From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

BREAST CANCER NATURAL HISTORY MODELS AND RISK PREDICTION IN MAMMOGRAPHY SCREENING COHORTS

Rickard Strandberg

Stockholm 2022

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2022

ISBN 978-91-8016-493-1
Cover illustration: ‘L-case’, by the author.

Breast cancer natural history models and risk prediction in mammography screening cohorts
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Rickard Strandberg

The thesis will be defended in public at lecture hall Petrén, Nobels väg 12B,
Karolinska Institutet, Solna,
Friday April 22, 2022, 09:00

Principal Supervisor:
Professor Keith Humphreys
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Co-supervisor(s):
Professor Kamila Czene
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Professor Per Hall
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Opponent:
Professor Marco Bonetti
Bocconi University
Department of Social and Political Sciences

Examination Board:
Associate Professor Pär-Ola Bendahl
Lund University
Department of Clinical Sciences
Division of oncology and pathology

Associate Professor My Catarina von Euler-Chelpin
University of Copenhagen
Department of Public Health
Division of Environmental Health

Associate Professor Nicola Orsini
Karolinska Institutet
Department of Global Public Health

For Viktoria…

(…and Peggy!)

“*Essentially, all models are wrong, but some are useful.*”

—George E. P. Box

ABSTRACT

In this thesis, the foundations are laid for a new natural history model for breast cancer—specifically designed to take advantage of detailed screening cohorts. Three diverse applications of this model are then presented.

Study I develops the statistical framework for the natural history model, and shows with simulations that the model parameters can be estimated based on only the information available at diagnosis. It also describes how to adjust for random left truncation—an important aspect to consider when studying prospective cohorts.

In **Study II**, the newly developed natural history model is applied to a Swedish mammography screening cohort. It estimates the population-level distributions of age at onset and tumor volume doubling time. As an extension, the tumor volume doubling time is allowed to depend on the age at onset. The study also estimates the rate of symptomatic detection and screening sensitivity as functions of tumor size. Simulations are used to validate the estimates.

Study III shifts the focus from inference to risk prediction. The natural history model is modified to incorporate risk factors separately in each of the four components of the model. Short-term risk prediction is then performed on the screening cohort and the results are compared to a conventional approach to breast cancer risk prediction. The study also develops novel predictions based on, for example, having experienced tumor onset, having a tumor detected at the next screening, and having a tumor detected before it reaches a certain size if attending the next screening.

In **Study IV**, the model is used to study the effect that certain acquisition parameters used in mammography have on the detectability of the breast cancer tumor. With the model, it is possible to more directly study the mammography screening sensitivity, compared to the ad hoc definition of sensitivity commonly seen in the screening literature. It was identified that the compressed breast thickness—in addition to the percent mammographic density and latent tumor size—was inversely associated with the screening sensitivity.

LIST OF SCIENTIFIC PAPERS

I. Rickard Strandberg & Keith Humphreys
Statistical Models of Tumour Onset and Growth for Modern Breast Cancer Screening Cohorts
Mathematical Biosciences, 2019, 318, 108270,
DOI: 10.1016/j.mbs.2019.108270

II. Rickard Strandberg, Kamila Czene, Mikael Eriksson, Per Hall, Keith Humphreys
Estimating Distributions of Breast Cancer Onset and Growth in a Swedish Mammography Screening Cohort
Cancer Epidemiology, Biomarkers and Prevention, 2022, 31 (3), 569-577,
DOI: 10.1158/1055-9965.EPI-21-1011

III. Rickard Strandberg, Kamila Czene, Per Hall, Keith Humphreys
Novel Predictions Of Breast Cancer Risk In Mammography Screening Cohorts
Manuscript

IV. Rickard Strandberg, Maya Alsheh Ali, Kamila Czene, Per Hall, Keith Humphreys
Modelling the effects of Mammographic Density and Compressed Breast Thickness on Mammographic Sensitivity: A Natural History Approach
Manuscript

CONTENTS

LIST OF ABBREVIATIONS

AQPD	Age-specific quartile of percent mammographic density
AUC	Area under curve
BC	Breast cancer
BI-RADS	Breast imaging-reporting and data system
BMI	Body mass index
BRCA	Breast cancer gene
CBT	Compressed breast thickness
CC	Craniocaudal view
CI	Confidence interval
CP	Compression pressure
ER	Estrogen receptor
EXP	Total x-ray exposure
HDI	Human development index
HER2	Human epidermal growth factor receptor 2
HRT	Hormone replacement therapy
KARMA	Karolinska mammography project for risk prediction of breast cancer
LR	Likelihood-ratio
MLE	Maximum likelihood estimate
MLO	Mediolateral oblique view
MST	Mean sojourn time
PD	Percent (mammographic) density
PR	Progesterone receptor
PRS	Polygenic risk score
ROC	Receiver operating characteristic
SNP	Single nucleotide polymorphism

1 INTRODUCTION

Breast cancer is the most common cancer in the world, and the incidence is increasing [1]. It is a disease that predominantly affects women and is the most common cause of cancer death for women.

