positive as recall rates are higher compared to Europe (105). When a woman is recalled for further assessment, she will often experience breast cancer-specific psychological stress that can endure for up to 3 years (106).

Some cancers are not detected at screening but detected prior to the next screening round. These cancers are called interval cancers (IC). A commonly used measure of sensitivity of a screening test is the program sensitivity, which is the ratio of screen detected cancers (SDC) over SDC plus IC (107). As the program sensitivity of mammographic screening is in the range 70 - 90 % (108–110), some women will have a negative mammography examination while harboring a breast cancer. These women could be falsely reassured by the negative mammography examination, leading to a delay in diagnosis (95).

Radiation dose can cause radiation induced cancer. However, the risk of radiation induced cancer is reduced with increasing age of the patient (111). As the women targeted by mammography screening are older than 40, the risk of inducing cancer is low. An estimate for

the U.S. screening program yielded 86 cancers due to radiation, 11 of which results in cancer deaths per 100,000 women. This is small compared to the estimated mortality benefit of earlier detection of breast cancer due to screening (112). For a European screening regimen with biennial screening the number of cancers induced and corresponding deaths will be lower.

A limitation of mammography is pain experienced by the woman during the required breast compression. The compression is necessary to distribute the breast tissue across the detector to minimize the overlapping breast tissue, and to reduce the breast thickness to reduce radiation dose. The pain experienced by the woman could deter her from returning to subsequent screening (113).

2.5.2 Breast density and risk.

Since Wolfe showed association between parenchymal patterns and breast cancer risk (59), breast density and breast cancer risk has been studied extensively. In the literature, breast cancer risk is often reported comparing almost entirely fatty- to extremely dense breasts. This yields a relative risk of about 3 – 6 (63,114–119). This description of risk, where the extremes of breast density is compared, is not meaningful to most of the women, which fall into the intermediate breast density categories (120–122). Comparing women with heterogeneously dense breasts (BI-RADS density C) to the average of women, have 1.2 times relative risk, and women with extremely dense breast (BI-RADS density D) have about 2 times relative risk (120,122–126). Recently, volumetric breast density assessment has been validated as an alternative for risk assessment for breast cancer due to breast density (127,128). Automatically calculated breast density can also be followed longitudinally, and high fluctuations in a woman's breast density has been shown to be associated with interval cancers (129).

2.5.3 Breast density and masking.

Since the masking hypothesis was generated (1) and confirmed (2), several studies have shown that dense breast tissue can mask breast cancer during screening mammography (32,130–134). Masking occurs when fibroglandular tissue exists in the vicinity of the lesion, making the lesion indistinguishable from the background. In other cases, lesions can only be partly obscured, masking certain features of the lesion, such as spiculations, leading the radiologist to erroneously mistake the dense patch as fibroglandular tissue.

Studies on mammography screening using SFM has shown program sensitivity of 80 – 88 % for non-dense breasts, 58 – 69 % for heterogeneously dense breasts, and 29 – 62 % for extremely dense breasts (32,132,133). DM perform better in dense breasts reducing the effect of masking compared to SFM (30,34,36,135–137), due to a wider dynamic range (35,138). Estimates of program sensitivities range from 75 – 100 % for women with non-dense breasts, 69 – 82 % for women with heterogeneously dense breasts, and 47 – 84 % for women with extremely dense breasts (3,137,139,140). A similar trend has also been shown using Volpara™ for density stratification (141).

Some cancers might be masked at the time of screening without becoming interval cancers. Consequently, program sensitivity might underestimate the effect of masking. Several studies have investigated the use of MRI as an adjunctive modality in high risk cohorts. As MRI is a more sensitive technique than DM, these studies can be used to investigate the extent of masking when using DM. Aggregating the results from five studies the mean sensitivity for SFM was 40 % compared to the sensitivity of MRI was 81 % (142–147). Research combining screening ultrasound with mammography found that 78 % of tumors found using ultrasound only was obscured by overlapping breast tissue and 19 % were due interpretive errors in mammography (4).

With the recent introduction of DBT which produces pseudo 3D images in which "out of slice" objects are blurred, there is a potential for reducing masking, if DBT is implemented in mammography screening.

2.5.4 Density and false positives

Dense breast tissue can be superimposed such that it mimics the appearance of a breast cancer. For instance, superimposed fibrous strands of normal tissue might seem to radiate from a common origin, which could be interpreted by the radiologist as spiculations without a central mass (an architectural distortion). Or if in addition, a patch of fibroglandular tissue was superimposed the origin, the radiologist might interpret the image as a spiculated mass. Super position of glandular tissue can also generate the appearance of an asymmetric density. Such false positive result would lead to a recall of the woman for additional imaging and/or needle biopsy.

Women with dense breasts have increased false-positive rate compared to women of lower density. In SFM the specificity is reduced from 91.2 – 96.5 % in non-dense breasts to 90.8 – 89.6 % in dense breasts in a U.S. screening program (32,148). Similarly, in DM it was

shown that false positive rate was higher for women with dense breasts compared to women with non-dense breasts (34,36,140,149). With specificity of 94.7, 91.2, 87.3 and 88.7 % for BI-RADS density I – IV, respectively (137). A similar reduction in specificity has been seen in Europe with specificities of 98.9, 98.5, 98.2 and 97.6 % for Volpara density grade 1 – 4, respectively (141).

The lower specificity in dense breasts leads to higher recall rates and more recommendations for biopsies for these women (149). A reduction in false positive rates for women with dense breasts after implementation of DBT, could have a major effect on the cost-effectiveness of a screening program (150),

2.5.5 Breast density legislation

Due to the focus on breast density, breast density legislation was implemented at a federal level in the U.S. 28 March 2019. The mammography quality standards act was updates so mammography facilities must include breast density information to the patient and healthcare provider. Since 2008, many individual states already introduced similar laws. This breast density information must include whether the woman has dense or non-dense breasts, breast density score on a four point scale (similar to BI-RADS density) and a summary of the significance of the density written in lay man's terms (5). The goal is that the woman can discuss, with her healthcare provider, whether she should opt for supplemental examinations such as ultrasound and MRI. A few states also reimburse the additional examinations if the woman has dense breasts. Many researchers are cautious as dense breasts are common, and the evidence surrounding supplemental screening is lacking (126).

2.5.6 Personalized screening

In population-based mammography screening, women are offered annual or biennial digital examinations with a frequency dependent on programmes offered in their country. Except for women of exceptionally high risk of breast cancer, mammography is implemented as "one size fit all". This approach has been challenged by interest groups (151,152) and researchers (118,153) promoting personalized screening, based on breast density and breast cancer risk. In this type of personalized screening the screened cohort is divided into subgroups, e.g. women with dense and non-dense breasts, where certain groups are screened using adjunctive modalities.

3 Aims

The aim of this thesis was to compare the density classification using BI-RADS density classification to automatic breast density assessment software. Then to compare digital mammography and digital breast tomosynthesis with respect to diagnostic accuracy and radiation dose for women of different breast density categories.

3.1 Specific aims

3.1.1 Paper I

The main scope of paper I was to find the volumetric density threshold that best classifies breasts as fatty and dense compared to an average BI-RADS density classification by multiple radiologists. The secondary aim was to analyze the interobserver variability in density assessment.

3.1.2 Paper II

The main scope of paper II was to evaluate whether areometric or volumetric breast density best matched the BI-RADS density categorization by radiologists. The secondary aim was to generate a set of areometric and volumetric threshold values to allow estimation of BI-RADS classification.

3.1.3 Paper III

The main purpose of paper III was to compare the average glandular dose in paired digital mammography and digital breast tomosynthesis acquisitions in screening stratified by automatically calculated density categories. The secondary purpose was to analyze the effects of incorporating breast density assessment and measurements of radiation dose and beam quality on the estimates of average glandular dose.

3.1.4 Paper IV

The aim of paper IV was to compare the true-positive and false-positive interpretations in digital mammography vs. digital breast tomosynthesis in subgroups of volumetric breast density, age and mammographic findings.

4 Methodological considerations

4.1 Oslo Tomosynthesis Screening Trial

The Oslo Tomosynthesis Screening Trial (OTST) was performed, as evidence on performance of DBT versus DM was needed. When OTST was designed, it was unclear whether DM would be performed with computer aided detection (CAD) and if DBT would be acquired along with a DM- or a synthetic DM image. Consequently, OTST was designed as a four armed trial with two 2D (2-Dimensional) - (conventional DM and DM with CAD) and two 2D plus DBT (DBT plus DM and DBT plus synthetic DM) arms. OTST was a prospective trial where 24,301 women age 50 to 70 were recruited from the population-based screening program BreastScreen Norway at the breast center in Oslo University Hospital. All women included were imaged using both DM and DBT under the same compression (combo mode). A thorough description of the study design and image interpretation has been reported elsewhere (41,108,154,155). All women with at least one positive interpretation were discussed at a consensus meeting, where the radiologists evaluated breast density using BI-RADS density 4th edition in consensus. 4th edition was used as this was the gold standard when the trial started. Radiologists had access to both DM and DBT images with associated C-view images at the consensus meeting. Breast density was assessed using Quantra^{TM} version 2.0, which is based on the raw DM projections. Quantra^{TM} results were not shown to the radiologists. Quantra^{TM} calculated volumetric- and area-based breast density and provides a BI-RADS density like score based on volumetric breast density.

The paired design of OTST was well suited to compare diagnostic performance and radiation dose between DM and DBT as all women were imaged prospectively using both modalities. Thus, this study design had greater statistical power to show a potential difference between modalities, compared to an unpaired randomized controlled trial (RCT). Still, the paired design had a significant drawback compared to RCT, as we could not know if missed cancers in one arm would have developed into interval cancers.

An important part of the study design in OTST was the common consensus meeting, where the decision to recall women was taken based on access to all images. This design was necessary due to practical reasons, as the workload of the OTST trial would have exceeded the capabilities of the breast center by having separate consensus meetings for DM and DM plus DBT arms. This is a weakness of the study design and must be kept in mind when interpreting the results, especially with respect to recall rate.

4.2 Mammography equipment

The mammography equipment used in OTST were three Hologic Selenia Dimensions units. These DM systems are capable of performing both DM and DBT acquisitions. In this study, DM and DBT acquisitions were performed during the same compression (combo mode). The unit was operated with the AEC setting 'auto filter' when used for screening, as recommended by the manufacturer. In DM mode, 50 µm Rh filter was used for breasts of thickness < 70 mm and a 50 µm Silver (Ag) filter for breasts ≥ 70 mm. The unit is equipped with a Tungsten (W) target. When performing a DBT acquisition, the tube moves in a ±7.5^{o} arc recording 15 images. In DBT mode, a 0.7 mm Al filter is used. The anti-scatter grid is retracted, and the detector elements are binned 2x2. The DBT acquisition is reconstructed using a filtered backprojection reconstruction algorithm.

4.3 Automatic breast density assessment

Breast density assessment using Quantra^{TM} version 2.0 was performed for all women in the trial (the algorithm is described in chapter 2.4.3). To avoid reader bias, the result of this assessment was stored on file, not shown to the readers in the study.

The reader study (paper I and II) revealed three (of 540 women) Quantra^{TM} density assessment outliers. These images were unanimously classified as almost entirely fatty breasts by the radiologists, and dense (category 3 or 4) by Quantra^{TM}. These large fatty breasts had an image artifact manifesting as a dark rim along the breast edge, in which Quantra^{TM} likely selected the fat reference pixel. Therefore, the inner part of the breast was erroneously estimated to contain large amounts of fibroglandular tissue (156). These cases were excluded from the laboratory study (paper I and II).

There were three breasts which were classified as dense breasts in agreement with the radiologists. However, they were reported to have 100 % volumetric breast density, which is unrealistic. All these were small dense breasts imaged using a small paddle. This paddle is visible on the image. The Quantra^{TM} team at Hologic reported this likely was a failure in the image segmentation algorithm (Hologic, Personal communication).

Thus about 1 % (6 of 540 women) of the Quantra^{TM} results contain errors, with half of these result in obvious misclassification of breast density compared to radiologist's assessment.

4.4 Reader study (Paper I and II)

Rationale for performing the reader study

The goal of paper I and II was to compare the radiologist- to automatic breast density assessment performed using Quantra™ version 2.0 and to assess the inter-observer variability of radiologist's assessment. In order to minimize potential bias compared to density assessment in population-based screening, such comparison should have been performed within a population-based study, such as OTST. Unfortunately, this was not possible in OTST due to some elements of the study design, such as:

- The BI-RADS score was set in consensus producing only one score, making evaluation of inter-observer variability impossible. Single reader assessment would likely produce a higher number of outliers than a consensus assessment.
- The BI-RADS density assessment was only performed on women with positive interpretations. Therefore, they will represent women with cancers, benign findings or women that are difficult to assess. This group has a higher breast density on average than all women included in OTST (155).
- At the consensus meeting the readers had access to DBT and C-view as well as DM, which might influence the scores. As Quantra™ version 2.0 evaluates DM images we chose to limit the comparison to DM only.

Reader study cohort

We chose to compare Quantra™ and BI-RADS density assessment in a reader study with 537 women randomly selected from the OTST trial. Comparing the Quantized density distribution in the reader study and for all women included in OTST using Chi squared statistics (χ^2) ($p = 0.93$) indicate successful randomization and consequently minimal selection bias. Thus, this cohort is therefore representative of population-based mammography screening.

Using BI-RADS density 4th edition

Although BI-RADS 5th edition was released at the time of the reader study we chose to use 4th edition as Quantra version 2.0 uses mapping based on BI-RADS density 4th edition.

Experience and training of radiologists

Five radiologists from Oslo University Hospital participated in the trial. This allowed us to compare the assessment from several radiologists for the same women. However, as the

radiologists came from the same center, their inter-observer variability could potentially be lower, compared to a study performed in a multicenter setting, since radiologists in one center potentially learn from each other and develop a local breast density assessment "culture". As all radiologists were recruited from a large university hospital there is a possibility that the reader study has some reader selection bias (157). The radiologists had a wide span of experience in screening mammography, 1, 3, 11, 24 and 34 years. The reading order was randomized for each radiologist, to avoid reading order bias (157).

Prior to density assessment the radiologists participated in training to familiarize with the BI-RADS density 4th ed. scale (156). This was done to ensure radiologists were recently trained on density assessment on women of all densities. The radiologists were blinded to the Quantized density score when they were interpreting the 537 cases to avoid review bias (157). The fact that the radiologists trained with knowledge on Quantized density scores prior to the study might have introduced some minor review bias. Another factor potentially influencing breast density assessment is that this is a reader study setting, rather than a screening setting. When the radiologist evaluated the images in this study, breast density was the most important characteristic of the image, not whether cancer was present in the breasts. Additionally, when interpreting images in a reader study there is no consequence to the patient, which might lead radiologists to change their reporting patterns (158).

Comparison of density assessment in the reader study and OTST

In this section we will attempt to assess potential differences between BI-RADS density assessment in the reader study and OTST. Table 7 shows the distributions of Quantized- and BI-RADS density for the two study settings. First, the Quantized density in the reader study was comparable to the corresponding density distribution for all women included in OTST. Comparing Quantized density distributions for all women and women evaluated at consensus, the average breast density is higher for women evaluated at consensus. However, the proportion of women classified with dense breasts using BI-RADS density are similar (45.9 % in OTST at consensus and 45.4 % in the reader study) and more women were considered to have extremely dense breasts in the reader study. This indicates that the radiologists were more likely to classify breasts as dense and extremely dense in the reader study than in a screening setting.

Table 7: Density distributions in OTST and the reader study.

Cohort	Quantized density				BI-RADS density (%)			
	1	2	3	4	I	II	III	IV
OTST all women	11.8 %	53.3 %	27.4 %	7.5 %				
OTST consensus	7.8 %	49.4 %	33.3 %	9.6 %	8.2 %	45.9 %	38.9 %	7.0 %
Reader study	12.1 %	54.2 %	26.8 %	6.9 %	13.6 %	41.0 %	35.0 %	10.4 %

The correlation and inter-observer agreement between Quantized density and the radiologist's assessment was also slightly higher in the reader study. Spearman's correlation was 0.73 [95 % Confidence Interval (CI): 0.69 – 0.77] in the reader study versus 0.70 [95 % CI: 0.68 – 0.72] in OTST at consensus. Cohen's kappa using four categories with quadratic weights was 0.72 [95 % CI: 0.68 – 0.76] in the reader study versus 0.69 [95 % CI: 0.67 – 0.71] in OTST at consensus (155). As the confidence intervals overlap, the difference was not significant.

4.4.1 Paper I

In paper I the inter-observer variability in radiologist's breast density assessment was assessed. This can be performed in several ways: For instance, comparing the results of each radiologist or comparing each radiologist to a reference score. As a previous study used comparison to majority score, we opted to use a similar method to facilitate comparison of results (43). As we used five readers, we opted to use the median score, rather than majority as two scores potentially could tie for the majority score.

Although inter-observer variability described using kappa is a well-established method, its consequence can be hard to interpret. Therefore, we employed a novel method of describing the inter-observer variability in density assessment. We found the volumetric threshold for each radiologist which separates dense and non-dense breasts with the highest accuracy. Comparing these thresholds between radiologists and the density distribution of the cohort, facilitates a comparison of radiologists which are easier to interpret. A clinically relevant metric describing the inter-observer variability is the number of women classified as dense and non-dense by individual radiologist. As this would have a direct consequence in

personalized mammography screening. However, this proportion is greatly influenced by the number of readers, as outlier interpretations add to this metric.

In paper I Table 2, the inter observer agreement between Quantized density and the median score of the radiologists was not reported. These values are reported here: Four categories with quadratic weights: 0.719 (95 % Confidence interval (CI): 0.678 – 0.760), Four categories unweighted: 0.453 (95 % CI: 0.391 – 0.515) and dichotomized 0.635 (95 % CI: 0.570 – 0.700).

4.4.2 Paper II

The main scope of paper II was to evaluate whether volumetric or areometric breast density assessment best matched the radiologists' assessment, and to provide threshold values for generating BI-RADS density distributions from automatically performed breast density assessment. We also wanted to compare the threshold values found with the default classification in Quantra™; Quantized density.

To evaluate the assessment method that best matched the radiologist assessment, we used Receiver Operating Characteristics (ROC) methodology and the metric; area under the curve (AUC). This metric incorporates all possible thresholds for density classification and is therefore not dependent on the choice of threshold. Interpretation of the results must be performed with care, especially if the ROC curves for the two methods cross at some important location in the ROC curve. In paper II, none of the curves cross near a clinically relevant location, strengthening the conclusions derived from the ROC analysis.

In order to extract a threshold value, a single ROC point had to be selected. For this, we used the Youden's index, which is the point farthest from the chance diagonal. These thresholds resulted in the best sensitivity and specificity to classify dense breasts compared to BI-RADS density assessment. Still, it resulted in density distributions which were very different from traditional BI-RADS density assessment (67). We also selected density thresholds to enforce the distribution generated by the median radiologist's score. This method changed the thresholds for the extreme densities the most, resulting in less agreement. Using the default categorization in Quantra™; Quantized density generated even lower agreement than the two previous methods. Especially, with respect to the threshold between women with dense and non-dense breasts. However, this is not surprising as the former thresholds were generated by optimizing agreement with radiologists.

An important question is how the density categorization would change if Quantized density was implemented. Table 8 shows the correlation between Quantized density and BI-RADS density using the latter as reference. Almost all non-dense breasts categorized using BI-RADS density is correspondingly categorized using Quantized density. Almost a third of the dense breasts are downgraded to non-dense using Quantized density. In the four-category classification, much overlap between the extreme density categories and neighboring categories can be seen.

Table 8: Correlation of Quantized density with BI-RADS density in the reader study, where BI-RADS density is the reference.

Reader study		Quantized density					Quantized density	
		1	2	3	4		Non-dense	dense
BI-RADS density (reference)	I	46.6 %	53.4 %	0 %	0 %	Non-dense	94.5 %	5.5 %
	II	14.1 %	78.6 %	7.3 %	0 %			
	III	0 %	41.1%	55.9 %	3.2 %	Dense	32.4 %	67.6 %
	IV	0 %	3.6 %	41.1 %	55.4 %			

4.4.3 Validation of Quantra™

An important aspect of choosing volumetric or areometric assessment for density assessment was the validation of the two methods. Paper II addresses this by investigating the correlation between the two. The measures were well correlated ($r^2 = 0.76$), indicating they provide similar measurements. As the left and right breast of a woman tend to be similar, a validation of the density assessment is the correlation between the density of the left and right breast. Both areometric and volumetric assessment have similar correlation between breast density of contralateral breasts.

4.5 Dosimetry study (paper III)

In mammography screening trials average glandular dose (AGD) is frequently reported as the AGD estimated by the modality which can be obtained from DICOM metadata. We suspected the AGD to have considerable inaccuracies, particularly since it does not account for the breast density. Therefore, we wanted to perform an AGD comparison of DM and DBT in OTST incorporating breast density and on site measurements of the radiation output.

The "Dance model"

The dosimetry for paper III was performed using the European protocol for breast dosimetry (159,160), which uses a model published by Dance et al. (161–164).

In the "Dance model" the breast is modelled as a semi-cylinder with 16 cm cross section. The central region of the breast contains a homogeneous mixture of fat and fibroglandular tissue with a 5 mm outer rim of adipose tissue, except on the chest wall side. The ratio of fibroglandular tissue by weight to the total weight of the central region is referred to as the glandularity. AGD to this breast model has been estimated using Monte Carlo simulations (161–164). Identical exposures are modelled with an ionization chamber (without a breast) positioned at the entrance point of the breast, to simulate a measurement of entrance Air KERMA. The ratio of the simulated AGD and the entrance air KERMA is the g-factor in the "Dance model", can be used to convert entrance air KERMA to AGD. Since 1990 the "Dance model" has been refined by adding various correction factors to account for anode/filter combinations other than Mo/Mo (eq. 11); the s-factor (162,163). The c-factor was also added to account for glandularities other than 50 % (162). And a T-factor has been added to allow for AGD estimates in DBT (164).

$$D = K \cdot g \cdot c \cdot s \cdot T \tag{11}$$

Why measure half value layer

The g-factor is tabulated by breast thickness and the HVL of the radiation beam (161). The breast thickness is obtained with sufficient accuracy from the DICOM metadata of each image. The HVL can also be obtained from the DICOM metadata. But the value reported in the DICOM header showed to have limited accuracy, resulting in an error between 2 and 12 % in the final estimate of AGD compared to measuring HVL (165). We measured HVL for all clinically used beam qualities on all three mammography units used. The measurement of HVL was performed according to the International Atomic Energy Agency (IAEA) protocol (160). A 10X6-6M ionization chamber (RadCal Corporation, Moravia, CA, U.S.) was used with an AccuPro digitizer (RadCal Corporation) which automatically corrects for ambient temperature and pressure. The paddle was elevated to the maximum height, and the beam well collimated with a lead diaphragm, to minimize the influence of scattered radiation. The ionization chamber was positioned 4.5 cm above the breast support, which was covered by a thick steel plate to minimize backscatter and protect the detector. Sheets of high purity

aluminum (RadCal Corporation) was inserted between the lead diaphragm and the ionization chamber. And the HVL was calculated according to the standard formula (160).

Why measure air KERMA

The entrance Air KERMA of a mammogram can also be found in the DICOM metadata. However, paper III showed that the errors using the DICOM header entrance Air KERMA was up to 7.5 %, for the unit with the largest deviations compared to measuring entrance air KERMA. Therefore, Air KERMA was measured in a similar setup as for HVL, but without the lead diaphragm and aluminum sheets, and with the paddle in contact with the ionization chamber. Air KERMA for a 100 mAs exposure was measured for all clinically used beam qualities on all mammography units. As it is infeasible to perform this measurement for all clinically used mAs values, the air KERMA value was scaled according to the clinically used mAs value for the respective mammographic view. This approach assumes linearity between mAs and air KERMA, which was verified in the mAs range 40 – 400 with linear regression producing an r^2 value of 0.9999. The reproducibility of the ion chamber measurements was also tested for 10 identical exposures, resulting in a relative standard deviation of 0.25 %.

Measurements on DBT spectra

Dosimetry in DBT acquisitions are complicated by tube movement and the angular response by the dosimeter, which can lead to underestimation of the radiation dose. Hologic has implemented a specific protocol used to perform dose measurements called “zero-degree tomo”. In this mode, the tube is stationary perpendicular to the detector but otherwise performs acquisitions identical to tomosynthesis projections. In the “Dance model” of the breast at steeper DBT angles the x-ray beam must traverse more breast tissue compared to the perpendicular position of the tube. This reduces radiation dose. The T-factor in the Dance model is dependent on the DBT angle and accounts for this dose difference.

The choice of using an ionization chamber

We chose to perform the measurements using an ionization chamber (10X6-6M, RadCal Corporation), as the effective atomic number of air is similar to tissue. This eliminates the need for correction factors accounting for the difference in dose deposition in solid state dosimeters and soft tissue. Similar measurements could be performed, measuring HVL and air KERMA simultaneously (without needing aluminum), using solid state dosimeters with built in corrections for mammographic beam energies. However, preliminary tests revealed

deviations in the order of 8 % when measuring air KERMA and 2 – 8 % when measuring HVL.

Incorporating glandularity into dose estimates

There is a subtle, but important difference between the glandularity in the "Dance model" and the volumetric breast density measured by Quantra^{TM}. Glandularity in the "Dance model" is measured by weight rather than volume. In addition, glandularity refers to the faction of fibroglandular tissue only in the central core of the breast (not the 5 mm thick adipose tissue modelling subcutaneous fat). As the DICOM metadata contain the breast thickness and Quantra^{TM} report the fibroglandular and total volume of the breast, it is possible to convert volumetric density to glandularity using the method described in paper III (165). We chose to include the distal 5 mm semi-circle shell into the core volume, even though this was not included in the original "Dance model" (161). We did this as the glandularity versus age model in Dance's paper introducing the c-factor was based on estimates of glandularity based on clinical technique factors on phantoms with tissue known equivalent composition without the distal adipose tissue (166). Therefore, including this region would bias our glandularity estimates compared to the original estimations by Young which was applied by Dance (162).

Obtaining the DICOM metadata

As paper III involves dose calculations for numerous women, automatic collection of DICOM metadata was necessary. A script obtained from Hologic was used to extract relevant DICOM metadata from the examinations at a workstation.

Limiting the cohort for dose estimation

We performed dosimetry measurements on the three mammography units in November 2012. As beam energy and radiation output potentially can be adjusted at periodic service, we limited our cohort to within the service interval in which we had measurements. Additionally, a few months into the OTST trial, the manufacturer adjusted the AEC response for all similar systems worldwide. We judged it important that our analysis reflected the most current system design, which also was a reason for limiting the cohort for the OTST dosimetry. Thus, by excluding women images in other service intervals we chose quality over quantity of data, as 3,819 women with 15,276 paired views were considered sufficient.

Calculation of AGD

The DICOM metadata, Quantra™ data and the measurement data were combined on a per view basis. Then a Matlab (Mathworks, Natick, MA, U.S.) script calculated the AGD using the "Dance model".

4.6 Diagnostic accuracy stratified by breast density (paper IV)

Choice of outcome variables

In paper IV, the aim was to evaluate the diagnostic accuracy stratified by breast density and age. Usually evaluation of mammographic screening with adequate follow-up report sensitivity, specificity, cancer detection rate or recall rate. Screening detected cancers count as true positives and the interval cancers as false negatives. Our focus was to compare DM and DBT, not an evaluation of the screening program. Therefore, we opted to use outcome measures which were decided prior to the consensus meeting: positive and negative interpretations. This was done to minimize potential bias due to the common consensus meeting and diagnostic workup, and to ease the interpretation of the results of the analysis. In our analysis of 48 451 breasts there were 234 breasts with screen detected cancers and 52 breasts with interval cancers. Interval cancers were diagnosed in eight breasts (8 in the DM arms and 7 in the DBT arm) which had positive scores. As the scores in OTST were given breast based and not lesion based, it was not possible to determine whether the positive score represented detection of the interval cancer. In our analysis, these cases were therefore considered as false positives.

Many population-based studies reports recall rate. As the decision to recall the woman was performed at the consensus meeting with access to both DM and DBT images, DBT likely had influence on the decision to recall women. This can be seen in an analysis present by Skaane et al. (167), where the number of false positives using DM was higher, but the recall rate was lower compared to DM plus DBT. The reason was that many suspicious cases in DM were dismissed at consensus, when the readers had access to DBT. Therefore, a comparison of recall rate between DM and DBT would be too biased to be meaningful.

Classification of findings

When giving a positive score at interpretation, the radiologists had to classify the score as a circumscribed mass, spiculated mass, architectural distortion, asymmetric density, calcification or calcification plus density. The goal of this sub-analysis was to evaluate

difference in findings between DM and DBT. This classification has considerable inter-observer variation as the interpretation between categories is subjective. For instance, a spiculated mass is a mass with radial spiculations, while a distortion is spiculations without a central mass. Thus, many cases could be classified as either. Therefore, we chose to group these categories in our analysis. Calcification and calcification plus mass was grouped in a similar manner. After the OTST was finished, the screen detected cancers were reanalyzed and classified into a finding category by a consensus of radiologists to ensure thorough classification of the true positive findings.

Double reading

At the start of OTST several possibilities of future screening modalities were considered. The performance of computer aided detection (CAD) in DM and performance of synthetic DM images were not clear. Therefore, OTST was designed with two DM arms, one with CAD and two DBT plus DM arms with one where synthetic DM replaced conventional DM. As mammography screening is performed using double reading in Norway, we chose to perform our analysis double reading. Where 2D double reading consists of Arm A (DM) and Arm B (DM plus CAD) and 2D plus 3D double reading consists of Arm C (DM plus DBT) and Arm D (synthetic DM plus DBT). If one of the arms had a positive score, the double reading was considered positive. This approach is feasible only if the two arms making up the double reading has similar diagnostic accuracy, which was reported by Skaane et al. (41).

Choice of density assessment method

We stratified our analysis primarily on Quantra™. We only had BI-RADS density scores for the women evaluated at consensus, limiting our ability to stratify the analysis of the negative cases. The evaluation was done in consensus, which potentially could yield a more stable breast density evaluation compared to a single reader evaluation which is most common in the U.S. However, we reported the BI-RADS density stratified results in the appendix to allow for comparison of the two methods of assessment.

Transition to Quantra™ in population-based screening

The correlation between Quantized- and BI-RADS density for the reader study is shown in Table 8. A similar table for women evaluated at the consensus meeting is shown in Table 9.

Table 9: Correlation of Quantized density with BI-RADS density for women evaluated at the consensus meeting in the OTST, where BI-RADS density is the reference.

OTST trial (women at consensus)		Quantized density					Quantized density	
		1	2	3	4		Non-dense	Dense
BI-RADS density (Reference)	I	36.9 %	60.6 %	1.9 %	0.6 %	Non-dense	87.1 %	12.9 %
	II	10.2 %	75.0 %	14.1 %	0.6 %			
	III	0.2 %	25.6 %	62.3 %	11.9 %	Dense	21.9 %	78.1 %
	IV	0 %	0 %	34.6 %	65.4 %			

Comparing Table 8 and Table 9 the number of women downgraded to non-dense from dense is reduced. Furthermore, the agreement for women with extremely dense breasts are higher in OTST compared to the reader study. This might be a consequence of the thresholds in volumetric density for determining Quantized density are based on a density assessment in a population-based screening trial, the DMIST trial (30), rather than reader studies. Nevertheless, a transition to using Quantra™ would decrease the number of women classified as having dense and extremely dense breasts compared to BI-RADS density assessment.

4.7 Statistical considerations

4.7.1 Inter observer variability

There are numerous statistical methods that can be used to assess inter observer variability. The method used in paper I is Cohen's kappa (κ), which measures the agreement between two observers corrected for the expected agreement by chance. We chose to this method as it better facilitates comparison with previous studies. The agreement is often classified using an arbitrary scale published by Landis and Koch (168), where: $\kappa < 0$ is poor, 0 – 0.2 is slight, 0.2 – 0.4 is fair, 0.4 – 0.6 is moderate, 0.6 – 0.8 is substantial and 0.8 – 1 is excellent. As breast density is scored on an ordinal scale, we chose to use quadratically weighted kappa. This method penalizes disagreement between observers more if the discrepancy is larger and generally yields higher kappa values than unweighted and linearly weighted alternatives.

4.7.2 Accuracy

In paper I, Fig. 2, the performance of each radiologist's threshold is measured using accuracy. Accuracy is generally a poor measure for diagnostic accuracy, since for a common condition, good accuracy can be obtained using a test with high sensitivity but very low specificity (or

vica versa with a rare disease and specificity). Therefore, ideally diagnostic performance should be quoted using both sensitivity and specificity. We opted to use diagnostic accuracy as the proportion of dense and non-dense breasts are about 50 % each. Additionally, quoting a single number makes the figure easier to interpret.

4.7.3 p-values

A statistical comparison of two measures is usually accompanied by a p-value. If the null hypothesis is that the modalities have equal performance, the p-value represents the probability of the two measures being equal given the observed difference in measures. The most common threshold for rejecting the null hypothesis is a p-value of 0.05 and was employed throughout the papers in this thesis.

4.7.4 Confidence intervals

Often confidence intervals are calculated with assumptions of normal distributions. In mammography true- and false positive rate are often close to one or zero, where this assumption fails. Instead, one has to use a method based on binomial distributions, such as the Wilson method (169,170). The CI for the difference between two measures can be estimated using Newcombe's paired or unpaired method (170–172). Bootstrapping was also used to estimate CI's (170).

4.7.5 Paired statistics

OTST had a paired design. Therefore, measures comparing DM and DBT can often be performed using paired statistics. Similarly, density assessment was performed using BI-RADS density and Quantra™ on the same women, which also allows for paired statistics.

For this thesis we have used McNemar's test instead of Chi squared (χ^2) whenever possible. For confidence intervals evaluations are performed using the difference between performance measures, rather than their value if possible. This approach requires less study participants to achieve high statistical power compared to an unpaired approach.

4.7.6 Age and density adjustment

In the supplemental material of paper IV age and density adjusted true- and false positive rates are reported. As breast density decreases with age, one must adjust for one confounding parameter in order to estimate the isolated effect of the other. To perform this adjustment and calculate corresponding 95 % CI's a cluster bootstrap approach was used (173,174). For age adjustment we divided the cohort into four age groups: 50 – 54, 55 – 59, 60 – 64 and 65 - 69. Then performance measures were calculated for each density group within each age group

using bootstrapping with women as resampling unit. Then, performance measures were averaged for all age groups, which was the age adjusted estimate. This approach mitigates the effect of the different age distribution among each density as each age group is weighted equally, regardless of group size. The density adjusted estimates was performed similarly by grouping the women by volumetric density.

4.8 Limitations and challenges

Due to practical considerations and choices made during planning, study design and analysis the papers and thesis has some limitations.

As previously discussed, in the OTST there was a common consensus meeting where radiologists had access to all images when evaluating BI-RADS density. This evaluation was performed in consensus producing only one score. Our comparison of Quantra™ and BI-RADS density therefore was conducted in a reader study.

OTST was a single center study performed at a large breast center within a university hospital. Therefore, the interpretation of images represents those of a large university hospital enrolled in BreastScreen Norway. There might be differences in smaller rural centers. Some findings might have limited generalizability to other screening programs. For instance, in the U.S. single reading is used, and the recall rates are substantially higher.

We had access to Quantra™ version 2.0 which performs density classification based on BI-RADS density 4th edition, as this was the standard when OTST was initiated. Presently BI-RADS density 5th edition is the standard. This edition has less focus on the volumetric breast density and more focus on whether breast cancers can be masked by fibroglandular tissue. Even thought DBT was used in OTST, Quantra™ calculated breast density based on DM projections. Newer versions of Quantra™, can calculate density in DBT acquisitions. In DBT, Quantra™ only uses the central DBT projection for calculation. Therefore, the results should be very similar to calculations based on DM images. As the main difference is a noisier input image, the DM version should provide better estimates if there was a difference between the two.

Prior to OTST, BI-RADS density classification was not used routinely by the radiologists. But all radiologists had experience using BI-RADS density in OTST, and they were specifically trained prior to the reader study.

The study population was limited to women between 50 – 70 years as the study was a part of BreastScreen Norway. Therefore, we cannot generalize our conclusions to older or younger women. The generalizability is especially low towards younger women as these are more likely to be pre-menopausal, which affects breast density. As OTST was performed in BreastScreen Norway it is representative of the Norwegian screening population. Other locations might have different characteristics, such as breast density. An example is Italian women having less proportion of dense breasts compared to U.S. due to the Mediterranean diet and lifestyle (175).

We calculated optimal thresholds in percent density to estimate BI-RADS density distributions. A limitation for this analysis is that the same data was used to optimize the thresholds and to generate the synthetic distributions. This approach might lead to overfitting the data. Ideally, the dataset used for generating the thresholds should be separate from the dataset used to generate the synthetic distribution.

OTST used Hologic DBT equipment. Thus, there might limit the generalizability of our results compared to a study using different equipment. As radiation dose is a consequence of the vendor's choices of radiation spectrum and AEC operation, these results are vendor specific. Other aspects of the equipment, such as tomo angle and reconstruction algorithm might greatly influence the image quality which potentially could affect diagnostic accuracy of the examination.

The estimates of radiation dose are based on the "Dance model", which assumes an even distribution of fibroglandular tissue. In reality, the fibroglandular tissue can be located only in certain quadrants of the breast. If this quadrant is close to the detector, the beam will be highly attenuated compared to the "Dance model", thus we would overestimate the dose. Similarly, we could underestimate the dose if the fibroglandular tissue was located towards the x-ray tube. This uncertainty is most relevant when considering dose estimated for individual women. For paper III the estimates are reported for different strata of the cohort, in which these differences would be considerably lower. Individualized dose estimates based on DBT acquisitions and Monte Carlo simulations were out of the scope of the dose estimations for paper III.

Many of the papers report numerous p-values. As the number of p-values reported increases, the probability that some of them are a type I error (false positive hypothesis test). One can address this using a Bonferroni correction, where the significance criterion for p-

values of confidence intervals are reduced according to the number of statistical tests performed. In the OTST trial this was performed with respect to the main outcome of the trial, where the significance threshold was set to $p > 0.0264$ for the primary comparison (due to an interim analysis) (41). However, for secondary assessments the $p > 0.05$ was maintained.

In paper IV we found a significant correlation between age and true positive rate. This specific statistical comparison was not preplanned and should therefore be treated with caution. Significance in such ad-hoc analysis is often the result of random outcomes of the data. We chose to include it in the paper to make other researchers aware of a potential effect and use this result to generate an a priori hypothesis.

5 Summary of papers

5.1 Paper I: Classification of fatty and dense breast parenchyma: comparison of automatic volumetric density measurement and radiologists' classification and their inter-observer variation

Before a new method of breast density classification is implemented in mammography screening, it is important to evaluate the performance of the new method compared to the gold standard. Currently radiologists use the BI-RADS density scale, which is a subjective evaluation based on semi-quantitative criteria. In this study we found the volumetric density cutoff value that results in highest accuracy of radiologists' classification of fatty and dense breasts. Inter-observer variability is a major issue when radiologists use BI-RADS density. By calculating and comparing the best fit volumetric threshold for each individual radiologist, we can investigate inter-observer variability.

In this study 537 women were randomly selected from the women included in the Oslo Tomosynthesis screening trial. Five radiologists individually assessed BI-RADS density for all cases. Furthermore, volumetric density was automatically calculated for all cases. For each case a median radiologist score was calculated.

The volumetric threshold that best fit the median score was 10 % (i.e. 10 % or lower is fatty and 11 % or higher is dense). Using this threshold, the classification accuracy would be 87 % compared to the median radiologist score. The default threshold in Quantra is 13 %.

Looking at individual radiologists', their thresholds varied between 8 and 15 % with comparable accuracy. This interval includes about 40 % of the women. 36 % of the women were classified as both fatty and dense by individual radiologists. Comparison of inter-observer variability (kappa) between the median score versus the individual radiologists and Quantra operating at the 10 % threshold was comparable.

This study shows that current BI-RADS density assessment has considerable inter-observer variability, and about 40 % of women are at risk of being classified as either fatty or dense depending on the radiologist. If this type of assessment is replaced by automatic objective software, breasts would be classified with high accuracy compared to BI-RADS density assessment.

5.2 Paper II: BI-RADS Density Classification From Areometric and Volumetric Automatic Breast Density Measurements

Quantra™ calculates volumetric breast density by computing the fractional amount of fatty and glandular tissue in a column above each pixel and the compression paddle. Volumetric density is computed by aggregating density across all pixels. Furthermore, Quantra™ computes area-based (areometric) density by applying a threshold in fibroglandular fraction to each pixel, classifying the pixel as fatty of dense. Areometric density is computed by calculating the fraction of dense pixels. It is still under discussion whether areometric or volumetric density classification best fit the radiologists' BI-RADS density assessment. By applying thresholds in areometric and volumetric densities one can approximate BI-RADS density distribution from either measure. The default categorization in Quantra™ is called Quantized density and is based on volumetric density. This study was based on the same reader-study data as paper I.

There was a substantial overlap in both volumetric- and areometric density for different BI-RADS density categories. Areometric- and volumetric density was compared to the mean radiologist BI-RADS density using the AUC of the ROC curve. Areometric density was significantly better for the BI-RADS density I to II threshold. Areometric and volumetric was comparable for the II and III threshold and volumetric was significantly better for the III to IV threshold. The thresholds and the corresponding density distributions are shown in Table 10.

Table 10: The density thresholds and the corresponding density distributions.

Density measure	Density thresholds			Density distribution			
	1 and 2	2 and 3	3 and 4	1	2	3	4
Areometric	6	14	30	26.6 %	22.2 %	31.7 %	19.6 %
Volumetric	7	10	16	37.4 %	21.4 %	22.7 %	18.4 %
Quantized density	5	13	26	12.1 %	54.2 %	26.8 %	6.9 %
BI-RADS density	-	-	-	13.6 %	41.0 %	35.0 %	10.4 %

The distributions show high numbers of women in the almost entirely fatty- and extremely dense breasts. The Quantized density distribution showed similar proportions in the extreme density categories, but the proportion of dense breasts is lower than the radiologist' classification by about 12 %. Another method to define thresholds is to select density values

which results in similar four-category density distribution generated by the median radiologists score.

In paper II sensitivity is defined as the proportion of women categorized above a certain BI-RADS density threshold receiving corresponding categorization by the automatic assessment. Similarly, specificity is defined as the proportion of women categorized below a certain BI-RADS density threshold receiving a corresponding categorization by the automatic assessment. Table 11 shows the proportion of women classified into the respective BI-RADS density categories also receiving corresponding classification using automatic assessment.

Table 11: The proportion of women with automatic assessment in congruence within the respective radiologists median BI-RADS density. AUC represents the thresholds found using the ROC method (Youden's index). Distribution represents the thresholds selected to most closely generate the radiologist's density distribution.

Density measure	Proportion of women in the respective BI-RADS density category receiving corresponding automatic assessment		
	BI-RADS density I	BI-RADS density III and IV	BI-RADS density IV
Areometric (AUC)	87.7 %	89.3 %	87.5 %
Areometric (Distribution)	61.6 %	84.0 %	60.7 %
Volumetric (AUC)	89.0 %	80.7 %	91.1 %
Volumetric (Distribution)	54.8 %	86.5 %	67.9 %
Quantized density	46.6 %	67.6 %	55.4 %

When selecting thresholds to match the radiologist distribution, the agreement in the extreme density categories is limited. Quantized density, which has a distribution with more non-dense breasts has lower agreement with the radiologists.

This study showed that volumetric and areometric density was equally suited to classify breasts as dense and non-dense compare to BI-RADS density. Areometric density was performing better for non-dense breasts and volumetric density was performing better for dense breasts, which is the density of most clinical relevance. This study also shows that the agreement with radiologists is limited in the extreme density categories.

5.3 Paper III: Average glandular dose in paired digital mammography and digital breast tomosynthesis acquisitions in a population based screening program: effects of measuring breast density, air kerma and beam quality

Radiation doses in mammography screening is associated with relatively low risk as the radiation dose is low and the population is older. In a screening, many asymptomatic women are exposed to radiation. Therefore, it is important to quantify the radiation dose given to these women. Breast density and measurements of radiation output and beam quality are factors that affect dose estimates. In this study we quantified the radiation dose for DM and DBT for women with dense and non-dense breasts and quantified the effect of incorporating breast density and measurements of radiation output and beam quality.

Radiation dose was calculated for 3819 women using the model published by Dance et al. (161–164). The radiation output and half value layer were measured using an ion chamber (10x6-6M, RadCal Corporation) and aluminum sheets for all beam qualities on the three units used in this study. Imaging parameters were acquired from the DICOM metadata and breast density was measured using Quantra^{TM}.

The mean AGD was 1.74 and 2.10 for DM and DBT with an average increase of 24 % in DBT. For non-dense breasts the corresponding AGD were 1.74 and 2.27 with an increase of 33 %, while for dense breasts; 1.73 and 1.79 with an increase of 8 %. This difference reflects the system design of the automatic exposure control in DM and DBT for the Hologic mammography units used in OTST. For DM the AEC is sensitive to breast density and will increase exposure in dense breasts. For DBT the AEC is mainly controlled by the breast thickness and is therefore not sensitive to breast density. Consequently, the increase in dose when using DBT is lower in dense compared to non-dense breasts.

Accounting for breast density, AGD estimates increased 16 %. Including measured radiation output and bream quality for dose estimates, resulted in a mean change in AGD estimates of 3.8 %, but for one unit 7.9 %. Accounting for all measurements shows the AGD reported in the DICOM header is underestimated by an average of about 11 %.

This study showed that DBT increased dose by about 24 % in mammography screening using Hologic equipment. This increase was 33 % for non-dense and 8 % dense breasts. When the corrected for breast density, dose estimates increased about 16 %. AGD reported by the system underestimated dose by about 11 % compared to estimates based on measurements.

5.4 Paper IV: Digital Mammography versus Breast Tomosynthesis: Impact of Breast Density on Diagnostic Performance in Population-based Screening

DBT is being used more and more in mammography screening. As DBT acquires pseudo 3-dimensional images, it has potential to resolve some of the issue as out of plane dense breast tissue will be blurred. This has the possibility to enable the radiologist to detect lesions otherwise masked by dense breast tissue. In this study we compared the true-positive and false-positive interpretations in paired DM and DBT acquisitions in the OTST cohort, stratified by automatically calculated breast density and age.

DBT improved the true-positive rate for all density groups by 12 - 24 % and in all age groups by 15 – 35 %. The improvement was of similar magnitude across all density groups and increased with age. The improvement was significant for all age groups, scattered fibroglandular- and heterogeneously dense breasts. Women with almost entirely fatty- and extremely dense breasts contained about 10 % of the women each, thus the analysis lacked statistical power. DBT improved false-positive rate significantly for all age and breast density categories, except for extremely dense breasts, where it was comparable.

The main contribution to improved true-positive rate for DBT was more findings classified as spiculated masses. As the improvement is about even across all densities, the results suggest that DBT does not resolve the issue with dense breasts by allowing the readers to “see through” dense tissue. However, our results suggest DBT blurs out of plane textures from structures like vessels and fibers, which is present in all types of breasts. This leads to improved performance in all breast densities, not just dense.

False-positive rate were reduced mostly due to less findings classified as asymmetric densities. These findings are usually the result of superposition of normal tissue structures, mimicking the appearance of a lesion on a mammogram. In DBT, the radiologists might have an improved visualization of the 3-dimensional structure of the tissue as out of plane structures are blurred. This allows the radiologists to downgrade the finding to a “pseudo lesion”. Our results indicate that for extremely dense breasts, the density might be too high for the radiologist to resolve the 3-dimensional structure of the superimposed breast tissue.

This study showed that DBT improved the diagnostic performance in all breast densities and all age groups. However, as the improvement was similar in all breast density categories the results indicated that masking in dense breasts is still present in mammography screening using DBT.

6 Ethical considerations

All data acquisition and analysis required for the papers included in this thesis is covered by the ethical approval of OTST (Regional Ethics Committee Ref: 2010/144). All women gave written consent before being enrolled in the study. OTST was registered on clinicaltrials.gov: NCT01248546.

An ethical issue with OTST is that healthy women are exposed to double the radiation dose compared to a conventional screening mammography examination. This additional radiation dose is a one-time event which occurs in women 50 – 70 years old. Therefore, the added risk due to additional radiation was considered extremely small compared to the expected mortality reduction due to DBT screening in the OTST (112,176).

7 Discussion

7.1 Context and summary of main findings

Currently, major changes are being considered with regards to personalized mammography screening based on breast density. Breast density legislation was implemented in the U.S., where women are being informed of their breast density and the associated limited sensitivity of mammography screening (5). Personalized screening is debated, and several studies have been initiated investigating the potential supplemental screening using adjunct modalities (82,177). In Austria women with dense breasts are offered supplemental ultrasound (178). Two recently introduced technologies, automatic breast density assessment and digital breast tomosynthesis, may have major impact on breast density stratified mammography screening.

The goal of this PhD thesis was twofold. First, to compare the current standard of breast density assessment, BI-RADS density, to a new commercial objective automatic software-based calculation called Quantra^{TM}. Secondly, to investigate how implementation of DBT in mammography screening impacts radiation dose and diagnostic accuracy for women of different breast densities.

The main findings of the studies relevant for the specific aims of the PhD thesis were:

- For about 40 % of the women included in this study, there was a risk of breast density classification as both dense and non-dense using the current subjective BI-RADS density assessment. New objective software-based automatic assessment can classify breasts with high accuracy compared to a median radiologist score.

- Automatic density classification is best performed using volumetric rather than area-based assessment, due to better agreement with radiologists for the most dense breasts. Still, classification into the extreme density categories had limited agreement with radiologist.

- Average glandular dose increases about 33 % for non-dense- and 8 % in dense breasts using DBT compared to DM for the mammography units used in the Oslo Tomosynthesis Screening Trial. Incorporating breast density into the dose calculations increases the dose estimates by 16 % on average.

- True-positive rate improved significantly when using DBT compared to DM for all age groups and women with scattered fibroglandular- and heterogeneously dense breasts, due to more findings classified as architectural distortions or spiculated masses. The improvement was not significant for women with almost entirely fatty- and extremely dense breasts. False-positive rate improved significantly for all age and breast density groups, except for extremely dense breasts, where false positive rate was comparable. The improvement was mainly due to less false positives classified as asymmetric density.

7.2 Moving from subjective to objective breast density assessment

A key element in personalized screening and breast density legislation is the breast density assessment. In this chapter the consequences of variability of subjective assessment, choice of density assessment method and the changes introduced by a potential transition to automatic assessment will be discussed.

Ng and Lau published criteria (or "sanity checks") based on physics and common sense which can be useful for evaluating breast density assessment methods (Table 12) **Feil! Fant ikke referansekilden.**(61). With the introduction of personalized screening, where breast density has an effect on the screening process for individual women, there is a need for these criteria not only to be valid for a population level, but also on an individual level (61,179).

Table 12: Criteria for evaluation of breast density assessment (61).

1	"Density should be the same for the identical image of the breast"
2	"Density should be similar for a breast no matter what the view, in particular CC and MLO views"
3	"Density should be similar for the same breast no matter the imaging equipment, in particular it should not matter if the equipment is GE, Siemens, Hologic, or if the imaging is done on mammography, tomosynthesis, MRI or CT"
4	"Density should be invariant to breast compression"
5	"Left and right breast densities should be highly correlated but not identical"
6	"Density should, over a population, generally reduce with age"

7.2.1 Inter-observer variability in BI-RADS density categorization inter-observer agreement

Since disagreement among readers could lead to the same breast receiving different density scores, inter-observer variability will impact criterion 1 in Table 12. In a personalized screening setting inter-observer variability could have major consequences, due to resulting in different paths in the screening program. Inter-observer variability of BI-RADS density assessment has been investigated in many studies. Among them the first paper included in this thesis, showing a weighted kappa between radiologists to the median radiologist classification of 0.86 (range of individual scores, 0.76 – 0.93) (156). Other authors have reported mean study kappa ranging from 0.43 to 0.90. For individual radiologists compared to a reference standard, it has been shown weighted kappa ranges from 0.02 to 0.93 (180,181,190–195,182–189). Estimates of inter-observer variability varies considerably among studies, and the estimates from paper I are among the studies with the highest agreement.

Most of the studies focusing on inter-observer variability were reader studies, rather than studies in a screening setting. The lowest agreement reported in reader studies are from early studies including radiologists with little or no previous experience using BI-RADS density. They report a mean study weighted kappa of 0.43 and 0.59 (183,186).

In other reader studies, radiologists had previous experience using BI-RADS density or were given prior oral instructions. These studies result in better agreement, with a mean study weighted kappa between 0.68 and 0.79 (180–182,184,190). This is consistent with an investigation by Gard et al., which reported better agreement among radiologists with more than 10 years' experience compared to those with less experience. The mean weighted kappa difference was 0.10 (95 % CI: 0.01 – 0.24) (190).

Similar to paper I in this thesis, many studies trained the radiologists prior to the study. These studies generally report the highest kappa values, with mean study kappa of 0.73 – 0.88 (156,185,189,191,193). It has been reported that training in BI-RADS feature analysis and assessment improved consistency (196) and it is likely this also extends to breast density assessment (182). Therefore, if subjective assessment is to be used in a personalized screening setting, retraining of radiologists is likely important in order to increase consistency of density assessment (191).