At this stage, most countries have established nation-wide mammography screening programmes [2]. These programmes are estimated to have reduced the mortality of breast cancer by around 20% [3].

To understand the extent of the benefits (and harms) that breast cancer screening brings, we need to first understand the mechanics which drive the disease. To that end, researchers have studied the natural history of breast cancer by using statistical models. A wide range—including multi-state Markov models, continuous growth models, and simulation models—have been developed and applied to a wide array of breast cancer data.

While the current screening programmes are age-based, there has recently been considerable interest in adapting and personalizing the screening based on individual risk factors [4,5]. For this purpose, the various breast cancer risk prediction models developed over the last 30 years [6] are being brought to bear, and multiple trials are underway [7,8]. Many questions still remain surrounding who should be screened, and when. By better understanding the natural history of breast cancer, and what determines its detectability at mammography, some of these questions might find an answer.

2 BACKGROUND

2.1 THE BREAST

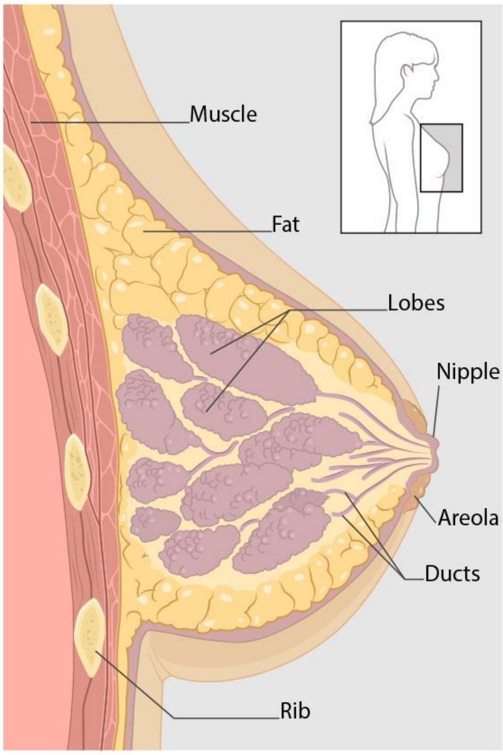


Figure 1: An anatomical depiction of the breast. Source: Centers for Disease Control and Prevention[1].

The breast is the mammary gland in humans. It is positioned over the pectoral muscles and houses potentially milk-producing glandular units called lobes. Each lobe consists of clusters of alveoli (also called lobules) which secrete the milk and are connected to the nipple by branching ducts that transport the milk. Each breast contains 15-20 lobes [9]. Enveloping the lobes is adipose tissue (fat cells). The rest of the breast interior consists of stroma in the form of mainly blood vessels and connective tissue (mostly collagen) which provides the structure and shape of the breast. See Figure 1.

The development of the breast begins in utero where the primitive structures of the lobes are formed with short ducts connecting to the nipple [10]. The breast tissue lies dormant until puberty, when the female breast undergoes significant changes. Estrogen is an important hormone that promotes the growth of the ducts and the stroma, while the hormone progesterone drives the additional formation and differentiation of the lobules.

After puberty, the composition of the breast experiences some fluctuations as the fluctuating hormone levels of estrogen and progesterone cause growth and apoptosis (programmed cell death) in the lobes [11]. Over time and repeated menstrual cycles the breast accumulates some of the growth and differentiation caused by these cycles [12].

[1] https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm, accessed 2022-01-20.

Another phase of development in the breast occurs during pregnancy. Increased levels of estrogen cause further growth and branching of the ducts and blood vessels, and progesterone causes additional growth and differentiation of the lobules. Another hormone, prolactin, is responsible for the final (terminal) differentiation of the lobules which enables the production of milk. After pregnancy, the breast experiences a reversal of some these changes when the prolactin production ceases (a process called *post-lactational involution*) [13]. Overall, women who experience one or more full-term pregnancies (*parous* women) maintain a higher differentiation in the breast compared to those who do not experience full-term pregnancy (*nulliparous* women) [12,14].

During menopause, the ovaries stop producing estrogen and progesterone. The decline in these hormone levels causes the lobes to shrink and causes a substantial decrease in the number and differentiation of the lobules [10,14]. This process is called the *lobular involution*. The stroma is also reduced, with more fatty tissue taking its place [13]. After menopause, the breast composition is much the same for both parous and nulliparous women [14,15].

2.2 BREAST CANCER

In 2020, female breast cancer overtook lung cancer as the most common cancer in the world—for both sexes combined [1]. An estimated 2.3 million new cases of breast cancer were reported, constituting 12% of all new cancer cases and 25% of female cancer cases. Despite having the highest incidence, breast cancer ranked 5th in mortality (6.9% of all cancer deaths) with an estimated 685,000 deaths. It is however still the most common cause of cancer death in women (16%).