Two studies have assessed inter-observer agreement in a screening setting, comparing density assessments from two separate screening sessions separated by more than a year, reporting mean study kappa of 0.54 – 0.70 (192,194). These results indicate that density assessment in a screening setting will have lower inter observer agreement than in a reader study. Therefore, the impact of inter-observer variability in breast density assessment is likely larger in a screening setting than estimates reported from reader studies.

It has been suggested that radiographers could do breast density assessment, but this has been shown to result in even less agreement (kappa 0.69 versus 0.62) (184). Studies comparing breast BI-RADS density assessment un U.S. and United Kingdom (U.K.), has reported higher observer agreement in U.S. compared to U.K. since the U.K. radiologists use a different scale in daily practice (197). Still, even when using their native scales, readers from U.K. had worse agreement compared to readers from U.S., with even lower agreement among Australian radiologists (198). This indicates that BI-RADS density has relatively higher agreement compared to alternative subjective scales.

Intra-observer agreement

Even the same radiologist might not score the same breast consistently. Studies have found intra-observer variability using BI-RADS density ranging from 0.69 – 0.90 (183–185,187,188,190,193,199). Reported intra-observer variability show an increase in kappa by 0.08 - 0.20 compared to the corresponding inter-observer variability. This relatively consistent increase indicates that intra-observer variability is influenced by similar factors as experience and training in the use of the BI-RADS density scale.

Consequence of inter-observer variability in personalized screening

The authors of studies on inter-observer variability using BI-RADS density generally conclude the inter-observer variation is low, as mean kappa usually falls into the substantial or almost perfect category as stated by Landis and Koch (168). One interpretation of this, is that the majority score is a reliable reference standard for visual classification (200) and that BI-RADS density is a suitable and reproducible measure for application on a population basis. But for personalized screening and breast density legislation, the assessment must be reproducible on an individual level (61).

One statistical reason for the high agreement is use of quadratically weighted kappa, which penalizes a discrepancy of two or more categories more. Most discrepancies are by one

category (156,181), with BI-RADS density III as the category with the highest discrepancy between readers (8,181,182,186). Therefore, the highest disagreement is in the interface between BI-RADS density II and III (180,181,183,201). Approximately 80 % of the women are included in the density categories II and III (66,202), which is the currently the relevant threshold for breast density legislation. Therefore, kappa values can be high at the same time as the variation in number of women considered dense could be substantial.

An effect of the interobserver variability is that there will be a variation in proportion of women radiologists will consider as having dense breasts. From the reader studies, the proportion of women considered dense by individual radiologists varies by about 15 – 30 % from radiologists classifying least to most women with dense breasts (156,180–183,190,191,193,200). In two of the studies the variation was larger, with one radiologist classifying about 10 % of the women with dense breasts and another in excess of 60 % (183,190). A multicenter screening study including 83 radiologists, reported a range of 6.3 to 84.5 % dense women (8). This illustrates the great potential for variation in extreme radiologists. The interquartile range was 28.9 – 50.9 % (a 22 % difference), which puts a great number of women at risk of receiving either dense or non-dense classification, even by non-extreme radiologists (8). The variation in the proportion of women considered dense reflects the individual radiologist's threshold in density for categorizing a woman as dense, which we quantified using volumetric density (156).

Paper I reported that 35.6 % of the women was scored both dense and non-dense by individual radiologists (156). Furthermore, another study comparing 19 radiologists found that more than 80 % of the women who had non-dense breasts were considered to have dense by at least one radiologist. Similarly, almost 50 % of the women having dense breasts were considered non-dense (190).

An important aspect in density assessment is consistency in density interpretation over time for the same woman. A large study found that in two subsequent screening rounds read by different radiologists about 1.2 years apart resulted in a different density assessment 32.6 % of the times, where 17.2 % (more than 1 in 6) had a discordance resulting in a change in dense, non-dense category (8). A meta-analysis found about 1 in 5 (23 %) of women changed their BI-RADS density category when the subsequent screening examination was assessed by the same radiologist. This increased to about 1 in 3 was categorized differently when the subsequent examination was read by a different reader (201).

Introduction of 5th edition of BI-RADS density

The BI-RADS density 5th edition scores are more focused on potential masking of cancers (67,191,203). A breast with less than 50 % dense tissue, with a very dense area posterior to the nipple would get a BI-RADS density score C in the 5th edition rather than II in the 4th (66,67). This makes the 5th edition more suitable to identify in which breasts cancers could be missed and might benefit from supplemental imaging, which better facilitates personalized screening (191,204). However, the removal of the percent density from the 4th edition could potentially introduce additional variability in assessment (203).

Lower inter-observer variability using 5th edition compared to 4th edition has been reported in one reader study (193), while two other report no significant difference (189,195). Furthermore, two reader and one screening study report a significant shift towards more dense breasts (193,195,205). However, a large screening study including more than 3 million examinations interpreted by 722 radiologists, across 144 facilities, before and after the implantation of the 5th edition of BI-RADS density, showed no increase in the proportion of dense breasts (206). This is also consistent with the observed lack of change when introducing percentage area in the transition from the 3rd to the 4th edition BI-RADS density (67,194).

BI-RADS density 4th edition was used in the reader study and OTST (155,156,207). As the most comprehensive study comparing 4th and 5th edition BI-RADS density concludes that the density assessment is consistent when transitioning to 5th edition, our results obtained using the 4th edition scale should still be relevant after transition to the 5th edition scale.

Subjective breast density assessment in digital breast tomosynthesis

Potential changes of the BI-RADS density classification using BDT has been investigated in several studies, with conflicting results. Three large screening studies (including 24.756, 78.810 and 15.571 women) showed that women screened using DBT had a lower likelihood of being classified with dense breasts compared to those screened with DM alone (208,209), with even lower likelihood was found for women screened using DBT and synthetic mammograms, compared to those screened using DBT and DM (208–210). A smaller reader study comparing DBT and DM density assessment found similar results (211). Indicating, implementation of DBT screening may affect breast density estimation.

A large study comparing about 220,000 DBT screening examinations with about 750,000 DM examinations, showed no difference in proportions of women with dense breasts, indicating consistent BI-RADS density evaluation regardless of whether the evaluation is performed on DBT or DM acquisitions (206).

Two laboratory studies have shown comparable breast density assessment in synthetic 2D images and conventional DM images (212,213).

Whether or not DBT and/or synthetic 2D images change BI-RADS density categorization is still not clear, as large well-designed screening studies show conflicting results. If BI-RADS density assessment is shown to change using DBT, it could affect the comparability of the BI-RADS density assessment in OTST at consensus versus the reader study, as the radiologists had access to DBT images in OTST.

Subjective assessment

Studies have indicated that subjective breast density assessment will put a lot of women at risk of being classified as both dense or non-dense. The subjective nature of BI-RADS density and the following recommendation for supplemental screening may be more dependent on the interpreter of the mammogram rather than the amount and distribution of dense tissue in the breast (203). This has led to the ACR releasing a statement addressing the subjective breast density assessment (214). Thus, it is clear that an objective alternative for breast density assessment could be beneficial, reducing the inter-observer variability.

7.2.2 Automatic assessment of breast density

A potential improvement to solve the inter-observer variability using radiologist's classification, is to use automatic breast density assessment software. Since radiologists interpret projection images and not a 3D model of the physical breast, it is not clear whether volumetric or area-based breast density corresponds best with the radiologist's assessment. In this thesis we have investigated cut-off values using both area-based and volumetric measures as reported by Quantra™.

In paper II, two different methods for determining cutoff values were investigated (207). One where sensitivity and specificity compared to BI-RADS density assessment was maximized (using Youden's index). And one where the cutoff values were set to reproduce the radiologist's density distribution. For both methods of obtaining cutoff values, the performance was better using volumetric assessment for the threshold between dense and

extremely dense breasts. The performance was approximately equal performance for the threshold between dense and non-dense breasts. The performance was improved, using area-based assessment for the threshold between almost entirely fatty and scattered fibroglandular breasts. Since it is the denser breasts that are potentially interesting with respect to personalized screening and breast density legislation, volumetric density was concluded as preferable (207). The ROC method produced density distributions which were different from the radiologist's distribution, due to the excessive numbers of women classified as almost entirely fatty and extremely dense. These thresholds are therefore not likely to be clinically relevant. Therefore, the method of setting cutoff values to generate the radiologist's distribution is preferable. This method produces a comparable distribution as the default categorization in Quantra™ (Quantized density), and it is likely the approach used by the manufacturer. The thresholds used for Quantized density are not identical as the ones obtained in paper II. This could be due to Quantra™ internally uses decimal precision and potential bias due to the study limitations in the reader study in this thesis.

Another reason for using volumetric assessment can be found using criteria 2 in Table 12; the density should be the same for CC and MLO views. This is not necessarily true for area-based assessment. If a hypothetical breast has 100 % dense tissue in both upper quadrants and 0 % dense tissue in the lower quadrants, the volumetric density would be 50 %.

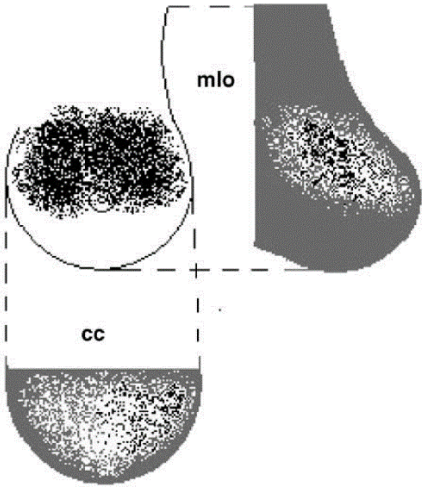

Figure 9: An illustration of the difference in areometric density assessment in CC and MLO views of a hypothetical breast with dense fibroglandular tissue in the two upper quadrants. Source: Elsevier (183) *(with permission).*

Using areometric assessment, the CC density could be 100 % as all pixels representing the breast show fibroglandular tissue, while the MLO view would show 50 % density (Figure 9) (65,183). BI-RADS density 4th edition is a measure which represents the volume of fibroglandular tissue in the breast (65). Therefore, volumetric breast density assessment is preferable to areometric as it best represents the physical properties of the breast. The correlation between breast density for CC and MLO views using area-based (Figure 10a) and volumetric (Figure 10b) breast density in OTST shows higher r^2 for volumetric breast density. This confirms that volumetric breast density is a better measure of breast density according to criteria 2 of Table 12.

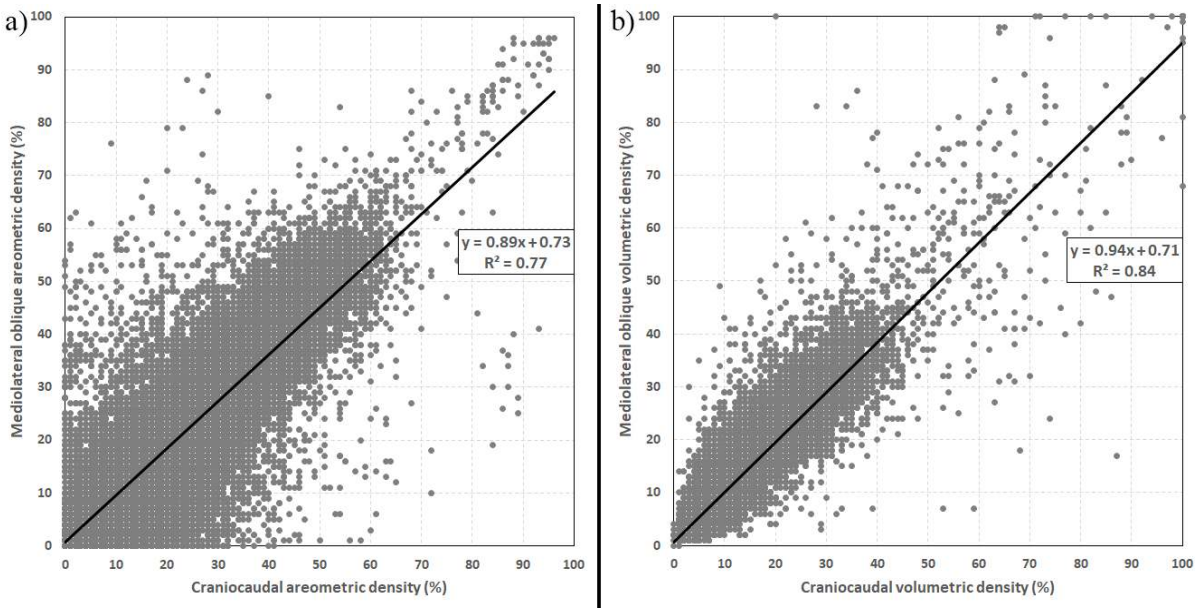

Figure 10: Scatterplot of craniocaudal and mediolateral oblique views of the same breast using areometric breast density a), and volumetric breast density b).

The results from this thesis showed that volumetric breast density assessment was preferable to area-based assessment which correlates better with BI-RADS 4th edition assessment in clinically relevant density ranges. It is also a better measure of breast density, since measurements of two different projections of the same breast is in better agreement than area-based assessment.

Potential transition to volumetric breast density assessment

When a new method of breast density classification is introduced, it will likely introduce some systematic changes in breast density classification. One method of quantifying a potential change, is to assess the inter-observer variability between the radiologist's score and

Quantized density. This measure can be compared to the inter-observer variability between individual radiologists and the majority score, to assess whether the systematic shift is larger than the current variability in subjective assessment. Paper I showed that mean kappa between radiologists and the median score for four category classification was 0.86 with the lowest at 0.76 (156). A similar result for Quantized density was 0.72. For classification of women with dense and non-dense breasts, mean radiologist kappa was 0.80 with the lowest at 0.62. The similar categorization for Quantized density was 0.64. These results indicate that the systematic shift introduced by introducing Quantized density, would be of the same magnitude as the inter-observer variation among some radiologists. Youk et al. reported similar results (86), while Singh et al. reported better agreement between Quantized density and radiologists than among the radiologists themselves in a small sample (215). Brandt et al. reported lower kappa between Quantized density and radiologists in a retrospective study in a screening setting (0.46 for four categories and 0.59 for two category classification) (91). Once the shift to automatic assessment has been performed, the reproducibility in breast density assessment going forward will be very high (90,215–217).

The systematic shift in density assessment introduced by automatic assessment could also affect the distribution of breast density categories. Most notably the number of women with scattered fibroglandular densities was 13.2 % higher using Quantra™ compared to radiologist assessment. Similarly, the number of women with heterogeneously dense was reduced by 8.2 % and the number of women with extremely dense breasts were reduced by 3.5 % (a 34 % decrease in number of extremely dense breast compared to BI-RADS density assessment). Consequently, about a third of the women are considered dense compared to about half using BI-RADS density assessment (207). Similar changes in density distribution has been reported by other authors (86,91). If Quantized density was implemented in personalized screening, fewer women would be subject to receiving breast density notification or supplemental screening. For the extreme density categories, the agreement with radiologists was poor. Only 37 – 47 % of women classified as almost entirely fatty were correspondingly classified using Quantized density (Table 8 and Table 9). For extremely dense breasts the respective corresponding classification was 55 – 65 % (Table 8 and Table 9). This has been noted by other authors, suggesting Quantra™ ver. 2.0 should only be used for two category classification; dense and non-dense (218). The conclusion by Ekpo et al. is logical considering the poor agreement. But as the distributions (Figure 1a) in paper II shows, most of the discrepant cases are neighboring categories with volumetric density close to the

cutoff value. When comparing diagnostic performance stratified using both density measure (paper IV), BI-RADS density assessment and Quantra™ produce similar results (155).

Different manufacturers of breast density assessment software can result in different density distributions. Most notably the proportion of extremely dense breasts have been reported substantially higher using Volpara™ compared to BI-RADS density and Quantra™ (86,91,219). This can have a substantial effect on personalized screening, especially if only women with extremely dense breasts are targeted for supplemental modalities. Therefore, a standardization of automatic breast density algorithms prior to implementation in personalized breast cancer screening would be beneficial. Researchers evaluating such screening programs in the future will have to account for the method of automatic breast density assessment used.

Variation in automatic assessment

Even if breast density is evaluated using a computer algorithm, variation in assessment can occur due to several reasons. Some examples are; paddle tilt (204,220) and positioning (204,221), which cause variation in breast thickness and tissue being evaluated. If women with implants are not correctly tagged using DICOM, the algorithm might not calculate density correctly (92). The radiation dose used to acquire the image has also been shown to potentially affect automatic breast density estimates (222). However, a clinical study showed that breast density assessment was reproducible using low dose mammograms (223). It is also important to ensure the mammography unit performs consistently as irregularities in the dose control and breast thickness calibration could cause errors (224). Therefore, regular monitoring of the mammography units through quality assurance is important. Changes to the breast density assessment algorithm could also potentially affect density values, such as when Hologic implemented a skin tissue correction in Quantra™ ver. 2.0 (156).

Automated breast density assessment in DBT

After OTST was finished Quantra™ was updated to facilitate breast density assessment for DBT acquisitions (225). Quantra™ calculates the breast density in a very similar manner in DBT as in DM. In DBT assessment, only the central (0^{o}) projection is used. The input data of the DBT algorithm is therefore a low dose image produced using a W/Al anode/filter combination with slightly higher kVp than the corresponding DM image. The images are also obtained without using a grid, which increases the amount of scattered radiation to the detector, complicating the analysis (61). Preliminary tests on phantoms found that the breast density was about 10 % higher using the DBT algorithm compared to the DM algorithm

(225,226). Quantra™ version 2.1.1, capable of breast density assessment in DBT uses a similar algorithm as Quantra™ version 2.0 used in this thesis (227). The cut-off values for BI-RADS like density categories are slightly different as it was calibrated to the 5th edition of BI-RADS. Compared to radiologists majority score, Quantra™ used on DBT showed comparable inter-observer variability (227). Other algorithms than Quantra™ has been shown to perform comparable breast density estimation in DBT and DM (228,229) and between DM and synthetic DM (230).

Automatic assessment and BI-RADS density 5th edition

In BI-RADS 5th edition the breast density categories are not entirely dependent on volumetric breast density. Dense patches of breast tissue can mask lesions, and thereby fulfilling the criteria for a dense breast, even without high volumetric density. This change is important as localized densities have been shown to be associated with interval- and large cancers (231). Consequently, Hologic has updated Quantra™ (version 2.2), and replaced volumetric density as the input to BI-RADS like classification with machine learning algorithms using pattern and texture analysis as input (82). In this version of Quantra™ only a BI-RADS 5th edition category is shown, not volumetric breast density. Quantra™ version 2.2 is still likely capable of calculating and exporting volumetric breast density for research and other purposes. Still, in order to receive FDA approval, they were obliged to hide the volumetric density to avoid confusing the clinical users (Personal communication, Ashwini Kirshagari, Hologic Inc.).

Automatic volumetric assessment

Automatic volumetric breast density has proved to be feasible in both DM and DBT mammographic screening, producing density scores similar to 4th and 5th edition of BI-RADS density (86,91,207). Still, a transition to implementing automatic assessment would introduce a shift in breast density assessment comparable to the inter-observer variation currently present in subjective interpretation for some radiologists. This transition would also potentially change the proportion of dense and extremely dense breasts in the screening population. This change must be taken into account when planning a potential personalized mammography screening. In case of Quantra™ version 2.0 the proportion of dense and extremely dense breasts would be reduced compared to the current subjective assessment.

7.3 Radiation dose and the potential transition from DM to DBT

Radiation exposure can lead to radiation induced cancer (232). Therefore, all use of medical radiation should be justified and as low as reasonably achievable (233). In medical screening the women attending are not known to have a disease justifying the exposure. Therefore, the tolerance exposing women to radiation is lower in screening- compared to clinical mammography. However, the risk of radiation induced cancer is reduced with age (111), and as mammography screening targets older women, the risk of radiation induced cancer in this cohort is low (112). One potential drawback of using DBT is increased radiation dose due to the acquisition of multiple projections. If this dose difference is large, it would require a greater difference in diagnostic accuracy to justify the additional radiation exposure.

Radiation dose estimates and volumetric breast density

Volumetric breast density assessment provides not only a mean for stratifying women as having dense and non-dense breasts. It also provides information on the breast content, which can be used to increase the precision of breast density assessment. Breast dosimetry is usually performed assuming 50 % glandularity (162). As seen by the density distribution in paper I and II most women fall between 4 – 12 % volumetric breast density, resulting in a mean glandularity of 15.9 % (165). Since fat is more radiolucent than fibroglandular tissue, there will be less attenuation compared to less glandular breasts, resulting in higher AGD for the same exposure. In OTST, there would have been an underestimation of mean glandular dose by about 16 % using the Dance method (165). VolparaDose™ (Volpara) incudes volumetric breast density into dose calculations in a similar manner as the method presented in paper III. A comparison of the mammography unit's AGD estimate to that of VolparaDose™ showed similar results as our study, underestimating AGD when not accounting for breast density (234).

Report no. 457 of the IAEA provides estimates on appropriate uncertainties in radiation dose estimates in diagnostic radiology. If doses are used for estimates of risk due to radiation exposure an accuracy of 20 % at a 95 % confidence interval is suggested. If dose estimates are to be used to compare procedures the corresponding accuracy is 7 % (159). With the mean effect of breast density being 16 %, it is necessary to incorporate volumetric density into the radiation dose estimates in order to reach the accuracy recommended by IAEA. This means that even if vendors transition to machine learning-based categorization of breast

density, volumetric breast density estimation should be included as a part of the radiation dosimetry estimation provided by the mammography system.

Another obstacle for reaching the desired accuracy is that the simple models by Dance et al. is based on assumption of homogeneous glandular tissue. This model has been shown to overestimate radiation dose compared to simulations using realistic breast models. The average overestimation was about 30 % (235,236), with individual errors as large as 120 % (235). This systematic difference indicates that new conversion factors representing more realistic breast compositions in addition to accounting for volumetric breast density are needed. Alternatively, the imaged breast could be modelled using tissue classification from DBT series, followed by radiation dose simulation immediately after imaging (237). To avoid delays in the imaging workflow, the process of providing this information in the DICOM metadata should be fast. It is therefore unclear whether this level of dosimetry will become feasible to implement.

Radiation dose estimates for DM and DBT

Paper III showed that when using DBT, the AGD was about 24 % higher compared to DM, increasing from 1.74 to 2.10 mGy per view on average (165). Both the AGD from DM and DBT were within the limits set in the European Guidelines (238). It was also shown that the AGD from DBT was 33 % higher for women with non-dense breasts and 8 % higher for women with dense breasts (165). Therefore, the increase in AGD for DBT is lower for women with dense breasts. The reason for this difference is that the AEC is controlled primarily by breast thickness in DBT mode, while it compensates more for breast density in DM mode (239). For other vendors such as GE, DBT and DM provides a more equal dose level (240), as the radiation dose requirement for DBT and DM is calculated using the same method, and the dose is divided evenly among the DBT projections. Thus, it is important to be aware that the difference in dose between DM and DBT will vary between vendors due to the specific vendors choices of system design.

Other authors have compared DM and DBT doses using Hologic equipment (241–246). Their estimates agree well with our estimates from paper III, and differences can be attributed to differences in methodology, upgrades of the Hologic dose tables and differences in the study cohort (165).

The increase in mortality due to the increased AGD due to using DBT (based on data from paper III) was estimated by Brown and Covington to be extremely small (247). Their

estimates of risk for DM examinations are consistent with previous estimates from Yaffe and Mainprize for conventional screening mammography (112).

7.4 True- and false positives and the potential transition from DM to DBT

Cancer detection

Paper IV reported higher true positive rate for DBT compared to DM for all breast densities. The true positive rate increased by 12 – 24 %, with the results being significant for dichotomized breast density (dense and non-dense), scattered fibroglandular- and heterogeneously dense breasts (155). Results were not significant for almost entirely fatty breasts and extremely dense breasts, which could be due to OTST being underpowered for stratified analysis in these small subgroups. Consistent results have been reported in prospective and retrospective studies (40,43,44,97,175,248–251).

These results indicate that DBT reveals more cancers than DM for women of all breast densities, although large studies are still needed for conclusive results in the smaller almost entirely fatty and extremely dense categories. A meta-analysis found greater improvement in CDR in European studies than in U.S. studies, which might be due to lower CDR in U.S. because of annual screening (46). Most of the additional cancers detected by DBT was classified as spiculated mass or distortions, which has also been seen by other authors (40,252). These types of tumors are typically slow growing (248). The characteristics of these tumors in OTST were typically small invasive cancers with excellent prognosis (41). A meta-analysis concluded similarly that DBT has superior sensitivity for soft tissue masses, but results regarding improved detection of lesions containing calcifications are inconsistent (45). This could be due differences in image acquisition or reconstruction, as clusters of microcalcifications might be easier to visualize in thicker slabs (45). In the OTST, more true positives classified as calcifications were found using C-view, which highlights calcs, compared to DBT plus DM (41). The results from OTST indicates that DBT reduces masking compared to DM, primarily for tumors manifesting as spiculated masses or architectural distortions and that this reduction of masking is valid for breasts of all densities. Although cancer detection is improved using DBT, especially for early invasive cancers (252), results from paper IV and other studies does not show a consistent trend in improvement in cancer detection with breast density.

The reduction in masking using DBT for small spiculated masses for breasts of all densities can be understood by considering the images produced by the DBT system. The first important property of a DBT image is that it is not truly 3-dimensional like a computed tomography (CT) image. Every image in a DBT image stack contains image information from the entire breast. The DBT images differ in the location of the plane of focus. The remaining anatomy of the breast is blurred increasingly with distance from the focus plane. The contrast of an object relative to a background is therefore lost according to the artifact spread function (38). For a system like Selenia Dimensions with a relatively narrow tomo angle of ±7.5°, the out of focus signal will be blurred less compared to a system with a larger angle. The out of focus blurring results small structures such as fibers and vessels are being suppressed when out of focus. Our results indicate that this suppression removes distracting image texture which masks spiculations in DM. An example of this is shown in Figure 3 in paper IV (155). Although the tumor (Figure 3 in paper IV) is visible in DM, it is not presented sufficiently suspicious to the radiologists at DM as the overlying fibers mask spiculations. Therefore, the cancer was missed in a busy screening setting. Larger structures such as patches of glandular tissue will only have their edges blurred when out of focus. The signal centrally in the patch will remain similar as blurring simply will mix the signal from areas of similar density. This has been seen in a comparison of anatomical noise of structures of size greater than 2 mm in DM, DBT and dedicated breast CT. DM and DBT were shown to have similar anatomical noise, while CT images showed a reduction due to true 3D reconstruction (253). As DBT fundamentally relies on the same image contrast as DM, DBT must have peritumoral fat in order to visualize the tumor (254,255). Therefore, DBT improved cancer detection compared to DM similarly for all breast densities, including non-dense breasts, by finding smaller spiculated tumors or distortions.

False positives

Paper IV shows a significant reduction in false positives, primarily due to reduction in asymmetric densities, except for women with extremely dense breasts. A reduction in false positive rate results in a reduction in recall rate, which is the performance measure usually reported in studies.

Numerous retrospective studies from the U.S. has shown reduction in recall rate for women with dense and non-dense breasts (97,249–251,256–260). For almost entirely fatty breasts none of the U.S studies showed significant reduction in recall rate (249,257,258), while two studies showed a significant decrease in recall rate for extremely dense breasts

(249,257). The European studies were performed in screening programs with much lower recall rate (104). Three European studies reported increased recall rate (43,44,261). However, for the Screening with Tomosynthesis OR standard Mammography (STORM) trail, this was due to the study design, and implementation of DBT would have reduced the number of false positives (262). A recent published Norwegian trial found significant reduction in recall rate, but only in non-dense breasts (240).

These results show that DBT improves the false positive rate compared to DM. However, the magnitude of this improvement depends on the screening setting. In the U.S., where recall rates are higher, implementation of DBT reduces recall rate more compared to in Europe. Still, a meta-analysis concluded that implementation of DBT provides a benefit with respect to reducing recall rate in Europe (46). Differences in results in European and U.S. studies in almost entirely fatty- and extremely dense breast, indicate that women with very dense breasts benefit from recall reduction in the U.S., but not in Europe (155,240,249,257). Conversely, women with very low breast density seem to benefit more from a reduction in recall rate in Europe compared to the U.S (155,240,249,257).

The main contribution to the reduction of false positives using DBT, was less asymmetric densities. This can be explained by the 3D information provided by the DBT image stack. In conventional DM, glandular tissue can be superimposed creating a pseudo-lesion (manifesting as asymmetric densities). By blurring out of focus tissue, this superposition is minimized, allowing DBT to resolve many of the pseudo-lesions otherwise seen in DM. An example of this can been seen in Figure E3 in paper IV (155). A possible reason for the lack of reduction in recall rate in extremely dense breasts in our study, is that in these breasts, there might often be too much glandular tissue in the vicinity of the lesion for out of focus blurring to resolve the lesion. False positives were also reduced for spiculated masses and distortions, which might be due to DBT's improved ability to characterize small spiculated masses. However, in contrast to OTST, the Malmö trial found an increase in false positives, mostly due to stellate distortions (261). For findings classified as calcifications, there was an increase in false positives, due to more false positives classified as calcifications from Arm D in OTST, which used DBT plus synthetic DM (C-view) (41). As C-view highlights calcifications, they may become more conspicuous and more difficult to characterize (155).

7.5 Consequence of density assessment method

In the previous part, the change in density distribution due to the shift from subjective to objective breast density assessment has been discussed. A major benefit of objective assessment is the reproducibility. A potential drawback is a possible change in density distribution when compared to BI-RADS density, and the low agreement in almost entirely fatty and extremely dense breasts. In paper IV, the relative performance of DM and DBT using either density stratification was similar (155). This indicates that both BI-RADS and Quantra™ categorizes breasts with similar properties with respect to true and false positive rate in the same breast density categories, even though the density distribution was changed, and some categories had limited agreement (155). This is supported by another European study reporting sensitivity and specificity for DM stratified using Volpara™, which shows results consistent with changes observed using BI-RADS stratification (141). A U.S. study even showed that volumetric breast density using Volpara™, captures the potential for masking better than BI-RADS density (263). This strengthens the argument for using objective density assessment in mammography screening. Furthermore, indicating that volumetric breast density has the potential of replacing BI-RADS density to assess the potential for masking.

7.6 DBT in population-based and personalized screening

Recent evaluations of DBT as a screening modality concluded that the evidence was insufficient for recommending implementation of DBT (45,46). Such evaluation was beyond the scope of this thesis. Still, the results in this thesis and other studies contribute to the scientific evidence, which will continue to grow with large trials such as the Tomosynthesis Mammographic Imaging Screening Trial (TMIST) trial (264). The result from OTST and many other studies shows improved cancer detection and a reduction in false positive findings using DBT when compared to DM, indicating a benefit for women of all breast densities. This improvement comes at a cost of a minimal increase in radiation dose.

Does this mean the need for breast density legislation and personalized screening is minimized? Even though DBT revealed more cancers, the interval cancer rate in OTST was not reduced. Furthermore, the cancers detected using DBT only, had different characteristics compared to the interval cancers (108). Similar results have been found by other researchers (265). This indicates that the cancers appearing as interval cancers, still might be masked in DBT. If this is shown to be true, the need for breast density legislation and personalized

screening would be unchanged after a potential implementation of DBT as a primary mammography screening modality. As results indicate that DBT increases cancer detection without associated reduction in interval cancer rates, the DBT only detected cancers could potentially represent overdiagnosis (108). Thus, there is a need for further research on DBT to assess whether DBT detected cancers represent potential overdiagnosis or earlier detection of clinically relevant cancer compared to DM.

8 Conclusion and future aspects

8.1 Conclusion

BI-RADS breast density assessment results in considerable inter-observer variability, which leads to uncertainty concerning breast density notification and personalized screening. Automatically calculated breast density assessment potentially resolves this issue. This thesis showed that volumetric assessment would be preferable to area-based, due to better agreement with radiologists' assessment in dense breasts, which is the most clinically relevant. In addition, volumetric density was more reproducible for different mammographic views.

A transition from BI-RADS density to volumetric assessment using Quantra™ would introduce a shift in breast density assessment comparable to the inter-observer variability of some radiologists already performing breast density assessment today. In case of Quantra™ the number of women considered dense would also be reduced from about half the women in subjective assessment to a third. Additionally, the agreement between Quantra™ and BI-RADS density is limited in the extreme density categories. However, the relative diagnostic performance between DM and DBT is similar using either density measure. This indicates that transitioning from subjective to volumetric assessment would capture women with similar diagnostic performance in the corresponding categories.

Incorporation of volumetric breast density into breast dosimetry increases the mean estimate of AGD. A potential transition from DM to DBT would increase the radiation dose using Hologic equipment. This increase is lower for women with women with dense compared to non-dense breasts, due to the design of the AEC system. This increase in dose is estimated to have a minimal impact on risk of mortality due to radiation compared to the expected benefit if DBT in mammography screening.

The introduction of DBT in mammography screening improved the true- and false positive rate compared to DM, improving diagnostic performance for women of all breast densities. These results indicate that DBT could lead to earlier diagnosis of typically slow growing cancers with excellent prognosis, which manifests as spiculated masses or distortions. The reduction in false positives would reduce the number of women recalled for further assessment due to superposition of breast tissue, except for women with extremely dense breasts.

8.2 Future aspects

Implementation of automatic breast density assessment in population-based screening and personalized screening is in its infancy. Additionally, new deep learning algorithms such as Transpara™ have been introduced. This software is not based on breast density evaluation but quantifies the risk of a cancer being present based on the information in the image. Such software could also play a major role in the screening workflow and personalization of screening in the future (266). To fully assess the potential of software augmented screening, further research is needed, especially with respect to new machine learning-based methods (82).

DBT is gradually being implemented in mammography screening. Still, there is a need for larger screening trials to assess performance in certain subgroups. There is also a need for long follow-up studies on women screened with DBT, to assess whether the small cancers detected represent an earlier detection of what would be DM detected cancers at a later stage, or if they could represent overdiagnosis.

And finally, if personalized mammography screening is implemented for women with dense breasts there is a major need for future studies on adjunct modalities such as Ultrasound, MRI and contrast enhanced DM. These have considerable disadvantages compared to conventional mammography related to workload, cost and false positives. The benefit and risks associated with adjunct screening needs further evaluation in order to decide which women should be selected for such modalities. Fortunately, several studies are in progress, performing such evaluations (82,177). In the future, further technical developments in x-ray mammography, such as energy discriminating detectors, might facilitate spectral mammography or tomosynthesis, allowing for further improvements in detection, characterization of lesions and breast density assessment using mammographic equipment (267). Maybe the opportunities provided by new image analysis, and advances in medical imaging technology, will save and improve lives by detecting breast cancers early, allowing treatment with minimal morbidity. This benefit will only materialize following well designed studies and evidence-based recommendations.

References

1. Egan RL, Mosteller RC. Breast cancer mammography patterns. Cancer. 1977;40(5):2087–90.

2. Whitehead J, Carlile T, Kopecky KJ, Thompson DJ, Gilbert FI, Present AJ, et al. Wolfe mammographic parenchymal patterns. A study of the masking hypothesis of Egan and Mosteller. Cancer. 1985;56(6):1280–6.

3. Weigel S, Heindel W, Heidrich J, Hense HW, Heidinger O. Digital mammography screening: sensitivity of the programme dependent on breast density. Eur Radiol. 2017;27(7):2744–51.

4. Bae MS, Moon WK, Chang JM, Koo HR, Kim WH, Cho N, et al. Breast cancer detected with screening US: reasons for nondetection at mammography. Radiology. 2014;270(2):369–77.

5. Public law:
   U.S. Food and Drug Administration. Mammography Quality and Standards Act (2019). Available from: https://www.govinfo.gov/content/pkg/FR-2019-03-28/pdf/2019-05803.pdf

6. Hooley RJ. Breast Density Legislation and Clinical Evidence. Radiol Clin North Am. 2017 May;55(3):513–26.

7. Oberaigner W, Daniaux M, Geiger-Gritsch S, Knapp R, Siebert U, Buchberger W. Introduction of organised mammography screening in Tyrol: results following first year of complete rollout. BMC Public Health. 2011;11(1):673-80.

8. Sprague BL, Conant EF, Onega T, Garcia MP, Beaber EF, Herschorn SD, et al. Variation in Mammographic Breast Density Assessments among Radiologists in Clinical Practice: A Multicenter Observational Study. Ann Intern Med. 2016;165(7):457–64.

9. Niklason LT, Christian BT, Niklason LE, Kopans DB, Castleberry DE, Opsahl-Ong BH, et al. Digital tomosynthesis in breast imaging. Radiology. 1997;205(2):399–406.

10. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.

11. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al. Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer, 2018. Available from: https://gco.iarc.fr/today, accessed [11 Sept 2019].

12. Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Piñeros M, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J cancer. 2019;144(8):1941–53.

13. Cancer Registry of Norway. Cancer in Norway 2016 - Cancer incidence, mortality, survival and prevalence in Norway, 2017. Available from: https://www.kreftregisteret.no/globalassets/cancer-in-norway/2016/cin-2106.pdf

14. Oeffinger KC, Fontham ETH, Etzioni R, Herzig A, Michaelson JS, Shih Y-CT, et al. Breast Cancer Screening for Women at Average Risk. JAMA. 2015;314(15):1599-614.

15. American Cancer Society. Survival Rates for Breast Cancer. Atlanta, GA: American Cancer Society, 2018. Available from: https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-survival-rates.html, accessed [03 Dec 2018].

16. Ellis H. Anatomy of the breast. Surg. 2010 Mar;28(3):114–6.

17. Boyd NF, Dite GS, Stone J, Gunasekara A, English DR, McCredie MRE, et al. Heritability of mammographic density, a risk factor for breast cancer. N Engl J Med. 2002;347(12):886–94.

18. Vachon CM, Kuni CC, Anderson K, Anderson VE, Sellers TA. Association of mammographically defined percent breast density with epidemiologic risk factors for breast cancer (United States). Cancer Causes Control [Internet]. 2000 Aug;11(7):653–62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8911192

19. Moshina N. Understanding the role of mammographic density in a population based breast cancer screening program : A step towards stratified screening for breast cancer in Norway? [dissertation]. Oslo, Norway: University of Oslo; 2017. Available from: https://www.duo.uio.no/bitstream/handle/10852/59397/PhD-Moshina-DUO.pdf

20. McDonald S, Saslow D, Alciati MH. Performance and reporting of clinical breast examination: a review of the literature. CA Cancer J Clin. 2004;54(6):345–61.

21. Provencher L, Hogue JC, Desbiens C, Poirier B, Poirier E, Boudreau D, et al. Is

clinical breast examination important for breast cancer detection? Curr Oncol. 2016;23(4):e332–9.

22. Swann CA, Kopans DB, McCarthy KA, White G, Hall DA. Mammographic density and physical assessment of the breast. Am J Roentgenol. 1987;148(3):525–6.

23. Bushberg JT, Seibert JA, Leidholdt Jr EM, Boone JM. The Essential Physics of Medical Imaging. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2013.

24. Haus AG. Historical technical developments in mammography. Technol Cancer Res Treat. 2002;1(2):119–26.

25. Gold RH, Bassett LW, Widoff BE. Highlights from the History of Mammography. RadioGraphics. 1990;10(6):1111–31.

26. Yaffe MJ, Mainprize JG, Jong RA. Technical developments in mammography. Health Phys. 2008;95(5):599–611.

27. Lewin JM, Hendrick RE, D'Orsi CJ, Isaacs PK, Moss LJ, Karellas A, et al. Comparison of full-field digital mammography with screen-film mammography for cancer detection: results of 4,945 paired examinations. Radiology. 2001;218(3):873–80.

28. Skaane P, Young K, Skjennald A. Population-based mammography screening: comparison of screen-film and full-field digital mammography with soft-copy reading--Oslo I study. Radiology. 2003;229(3):877–84.

29. Skaane P, Skjennald A. Screen-film mammography versus full-field digital mammography with soft-copy reading: randomized trial in a population-based screening program--the Oslo II Study. Radiology. 2004;232(1):197–204.

30. Pisano ED, Gatsonis CA, Hendrick E, Yaffe MJ, Baum JK, Acharyya S, et al. Diagnostic Performance of Digital Versus Film Mammography for Breast-Cancer Screening. N Engl J Med. 2005;353(17):1774–83.

31. Sprague BL, Arao RF, Miglioretti DL, Henderson LM, Buist DSM, Onega T, et al. National Performance Benchmarks for Modern Diagnostic Digital Mammography: Update from the Breast Cancer Surveillance Consortium. Radiology. 2017;283(1):59–69.

32. Carney PA, Miglioretti DL, Yankaskas BC, Kerlikowske K, Rosenberg R, Rutter CM, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. Ann Intern Med. 2003;138(3):168–75.

33. Yaffe MJ, Bloomquist AK, Hunter DM, Mawdsley GE, Chiarelli AM, Muradali D, et al. Comparative performance of modern digital mammography systems in a large breast screening program. Med Phys. 2013;40(12):121915.

34. Prummel M V., Muradali D, Shumak R, Majpruz V, Brown P, Jiang H, et al. Digital Compared with Screen-Film Mammography: Measures of Diagnostic Accuracy among Women Screened in the Ontario Breast Screening Program. Radiology. 2016;278(2):365–73.

35. Skaane P. Studies comparing screen-film mammography and full-field digital mammography in breast cancer screening: updated review. Acta Radiol. 2009;50(1):3–14.

36. Pisano ED, Hendrick RE, Yaffe MJ, Baum JK, Acharyya S, Cormack JB, et al. Diagnostic Accuracy of Digital versus Film Mammography: Exploratory Analysis of Selected Population Subgroups in DMIST. Radiology. 2008;246(2):376–83.

37. Skaane P, Hofvind S, Skjennald A. Randomized trial of screen-film versus full-field digital mammography with soft-copy reading in population-based screening program: follow-up and final results of Oslo II study. Radiology. 2007;244(3):708–17.

38. Hu YH, Zhao B, Zhao W. Image artifacts in digital breast tomosynthesis: Investigation of the effects of system geometry and reconstruction parameters using a linear system approach. Med Phys. 2008;35(12):5242–52.

39. Ruth C, Smith A, Stein J, inventor; System and method for generating a 2D image from a tomosynthesis data set. United States patent US 7760924B2. 2008.

40. Zackrisson S, Lång K, Rosso A, Johnson K, Dustler M, Förnvik D, 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. Lancet Oncol. 2018;19(11):1493–503.

41. Skaane P, Bandos AI, Niklason LT, Sebuødegård S, Østerås BH, Gullien R, et al.

Digital Mammography versus Digital Mammography Plus Tomosynthesis in Breast Cancer Screening: The Oslo Tomosynthesis Screening Trial. Radiology. 2019;291(1):23–30.

42. Hofvind S, Holen ÅS, Aase HS, Houssami N, Sebuødegård S, Moger TA, et al. Two-view digital breast tomosynthesis versus digital mammography in a population-based breast cancer screening programme (To-Be): a randomised, controlled trial. Lancet Oncol. 2019;20(6):795-805.

43. Ciatto S, Houssami N, Bernardi D, Caumo F, Pellegrini M, Brunelli S, et al. Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study. Lancet Oncol. 2013;14(7):583–9.

44. Bernardi D, Macaskill P, Pellegrini M, Valentini M, Fantò C, Ostillio L, et al. Breast cancer screening with tomosynthesis (3D mammography) with acquired or synthetic 2D mammography compared with 2D mammography alone (STORM-2): a population-based prospective study. Lancet Oncol. 2016;17(8):1105–13.

45. Gilbert FJ, Tucker L, Young KC. Digital breast tomosynthesis (DBT): A review of the evidence for use as a screening tool. Clin Radiol. 2016;71(2):141–50.

46. Marinovich ML, Hunter KE, Macaskill P, Houssami N. Breast Cancer Screening Using Tomosynthesis or Mammography: A Meta-analysis of Cancer Detection and Recall. J Natl Cancer Inst. 2018;110(9):942–9.

47. Hofvind S, Geller B, Vacek PM, Thoresen S, Skaane P. Using the European guidelines to evaluate the Norwegian Breast Cancer Screening Program. Eur J Epidemiol. 2007;22(7):447–55.

48. Perry H, Phillips J, Dialani V, Slanetz PJ, Fein-Zachary VJ, Karimova EJ, et al. Contrast-Enhanced Mammography: A Systematic Guide to Interpretation and Reporting. Am J Roentgenol. 2019;212(1):222–31.

49. Sung JS, Lebron L, Keating D, D'Alessio D, Comstock CE, Lee CH, et al. Performance of Dual-Energy Contrast-enhanced Digital Mammography for Screening Women at Increased Risk of Breast Cancer. Radiology. 2019;293(1):81-8.

50. Mori M, Akashi-Tanaka S, Suzuki S, Daniels MI, Watanabe C, Hirose M, et al. Diagnostic accuracy of contrast-enhanced spectral mammography in comparison to

conventional full-field digital mammography in a population of women with dense breasts. Breast Cancer. 2017;24(1):104–10.

51. Brem RF, Tabár L, Duffy SW, Inciardi MF, Guingrich JA, Hashimoto BE, et al. Assessing improvement in detection of breast cancer with three-dimensional automated breast US in women with dense breast tissue: the SomoInsight Study. Radiology. 2015;274(3):663–73.

52. Ho JM, Jafferjee N, Covarrubias GM, Ghesani M, Handler B. Dense breasts: a review of reporting legislation and available supplemental screening options. Am J Roentgenol. 2014;203(2):449–56.

53. Skaane P, Gullien R, Eben EB, Sandhaug M, Schulz-Wendtland R, Stoeblen F. Interpretation of automated breast ultrasound (ABUS) with and without knowledge of mammography: a reader performance study. Acta Radiol. 2015;56(4):404-12.

54. Kuhl CK, Schrading S, Strobel K, Schild HH, Hilgers RD, Bieling HB. Abbreviated breast Magnetic Resonance Imaging (MRI): First postcontrast subtracted images and maximum-intensity projection - A novel approach to breast cancer screening with MRI. J Clin Oncol. 2014;32(22):2304–10.

55. ICRU. Tissue substitutes in radiation dosimetry and measurement. ICRU report 44. Bethesta, MD; 1989.

56. Nowotny R. XMuDat: Photon Attenuation Data on PC Version 1.0.1. Vienna, Austria: International Atomic Energy Agency, 1998. Available from: https://www-nds.iaea.org/publications/iaea-nds/iaea-nds-0195.htm, accessed [28 Nov 2019]

57. Boone JM, Chavez AE. Comparison of x-ray cross sections for diagnostic and therapeutic medical physics. Med Phys. 1996;23(12):1997–2005.

58. Rose A. The sensitivity performance of the human eye on an absolute scale. J Opt Soc Am. 1948;38(2):196–208.

59. Wolfe JN. Risk for breast cancer development determined by mammographic parenchymal pattern. Cancer. 1976;37(5):2486–92.

60. Wolfe JN. Breast patterns as an index of risk for developing breast cancer. Am J Roentgenol. 1976;126(6):1130–7.

61. Ng K-H, Lau S. Vision 20/20: Mammographic breast density and its clinical applications. Med Phys. 2015;42(12):7059–77.

62. Gram IT, Funkhouser E, Tabár L. The Tabár classification of mammographic parenchymal patterns. Eur J Radiol. 1997;24(2):131–6.

63. Boyd NF, Byng JW, Jong R a, Fishell EK, Little LE, Miller a B, et al. Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study. J Natl Cancer Inst. 1995;87(9):670–5.

64. The American College of Radiology. ACR BI-RADS® Atlas 3rd edition, Breast Imaging Reporting and Data System. Reston, VA: The American College of Radiology; 1993.

65. Kopans DB. Basic physics and doubts about relationship between mammographically determined tissue density and breast cancer risk. Radiology. 2008;246(2):348–53.

66. The American College of Radiology. ACR BI-RADS® Atlas 4th edition, Breast Imaging Reporting and Data System. Reston, VA: The American College of Radiology; 2003.

67. The American College of Radiology. ACR BI-RADS® Atlas 5th edition, Breast Imaging Reporting and Data System. Reston, VA: The American College of Radiology; 2013.

68. Byng JW, Boyd NF, Fishell E, Jong R a, Yaffe MJ. The quantitative analysis of mammographic densities. Phys Med Biol. 1994;39(10):1629–38.

69. Highnam RP. Model-based enhancement of mammographic images. Oxford, United Kingdom: Oxford University; 1992. Available from: http://www.cs.ox.ac.uk/files/3431/PRG105.pdf

70. Highnam R, Brady M, Shepstone B. A representation for mammographic image processing. Med Image Anal. 1996;1(1):1–18.

71. Highnam R, Brady M. Mammographic Image Analysis. Dordrecht, Netherlands: Springer; 1999.

72. Boone JM, Fewell TR, Jennings RJ. Molybdenum, rhodium, and tungsten anode spectral models using interpolating polynomials with application to mammography.

Med Phys. 1997;24(12):1863–74.

73. Johns PC, Yaffe MJ. X-ray characterisation of normal and neoplastic breast tissues. Phys Med Biol. 1987;32(6):675–95.

74. Blot L, Zwiggelaar R. A volumetric approach to glandularity estimation in mammography: A feasibility study. Phys Med Biol. 2005;50(4):695–708.

75. Pawluczyk O, Augustine BJ, Yaffe MJ, Rico D, Yang J, Mawdsley GE, et al. A volumetric method for estimation of breast density on digitized screen-film mammograms. Med Phys. 2003;30(3):352–64.

76. Highnam R, Pan X, Warren R, Jeffreys M, Davey Smith G, Brady M. Breast composition measurements using retrospective standard mammogram form (SMF). Phys Med Biol. 2006;51(11):2695–713.

77. van Engeland S, Snoeren PR, Huisman H, Boetes C, Karssemeijer N. Volumetric breast density estimation from full-field digital mammograms. IEEE Trans Med Imaging. 2006;25(3):273–82.

78. Hartman K, Highnam R, Warren R, Jackson V. Volumetric Assessmen of Breast Tissue Composition from FFDM Images. In: Krupinski EA. (ed.) Springer, Heidelberg. IWDM 2008. LNCS, vol. 5116:33-9.

79. Highnam R, Brady M, Yaffe M. Robust breast composition measurement-volpara™. In: Marti et al. (Eds.) Springer, Heidelberg. IWDM 2010. LNCS, vol. 6136:342-9.

80. Destounis S, Arieno A, Morgan R, Roberts C, Chan A. Qualitative Versus Quantitative Mammographic Breast Density Assessment: Applications for the US and Abroad. Diagnostics. 2017;7(2):30.

81. Hologic. Understanding Quantra™ 2.0 User Manual. Man-02004 Rev 004. Bedford, Hologic: 2012. Available from: www.hologic.com/sites/default/files/package%20inserts/Understanding%20Quantra%202.0%20User%20Manual.%20English.pdf, accessed [9 May 2017]

82. Arieno A, Chan A, Destounis S V. A review of the role of augmented intelligence in breast imaging: From automated breast density assessment to risk stratification. Am J Roentgenol. 2019;212(2):259–70.

83. Ko ES, Kim RB, Han BK. Reproducibility of automated volumetric breast density assessment in short-term digital mammography reimaging. Clin Imaging. 2015;39(4):582–6.

84. Schmachtenberg C, Hammann-Kloss S, Bick U, Engelken F. Intraindividual comparison of two methods of volumetric breast composition assessment. Acad Radiol. 2015;22(4):447–52.

85. van der Waal D, den Heeten GJ, Pijnappel RM, Schuur KH, Timmers JMH, Verbeek ALM, et al. Comparing Visually Assessed BI-RADS Breast Density and Automated Volumetric Breast Density Software: A Cross-Sectional Study in a Breast Cancer Screening Setting. PLoS One. 2015;10(9):e0136667.

86. Youk JH, Gweon HM, Son EJ, Kim JA. Automated volumetric breast density measurements in the era of the BI-RADS fifth edition: A comparison with visual assessment. Am J Roentgenol. 2016;206(5):1056–62.

87. Alonzo-Proulx O, Packard N, Boone JM, Al-Mayah a, Brock KK, Shen SZ, et al. Validation of a method for measuring the volumetric breast density from digital mammograms. Phys Med Biol. 2010;55(11):3027–44.

88. Wang J, Azziz A, Fan B, Malkov S, Klifa C, Newitt D, et al. Agreement of mammographic measures of volumetric breast density to MRI. PLoS One. 2013;8(12):e81653.

89. Rahbar K, Gubern-Merida A, Patrie JT, Harvey JA. Automated Volumetric Mammographic Breast Density Measurements May Underestimate Percent Breast Density for High-density Breasts. Acad Radiol. 2017;24(12):1561–9.

90. Alonzo-Proulx O, Mawdsley GE, Patrie JT, Yaffe MJ, Harvey JA. Reliability of automated breast density measurements. Radiology. 2015;275(2):366–76

91. Brandt KR, Scott CG, Ma L, Mahmoudzadeh AP, Jensen MR, Whaley DH, et al. Comparison of Clinical and Automated Breast Density Measurements: Implications for Risk Prediction and Supplemental Screening. Radiology. 2016;279(3):710–9.

92. Morrish OWE, Tucker L, Black R, Willsher P, Duffy SW, Gilbert FJ. Mammographic breast density: comparison of methods for quantitative evaluation. Radiology. 2015;275(2):356–65.