Countries with high or very high human development index (HDI) have an average 88% higher incidence rate of breast cancer (55.9 cases per 100,000) than countries with low or medium HDI (29.7 cases per 100,000). The difference has been attributed to longer life expectancy and overall older demographics in the developed countries, plus a higher prevalence of various risk factors of breast cancer. It is the opposite case for breast cancer mortality, however; developed countries have 15% lower mortality rates (12.8 deaths per 100,000) than developing countries (15.0 deaths per 100,000) [1].

In Sweden, 7361 women were diagnosed with breast cancer per year, on average, between 2015-2019, and 1399 women died from the disease per year [16]. The estimated risk of being diagnosed with breast cancer before age 75 was 9.4% on average.

Figure 2 displays the trends of breast cancer incidence and mortality in Sweden. Panel **(A)** shows the annual rates from 1960-2019. The incidence rate has steadily increased over time, while the mortality rate is now lower than it was in the 1960s. Panel **(B)** shows the incidence and mortality rates in 2019 separated into age groups. It shows that breast cancer becomes more prevalent with age. There is a sudden decrease in the incidence rate between age 75-79. This is the age where the Swedish screening programme ends.

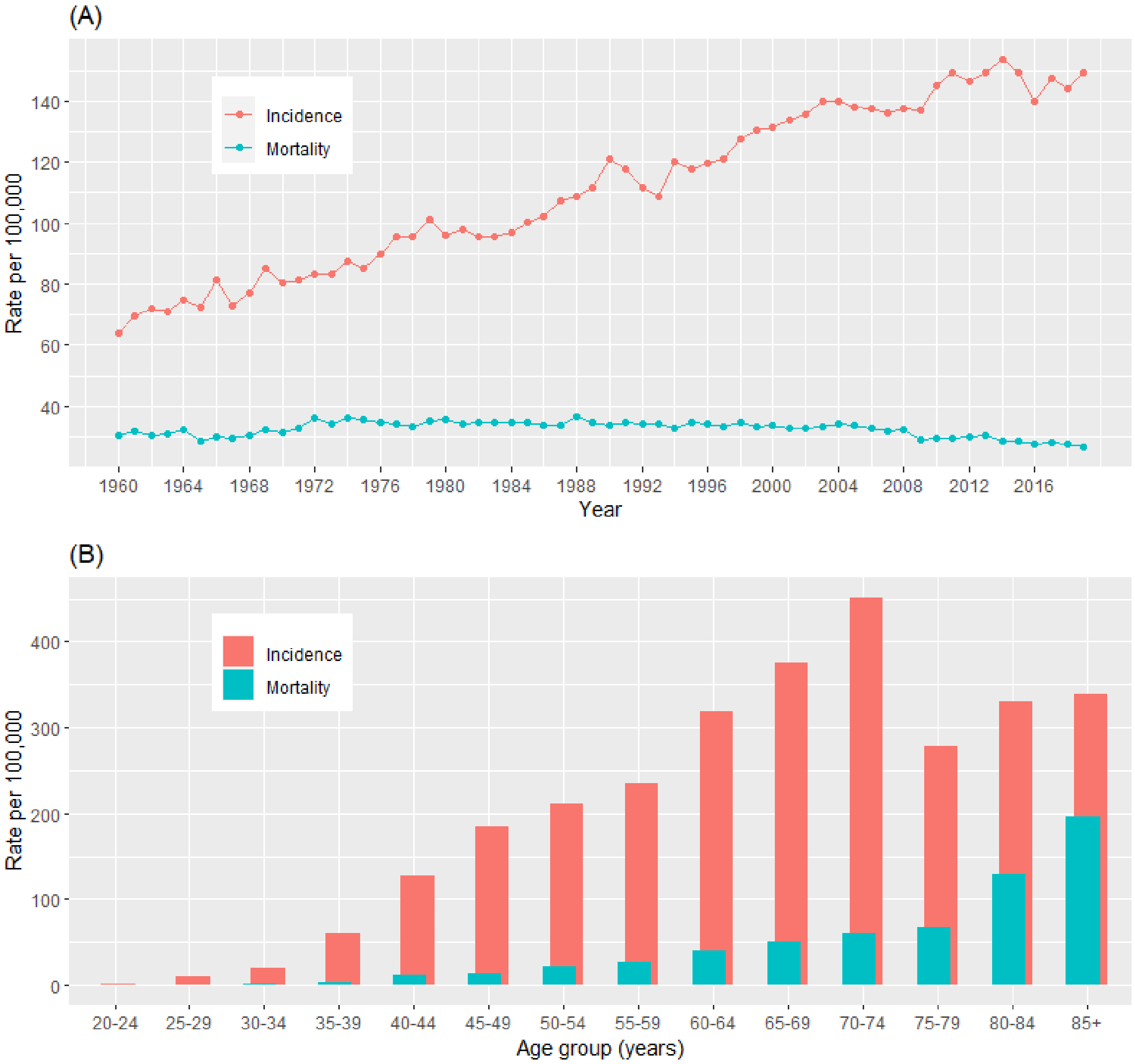


Figure 2: The incidence and mortality rates of breast cancer in Sweden. **(A)** The year-specific rates between 1960-2019. **(B)** The age-specific rates from 2019.

2.2.1 Types of Breast Cancer

There are two major forms of breast cancer: breast carcinoma (cancer) *in situ* and invasive breast cancer. A breast carcinoma in situ means that the cancer is still contained within the epithelial layers of its origin. For breast cancer this means being contained inside the ducts or lobules, and the two types are called *ductal carcinoma in situ* (DCIS) and *lobular carcinoma in situ* (LCIS). Approximately 20% of diagnosed breast cancers are in situ, of which around 85% are DCIS [17].

If a cancer is an invasive breast carcinoma, then it has penetrated from the epithelial layers into the stroma and surrounding tissue. The most common sites for invasive breast cancer are also the ducts and the lobules, with the respective names *invasive ductal carcinoma* (IDC) and *invasive lobular carcinoma* (ILC). Around 80% of invasive breast cancers are IDC, 15% are ILC, and the remaining 5% are other types that cannot be easily classified as either ductal or lobular [18–20].

In situ cancers are considered precursors to invasive cancer, where the invasiveness is a property gained after the tumor has started growing [18]. According to one review [21], 14%-46% of detected DCIS (which were not treated due to originally being misdiagnosed as benign) later, over the next 10+ years, developed into invasive breast cancer. This indicates that the transformation into invasive breast cancer can occur later, even for in situ cancers that have progressed enough to be detected.

2.2.2 Breast Cancer Metastasis and Prognosis

Cancer metastasis refers to the spread of a cancer from the original site to other locations in body (meta- “next”, stasis- “placement”). In breast cancer, the most common metastasis occurs in the lymph nodes near the breast and arm pits. The lymph system then becomes a potential vehicle for further spread. The most common sites for distant metastasis outside the lymph nodes are the skeleton (50%), the lungs (24%), the liver (20%), and the brain (6%) [22]. Between 30-35% of invasive breast cancers show lymph node involvement [23], and 2-6% of diagnosed breast cancers have confirmed distant metastasis [24].

Breast cancer can be classified or staged based on the current status of the metastasis. A breast cancer is referred to as *localized* if there is no metastasis, *regional* if there is metastasis in the nearby lymph nodes, and *distant* if there is metastasis in other sites [25].

A more detailed type of staging is the TNM staging system [26]. It classifies breast cancer progression in three parts based on the primary tumor size (T-stage), the extent of the lymph node spread (N-stage) and the occurrence of distant metastasis (M-stage). Each of the three stages are both individually (sub-)classified and combined into an overall stage for the breast cancer. The simplified version of the five stages of breast cancer are as follows:

0. In situ only (Tis). By the definition of in situ, there is no spread (N0 & M0).
I. The primary invasive tumor is less than 20mm in diameter (T1), and there is no spread (N0 & M0).
II. Either the primary tumor is greater than 20mm (T2-3) and there is no lymph node spread (N0), or there is limited local lymph node spread (1-3 nodes, N1) and the primary tumor is less than 20mm (T0-1).
III. Either there is extensive lymph node spread (4+ nodes, N2-3) or limited spread (N1) and a tumor larger than 50mm (T3).
IV. If and only if there is distant metastasis (M1).

The stage of breast cancer is a strong indicator for prognosis. Death due to breast cancer occurs when it metastasizes in vital organs [27], which leads to death most commonly from heart failure, infarctions, or infections to affected organs (e.g. pneumonia infecting the lungs). Therefore, the extent of metastasis at the time of diagnosis is very important for the prognosis. In the U.K., the estimated 10-year survival probability is 96% for Stage I, 79% for Stage II, 53% for Stage III, and 12% for Stage IV [28]. For Stage 0, the 10-year survival is around 99% [29].

2.2.3 Histological Grade

Another type of breast cancer assessment factor is the histological grade. Based on microscopic investigation of the tumor cells, the cancer is given a grade from 1 to 3 depending on how much its cells resemble healthy cells (i.e. how well-differentiated they are), where a high grade signifies little resemblance [30,31]. A high tumor grade is associated with more malignant and aggressive breast cancer [30], and is associated with a worse prognosis—even when accounting for the cancer stage [32].

2.2.4 Molecular Subtypes

Breast cancer can also be classified on the molecular level. Using immunohistochemical staining, one can test for the presence of estrogen receptors (ER) and progesterone receptors (PR) on the tumor cells. As previously mentioned, estrogen and progesterone are two important hormones for regulating the proliferation of healthy breast tissue. If the staining reveals receptors in more than 10% of the tumor cells, the tumor is classified as positive for that hormone receptor (ER+ and PR+ respectively). Another important receptor to assess is the human epidermal growth factor receptor 2 (HER2). HER2 is overexpressed in some cancers as a means to proliferate more aggressively. The statuses of these three receptors give an indication of the cancer's aggressiveness, and have implications for treatment options (see below).