93. Chen L, Linden HM, Anderson BO, Li CI. Trends in 5-year survival rates among breast cancer patients by hormone receptor status and stage. Breast Cancer Res Treat. 2014;147(3):609–16.

94. Hofvind S, Tsuruda K, Mangerud G, Ertzaas AK, Holen ÅS, Pedersen K, et al. The Norwegian Breast Cancer Screening Program 1996-2016: Celebrating 20 years of organised mammographic screening. In: Cancer in Norway 2016 – Cancer incidence, mortality, survival and prevalence in Norway. Oslo, Norway: Cancer registry of Norway; 2017. Available from: https://www.kreftregisteret.no/globalassets/cancer-in-norway/2016/mammo_cin2016_special_issue_web.pdf

95. U.S. Preventive Services Task Force. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. Ann Intern Med. 2016;164(4):279–96.

96. European commission initiative on breast cancer [Internet]. Recommendations from European Breast Guidelines; 2019. Available from: http://ecibc.jrc.ec.europa.eu/recommendations/, accessed [28 Nov 2019].

97. Conant EF, Barlow WE, Herschorn SD, Weaver DL, Beaber EF, Tosteson ANA, et al. Association of Digital Breast Tomosynthesis vs Digital Mammography With Cancer Detection and Recall Rates by Age and Breast Density. JAMA Oncol. 2019;5(5):635-42.

98. Tabár L, Vitak B, Chen TH-H, Yen AM-F, Cohen A, Tot T, et al. Swedish Two-County Trial: Impact of Mammographic Screening on Breast Cancer Mortality during 3 Decades. Radiology. 2011;260(3):658–63.

99. Njor S, Nystrom L, Moss S, Paci E, Broeders M, Segnan N, et al. Breast cancer mortality in mammographic screening in Europe: A review of incidence-based mortality studies. J Med Screen. 2012;19(suppl. 1):33–41.

100. Massat NJ, Dibden A, Parmar D, Cuzick J, Sasieni PD, Duffy SW. Impact of screening on breast cancer mortality: The UK program 20 years on. Cancer Epidemiol Biomarkers Prev. 2016;25(3):455–62.

101. Smith RA, Duffy SW, Gabe R, Tabar L, Yen AMF, Chen THH. The randomized trials of breast cancer screening: What have we learned? Radiol Clin North Am. 2004;42(5):793–806.

102. The American College of Surgeons. AJCC Cancer Staging Manual, Eighth Edition. In: Hortobagyi GN, Connolly JL, D'Orsi CJ, Edge SB, Mittendorf EA, Rugo HS, et al., editors. 2017. p. 589–636.

103. International Agency for Research on Cancer. IARC hand- books of cancer prevention. Vol. 15. Breast cancer screening. Lyon, France; 2015. 469 p.

104. Hofvind S, Ponti A, Patnick J, Ascunce N, Njor S, Broeders M, et al. False-positive results in mammographic screening for breast cancer in Europe: A literature review and survey of service screening programmes. J Med Screen. 2012;19(suppl. 1):57–66.

105. Pace LE, Keating NL. A systematic assessment of benefits and risks to guide breast cancer screening decisions. JAMA. 2014;311(13):1327–35.

106. Bond M, Pavey T, Welch K, Cooper C, Garside R, Dean S, et al. Psychological consequences of false-positive screening mammograms in the UK. Evid Based Med. 2013;18(2):54–61.

107. Hakama M, Auvinen A, Day NE, Miller AB. Sensitivity in cancer screening. J Med Screen. 2007;14(4):174–7.

108. Skaane P, Sebuødegård S, Bandos AI, Gur D, Østerås BH, Gullien R, et al. Performance of breast cancer screening using digital breast tomosynthesis: results from the prospective population-based Oslo Tomosynthesis Screening Trial. Breast Cancer Res Treat. 2018;169(3):489–96.

109. Lehman CD, Arao RF, Sprague BL, Lee JM, Buist DSM, Kerlikowske K, et al. National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. Radiology. 2017;283(1):49–58.

110. Hofvind S, Geller BM, Skelly J, Vacek PM. Sensitivity and specificity of mammographic screening as practised in Vermont and Norway. Br J Radiol. 2012;85(1020):e1226-32.

111. Hall EJ, Brenner DJ. Cancer risks from diagnostic radiology. Br J Radiol. 2008;81(965):362–78.

112. Yaffe MJ, Mainprize JG. Risk of radiation-induced breast cancer from mammographic screening. Radiology. 2011;258(1):98–105.

113. Whelehan P, Evans A, Wells M, MacGillivray S. The effect of mammography pain on repeat participation in breast cancer screening: A systematic review. Breast. 2013;22(4):389–94.

114. Byrne C, Schairer C, Wolfe J, Parekh N, Salane M, Brinton LA, et al. Mammographic Features and Breast Cancer Risk: Effects With Time, Age, and Menopause Status. J Natl Cancer Inst. 1995;87(21):1622–9.

115. Ursin G, Ma H, Wu AH, Bernstein L, Salane M, Parisky YR, et al. Mammographic density and breast cancer in three ethnic groups. Cancer Epidemiol Biomarkers Prev. 2003;12(4):332–8.

116. Vacek PM, Geller BM. A prospective study of breast cancer risk using routine mammographic breast density measurements. Cancer EpidemiolBiomarkers Prev. 2004;13(5):715–22.

117. Boyd NF, Martin LJ, Sun L, Guo H, Chiarelli A, Hislop G, et al. Body size, mammographic density, and breast cancer risk. Cancer Epidemiol Biomarkers Prev. 2006;15(11):2086–92.

118. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, et al. Mammographic density and the risk and detection of breast cancer. N Engl J Med. 2007;356(3):227–36.

119. McCormack V a, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiol Biomarkers Prev. 2006;15(6):1159–69.

120. Price ER, Hargreaves J, Lipson JA, Sickles EA, Brenner RJ, Lindfors KK, et al. The california breast density information group: a collaborative response to the issues of breast density, breast cancer risk, and breast density notification legislation. Radiology. 2013;269(3):887–92.

121. Colin C, Schott A-M, Valette P-J. Mammographic density is not a worthwhile examination to distinguish high cancer risk women in screening. Eur Radiol. 2014;24(10):2412–6.

122. Freer PE. Mammographic breast density: impact on breast cancer risk and implications for screening. Radiographics. 2015;35(2):302–15.

123. Sickles EA. The Use of Breast Imaging to Screen Women at High Risk for Cancer.

Radiol Clin North Am. 2010;48(5):859–78.

124. Haas JS, Kaplan CP. The Divide Between Breast Density Notification Laws and Evidence-Based Guidelines for Breast Cancer Screening Legislating Practice. JAMA Intern Med. 2015;02120(9):6–7.

125. Feig SA. Personalized screening for breast cancer: A Wolf in sheep's clothing? Am J Roentgenol. 2015;205(6):1365–71.

126. Slanetz PJ, Freer PE, Birdwell RL. Breast-density legislation--practical considerations. N Engl J Med. 2015;372(7):593–5.

127. Eng A, Gallant Z, Shepherd J, McCormack V, Li J, Dowsett M, et al. Digital mammographic density and breast cancer risk: a case control study of six alternative density assessment methods. Breast Cancer Res. 2014;16(5):439-51.

128. Jeffers AM, Sieh W, Lipson JA, Rothstein JH, McGuire V, Whittemore AS, et al. Breast cancer risk and Mammographic Density assessed with semiautomated and Fully automated Methods and BI-RADS. Radiology. 2017;282(2):348–55.

129. Strand F, Humphreys K, Eriksson M, Li J, Andersson TML, Törnberg S, et al. Longitudinal fluctuation in mammographic percent density differentiates between interval and screen-detected breast cancer. Int J Cancer. 2017;140(1):34–40.

130. Jackson VP, Hendrick RE, Feig S a, Kopans DB. Imaging of the radiographically dense breast. Radiology. 1993;188(2):297–301.

131. van Gils CH, Otten JD, Verbeek AL, Hendriks JH. Mammographic breast density and risk of breast cancer: masking bias or causality? Eur J Epidemiol. 1998;14(4):315–20.

132. Mandelson MT, Oestreicher N, Porter PL, White D, Finder CA, Taplin SH, et al. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. J Natl Cancer Inst. 2000;92(13):1081–7.

133. Buist DSM, Porter PL, Lehman C, Taplin SH, White E. Factors contributing to mammography failure in women aged 40-49 years. J Natl Cancer Inst. 2004;96(19):1432–40.

134. Pinsky RW, Helvie MA. Mammographic breast density: effect on imaging and breast cancer risk. J Natl Compr Canc Netw. 2010;8(10):1157–64.

135. Del Turco MR, Mantellini P, Ciatto S, Bonardi R, Martinelli F, Lazzari B, et al. Full-field digital versus screen-film mammography: Comparative accuracy in concurrent screening cohorts. Am J Roentgenol. 2007;189(4):860–6.

136. Chiarelli AM, Edwards SA, Prummel M V., Muradali D, Majpruz V, Done SJ, et al. Digital Compared with Screen-Film Mammography: Performance Measures in Concurrent Cohorts within an Organized Breast Screening Program. Radiology. 2013;268(3):684–93.

137. Kerlikowske K, Hubbard R a., Miglioretti DL, Geller BM, Yankaskas BC, Lehman CD, et al. Comparative effectiveness of digital versus film-screen mammography in community practice in the United States: A cohort study. Ann Intern Med. 2011;155(8):493–502.

138. Dershaw DD. Status of mammography after the Digital Mammography Imaging Screening Trial: Digital versus film. Breast J. 2006;12(2):99–102.

139. Von Euler-Chelpin M, Lillholm M, Vejborg I, Nielsen M, Lynge E. Sensitivity of screening mammography by density and texture: A cohort study from a population-based screening program in Denmark. Breast Cancer Res. 2019;21(1):1–7.

140. Kerlikowske K, Zhu W, Tosteson ANA, Sprague BL, Tice JA, Lehman CD, et al. Identifying women with dense breasts at high risk for interval cancer a cohort study. Ann Intern Med. 2015;162(10):673–81.

141. Wanders JOP, Holland K, Veldhuis WB, Mann RM, Pijnappel RM, Peeters PHM, et al. Volumetric breast density affects performance of digital screening mammography. Breast Cancer Res Treat. 2017;162(1):95–103.

142. Kriege M, Brekelmans CTM, Boetes C, Besnard PE. Efficacy of MRI and mammography for breast-cancer screening in women with a familial or genetic predisposition. N Engl J Med. 2004;351(5):427–37.

143. Warner E, Plewes DB, Hill KA, Causer. Surveillance of BRCA1 and BRCA2 mutation carriers with magnetic resonance imaging, ultrasound, mammography, and clinical breast examination. JAMA. 2004;292(11):1317–25.

144. Leach MO, Boggis CRM, Dixon a K, Easton DF, Eeles R a, Evans DGR, et al. Screening with magnetic resonance imaging and mammography of a UK population at

high familial risk of breast cancer: a prospective multicentre cohort study (MARIBS). Lancet. 2005;365:1769–78.

145. Kuhl CK, Schrading S, Leutner CC, Morakkabati-Spitz N, Wardelmann E, Fimmers R, et al. Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer. J Clin Oncol. 2005;23(33):8469–76.

146. Sardanelli F, Podo F, D'Agnolo G, Verdecchia A, Santaquilani M, Musumeci R, et al. Multicenter comparative multimodality surveillance of women at genetic-familial high risk for breast cancer (HIBCRIT study): interim results. Radiology. 2007;242(3):698–715.

147. Sardanelli F, Podo F. Breast MR imaging in women at high-risk of breast cancer. Is something changing in early breast cancer detection? Eur Radiol. 2007;17:873–87.

148. Lehman CD, White E, Peacock S, Drucker MJ, Urban N. Effect of age and breast density on screening mammograms with false-positive findings. Am J Roentgenol. 1999;173(6):1651–5.

149. Nelson HD, O'Meara ES, Kerlikowske K, Balch S, Miglioretti D. Factors associated with rates of false-positive and false-negative results from digital mammography screening: An analysis of registry data. Ann Intern Med. 2016;164(4):226–35.

150. Lee CI, Cevik M, Alagoz O, Sprague BL, Tosteson ANA, Miglioretti DL, et al. Comparative effectiveness of combined digital mammography and tomosynthesis screening for women with dense breasts. Radiology. 2015;274(3):772–80.

151. Cappello NM. Decade of "normal" mammography reports--the happygram. J Am Coll Radiol. 2013;10(12):903–8.

152. Are You Dense? [Internet]. 2014. Available from: http://www.areyoudense.org/, accessed on [28 Nov 2019]

153. Schousboe JT, Kerlikowske K, Loh A, Cummings SR. Personalizing mammography by breast density and other risk factors for breast cancer: analysis of health benefits and cost-effectiveness. Ann Intern Med. 2011;155(1):10–20.

154. Skaane P, Bandos AI, Eben EB, Jebsen IN, Krager M, Haakenaasen U, et al. Two-view digital breast tomosynthesis screening with synthetically reconstructed projection

images: comparison with digital breast tomosynthesis with full-field digital mammographic images. Radiology. 2014;271(3):655–63.

155. Østerås BH, Martinsen ACT, Gullien R, Skaane P. Digital Mammography versus Breast Tomosynthesis: Impact of Breast Density on Diagnostic Performance in Population-based Screening. Radiology. 2019;293(1):60–8.

156. Østerås BH, Martinsen ACT, Brandal SHB, Chaudhry KN, Eben E, Haakenaasen U, et al. Classification of fatty and dense breast parenchyma: comparison of automatic volumetric density measurement and radiologists' classification and their inter-observer variation. Acta Radiol. 2016;57(10):1178–85.

157. Gennaro G. The "perfect" reader study. Eur J Radiol. 2018;103:139–46.

158. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, et al. The "laboratory" effect: Comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. Radiology. 2008;249(1):47–53.

159. International Atomic Energy Agency. Dosimetry in Diagnostic Radiology: An International Code of Practice. IAEA, Technical Report Series No. 457. Vienna, Austria: International Atomic Energy Agency; 2007.

160. IAEA. Quality assurance programme for digital mammography. Vienna, Austria: International Atomic Energy Agency; 2011. Available from: http://www-pub.iaea.org/MTCD/Publications/PDF/Pub1482_web.pdf

161. Dance DR. Monte Carlo calculation of conversion factors for the estimation of mean glandular breast dose. Phys Med Biol. 1990;35(9):1211–9.

162. Dance DR, Skinner CL, Young KC, Barrett JF, Kotre CJ. Additional factors for the estimation of mean glandular breast dose using the UK mammography dosimetry protocol. Phys Med Biol. 2000;45(11):3225–40.

163. Dance DR, Young KC, van Engen RE. Further factors for the estimation of mean glandular dose using the United Kingdom, European and IAEA breast dosimetry protocols. Phys Med Biol. 2009;54(14):4361–72.

164. Dance DR, Young KC, van Engen RE. Estimation of mean glandular dose for breast tomosynthesis: factors for use with the UK, European and IAEA breast dosimetry

protocols. Phys Med Biol. 2011;56(2):453–71.

165. Østerås BH, Skaane P, Gullien R, Martinsen ACT. Average glandular dose in paired digital mammography and digital breast tomosynthesis acquisitions in a population based screening program: effects of measuring breast density, air kerma and beam quality. Phys Med Biol. 2018;63(3):035006.

166. Young KC, Ramsdale ML, Bignell F. Review of Dosimetric Methods for Mammography in the UK Breast Screening Programme. Radiat Prot Dosimetry. 1998;80(1):183–6.

167. Skaane P, Bandos AI, Gullien R, Eben EB, Ekseth U, Haakenaasen U, et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. Radiology. 2013;267(1):47–56.

168. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.

169. Wilson EB. Probable Inference, the Law of Succession, and Statistical Inference. J Am Stat Assoc. 1927;22(158):209–12.

170. Altman DG, Machin D, Bryant TN, Gardner MJ. Statistics with confidence. 2nd ed. Altman D, Machin D, Byant T, Gardner M, editors. BMJ Books. 2000. 252 p.

171. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med. 1998;17(8):873–90.

172. Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. Stat Med. 1998;17(22):2635–50.

173. Wagstaff DA, Elek E, Kulis S, Marsiglia F. Using a nonparametric bootstrap to obtain a confidence interval for Pearson's r with cluster randomized data: a case study. J Prim Prev. 2009;30(5):497–512.

174. Field CA, Welsh AH. Bootstrapping clustered data. J R Stat Soc Ser B Stat Methodol. 2007;69(3):369–90.

175. Caumo F, Zorzi M, Brunelli S, Romanucci G, Rella R, Cugola L, et al. Digital Breast Tomosynthesis with Synthesized Two-Dimensional Images versus Full-Field Digital

Mammography for Population Screening: Outcomes from the Verona Screening Program. Radiology. 2018;287(1):37–46.

176. Skaane P. Reply to: 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. European Radiology Opinions. 2016. Available on: https://www.european-radiology.org/opinions/reply-rosengurtt-2016/, acccessed on [28 Nov 2019]

177. Emaus MJ, Bakker MF, Peeters PHM, Loo CE, Mann RM, De Jong MDF, et al. MR imaging as an additional screening modality for the detection of breast cancer in women aged 50-75 years with extremely dense breasts: The DENSE trial study design. Radiology. 2015;277(2):527–37.

178. Buchberger W, Geiger-Gritsch S, Knapp R, Gautsch K, Oberaigner W. Combined screening with mammography and ultrasound in a population-based screening program. Eur J Radiol. 2018;101:24–9.

179. Winkler NS, Raza S, Mackesy M, Birdwell RL. Breast Density: Clinical Implications and Assessment Methods. RadioGraphics. 2015;35:316–24.

180. Bernardi D, Pellegrini M, Di Michele S, Tuttobene P, Fantò C, Valentini M, et al. Interobserver agreement in breast radiological density attribution according to BI-RADS quantitative classification. Radiol Med. 2012;117(4):519–28.

181. Winkel RR, von Euler-Chelpin M, Nielsen M, Diao P, Nielsen MB, Uldall WY, et al. Inter-observer agreement according to three methods of evaluating mammographic density and parenchymal pattern in a case control study: Impact on relative risk of breast cancer. BMC Cancer. 2015;15(1):274–90.

182. Ooms E a, Zonderland HM, Eijkemans MJC, Kriege M, Mahdavian Delavary B, Burger CW, et al. Mammography: interobserver variability in breast density assessment. Breast. 2007;16(6):568–76.

183. Ciatto S, Houssami N, Apruzzese a, Bassetti E, Brancato B, Carozzi F, et al. Categorizing breast mammographic density: intra- and interobserver reproducibility of BI-RADS density categories. Breast. 2005;14(4):269–75.

184. Mazor RD, Savir A, Gheorghiu D, Weinstein Y, Abadi-Korek I, Shabshin N. The inter-

observer variability of breast density scoring between mammography technologists and breast radiologists and its effect on the rate of adjuvant ultrasound. Eur J Radiol. 2016;85(5):957–62.

185. Redondo A, Comas M, Macià F, Ferrer F, Murta-Nascimento C, Maristany MT, et al. Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms. Br J Radiol. 2012;85(1019):1465–70.

186. Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast Imaging Reporting and Data System: inter- and intraobserver variability in feature analysis and final assessment. Am J Roentgenol. 2000;174(6):1769–77.

187. Kerlikowske K, Grady D, Barclay J, Frankel SD, Ominsky SH, Sickles E a, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J Natl Cancer Inst. 1998;90(23):1801–9.

188. Nicholson BT, LoRusso AP, Smolkin M, Bovbjerg VE, Petroni GR, Harvey J a. Accuracy of assigned BI-RADS breast density category definitions. Acad Radiol. 2006;13(9):1143–9.

189. Alikhassi A, Esmaili Gourabi H, Baikpour M. Comparison of inter- and intra-observer variability of breast density assessments using the fourth and fifth editions of Breast Imaging Reporting and Data System. Eur J Radiol Open. 2018;5:67–72.

190. Gard CC, Aiello Bowles EJ, Miglioretti DL, Taplin SH, Rutter CM. Misclassification of breast imaging reporting and data system (BI-RADS) mammographic density and implications for breast density reporting legislation. Breast J. 2015;21(5):481–9.

191. Ekpo EU, Ujong UP, Mello-Thoms C, McEntee MF. Assessment of Interradiologist Agreement Regarding Mammographic Breast Density Classification Using the Fifth Edition of the BI-RADS Atlas. Am J Roentgenol. 2016;206(5):1119–23.

192. Spayne MC, Gard CC, Skelly J, Miglioretti DL, Vacek PM, Geller BM. Reproducibility of BI-RADS breast density measures among community radiologists: a prospective cohort study. Breast J. 2012;18(4):326–33.

193. Irshad A, Leddy R, Ackerman S, Cluver A, Pavic D, Abid A, et al. Effects of changes in BI-RADS density assessment guidelines (fourth versus fifth edition) on breast

density assessment: Intra-and interreader agreements and density distribution. Am J Roentgenol. 2016;207(6):1366–71.

194. Harvey JA, Gard CC, Miglioretti DL, Yankaskas BC, Kerlikowske K, Buist DSM, et al. Reported mammographic density: film-screen versus digital acquisition. Radiology. 2013;266(3):752–8.

195. Youk JH, Kim SJ, Son EJ, Gweon HM, Kim JA. Comparison of visual assessment of breast density in BI-RADS 4th and 5th editions with automated volumetric measurement. Am J Roentgenol. 2017;209(3):703–8.

196. Berg WA, D'Orsi CJ, Jackson VP, Bassett LW, Beam CA, Lewis RS, et al. Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography? Radiology. 2002;224(3):871–80.

197. Alomaim W, O'Leary D, Ryan J, Rainford L, Evanoff M, Foley S. Variability of Breast Density Classification Between US and UK Radiologists. J Med imaging Radiat Sci. 2019;50(1):53–61.

198. Damases CN, Hogg P, McEntee MF. Intercountry analysis of breast density classification using visual grading. Br J Radiol. 2017;90(1076):1–10.

199. Garrido-Estepa M, Ruiz-Perales F, Miranda J, Ascunce N, González-Román I, Sánchez-Contador C, et al. Evaluation of mammographic density patterns: reproducibility and concordance among scales. BMC Cancer. 2010;10(1):485.

200. Ciatto S, Bernardi D, Calabrese M, Durando M, Gentilini MA, Mariscotti G, et al. A first evaluation of breast radiological density assessment by QUANTRA software as compared to visual classification. Breast. 2012;21(4):503–6.

201. Melnikow J, Fenton JJ, Whitlock EP, Miglioretti DL, Weyrich MS, Thompson JH, et al. Supplemental Screening for Breast Cancer in Women With Dense Breasts: A Systematic Review for the U.S. Preventive Services Task Force. Ann Intern Med. 2016;164(4):268–78.

202. Sprague BL, Gangnon RE, Burt V, Trentham-Dietz A, Hampton JM, Wellman RD, et al. Prevalence of mammographically dense breasts in the United States. J Natl Cancer Inst. 2014;106(10).

203. Conant EF, Sprague BL, Kontos D. Beyond BI-RADS Density: A Call for Quantification in the Breast Imaging Clinic. Radiology. 2018;286(2):401–4.

204. Ekpo EU, Hogg P, Highnam R, McEntee MF. Breast composition: Measurement and clinical use. Radiography. 2015;21(4):324–33.

205. Irshad A, Leddy R, Lewis M, Cluver A, Ackerman S, Pavic D, et al. Changes in breast density reporting patterns of radiologists after publication of the 5th edition BI-RADS guidelines: A single institution experience. Am J Roentgenol. 2017;209(4):943–8.

206. Sprague BL, Kerlikowske K, Bowles EJA, Rauscher GH, Lee CI, Tosteson ANA, et al. Trends in Clinical Breast Density Assessment From the Breast Cancer Surveillance Consortium. J Natl Cancer Inst. 2019;111:1–4.

207. Østerås BH, Martinsen ACT, Brandal SHB, Chaudhry KN, Eben E, Haakenaasen U, et al. BI-RADS Density Classification From Areometric and Volumetric Automatic Breast Density Measurements. Acad Radiol. 2016;23(4):468–78

208. Gastounioti A, McCarthy AM, Pantalone L, Synnestvedt M, Kontos D, Conant EF. Effect of Mammographic Screening Modality on Breast Density Assessment: Digital Mammography versus Digital Breast Tomosynthesis. Radiology. 2019;291(2):320–7.

209. Aujero MP, Gavenonis SC, Benjamin R, Zhang Z, Holt JS. Clinical Performance of Synthesized Two-dimensional Mammography Combined with Tomosynthesis in a Large Screening Population. Radiology. 2017;283(1):70–6.

210. Zuckerman SP, Conant EF, Keller BM, Maidment ADA, Barufaldi B, Weinstein SP, et al. Implementation of Synthesized Two-dimensional Mammography in a Population-based Digital Breast Tomosynthesis Screening Program. Radiology. 2017;281(3):730–6.

211. Tagliafico AS, Tagliafico G, Cavagnetto F, Calabrese M, Houssami N. Estimation of percentage breast tissue density: comparison between digital mammography (2D full field digital mammography) and digital breast tomosynthesis according to different BI-RADS categories. Br J Radiol. 2013;86(1031):20130255.

212. Alshafeiy TI, Wadih A, Nicholson BT, Rochman CM, Peppard HR, Patrie JT, et al. Comparison between digital and synthetic 2D mammograms in breast density interpretation. Am J Roentgenol. 2017;209(1):W36–41.

213. Haider I, Morgan M, McGow A, Stein M, Rezvani M, Freer P, et al. Comparison of Breast Density Between Synthesized Versus Standard Digital Mammography. J Am Coll Radiol. 2018;15(10):1430–6.

214. The American College of Radiology. ACR Statement on Reporting Breast Density in Mammography Reports and Patient Summaries, 2017. Available from: https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Reporting-Breast-Density accessed on, accessed on [28 Nov 2019]

215. Singh JM, Fallenberg EM, Diekmann F, Renz DM, Witlandt R, Bick U, et al. Volumetric breast density assessment: Reproducibility in serial examinations and comparison with visual assessment. RoFo. 2013;185(9):844–8.

216. Engelken F, Singh J-M, Fallenberg E-M, Bick U, Böttcher J, Renz DM. Volumetric breast composition analysis: reproducibility of breast percent density and fibroglandular tissue volume measurements in serial mammograms. Acta Radiol. 2014;55(1):32–8.

217. Holland K, van Zelst J, den Heeten GJ, Imhof-Tas M, Mann RM, van Gils CH, et al. Consistency of breast density categories in serial screening mammograms: A comparison between automated and human assessment. Breast. 2016;29:49–54.

218. Ekpo EU, McEntee MF, Rickard M, Brennan PC, Kunduri J, Demchig D, et al. Quantra™ should be considered a tool for two-grade scale mammographic breast density classification. Br J Radiol. 2016;89(1060).