Global genetic profiling of breast cancer tumors have shown that the molecular heterogeneity can be summarized by four main intrinsic subtypes [33]. The four intrinsic subtypes approximately correspond to the amount of expression of the three receptors and to tumor grade according to the following (simplified) table [34]:

INTRINSIC SUBTYPE	**ER**	**PR**	**HER2**	**GRADE**
LUMINAL A	High	Some/high	None	Low
LUMINAL B	Low	Low/none	Some	High
HER2 ENRICHED	None	None	High	High
BASAL-LIKE	None	None	None	High

The basal-like subtype is often called *triple negative* breast cancer, referring to testing negative for all three receptors. The majority of diagnosed invasive breast cancers are of the Luminal A subtype (70-75%), followed by Luminal B and Basal-like (10-12% each), and HER2 enriched (4-5%) [35,36].

2.2.5 Symptoms of Breast Cancer

The most common symptom of breast cancer is finding a palpable lump in the breast—the primary tumor—which occurs in 82% of breast cancers at the time of diagnosis. 17% of patients show other breast-related symptoms, such as nipple abnormalities, pain, swelling, rashes, or ulcerations. 6% experience symptoms outside of the breast that are related to metastasis, such as lumps in the armpits or neck, and non-specific symptoms like muscular or skeletal pain, fatigue, or weakness [37].

2.2.6 Treatment

The treatment of a breast cancer depends on its attributes, both in terms of stage and subtype [38]. The primary treatment is surgery, whereby the affected breast is either entirely removed (mastectomy), or just the tumor and a margin around it (breast-conserving surgery). This removes the primary tumor and any small surrounding tumors. Afterwards, radiation therapy is standard—especially after breast-conserving surgery.

Depending on which receptors test positive, targeted treatment is recommended. In the case of an ER+ tumor, an estrogen receptor modulator such as tamoxifen is offered. If there is positive staining for HER2, the monoclonal antibody Trastuzumab is used to block the receptors, and has been shown to be an effective treatment [39]. The purpose of these targeted treatments is to limit the probability of recurrence or spread by blocking the mechanisms the cancer (at least partly) depends on for its growth and proliferation.

Additional treatment with chemotherapy is recommended if the risk of recurrence or additional spread is high (i.e. all subtypes but Luminal A), or if there is already confirmed lymph node metastasis (regional Stage II or higher), or if the patient is under 35 years of age [38].

2.3 MAMMOGRAPHY SCREENING

The most prominent method for detecting and diagnosing breast cancer is called mammography. Mammography is an x-ray imaging technique for exposing breast cancer tumors. The breast is compressed between two paddles—one transparent which the x-rays can pass through, and one with a receptor for the x-rays. The compression is done to flatten and spread the breast tissue, hold the breast in place, and to reduce the x-ray dose required to penetrate the breast.

Images are taken of each breast from two different views: the craniocaudal (CC) view taken horizontally from above; and the mediolateral oblique (MLO) view taken diagonally along a line from the armpit to the lower sternum. Figure 3 below shows examples taken from the MLO view.

2.3.1 Screening Programmes

To detect and treat breast cancer earlier, many countries have implemented nation-wide mammography screening programmes [2]. In these screening programmes, women within certain age ranges are invited to attend a mammography at regular intervals (typically 1 to 3 years apart). The estimated reduction in breast cancer mortality is around 20% for women invited to screening [40,41].

In Sweden, the first pilot study began in 1974 in Gävleborg, and nation-wide coverage was achieved in 1997 [42]. Today, women between ages 40 and 74 are invited to attend screening [43]. Depending on the county (landsting), the screening interval is between 18 and 24 months. The participation rate in Stockholm county is approximately 75% [44].

2.3.2 Mammographic Density

As previously described, the breast tissue can be classified into three different types: the epithelial tissue, consisting of the lobules and ducts responsible for lactation; stromal tissue, which is the connective tissue and vessels giving the breast its structure; and adipose tissue (fat cells). The breast tissue composition varies greatly between women, in terms of both the amount and the organization of the tissue [45].

In the context of mammography, the epithelial tissue and the stroma are often referred to together as *fibroglandular tissue*. The distinction between the fibroglandular tissue and the adipose tissue is important in mammography. This is because—like breast cancer tumors—fibroglandular tissue is opaque to x-rays. If a woman has a high amount of such tissue, there is a risk that a tumor will be masked during mammography.

The American College of Radiology has developed a system to classify a woman's breast tissue composition, called the BI-RADS (Breast Imaging-Reporting And Data System) score [46]. The classification is done through a qualitative assessment by radiologists. They divide mammographic density into four categories from the least to the most dense:

a. The breast is almost entirely fatty,
b. There are scattered areas of fibroglandular density,
c. The breast is heterogeneously dense, which may obscure small masses,
d. The breast is extremely dense, which lowers the sensitivity of mammography.

A quantitative measure of mammographic density is percent density (PD) [47–49]. PD is an estimate of the proportion of dense tissue in the breast. The estimation is done with image analysis software by counting the proportion of bright pixels on the mammogram.

In Figure 3 are two examples of mammograms taken from the MLO view with different mammographic densities.

It is known that older women have, on average, lower mammographic density than younger women [50,51]. Longitudinal studies of mammographic density have shown a decrease in PD with age between 40 and 65 [52,53]. In addition to the age trend, there is a further reduction in density during menopause [54]. These changes are tied to the hormonal changes in the breast.

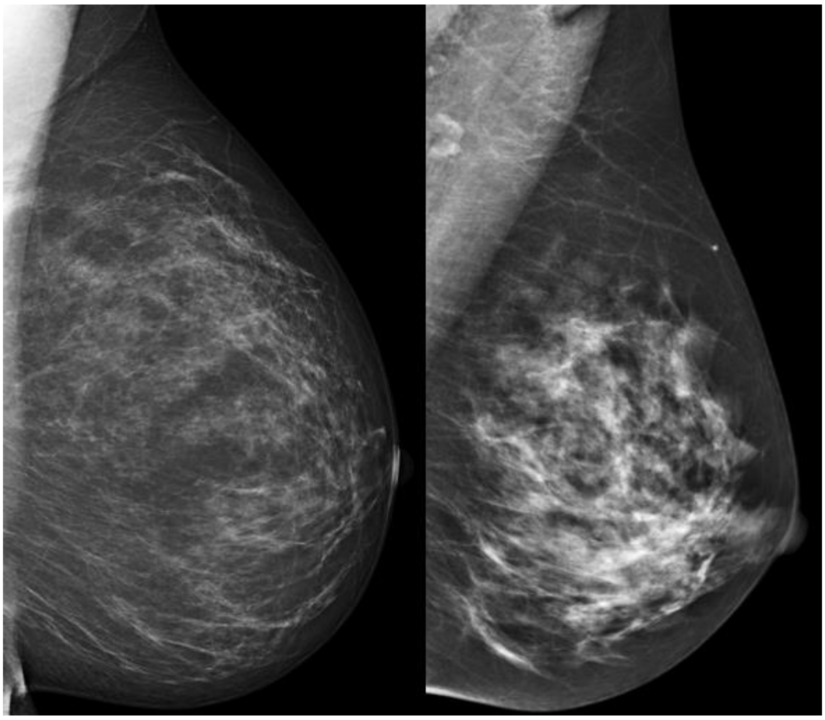


Figure 3 Examples of mammograms (MLO view) of two breasts with different mammographic density. *Left:* low density corresponding to BI-RADS category *a*, with estimated PD of 8%. *Right:* high density corresponding to BI-RADS category *c*, with estimated PD of 42%.

2.4 DETERMINANTS OF BREAST CANCER

There are many factors which determine someone's risk of developing breast cancer. Sex is the largest by far, as only 1% of breast cancers occur in men [55]. The risk also increases with age [56,57] as can be seen in Figure 2, and a high mammographic density has been shown to be a very important risk factor [54,58,59].

A large group of risk factors are the various reproductive factors. Having a late age at menopause, or an early age at menarche will increase the risk of breast cancer [59–61]. Using hormone replacement therapy to treat menopausal symptoms is also known to increase the risk [59,61,62]. Breastfeeding, on the other hand, reduces the risk of breast cancer [59,62]. These factors are believed to be partly mediated by mammographic density [63,64] and the differences in the breast tissue during these events.

Researchers have found that parity (childbirth) has a "dual effect" on breast cancer risk [65–68], where childbirth confers a short term increase to risk, believed to be due to hormonal changes during the pregnancy; but a long term protective effect, due to increased differentiation of the breast tissue [69,70]. The timing of the births is also a factor, as an older age at both first and last birth have been shown to increase the risk further, particularly for childbirths after age 30 [57,71].

Family history is a significant risk factor for breast cancer. Women who have a first-degree relative with breast cancer approximately have a two-fold risk of breast cancer compared to those without; and there is a 50% higher risk among women with affected second-degree relatives [59,60,72]. This heritability of breast cancer can partly be explained by genetic mutations which are inherited. The most significant single mutations are of the BRCA1 and BRCA2 genes. These mutations are estimated to increase risk 10- to 20-fold, and account for 16% of familial breast cancer risk [73]. The estimated cumulative risk of breast cancer by age 70 is 71% for BRCA1 mutation carriers and 84% for BRCA2 mutation carriers [74]. It is estimated that 0.05% of people carry BRCA1 mutations and 0.07% carry BRCA2 mutations [75].

Genome-wide association studies (GWAS) have so far found over 180 other single nucleotide polymorphism (SNP) alleles which increase the risk of breast cancer [76,77]. These SNPs/mutations are more common in the population, but confer a much lower risk increase, and together account for an additional 18% of the familial risk [77].

Since each individual SNP confers little risk by itself, they are often combined into a *polygenic risk score* (PRS) to determine an individual's genetic risk due to a collection of SNPs [78,79]. One study based on a PRS with 77 SNPs found that women who scored in the top 1% had a three-fold increased risk compared to the median [80]. A more recent study [81] found a four-fold increased risk based on 313 SNPs. The study also found that family history of breast cancer was still strongly associated with risk even when adjusting for the PRS.

Several lifestyle factors have been studied. Having a high BMI after menopause increases the risk of breast cancer [62,82,83], while regular physical activity reduces it [84]. Similarly to other cancers, smoking [85] and alcohol use [86,87] are known to increase risk.