219. Ko SY, Kim E-K, Kim MJ, Moon HJ. Mammographic density estimation with automated volumetric breast density measurement. Korean J Radiol. 2014;15(3):313–21.

220. Kallenberg MGJ, van Gils CH, Lokate M, den Heeten GJ, Karssemeijer N. Effect of compression paddle tilt correction on volumetric breast density estimation. Phys Med Biol. 2012;57(16):5155–68.

221. Bakic PR, Carton AK, Kontos D, Zhang C, Troxel AB, Maidment ADA. Breast percent density: Estimation on digital mammograms and central tomosynthesis projections. Radiology. 2009;252(1):40–9.

222. Jing H, Keller B, Choi JY, Crescenzi R, Conant E, Maidment A, et al. Dependence of

radiation dose on area and volumetric mammographic breast density estimation. In: Nishikawa RM, Whiting BR, eds. Proc SPIE 8668, Medical Imaging 2013. 866827.

223. Chen L, Ray S, Keller BM, Pertuz S, McDonald ES, Conant EF, et al. The impact of acquisition dose on quantitative breast density estimation with digital mammography: Results from ACRIN PA 4006. Radiology. 2016;280(3):693–700.

224. Heine JJ, Cao K, Beam C. Cumulative sum quality control for calibrated breast density measurements. Med Phys. 2009;36(12):5380–90.

225. Ren B, Smith A, Jing Z. Measurement of breast density with digital breast tomosynthesis. In: Pelc NJ, Nishikawa RM, Whiting BR, eds. Proc SPIE 8313, Medical Imaging 2012. 83134Q.

226. Ekpo EU, McEntee MF. Measurement of breast density with digital breast tomosynthesis-a systematic review. Br J Radiol. 2014;87(1043):1–9.

227. Ekpo EU, Mello-Thoms C, Rickard M, Brennan PC, McEntee MF. Breast density (BD) assessment with digital breast tomosynthesis (DBT): Agreement between Quantra™ and 5th edition BI-RADS(®). Breast. 2016;30:185–90.

228. Pertuz S, McDonald ES, Weinstein SP, Conant EF, Kontos D. Fully Automated Quantitaive Estimation of Volumetric Breast Density from Digital Breast Tomosynthesis Images : Preliminary Results and Comparison with Digital Mammography and MR Imaging. Radiology. 2016;279(1):65-74.

229. 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(1):330–6.

230. Conant EF, McDonald ES, Gastounioti A, Keller BM, Pantalone L, Kontos D. Agreement between Breast Percentage Density Estimations from Standard-Dose versus Synthetic Digital Mammograms: Results from a Large Screening Cohort Using Automated Measures. Radiology. 2017;283(3):673–80.

231. Strand F, Azavedo E, Hellgren R, Humphreys K, Eriksson M, Shepherd J, et al. Localized mammographic density is associated with interval cancer and large breast cancer: A nested case-control study. Breast Cancer Res. 2019;21(1):1–9.

232. Shah DJ, Sachs RK, Wilson DJ. Radiation-induced cancer: A modern view. Br J

Radiol. 2012;85(1020):1166–73.

233. ICRP, 2007. The 2007 Recommendations of the International Commission on Radiological Protection. ICRP publication 103. Ann ICRP. 2007;37(2–4):1–332.

234. Tromans CE, Highnam R, Morrish O, Black R, Tucker L, Gilbert F, et al. Patient specific dose calculation using volumetric breast density for mammography and tomosynthesis. In: Fujita H., Hara T., Muramatsu C. (eds.) Springer, Heidelberg. IWDM 2014. LNCS, vol. 8539:158-65.

235. Sechopoulos I, Bliznakova K, Qin X, Fei B, Feng SSJ. Characterization of the homogeneous tissue mixture approximation in breast imaging dosimetry. Med Phys. 2012;39(8):5050.

236. Hernandez AM, Seibert JA, Boone JM. Breast dose in mammography is about 30% lower when realistic heterogeneous glandular distributions are considered. Med Phys Med Phys. 2015;42(101):6337–48.

237. Dance DR, Sechopoulos I. Dosimetry in x-ray-based breast imaging. Phys Med Biol. 2016;61(19):R271–304.

238. 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. European Commision Guidelines. Brussels, Belgium; 2006.

239. Ren B, Smith AP, Jing Z. Local versus whole breast volumetric breast density assessments and implications. In: Maidment A.D.A., Bakic P.R., Gavenonis S. (eds.) Springer, Heidelberg. IWDM 2012. LNCS, vol. 7361:775–82.

240. Aase HS, Holen ÅS, Pedersen K, Houssami N, Haldorsen IS, Sebuødegård S, et al. A randomized controlled trial of digital breast tomosynthesis versus digital mammography in population-based screening in Bergen: interim analysis of performance indicators from the To-Be trial. Eur Radiol. 2019;29(3):1175–86.

241. Olgar T, Kahn T, Gosch D. Average glandular dose in digital mammography and breast tomosynthesis. RoFo. 2012;184(10):911–8.

242. Cavagnetto F, Taccini G, Rosasco R, Bampi R, Calabrese M, Tagliafico A. “In vivo” average glandular dose evaluation: one-to-one comparison between digital breast tomosynthesis and full-field digital mammography. Radiat Prot Dosimetry.

2013;157(1):53–61.

243. Shin SU, Chang JM, Bae MS, Lee SH, Cho N, Seo M, et al. Comparative evaluation of average glandular dose and breast cancer detection between single-view digital breast tomosynthesis (DBT) plus single-view digital mammography (DM) and two-view DM: correlation with breast thickness and density. Eur Radiol. 2015;25(1):1–8.

244. Bouwman RW, van Engen RE, Young KC, den Heeten GJ, Broeders MJM, Schopphoven S, et al. Average glandular dose in digital mammography and digital breast tomosynthesis: comparison of phantom and patient data. Phys Med Biol. 2015;60(20):7893–907.

245. Castillo-García M, Chevalier M, Garayoa J, Rodriguez-Ruiz A, García-Pinto D, Valverde J. Automated Breast Density Computation in Digital Mammography and Digital Breast Tomosynthesis: Influence on Mean Glandular Dose and BIRADS Density Categorization. Acad Radiol. 2017;24(7):802–10.

246. Gennaro G, Bernardi D, Houssami N. Radiation dose with digital breast tomosynthesis compared to digital mammography: per-view analysis. Eur Radiol. 2017;28(2):573-81.

247. Brown M, Covington MF. Comparative Benefit-to–Radiation Risk Ratio of Molecular Breast Imaging, Two-Dimensional Full-Field Digital Mammography with and without Tomosynthesis, and Synthetic Mammography with Tomosynthesis. Radiol Imaging Cancer. 2019;1(1):e190005.

248. 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. Eur Radiol. 2016;26(1):184–90.

249. Rafferty EA, Durand MA, Conant EF, Copit DS, Friedewald SM, Plecha DM, et al. Breast Cancer Screening Using Tomosynthesis and Digital Mammography in Dense and Nondense Breasts. JAMA. 2016;315(16):1784–6.

250. Conant EF, Beaber EF, Sprague BL, Herschorn SD, Weaver DL, Onega T, et al. Breast cancer screening using tomosynthesis in combination with digital mammography compared to digital mammography alone: a cohort study within the PROSPR consortium. Breast Cancer Res Treat. 2016;156(1):109–16.

251. McCarthy AM, Kontos D, Synnestvedt M, Tan KS, Heitjan DF, Schnall M, et al. Screening outcomes following implementation of digital breast tomosynthesis in a general-population screening program. J Natl Cancer Inst. 2014;106(11).

252. Yun SJ, Ryu CW, Rhee SJ, Ryu JK, Oh JY. Benefit of adding digital breast tomosynthesis to digital mammography for breast cancer screening focused on cancer characteristics: a meta-analysis. Breast Cancer Res Treat. 2017;164(3):557–69.

253. Chen L, Abbey CK, Nosrateih A, Lindfors KK, Boone JM. Anatomical complexity in breast parenchyma and its implications for optimal breast imaging strategies. Med Phys. 2012;39(3):1435–41.

254. García-Barquín P, Páramo M, Elizalde A, Pina L, Etxano J, Fernandez-Montero A, et al. The effect of the amount of peritumoral adipose tissue in the detection of additional tumors with digital breast tomosynthesis and ultrasound. Acta radiol. 2017;58(6):645–51.

255. Lee SH, Jang MJ, Kim SM, Yun B La, Rim J, Chang JM, et al. Factors affecting breast cancer detectability on digital breast tomosynthesis and two-dimensional digital mammography in patients with dense breasts. Korean J Radiol. 2019;20(1):58–68.

256. Starikov A, Drotman M, Hentel K, Katzen J, Min RJ, Arleo EK. 2D mammography, digital breast tomosynthesis, and ultrasound: Which should be used for the different breast densities in breast cancer screening? Clin Imaging. 2015;40(1):68–71.

257. Haas BM, Kalra V, Geisel J, Raghu M, Durand M, Philpotts LE. Comparison of Tomosynthesis Plus Digital Mammography and Digital Mammography Alone for Breast Cancer Screening. Radiology. 2013;269(3):694–700.

258. Sharpe RE, Venkataraman S, Phillips J, Dialani V, Fein-Zachary VJ, Prakash S, et al. Increased Cancer Detection Rate and Variations in the Recall Rate Resulting from Implementation of 3D Digital Breast Tomosynthesis into a Population-based Screening Program. Radiology. 2016;278(3):698–706.

259. Rose SL, Tidwell AL, Bujnoch LJ, Kushwaha AC, Nordmann AS, Sexton R. Implementation of breast tomosynthesis in a routine screening practice: An observational study. Am J Roentgenol. 2013;200(6):1401–8.

260. Alsheik NH, Dabbous F, Pohlman SK, Troeger KM, Gliklich RE, Donadio GM, et al.

Comparison of Resource Utilization and Clinical Outcomes Following Screening with Digital Breast Tomosynthesis Versus Digital Mammography: Findings From a Learning Health System. Acad Radiol. 2018;60068:1–9.

261. 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. Eur Radiol. 2016;26(11):3899–907.

262. Bernardi D, Caumo F, Macaskill P, Ciatto S, Pellegrini M, Brunelli S, et al. Effect of integrating 3D-mammography (digital breast tomosynthesis) with 2D-mammography on radiologists' true-positive and false-positive detection in a population breast screening trial. Eur J Cancer. 2014;50(7):1232–8.

263. 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. Am J Roentgenol. 2017;208(1):222–7.

264. ECOG-ACRIN. TMIST Breast Cancer Screening Trial [Internet] 2019. Available from: https://ecog-acrin.org/tmist, accessed on [28 Nov 2019]

265. Bernardi D, Gentilini MA, De Nisi M, Pellegrini M, Fantò C, Valentini M, et al. Effect of implementing digital breast tomosynthesis (DBT) instead of mammography on population screening outcomes including interval cancer rates: Results of the Trento DBT pilot evaluation. The Breast. 2019:10–5.

266. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Teuwen J, Broeders M, Gennaro G, 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(9):4825–32.

267. Berggren K. Spectral image quality and applications in breast tomosynthesis [dissertation]. Stockholm, Sweden: Royal Institute of Technology; 2018. Available from: http://kth.diva-portal.org/smash/record.jsf?pid=diva2%3A1209190&dswid=6988

Papers I-IV

IV

ORIGINAL RESEARCH • BREAST IMAGING

Radiology

Digital Mammography versus Breast Tomosynthesis: Impact of Breast Density on Diagnostic Performance in Population-based Screening

Bjørn Helge Østerås, MSc • Anne Catrine T. Martinsen, PhD • Randi Gullien, MSc • Per Skaane, MD, PhD

From the Department of Diagnostic Physics (B.H.Ø., A.C.T.M.) and Division of Radiology and Nuclear Medicine (R.G., P.S.), Oslo University Hospital, Building 20, Gaustad, PO Box 4959, Nydalen, 0424 Oslo, Norway; and Institute of Clinical Medicine (B.H.Ø., P.S.) and Department of Physics (A.C.T.M.), University of Oslo, Oslo, Norway. Received March 1, 2019; revision requested April 15; final revision received June 4; accepted June 14. **Address correspondence to** B.H.Ø. (e-mail: *bjorn.helge.osteras@ous-hf.no*).

P.S. received equipment and funding for additional case interpretations related to the Oslo Tomosynthesis Screening Trial from Hologic.

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

See also the editorial by Fuchsjäger and Adelsmayr in this issue.

Radiology 2019; 293:60–68 • https://doi.org/10.1148/radiol.2019190425 • Content code: BR

Background: Previous studies comparing digital breast tomosynthesis (DBT) to digital mammography (DM) have shown conflicting results regarding breast density and diagnostic performance.

Purpose: To compare true-positive and false-positive interpretations in DM versus DBT according to volumetric density, age, and mammographic findings.

Materials and Methods: From November 2010 to December 2012, 24 301 women aged 50–69 years (mean age, 59.1 years ± 5.7) were prospectively included in the Oslo Tomosynthesis Screening Trial. Participants received same-compression DM and DBT with independent double reading for both DM and DM plus DBT reading modes. Eight experienced radiologists rated the images by using a five-point scale for probability of malignancy. Participants were followed up for 2 years to assess for interval cancers. Breast density was assessed by using automatic volumetric software (scale, 1–4). Differences in true-positive rates, false-positive rates, and mammographic findings were assessed by using confidence intervals (Newcombe paired method) and *P* values (McNemar and χ^2 tests).

Results: The true-positive rate of DBT was higher than that of DM for density groups (range, 12%–24%; $P < .001$ for density scores of 2 and 3, and $P > .05$ for density scores of 1 and 4) and age groups (range, 15%–35%; $P < .05$ for all age groups), mainly due to the higher number of spiculated masses and architectural distortions found at DBT ($P < .001$ for density scores of 2 and 3; $P < .05$ for women aged 55–69 years). The false-positive rate was lower for DBT than for DM in all age groups (range, −0.6% to −1.2%; $P < .01$) and density groups (range, −0.7 to −1.0%; $P < .005$) owing to fewer asymmetric densities ($P \leq .001$), except for extremely dense breasts (0.1%, $P = .82$).

Conclusion: Digital breast tomosynthesis enabled the detection of more cancers in all density and age groups compared with digital mammography, especially cancers classified as spiculated masses and architectural distortions. The improvement in cancer detection rate showed a positive correlation with age. With use of digital breast tomosynthesis, false-positive findings were lower due to fewer asymmetric densities, except in extremely dense breasts.

Online supplemental material is available for this article.

The sensitivity of digital mammography (DM) is lower in women with dense breasts than in those with lower breast density (1). Breast density is also associated with higher false-positive rates and recall rates (2) due to superposition of normal glandular tissue that can mimic cancer. The woman's age has an impact on mammography screening as breast density decreases (3) and cancer incidence increases. The distribution of cancers shifts toward less-aggressive slower-growing cancers with increasing age (4). It has been shown that mammography screening has a lower sensitivity (1) and higher false-positive rate (2) among younger women.

Digital breast tomosynthesis (DBT) generates pseudo three-dimensional (3D) images where a single section of anatomy is in focus. The rest is blurred, with greater magnitude proportional to the distance from the focus plane. The screening performance of DBT for specific density and age groups may be different from that of DM, as DBT potentially can reduce masking and resolve superposition of breast tissue. Prospective (5–11) and retrospective (12–18) studies have shown that the integration of DBT improves the cancer detection or recall rates for both fatty and dense breasts and in age groups relevant for mammography screening. Data are limited in almost entirely fatty and extremely dense breasts. Two large studies compared DBT and DM in women with extremely dense breasts, with one study finding an increased cancer detection rate with DBT (5) and the other finding similar rates for DBT and DM (13). Therefore, there is a need for more data from large prospective trials.

Abbreviations

BI-RADS = Breast Imaging Reporting and Data System, CI = confidence interval, DBT = digital breast tomosynthesis, DM = digital mammography

Summary

For digital breast tomosynthesis compared with digital mammography, true-positive rates were higher and false-positive rates were lower for all volumetric breast density categories (except for extremely dense breasts) and age groups (ages 50–69 years).

Key Results

- The true-positive rate with digital breast tomosynthesis (DBT) was higher than that with digital mammography (DM) in all volumetric density groups (range, 12%–24%; $P < .001$ in women with scattered fibroglandular and heterogeneously dense breasts; $P > .05$ in women with almost entirely fatty and extremely dense breasts) and all age groups (range, 15%–35%; $P < .05$).
- The false-positive rate with DBT was lower than that with DM in all age groups (range, −0.6% to −1.2%; $P < .01$) and volumetric density groups (range, −0.7% to −1.0%; $P < .005$), except for women with extremely dense breasts (0.1%, $P = .82$).
- DBT showed a greater number of true-positive findings classified as spiculated masses or architectural distortions ($P < .001$ in women with scattered fibroglandular and heterogeneously dense breasts, $P < .05$ in women aged 55–69 years) and a reduction of false-positive findings classified as asymmetric densities ($P < .001$, except in women with extremely dense breasts).

Radiologists usually classify breasts into one of four Breast Imaging Reporting and Data System (BI-RADS) density categories (19). This method has considerable interobserver variability (20,21). Commercial software for automatic density classification has recently become available. Such software uses image processing and a physical model of the breast to calculate the woman's breast density objectively (22), thereby facilitating reproducible breast density stratification in the mammography screening.

The paired design of the prospective Oslo Tomosynthesis Screening Trial facilitates comparison of true- and false-positive interpretation between DM and DBT. Previous analysis showed an improvement in true- and false-positive rates with DBT (23). The benefit across density and age groups has not previously been analyzed in this cohort.

The aim of this study was to compare true- and false-positive interpretations in DM and DBT in prospective population-based screening according to volumetric density, age, and mammographic findings.

Materials and Methods

This prospective clinical trial (ClinicalTrials.gov NCT01248546) was approved by the regional ethics committee (reference number: 2010/144). Written informed consent was obtained from all participants. Hologic (Bedford, Mass) sponsored this study by providing equipment and financial support for additional readings. The authors had control of data and information submitted for publication. Five reports have been published on the Oslo Tomosynthesis Screening Trial, including two interim analyses comparing DM and DBT (n = 12631) (24,25). After inclusion of all women, a study comparing two versions of synthetic DM (n = 24901, some women were imaged twice) was published (26). After 2 years of follow-up to assess interval cancers, the sensitivity and specificity of DM plus DBT were compared with those from previous DM screening rounds (n = 24301) (27) and for all screening arms (23). None of these reports have stratified results according to breast density or age, which is the goal of this preplanned analysis.

Study Cohort

From November 2010 to December 2012, 24301 women (48451 breasts) imaged at one breast center were included in the Oslo Tomosynthesis Screening Trial (Fig 1). The mean participant age ± standard deviation was 59.1 years ± 5.7. Women with pacemakers, women unable to stand, and women with breast implants were not included. The selection of women was solely based on the availability of radiographers and imaging systems. Recruitment was part of the population-based screening program BreastScreen Norway, which invites women aged 50–69 years to undergo biennial two-view screening. Eight experienced radiologists (including P.S.) with 2–31 years of experience in screening mammography (average, 16 years) participated in the trial. Before the trial, each radiologist received intensive personalized training with a set of 100 cancer-enriched cases. Details regarding study population and radiologist training are reported elsewhere (23,26).

Mammographic Imaging

Imaging was performed with three mammography systems (Selenia Dimensions, Hologic) by using the "combo" mode (single breast compression for DM and DBT). Craniocaudal and mediolateral oblique projections were acquired of both breasts. A standard screening setting (auto filter) was used, resulting in a mean average glandular radiation dose of 1.74 and 2.10 mGy for DM and DBT views, respectively (28).

Image Evaluation

Radiologists categorized their findings by using the BreastScreen Norway scale for the probability of cancer, as follows: 1, negative or definitely benign; 2, probably benign; 3, indeterminate; 4, probably malignant; and 5, malignant. A score of 2 or higher was classified as a positive mammographic finding, and these examinations were discussed at a consensus meeting (Fig 1). Four readers independently interpreted four study arms: two DM arms and two DM plus DBT arms. The workstation for each arm was in different rooms, and the patient's score for the respective arm was locked after closing the reading session. More details regarding study arms are reported elsewhere (23,27). Scores from DM arms were combined into a single score: two-dimensional (2D) double reading. If at least one DM arm had a positive score, 2D double reading was considered positive. Similarly, scores from DM plus DBT arms were combined into double reading 2D plus 3D. For positive scores, the radiologist classified the mammographic finding as mass (round, oval, irregular), spiculated mass, architectural distortion, asymmetric density, calcification, or calcification with density. All screening-detected cancers were classified at consensus.

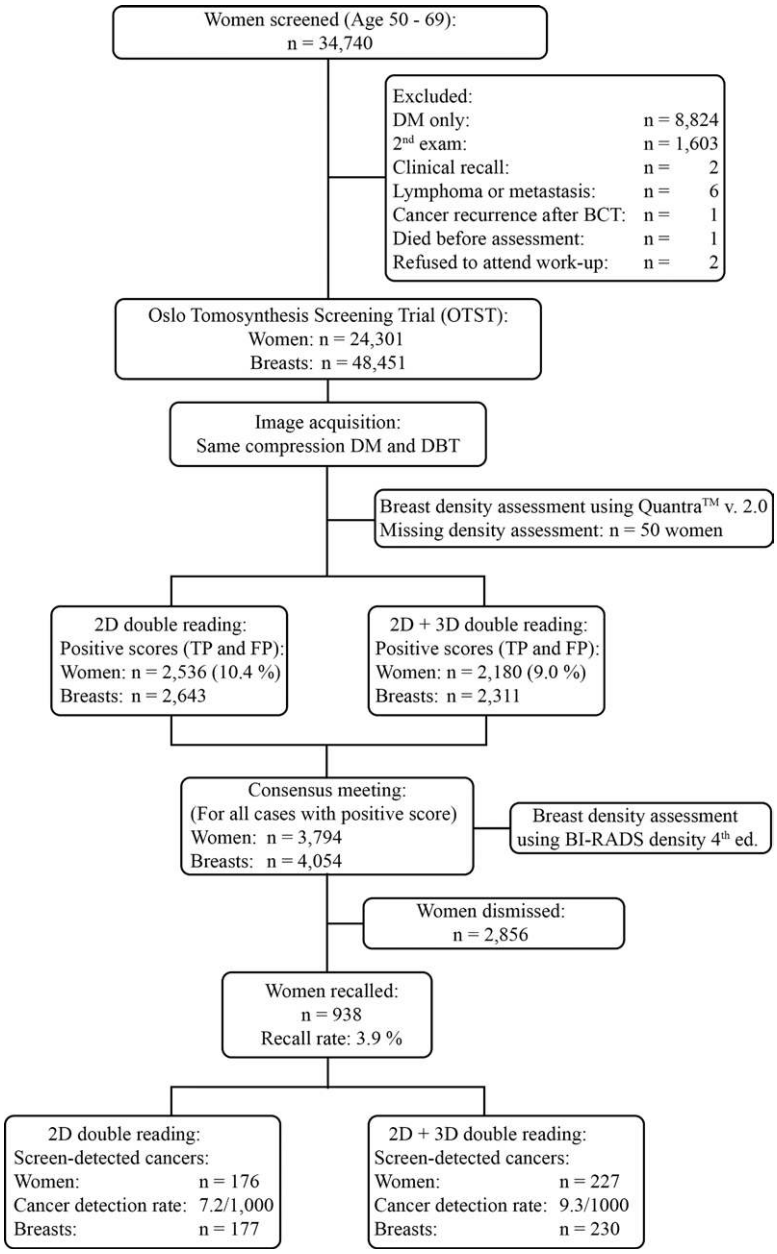

Figure 1: Flowchart of study population shows overall recall rate and cancer detection rate for independent two-dimensional (2D) and 2D plus three-dimensional (3D) double reading. BCT = breast-conserving therapy, BI-RADS = Breast Imaging Reporting and Data System, DBT = digital breast tomosynthesis, DM = digital mammography, FP = false positive, OTST = Oslo Tomosynthesis Screening Trial, TP = true positive, 2D = study arms using DM only, 2D+3D = study arms using DM plus DBT.

Breast Density Assessment

Volumetric breast density was calculated automatically by using commercial software (Quantra, version 2.0; Hologic). This information was not shown to readers (Fig 1). The software uses the raw DM image, physical model of the radiographic imaging chain, and attenuation in adipose and fibroglandular tissue to estimate volumetric breast density (ratio of fibroglandular to total breast volume) (22). Results for each view are aggregated into woman-based scores. Volumetric density was mapped to a quantized density score that was similar to BI-RADS 4th edition density scores, as follows: 1, almost entirely fatty; 2, scattered fibroglandular densities; 3, heterogeneously dense; and 4, extremely dense.

At the consensus meetings at least two or three of the eight participating radiologists assessed breast density in consensus according to BI-RADS 4th edition density (Fig 1) (19).

Statistical Analysis

Differences in distributions of breast density were assessed by using the χ^2 test (tabulate and chi2 commands, Stata, version 15.1; StataCorp, College Station, Tex). Agreement in density assessment between volumetric and BI-RADS density was assessed by using κ statistics with quadratic weights (kap, wgt [w2] commands; Stata) and Spearman correlation coefficients (Spearman command; Stata). The 95% confidence intervals (CIs) were estimated by using bootstrapping with 10 000 replacements (bootstrap command; Stata).

The 95% CIs in differences in proportions were estimated by using the Newcombe method for paired proportions (the Newcombe method for unpaired proportions was used as a conservative estimate where one modality found all screening-detected cancers).

The McNemar test (mcci command; Stata) was used when comparing differences in true- and false-positive rates for 2D and 2D plus 3D. For all analyses, $P < .05$ was considered indicative of a statistically significant difference.

We calculated the true-positive rate difference for 2D plus 3D and 2D with respect to age by using linear regression (regstat command; Matlab, Natick, Mass). Associated 95% CIs were calculated by using bootstrapping with 10 000 replacements (Matlab).

Differences in proportions of mammographic finding classifications in 2D and 2D plus 3D were evaluated by using the χ^2 test.

Results

Automatically Calculated Breast Density Distributions

With use of volumetric density, 65% of women (15 785 of 24 251) were considered to have nondense breasts (density 1 and 2) and 35% (8466 of 24 251) were considered to have dense breasts (density 3 and 4). The density distribution was different in all age groups ($P < .001$), shifting toward lower breast density with age (Table 1). The density distribution was different in women with positive scores ($P < .001$) and screening-detected cancers ($P = .002$) compared with all women. In addition, the density distribution was different between screening-detected and interval cancers when density was measured with use of volumetric (quantized) density ($P = .03$).