2.5 STATISTICAL MODELS OF BREAST CANCER

The purpose of breast cancer natural history models is to use data available at diagnosis to study the latent processes leading to a breast cancer diagnosis and beyond. These types of models can then be further used to study the impact of screening on, for example, breast cancer mortality at a population level. The type of data that these models can be used to analyze may include e.g. the age of the patient, the incidence and mortality over a calendar period, tumor characteristics (e.g. stage or type), and the mode of detection (during screening or symptomatically between/outside screening rounds). Although these models typically take as input, information limited to that collected at diagnosis (or during follow-up), one can—on a population level—estimate latent, dynamic processes.

2.5.1 Multi-state Markov models

The multi-state Markov model is historically the most common likelihood-based approach for modelling the natural history of breast cancer [88–91]. The most basic model consists of three states and is illustrated in Figure 4. The first state is having *no detectable cancer*. This includes both being cancer-free and having a tumor which is not (yet) detectable by (mammography) screening. Tumors then transition into the *pre-clinical* state. During this state, if the woman is screened for breast cancer, there is a probability that the tumor will be found. If it is not screen-detected in time, the tumor will transition to the third state, which is *clinical cancer*. This means that the tumor is detected clinically by displaying symptoms.

Since it is a Markov model, the state transition times are assumed to be exponentially distributed. An important quantity in multi-state Markov models is the *mean sojourn time* (MST), which is the average time spent in the pre-clinical state, which is the reciprocal of the transition rate from the pre-clinical state to the clinical state (for exponentially distributed times in the pre-clinical state).

The basic 3-state model can be extended in various ways to include additional states. Several researchers have modelled lymph node metastasis by including states for pre-clinical and clinical lymph node positive tumors [90,92,93]. Tan et al. [94] included DCIS as a possible precursor to invasive cancer. They also modelled tumor size using states and transitions for different size intervals. This resulted in a 13-state Markov model.

Another way of extending the Markov model is by regressing the state transition rates. For example, Chiu et al. [95] used a 3-state model to study the effect of mammographic density on both the pre-clinical incidence rate and the MST. Wu et al. [96] studied mammographic density, BMI, and age at first birth in the same way, and also included genetic factors (such as BRCA1/2 mutations) in the pre-clinical rate; and tumor markers in the MST. Taghipour et al. [97] studied the effect of age, menstruation length, and number of births on the pre-clinical rate. They also included death from other causes as a 4th state, with possible transitions from the 1st and 2nd state.

Multi-state Markov models have been used extensively to assess breast cancer screening. Wu et al. [91] for example combined a 5-state lymph node model with a microsimulation study to compare different screening intervals and the risk of lymph node spread. Schousboe et al. [5] used a Markov microsimulation model to explore the cost-effectiveness of personalized screening intervals based on age, mammographic density, and family history.

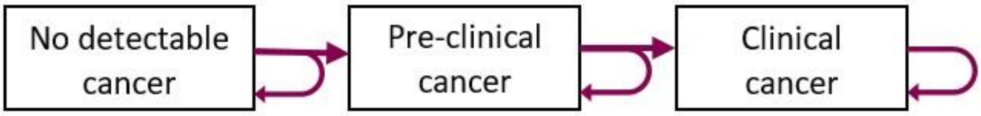


Figure 4: An illustration of the basic 3-state Markov model for breast cancer. The states are (1) no detectable cancer; (2) pre-clinical cancer detectable by screening, and (3) clinical cancer, detected through symptoms.

2.5.2 Continuous growth models

The continuous growth models for breast cancer represent an alternative approach to multi-state Markov models for studying the natural history of the disease. These models can be divided into different components, each representing separate (biological) processes. The four main components to consider are: i) the tumor onset, where the tumor is first formed (carcinogenesis); ii) the growth of the tumor according to some specified growth function, with an individual growth rate sometimes modelled as a random effect; iii) the process for symptomatic/clinical detection of the tumor; and iv) the possible early detection through mammography screening. Other processes which can be featured in natural history models include lymph node spread and distant metastasis, the promotion from in situ to invasive cancer, treatment, recurrence, and death.

Different models in the literature have included some selection of these components. For example, Bartoszyński et al. [98] defined a model for onset, growth, and symptomatic detection for a population without screening; Weedon-Fekjær et al. [99] modelled growth and sensitivity, but combined onset and symptomatic detection into a single process for breast cancer incidence between screenings; and Abrahamsson & Humphreys [100] modelled growth, symptomatic detection and screening sensitivity using the observed tumor sizes at diagnosis. Instead of a component for onset, they used a stable disease assumption, applicable to their cases-only design [101].