Table 1: Volumetric Breast Density according to Age, Positive Mammographic Score, and Screening-detected and Interval Cancers

Subgroup	No. of Women	No. of Women with Missing Density Measurements*	Volumetric Density Grade†			
			1	2	3	4
All included women	24301	50	11.8 (2863/24251)	53.3 (12 922/24251)	27.4 (6645/24251)	7.5 (1821/24251)
Age 50–54 y	6508	10	6.2 (401/6498)	45.2 (2939/6498)	35.7 (2318/6498)	12.9 (840/6498)
Age 55–59 y	6693	14	11.2 (751/6679)	53.0 (3538/6679)	28.7 (1917/6679)	7.1 (473/6679)
Age 60–64 y	5578	10	15.0 (837/5568)	56.7 (3157/5568)	22.9 (1275/5568)	5.4 (299/5568)
Age 65–69 y	5522	16	15.9 (874/5506)	59.7 (3288/5506)	20.6 (1135/5506)	3.8 (209/5506)
Women with positive mammographic score	3794	4	7.8 (295/3790)	49.4 (1871/3790)	33.3 (1262/3790)	9.6 (362/3790)
Women with screening-detected cancers‡	230	1	7.0 (16/229)	46.7 (107/229)	36.2 (83/229)	10.0 (23/229)
Women with interval cancers§	51	0	0.0 (0/51)	33 (17/51)	49 (25/51)	18 (9/51)

* Density measurement was missing for one woman with bilateral screening-detected cancer.

† Data are percentages, with raw data in parentheses. Volumetric density was obtained with software (Quantra, Hologic).

‡ Four screening detected cancers was bilateral.

§ One interval cancer was bilateral.

Table 2: Correlation between BI-RADS and Volumetric Density

BI-RADS Density	Volumetric Density			
	1	2	3	4
I	115	189	6	2
II	177	1305	246	11
III	3	377	918	175
IV	0	0	92	174

Note.—Data are numbers of women. Breast Imaging Reporting and Data System (BI-RADS) density was obtained with BI-RADS 4th edition (19). Volumetric ("quantized") density was obtained with software (Quantra, Hologic). Interobserver agreement between BI-RADS density and volumetric density was substantial (κ = 0.69 [95% confidence interval: 0.67, 0.71]; Spearman correlation coefficient: 0.70 [95% confidence interval: 0.68, 0.72]).

Table 2 shows the Spearman correlation and interobserver agreement between volumetric density and BI-RADS density for women with positive mammographic scores (n = 3790). The interobserver agreement (κ value) was substantial with correlation ($P < .001$) among the two measures of breast density assessment.

True- and False-Positive Interpretations according to Density

Radiologists detected more cancers using 2D plus 3D compared with 2D double reading for all breast densities (Table 3). The breast-based true-positive rate was higher by 13% (two of 15; $P = .50$) for almost entirely fatty breasts, 33% (26 of 79; $P < .001$) for breasts with scattered fibroglandular densities, 30% (19 of 64; $P < .001$) for heterogeneously dense breasts, and 28% (five of 15; $P = .06$) for extremely dense breasts. The 95% CIs for the difference in true-positive rate between 2D plus 3D and 2D overlap for all density categories.

Table E1 (online) shows true-positive interpretations stratified according to BI-RADS density, with similar results as stratification with volumetric ("quantized") density.

Radiologists reported fewer false-positive scores using 2D plus 3D compared with 2D double reading for women with all densities, except those with extremely dense breasts. The breast-based false-positive rate was lower by 23% (45 of 197; $P = .004$) for almost entirely fatty breasts, 21% (252 of 1224; $P < .001$) for breasts with scattered fibroglandular densities, and 12% (94 of 815; $P = .004$) for heterogeneously dense breasts. The breast-based false-positive rate was higher by 2% (five of 229; $P = .82$) for extremely dense breasts.

True- and False-Positive Interpretations according to Age

The number of true-positive scores was higher for all age strata for 2D plus 3D compared with 2D double reading (Table 3). The improvement in the true-positive rate was 18% (eight of 44 breasts; $P = .008$) for ages 50–54 years, 19% (nine of 48 breasts; $P = .02$) for ages 55–59 years, 33% (15 of 46 breasts; $P < .001$) for ages 60–64 years, and 54% (21 of 39 breasts; $P < .001$) for ages 65–69 years. The 95% CIs for the difference in the true-positive rate overlap for all age strata. Still, linear regression

Table 3: Breast-based True- and False-Positive Interpretations for 2D and 2D Plus 3D Double Reading according to Volumetric Density and Age

Parameter	No. of Breasts		True-Positive Interpretations				False-Positive Interpretations			
	With SDC*	Without SDC†	No. with 2D	No. with 2D plus 3D	Difference‡	*P* Value	No. with 2D	No. with 2D plus 3D	Difference‡	*P* Value
All women	234	48 217	177	230	53 (22.7) [17.0, 28.6]	<.001	2466	2081	−385 (−0.80) [−1.03, −0.57]	<.001
Volumetric density§										
1 and 2	125	31 334	94	122	28 (22.4) [14.3, 30.8]	<.001	1421	1124	−297 (−0.95) [−1.21, −0.69]	<.001
3 and 4	107	16 787	82	106	24 (22.4) [14.1, 31.4]	<.001	1044	955	−89 (−0.53) [−0.96, −0.10]	.02
1	17	5681	15	17	2 (11.8) [−8.5, 34.3]‖	.50	197	152	−45 (−0.79) [−1.33, −0.26]	.004
2	108	25 653	79	105	26 (24.1) [15.1, 33.3]	<.001	1224	972	−252 (−0.98) [−1.28, −0.69]	<.001
3	84	13 182	64	83	19 (22.6) [12.9, 32.9]	<.001	815	721	−94 (−0.71) [−1.19, −0.24]	.004
4	23	3605	18	23	5 (21.7) [3.0, 41.9]‖	.06	229	234	5 (0.14) [−0.83, 1.11]	.82
Age										
50–54 y	52	12 952	44	52	8 (15.4) [5.3, 27.5]‖	.008	901	800	−101 (−0.78) [−1.27, −0.29]	.002
55–59 y	59	13 296	48	57	9 (15.3) [3.4, 27.5]	.02	619	542	−77 (−0.58) [−1.01, −0.15]	.009
60–64 y	63	11 049	46	61	15 (23.8) [11.7, 36.1]	<.001	521	388	−133 (−1.20) [−1.66, −0.76]	<.001
65–69 y	60	10 920	39	60	21 (35.0) [22.6, 47.6]‖	<.001	425	351	−74 (−0.68) [−1.12, −0.24]	.003

Note.—There were 2643 true- and false-positive interpretations with two-dimensional (2D) double reading and 2311 with 2D and three-dimensional (3D) double reading. SCD = screening-detected cancer.

* One woman with bilateral cancer had missing density. Therefore, analysis stratified according to density is missing for two breasts.

† Density measurements were missing in 49 women (96 breasts). Therefore, analysis stratified according to density is missing for 96 breasts.

‡ Data are numbers of interpretations, with percentages in parentheses and 95% confidence intervals in brackets.

§ Volumetric ("quantized") density was obtained with software (Quantra, Hologic).

‖ Determined with the Newcombes method, unpaired.

showed an improvement of 6.7% (95% CI: 1.8%, 11.6%; $P <$ .01) for every 5 years, with an r^2 of 0.87 (95% CI: 0.14, 0.99).

The number of false-positive interpretations was lower for all age strata with use of 2D plus 3D compared with 2D double reading (Table 3). The false-positive rate was lower by 11% (101 of 901 breasts; $P = .001$) for ages 50–54 years, 12% (77 of 619 breasts; $P = .009$) for ages 55–59 years, 26% (133 of 521 breasts; $P < .001$) for ages 60–64 years, and 17% (74 of 425 breasts; $P = .003$) for ages 65–69 years.

Age and Density Adjustments

Tables E2–E4 (online) show the age-adjusted difference in true- and false-positive interpretations stratified according to volumetric (quantized) and BI-RADS density and the volumetric density–adjusted difference in true- and false-positive interpretations stratified according to age, respectively. The tables show only minor differences in estimates when adjusting for age or volumetric density.

True- and False-Positive Mammographic Features according to Density

The number of breast-based true-positive interpretations for 2D plus 3D and 2D double reading stratified according to mammographic finding and volumetric (quantized) density is shown in Figure 2. Most additional cancers found in the 2D plus 3D analysis were classified as a spiculated mass or architectural distortion (Figs 3, E1 [online]). For these lesions, differences were 25% in almost entirely fatty breasts (two of eight [$P = .50$], with 10 detected with 2D plus 3D and eight detected with 2D), 43% in breasts with scattered fibroglandular densities (20 of 46 [$P < .001$], with 66 detected with 2D plus 3D and 46 detected with 2D), 40% in heterogeneously dense

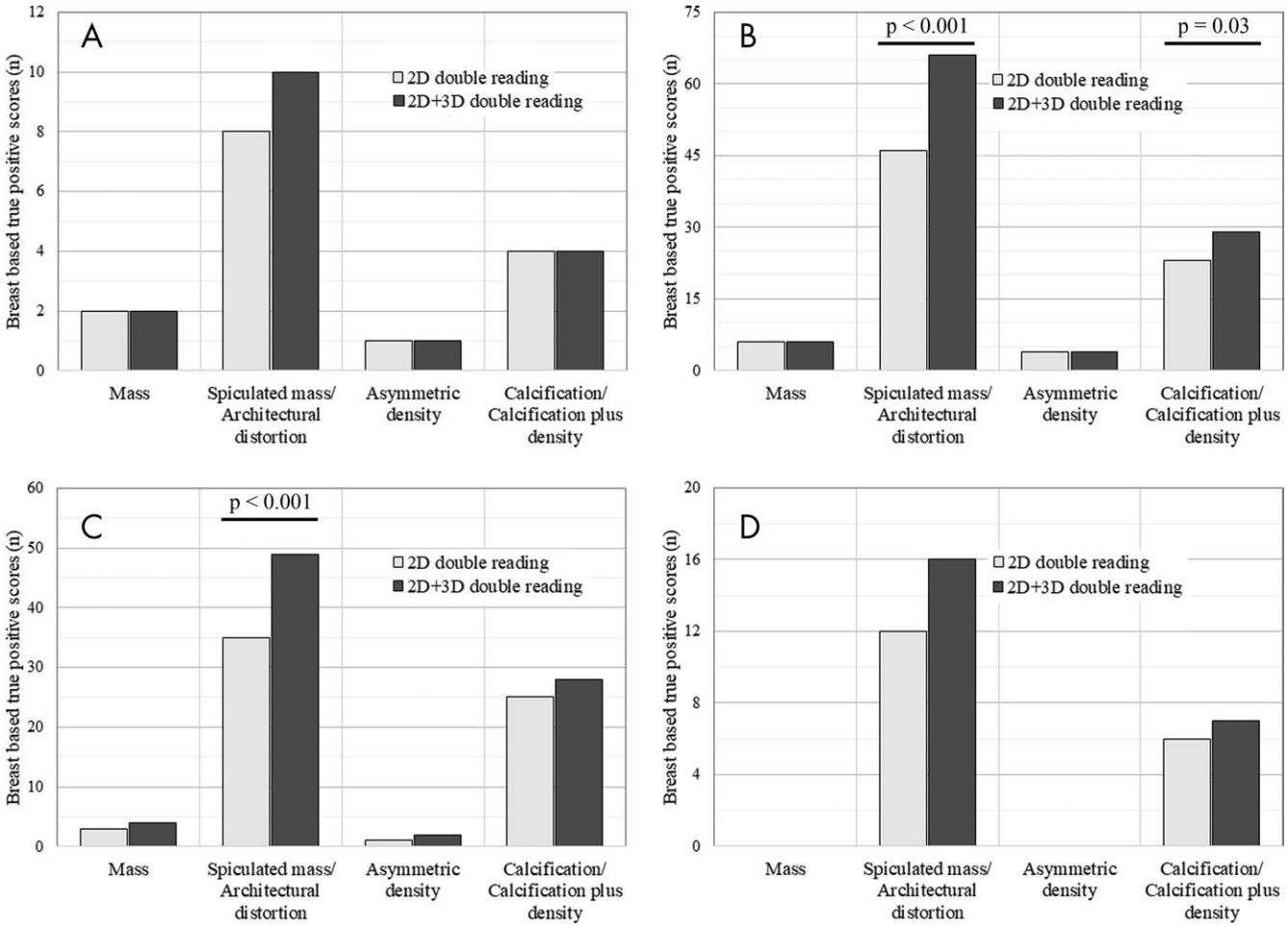

Figure 2: Bar charts show breast-based true-positive findings for two-dimensional (2D) and 2D plus three-dimensional (3D) double reading stratified according to volumetric ("quantized") density in, *A*, entirely fatty breasts, *B*, scattered fibroglandular breasts, *C*, heterogeneously dense breasts, and, *D*, extremely dense breasts. *P* values are given for significant differences. One woman with bilateral cancer had missing density values. Quantized density was obtained with software (Quantra, Hologic).

breasts (14 of 35 [$P < .001$], with 49 detected with 2D plus 3D and 35 detected with 2D), and 33% in extremely dense breasts (four of 12 [$P = .13$], with 16 detected with 2D plus 3D and 12 detected with 2D). In addition, more cancers classified as calcification and/or calcification plus density were found in women with scattered fibroglandular breasts (26%, six of 23 [$P = .03$], with 29 detected with 2D plus 3D and 23 detected with 2D).

The number of breast-based false-positive interpretations for 2D plus 3D and 2D double reading stratified according to the mammographic finding reported by the readers and volumetric (quantized) density is shown in Figure E2 (online). There was a reduction in false-positive interpretations classified as asymmetric densities with use of 2D plus 3D for women in all density categories (Fig E3 [online]), except for women with extremely dense breasts. The reduction was 51% in almost entirely fatty breasts (32 of 63 [$P = .001$], with 31 false-positive interpretations with 2D plus 3D and 63 with 2D), 58% in breasts with scattered fibroglandular densities (231 of 397 [$P < .001$], with 166 false-positive interpretations with 2D plus 3D and 397 with 2D), 52% in heterogeneously dense breasts (111 of 215 [$P < .001$], with 104 false-positive interpretations with 2D plus 3D and 215 with 2D), and 27% in extremely dense breasts (13 of 48 [$P = .15$], with 35 false-positive interpretations with 2D plus 3D and 48 with 2D). There was a lower amount of false-positive findings classified as spiculated mass and/or architectural distortions and a higher amount of false-positive findings classified as calcifications and/or calcification plus density for 2D plus 3D compared to 2D.

True- and False-Positive Mammographic Features according to Age

The number of breast-based true-positive interpretations for 2D plus 3D and 2D double reading stratified according to mammographic findings and age is shown in Figure E4 (online). The number of true-positive findings was higher in 2D plus 3D only for spiculated masses and/or architectural distortions for all age groups. The magnitude of the increase was 18% for age 50–54 years (five of 28 [$P = .06$], with 33 true-positive interpretations for 2D plus 3D and 28 with 2D), 21% for age 55–59 years (seven of 33 [$P = .04$], with 40 true-positive interpretations for 2D plus 3D and 33 for 2D), 60% for age 60–64 years (12 of 20 [$P = .002$], with 32 true-positive interpretations for 2D plus 3D and 20 for 2D), and 76% for age 65–69 years

(16 of 21 [$P < .001$], with 37 true-positive interpretations for 2D plus 3D and 21 with 2D). The number of breast-based false-positive interpretations for 2D plus 3D and 2D stratified according to age and mammographic finding reported by the readers is shown in Figure 4. There was a similar difference in type of false-positive findings across all age categories.

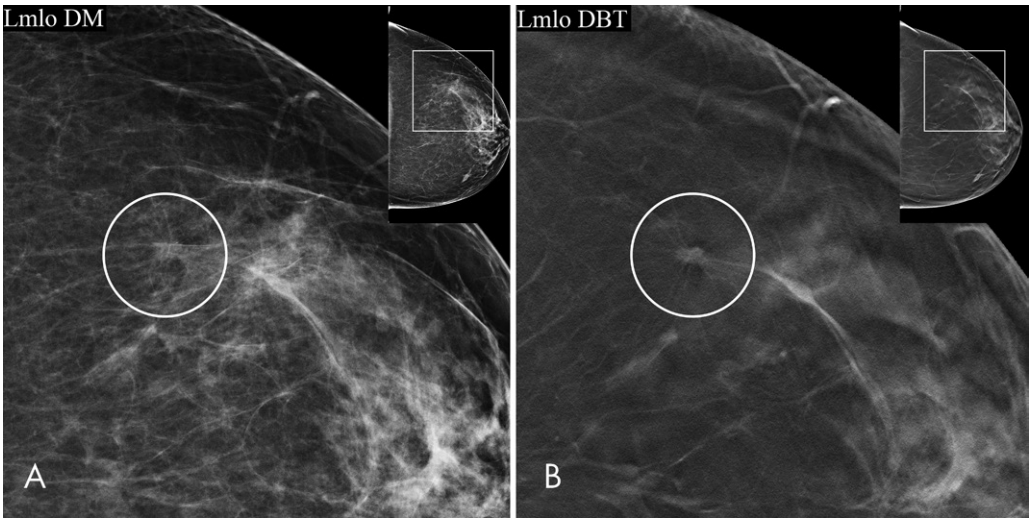

Figure 3: Screening mammograms (mediolateral oblique [Lmlo] views) obtained with, *A*, digital mammography (DM) and, *B*, digital breast tomosynthesis (DBT) in left breast of 68-year-old woman with fatty breast (volumetric density score of 2 with software [Quantra, Hologic] and Breast Imaging Reporting and Data System density category II) show a spiculated mass (in circle). Positive scores were given only by the DBT readers; the cancer was overlooked by both two-dimensional mammography readers. Histologic examination revealed an 8-mm tubular carcinoma. As the textures masking the tumor in the DM image are small, out-of-plane blurring effectively renders the masking textures invisible on the DBT images, leaving the in-plane tumor and spiculations visible. In addition, a dark halo image artifact in the tube movement direction helps highlight the tumor on the DBT image (29).

Discussion

We compared cancer detection rates and false-positive findings with digital breast tomosynthesis (DBT) and digital mammography (DM) stratified according to density and age, as previous studies have shown conflicting results. The results of our study show that adding DBT to DM in screening yields more cancers in women of all density categories (range, 12%–24%; $P < .001$ in women with scattered fibroglandular and heterogeneously dense breasts, $P > .05$ in women with almost entirely fatty and extremely dense breasts) and age groups (range, 15%–35%; $P < .05$). Most additional cancers manifest as spiculated masses or architectural distortions ($P < .001$ in women with scattered fibroglandular and heterogeneously dense breasts; $P < .05$ in women aged 55–69 years). Improvement in cancer detection with DBT showed a positive correlation ($P < .01$) with age (50–69 years). The false-positive rate was lower for women in most density categories (range, −0.7% to −1.0%; $P < .005$) and age groups (range, −0.6% to −1.2%, $P < .01$). The false-positive rate was not lower in women with extremely dense breasts (0.1%, $P = .82$), mostly due to lower number of asymmetric densities ($P < .001$).

Studies have shown improvement in cancer detection by using DBT over DM in fatty and dense breasts (5,9,11–13,15), in agreement with our results. The Malmö trial (5) indicated improvement using DBT for women with all densities, whereas another retrospective study (13) indicated similar performance, except for women with extremely dense breasts. Our results agree with those from the Malmö trial (5), which found improved cancer detection for women with extremely dense breasts and most additional cancers manifesting as spiculated masses or architectural distortions. Image texture from normal tissue (eg, fibroglandular tissue) can mask tumor spiculations and mass at DM (29). DBT removes out-of-plane fine structures, leaving in-plane tumor and spiculations visible. It has been shown that DBT requires at least some amount of peritumoral fat to be effective in dense breasts (30). In addition, the anatomic noise of structures larger than 2 mm is almost identical in DBT and DM images (31). This indicates that DBT does not allow radiologists to "see though" dense breast parenchyma but removes fine textures masking tumor spiculations in women of all breast densities. Our study showed that radiologists detected more spiculated masses with DBT compared with DM as the woman's age increases. An explanation is as age increases, the proportion of less-aggressive slower-growing cancers increases (4). Such small low-grade tumors tend to manifest as spiculated masses (6), which are better visualized with DBT. Other studies have shown increased cancer detection for all age groups in women aged 50–69 years (5,8,9,11,14,15); to our knowledge, no studies have shown positive correlation with age.

Studies have shown a reduction in the recall rate using DBT for women in all breast density categories (9,12,13,15,16,18). Our results and those from another Norwegian trial (10) differ from the results of these studies, with no difference in false-positive or recall rates for women with extremely dense breasts. However, we found a reduction in false-positive findings in heterogeneously dense breasts; the other study did not (10). The differences in the false-positive or recall rate in women with extremely dense breasts might be explained by the large difference in recall rate between countries. Other studies (32,33) have shown recall reduction mainly due to asymmetric densities, similar to our results. Superimposition of glandular tissue creates pseudo-lesions in DM, which DBT often resolves as glandular tissue is depicted in different sections. Unlike the Malmö trial (7), we detected fewer false-positive findings classified as spiculated mass or architectural distortions. We found more false-positive findings with calcifications, which might be due to highlighting of calcifications by using synthetic 2D imaging.

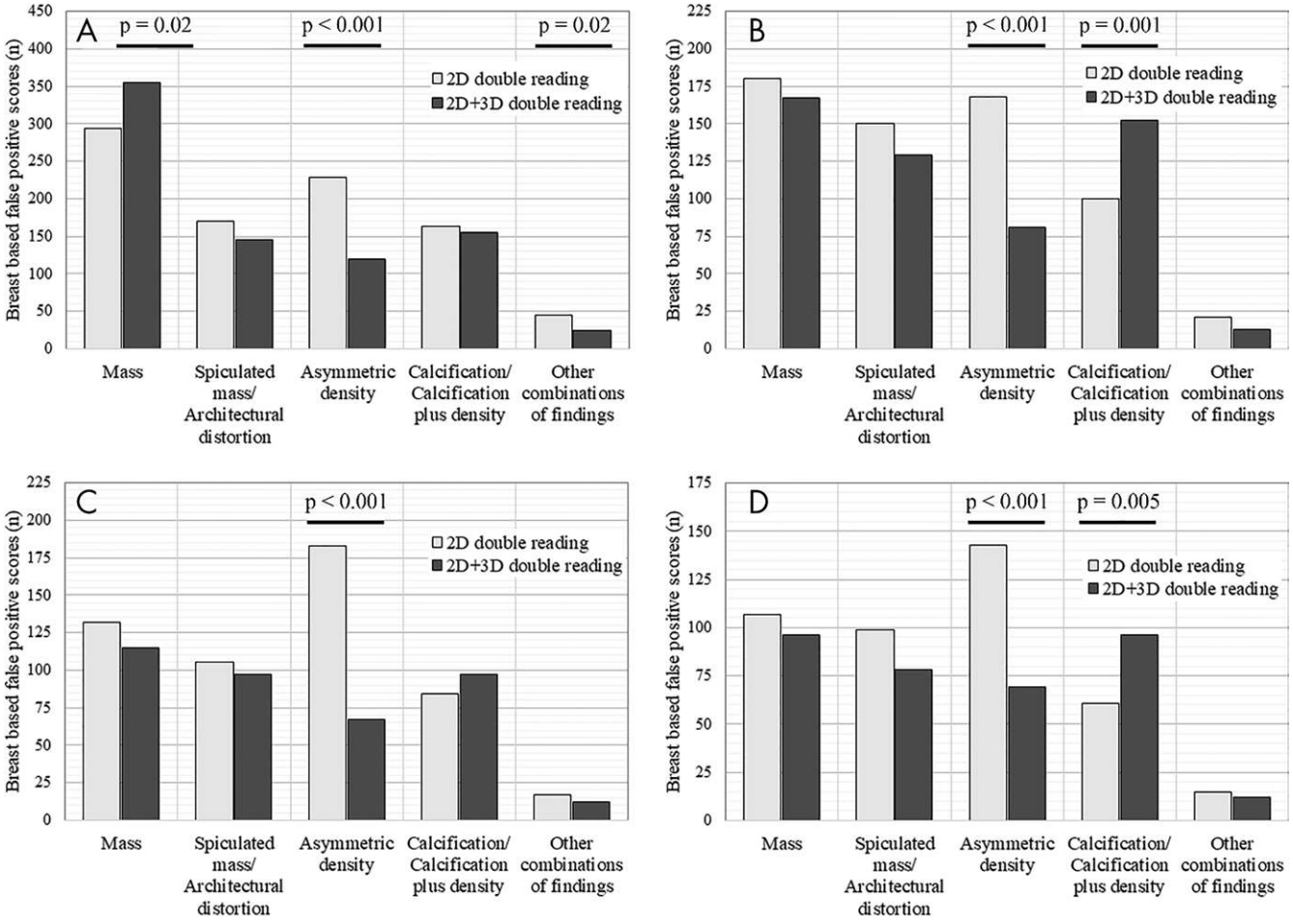

Figure 4: Bar charts show breast-based false-positive findings for two-dimensional (2D) and 2D plus three-dimensional (3D) double reading stratified according to patient age: *A*, 50–54 years, *B*, 55–59 years, *C*, 60–64 years, and, *D*, 65–69 years. *P* values are given for significant differences.

Density was assessed with use of volumetric and BI-RADS density. BI-RADS density has large interobserver variability, making it potentially advantageous to use volumetric density (20,21). Our study and a previous analysis (34) showed that agreement of BI-RADS and volumetric density is limited in almost entirely fatty and extremely dense breasts. Most discrepant classifications will be borderline cases (34), resulting in similar comparison of true- and false-positive rates by using either density measure. If volumetric density was used, two-thirds of the women would be classified as having fatty breasts, compared with half if using BI-RADS density (19,34).

Our study has limitations. It was a single-institution trial. We used the 4th edition of BI-RADS for breast density categorization because this was the standard scale when the Oslo Tomosynthesis Screening Trial started. In addition, Quantra version 2.0 maps volumetric density into categories according to the BI-RADS 4th edition density. Newer versions of Quantra use the BI-RADS 5th edition as reference. In the 5th edition, breasts are classified into a higher category if an area is dense and can obscure lesions. The 5th edition might be more associated with a reduction in sensitivity in both DM and DBT as very dense areas can obscure cancers.

In conclusion, mammography screening using digital breast tomosynthesis (DBT) depicted more cancers in all density and age groups compared with digital mammography (DM) owing to the higher number of cancers classified as spiculated masses and architectural distortions at DBT. Improvement in cancer detection showed a positive correlation with age. The number of false-positive findings with DBT was lower than that with DM due to fewer asymmetric densities, except in extremely dense breasts.

Author contributions: Guarantor of integrity of entire study, B.H.Ø.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, B.H.Ø., P.S.; clinical studies, B.H.Ø., R.G., P.S.; statistical analysis, B.H.Ø., A.C.T.M.; and manuscript editing, all authors.

Disclosures of Conflicts of Interest: **B.H.O.** disclosed no relevant relationships. **A.C.T.M.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: Oslo University Hospital has a research collaboration with GE Heathcare on CT. Other relationships: disclosed no relevant relationships. **R.G.** disclosed no relevant relationships. **P.S.** Activities related to the present article: institution received equipment and funding for additional case interpretations from Hologic. Activities not related to the present article: received payment for lectures including service on speakers bureaus from Hologic. Other relationships: disclosed no relevant relationships.

References

1. Carney PA, Miglioretti DL, Yankaskas BC, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. Ann Intern Med 2003;138(3):168–175.
2. Lehman CD, White E, Peacock S, Drucker MJ, Urban N. Effect of age and breast density on screening mammograms with false-positive findings. AJR Am J Roentgenol 1999;173(6):1651–1655.
3. Sprague BL, Gangnon RE, Burt V, et al. Prevalence of mammographically dense breasts in the United States. J Natl Cancer Inst 2014;106(10):1–6.
4. Howlader N, Altekruse SF, Li CI, et al. US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 status. J Natl Cancer Inst 2014;106(5):1–8.
5. 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. Lancet Oncol 2018;19(11):1493–1503.
6. 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. Eur Radiol 2016;26(1):184–190.
7. 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. Eur Radiol 2016;26(11):3899–3907.
8. Bernardi D, Macaskill P, Pellegrini M, et al. Breast cancer screening with tomosynthesis (3D mammography) with acquired or synthetic 2D mammography compared with 2D mammography alone (STORM-2): a population-based prospective study. Lancet Oncol 2016;17(8):1105–1113.
9. 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. Lancet Oncol 2013;14(7):583–589.
10. Aase HS, Holen ÅS, Pedersen K, et al. A randomized controlled trial of digital breast tomosynthesis versus digital mammography in population-based screening in Bergen: interim analysis of performance indicators from the To-Be trial. Eur Radiol 2019;29(3):1175–1186.
11. Caumo F, Zorzi M, Brunelli S, et al. Digital breast tomosynthesis with synthesized two-dimensional images versus full-field digital mammography for population screening: outcomes from the Verona Screening Program. Radiology 2018;287(1):37–46.
12. Conant EF, Barlow WE, Herschorn SD, et al. Association of digital breast tomosynthesis vs digital mammography with cancer detection and recall rates by age and breast density. JAMA Oncol 2019;5(5):635.
13. Rafferty EA, Durand MA, Conant EF, et al. Breast cancer screening using tomosynthesis and digital mammography in dense and nondense breasts. JAMA 2016;315(16):1784–1786.
14. Rafferty EA, Rose SL, Miller DP, et al. Effect of age on breast cancer screening using tomosynthesis in combination with digital mammography. Breast Cancer Res Treat 2017;164(3):659–666.
15. Conant EF, Beaber EF, Sprague BL, et al. Breast cancer screening using tomosynthesis in combination with digital mammography compared to digital mammography alone: a cohort study within the PROSPR consortium. Breast Cancer Res Treat 2016;156(1):109–116.
16. Haas BM, Kalra V, Geisel J, Raghu M, Durand M, Philpotts LE. Comparison of tomosynthesis plus digital mammography and digital mammography alone for breast cancer screening. Radiology 2013;269(3):694–700.
17. Sharpe RE Jr, Venkataraman S, Phillips J, et al. Increased cancer detection rate and variations in the recall rate resulting from implementation of 3D digital breast tomosynthesis into a population-based screening program. Radiology 2016;278(3):698–706.
18. Alsheik NH, Dabbous F, Pohlman SK, et al. Comparison of resource utilization and clinical outcomes following screening with digital breast tomosynthesis versus digital mammography: findings from a learning health system. Acad Radiol 2019;26(5):597–605.
19. American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS): mammography. 4th ed. Reston, Va: American College of Radiology, 2003.
20. Sprague BL, Conant EF, Onega T, et al. Variation in mammographic breast density assessments among radiologists in clinical practice: a multicenter observational study. Ann Intern Med 2016;165(7):457–464.
21. Østerås BH, Martinsen ACT, Brandal SHB, et al. Classification of fatty and dense breast parenchyma: comparison of automatic volumetric density measurement and radiologists' classification and their inter-observer variation. Acta Radiol 2016;57(10):1178–1185.
22. van Engeland S, Snoeren PR, Huisman H, Boetes C, Karssemeijer N. Volumetric breast density estimation from full-field digital mammograms. IEEE Trans Med Imaging 2006;25(3):273–282.
23. Skaane P, Bandos AI, Niklason LT, et al. Digital mammography versus digital mammography plus tomosynthesis in breast cancer screening: the Oslo Tomosynthesis Screening Trial. Radiology 2019;291(1):23–30.
24. Skaane P, Bandos AI, Gullien R, et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. Radiology 2013;267(1):47–56.
25. 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(8):2061–2071.
26. Skaane P, Bandos AI, Eben EB, et al. Two-view digital breast tomosynthesis screening with synthetically reconstructed projection images: comparison with digital breast tomosynthesis with full-field digital mammographic images. Radiology 2014;271(3):655–663.
27. Skaane P, Sebuødegård S, Bandos AI, et al. Performance of breast cancer screening using digital breast tomosynthesis: results from the prospective population-based Oslo Tomosynthesis Screening Trial. Breast Cancer Res Treat 2018;169(3):489–496.
28. Østerås BH, Skaane P, Gullien R, Martinsen ACT. Average glandular dose in paired digital mammography and digital breast tomosynthesis acquisitions in a population based screening program: effects of measuring breast density, air kerma and beam quality. Phys Med Biol 2018;63(3):035006.
29. Nakashima K, Uematsu T, Itoh T, et al. Comparison of visibility of circumscribed masses on Digital Breast Tomosynthesis (DBT) and 2D mammography: are circumscribed masses better visualized and assured of being benign on DBT? Eur Radiol 2017;27(2):570–577.
30. García-Barquín P, Páramo M, Elizalde A, et al. The effect of the amount of peritumoral adipose tissue in the detection of additional tumors with digital breast tomosynthesis and ultrasound. Acta Radiol 2017;58(6):645–651.
31. Chen L, Abbey CK, Nosratieh A, Lindfors KK, Boone JM. Anatomical complexity in breast parenchyma and its implications for optimal breast imaging strategies. Med Phys 2012;39(3):1435–1441.
32. Durand MA, Haas BM, Yao X, et al. Early clinical experience with digital breast tomosynthesis for screening mammography. Radiology 2015;274(1):85–92.
33. Lourenco AP, Barry-Brooks M, Baird GL, Tuttle A, Mainiero MB. Changes in recall type and patient treatment following implementation of screening digital breast tomosynthesis. Radiology 2015;274(2):337–342.
34. Østerås BH, Martinsen ACT, Brandal SHB, et al. BI-RADS Density Classification From Areometric and Volumetric Automatic Breast Density Measurements. Acad Radiol 2016;23(4):468–478.

10.1148/radiol.2019190425

Table E1: Breast-based true- and false-positive interpretations for 2D and 2D plus 3D double reading according to BI-RADS density

BI-RADS breast density	Breasts with SDC (n)	True-positive interpretations					False-positive interpretations		
		2D (n)	2D plus 3D (n)	Difference			2D (n)	2D plus 3D (n)	Difference (n)
				n	% [95% CI]	*P* value			
All women	234	177	230	53	22.7 [17.0–28.6]	< 0.001	2,466	2,081	–385
Density I and II	115	90	114	24	20.9 [13.5–29.2]	< 0.001	1,388	1,052	–336
Density III and IV	119	87	116	29	24.4 [15.5–33.3]	< 0.001	1,078	1,029	–49
Density I	19	16	19	3	15.8 [-3.9–37.6]*	0.25	195	166	–29
Density II	96	74	95	21	21.9 [13.5–31.2]	< 0.001	1,193	886	–307
Density III	98	72	95	23	23.5 [13.5–33.4]	< 0.001	915	863	–52
Density IV	21	15	21	6	28.6 [7.2–50.0]*	0.03	163	166	3

Note. – There were 2643 true- and false-positive interpretations with two-dimensional (2D) double reading and 2311 with 2D plus three-dimensional (3D) double reading. SDC = Screening detected cancer, CI = Confidence interval.

* Newcombes method, unpaired.

Table E2: Age adjusted difference in true-positive rate (TPR) and false-positive rate (FPR) for 2D plus 3D and 2D stratified by volumetric density*

	Difference TPR (2D plus 3D vs 2D) % [95% CI]	Difference FPR (2D plus 3D vs 2D) % [95% CI]
All women	22.7 [17.0–28.6]	–0.80 [-1.03--0.57]
Age adjusted	22.4 [16.7–28.0]	–0.81 [-1.04--0.57]
Density 1 and 2	22.4 [14.3–30.8]	–0.95 [-1.21--0.69]
Age adjusted	21.3 [13.5–29.4]	–0.96 [-1.22--0.69]
Density 3 and 4	22.4 [14.1–31.4]	–0.53 [-0.96--0.10]
Age adjusted	24.0 [15.4–32.8]	–0.50 [-0.94--0.05]
Density 1	11.8 [-8.5–34.3]†	–0.79 [-1.33--0.26]
Age adjusted	8.3 [0.0–18.8]	–0.78 [-1.42--0.17]
Density 2	24.1 [15.1–33.3]	–0.98 [-1.28--0.69]
Age adjusted	23.0 [14.6–32.2]	–1.00 [-1.30--0.70]
Density 3	22.6 [12.9–32.9]	–0.71 [-1.19--0.24]
Age adjusted	23.5 [14.2–33.3]	–0.69 [-1.18--0.20]
Density 4	21.7 [3.0–41.9]†	0.14 [-0.83–1.11]
Age adjusted	19.5 [5.0–38.3]	-0.23 [-0.85–1.34]

The age adjustment was calculated by averaging results in each age strata (age 50–54, 55–59, 60–64 and 65–69). This was done to correct for differences in age distributions within each density strata. 95% confidence intervals in the age adjusted estimates was calculated using bootstrapping with 10,000 resamples. The unadjusted estimates (shown in Table 3) are also shown. TPR = True positive rate, FPR = False positive rate, CI = Confidence interval.

* Quantized density (Quantra; Hologic).

† Newcombes method, unpaired.

Table E3: Age adjusted difference in true positive rate (TPR) for 2D plus 3D and 2D stratified by BI-RADS density

	Difference TPR (2D plus 3D vs 2D) % [95% CI]
BI-RADS density I and II	20.9 [13.5–29.2]
Age adjusted	19.9 [12.8–27.4]
BI-RADS density III and IV	24.4 [15.5–33.3]
Age adjusted	25.1 [16.3–33.8]
BI-RADS density 4th ed. I	15.8 [-3.9–37.6]*
Age adjusted	12.5 [0.0–27.5]
BI-RADS density 4th ed. II	21.9 [13.5–31.2]
Age adjusted	21.1 [13.3–29.7]
BI-RADS density 4th ed. III	23.5 [13.5–33.4]
Age adjusted	23.4 [13.9–32.6]
BI-RADS density 4th ed. IV	28.6 [7.2–50.0]*
Age adjusted	29.7 [7.1–57.7]

The age adjustment was calculated by averaging results in each age strata (age 50–54, 55–59, 60–64 and 65–69). This was done to correct for differences in age distributions within each BI-RADS density strata. 95% confidence intervals in the age adjusted estimates was calculated using bootstrapping with 10,000 resamples. The unadjusted estimates (shown in Table E1) are also shown. TPR = True positive rate, CI = Confidence interval.

* Newcombes method, unpaired.

Table E4: Volumetric density adjusted difference in true positive rate (TPR) and false positive rate (FPR) for 2D plus 3D and 2D stratified by age

	Difference TPR (2D plus 3D vs 2D) % [95% CI]	Difference FPR (2D plus 3D vs 2D) % [95% CI]
All women	22.7 [17.0–28.6]	–0.80 [-1.03--0.57]
Density adjusted	21.6 [16.1–27.2]	–0.80 [-1.03--0.56]
Age 50–54	15.4 [5.3–27.5]*	–0.78 [-1.27--0.29]
Density adjusted	16.2 [5.4–29.7]	–0.84 [-1.33--0.34]
Age 55–59	15.3 [3.4–27.5]	–0.58 [-1.01--0.15]
Density adjusted	13.4 [2.8–24.5]	–0.58 [-1.02--0.15]
Age 60–64	23.8 [11.7–36.1]	–1.20 [-1.66--0.76]
Density adjusted	23.4 [12.0–36.1]	–1.24 [-1.73--0.74]
Age 65–69	35.0 [22.6–47.6]	–0.68 [-1.12--0.24]
Density adjusted	35.1 [23.7–47.3]	–0.57 [-1.06--0.09]

The volumetric density adjustment was calculated by averaging results in each volumetric density (in percent) strata: 0–6, 7–9, 10–14 and 15–100%. These density strata were chosen to ensure the number of women in each density group was comparable. Density correction was done to correct for differences in density distributions within each age strata. 95% confidence intervals in the density adjusted estimates was calculated using bootstrapping. The unadjusted estimates (shown in Table 3) are also shown. TPR = True positive rate, FPR = False positive rate, CI = Confidence interval.

* Newcombes method, unpaired.

Supplementary figures (https://pubs.rsna.org/doi/suppl/10.1148/radiol.2019190425):

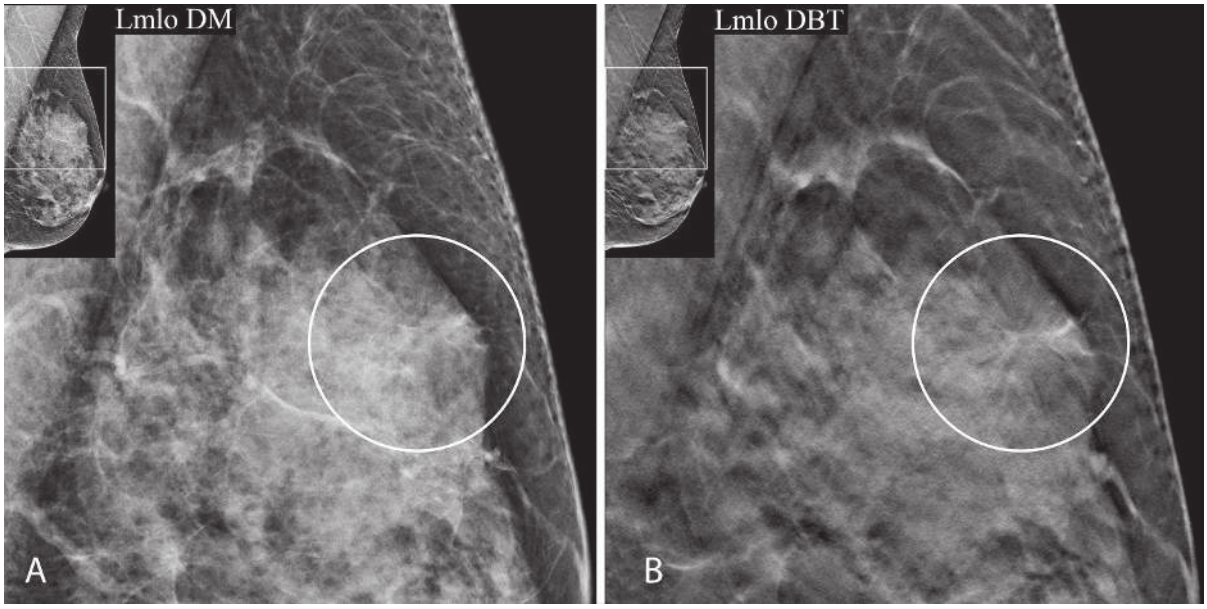

Figure E1: Screening mammogram mediolateral oblique views of a 66-year-old woman with a 6 mm invasive ductal carcinoma in the left breast. The cancer is not seen on digital mammography **(a)** and both readers had a normal score. Digital breast tomosynthesis **(b)** reveals a small spiculated lesion in the dense breast (Volumetric density[a] 3 and BI-RADS density III), and both readers gave a true-positive score. [a]Quantized density (Quantra; Hologic).

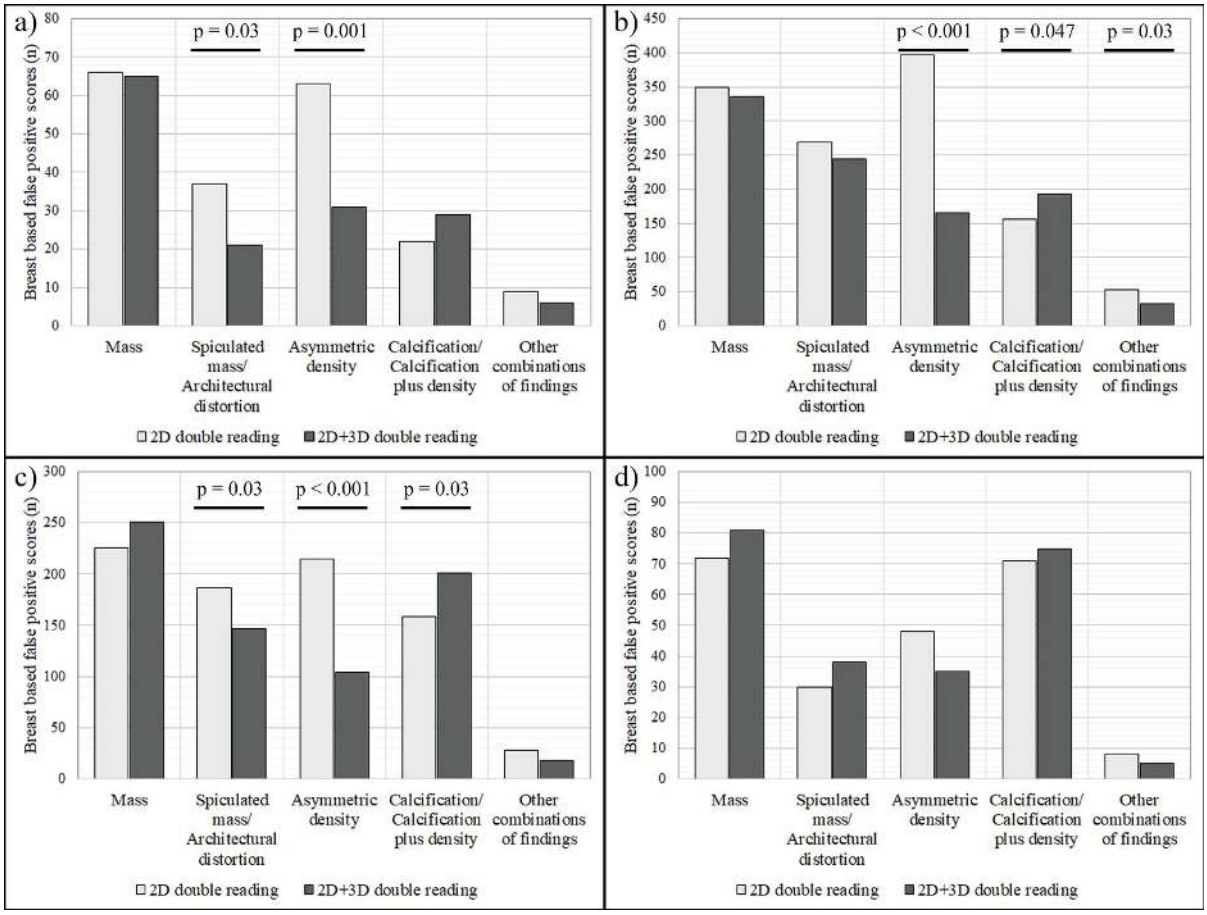

Figure E2: Breast based false positive findings for 2D and 2D plus 3D double reading stratified by volumetric (Quantized) density. (a) Almost entirely fatty breasts. (b) Scattered fibroglandular breasts. (c) Heterogeneously dense breasts. (d) Extremely dense breasts. *P* values is given for significant differences. Quantized density was obtained with software (Quantra, Hologic).

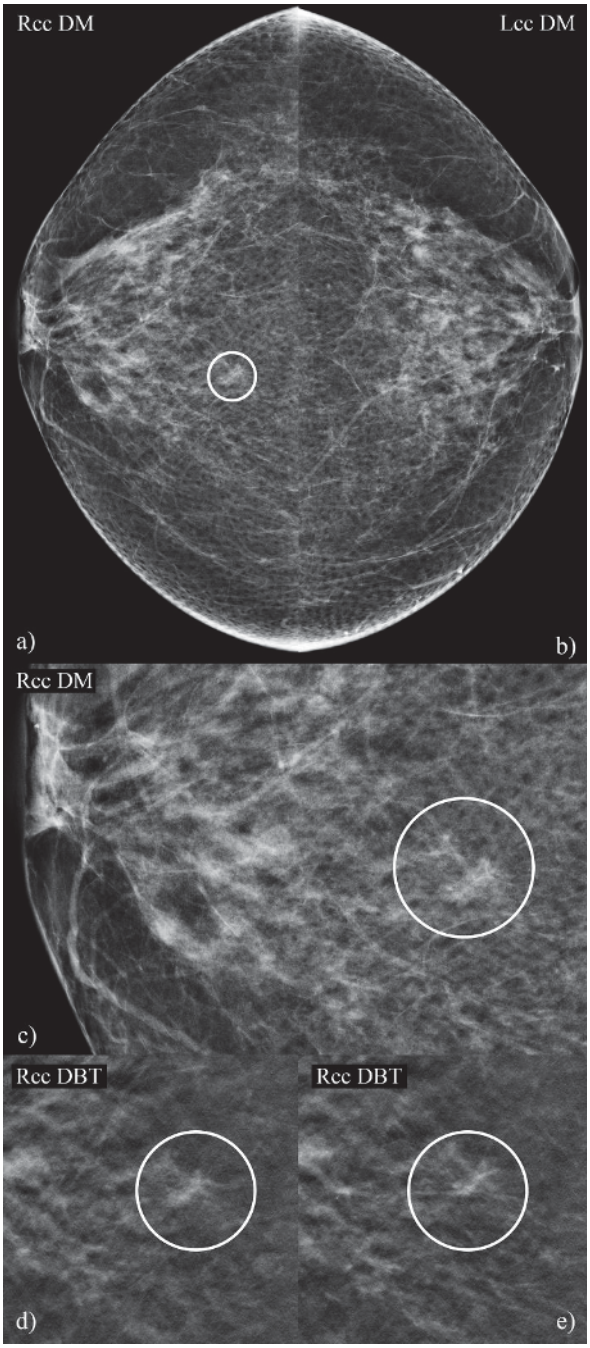

Figure E3: Screening mammogram craniocaudal (CC) views of a 68-year-old woman with fatty breasts (volumetric density[a] 2 and BI-RADS density II). Digital mammography (DM) shows a small focal asymmetric density (a) and (b) in the right breast (circle) suspicious to the readers of 2D, both giving a false-positive score of 2. Zooming the CC DM image (c) demonstrates a nonspecific focal density. Zoomed slices using digital breast tomosynthesis (18/47 and 23/47, (d) and (e), respectively) downgrade the finding as a "pseudo lesion" due to superposition of a small area of parenchyma and crossing connective tissue. Both 2D plus 3D readers gave the case a true-negative score. [a]Quantized density (Quantra; Hologic).

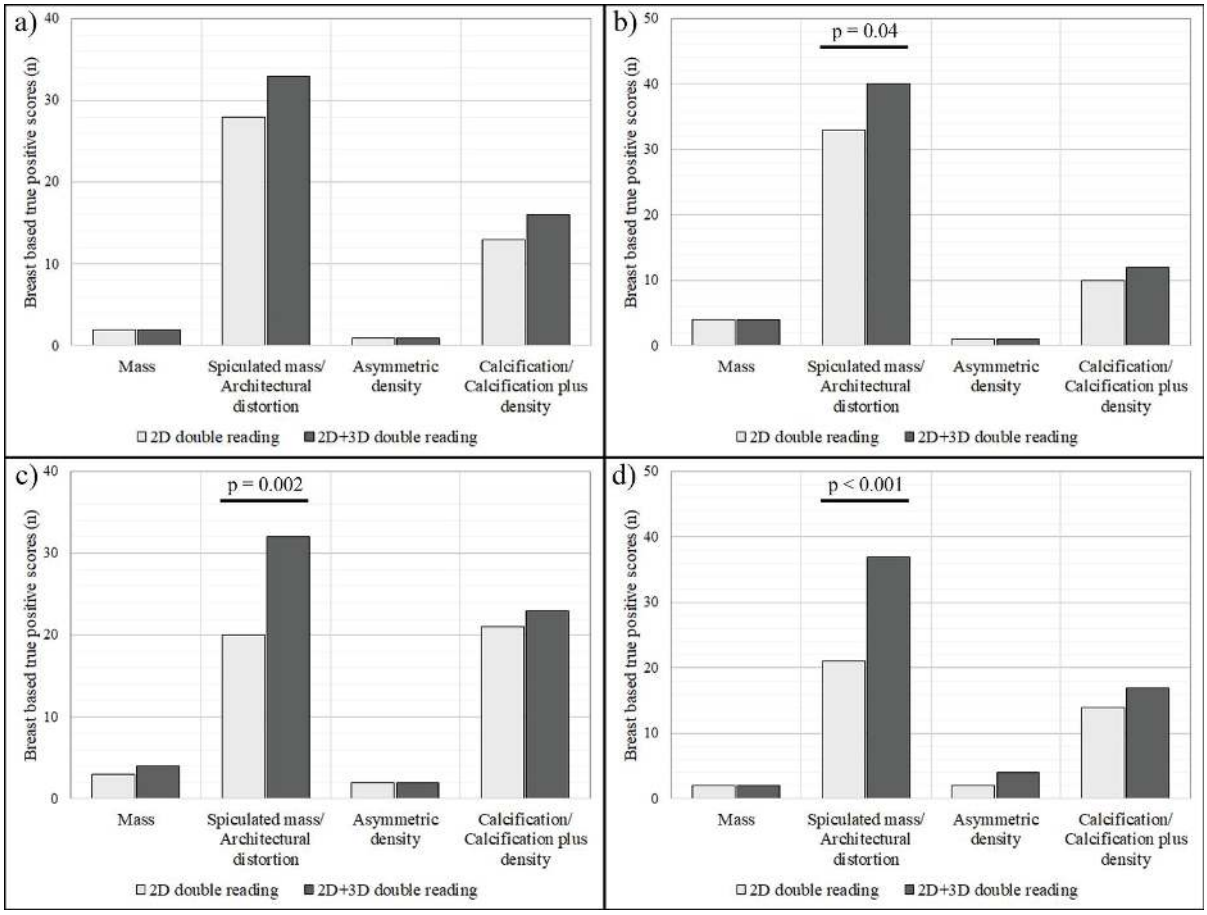

Figure E4: Breast based true positive findings for 2D and 2D plus 3D double reading stratified by age. (a) Age 50–54. (b) Age 55–59. (c) Age 60–64. (d) Age 65–69. *P* values is given for significant differences.