2.5.2.1 Onset/Carcinogenesis

The first event in the natural history of breast cancer—from the continuous growth point of view—is the onset of the tumor. The first mathematical model of cancer onset/carcinogenesis was proposed by Armitage & Doll [102] in 1954. They observed that cancer mortality rates looked approximately linear with age on the log-log scale, indicating a power law function for the rates. They postulated that a series of k mutations or events leads to the formation of cancer. Assuming that the events can occur in any order, they derived the hazard function at age *t*:

$$h(t) \approx \frac{p_1 p_2 \dots p_k t^{k-1}}{(k-1)!},$$

where the rate of each respective event is $p_1 p_2 \dots p_k$. This hazard leads to the familiar Weibull-distribution for age at onset [103].

Armitage & Doll [102] found that a slope *k-1* of magnitude between 5 and 6 gave a good fit to most cancers, except for the female reproductive cancers (breast, cervix, uterus), where there was a noticeable reduction after age 50.

A different model for onset is the Moolgavkar-Venson-Knudson (MVK) clonal expansion model [104–106]. Based on the Knudson two-hit hypothesis [107], it uses a Poisson process to model a cell population's transition from healthy to malignant, through an intermediate step. An illustration of the MVK model can be seen in Figure 5. From the healthy cell population, cells will experience the first event (at the rate $\tilde{\nu}$), creating "intermediate" cells. An intermediate cell can then either a) die or differentiate (at the rate $\tilde{\beta}$); b) divide into two intermediate cells (at the rate $\tilde{\alpha}$); or c) divide into one intermediate and one malignant cell (at the rate $\tilde{\mu}$). A malignant cell is then assumed to keep proliferating, eventually forming a tumor. Cancer onset is then defined as the time when the first malignant cell is formed.

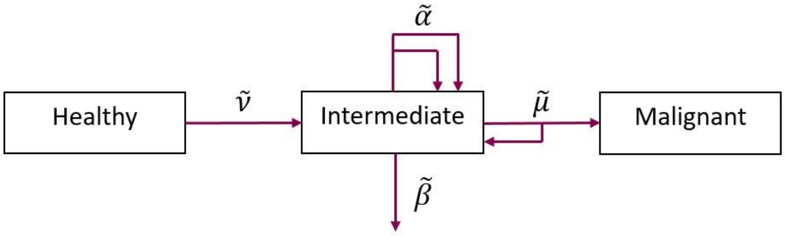


Figure 5: An illustration of the Moolgavkar-Venson-Knudson clonal expansion model. Onset is defined as the time when the first malignant cell is formed.

The problem with the MVK model is that its four parameters are not jointly identifiable from time-to-event data (e.g. incidence data). However, Heidenreich et al. [108] found a parameterization into three parameters which is:

$$A = \frac{1}{2}\left[\left(\tilde{\beta} + \tilde{\mu} - \tilde{\alpha}\right) - \sqrt{\left(\tilde{\beta} + \tilde{\mu} - \tilde{\alpha}\right)^2 + 4\tilde{\alpha}\tilde{\mu}}\right],$$

$$B = \frac{1}{2}\left[\left(\tilde{\beta} + \tilde{\mu} - \tilde{\alpha}\right) + \sqrt{\left(\tilde{\beta} + \tilde{\mu} - \tilde{\alpha}\right)^2 + 4\tilde{\alpha}\tilde{\mu}}\right],$$

$$\delta = \frac{\tilde{\nu}}{\tilde{\alpha}}.$$

With these parameters, the hazard function for onset is given by

$$h_T(t) = \frac{\delta AB\left(1 - e^{(B-A)t}\right)}{Be^{(B-A)t} - A},$$

and the survival function is

$$G_T(t) = P(T > t) = \left[\frac{(B-A)e^{Bt}}{Be^{(B-A)t} - A}\right]^{\delta}.$$

2.5.2.2 Tumor growth

Once breast cancer onset occurs, the cancer cells will proliferate, and the tumor will grow larger and larger. The simplest model for tumor growth is the exponential function given by

$$V_{exp}(x) = v_0 e^{\rho x}$$

at time *x* after onset, with a growth rate parameter $\rho > 0$ and starting volume of v_0 (usually a single cell, 10^{-6}mm^3). This model assumes a constant division rate for all tumour cells. The exponential model has been widely used to model breast cancer tumor growth [98,100,109,110].

The next tumor growth model is the Gompertz model. While it was first developed for modelling mortality in life insurance, it was later used by A.K. Laird [111] to model tumor cell proliferation. Notably, it has been used by L. Norton to specifically model breast cancer tumor growth [112–114]. This model differs from the exponential in that the growth rate decelerates over time, asymptotically reaching a maximum volume. The growth function is defined as

$$V_{Gz}(x) = v_0 \exp\left[\ln\left(\frac{v_{max}}{v_0}\right)(1 - e^{-\rho x})\right],$$

where v_{max} is the asymptotic maximum tumor volume. Norton [113] estimated v_{max} to have an approximate value of 10^6mm^3 (corresponding to 118mm in diameter).

Another model is a generalized logistic growth model, introduced for modelling breast cancer growth by Spratt et al. [115,116]. Like the Gompertz model, the growth rate decelerates until it reaches an asymptotic size v_{max}. However, the deceleration begins later than for the Gompertz function. The model is represented by the formula

$$V_{gl}(x) = v_{max}\left[1 + \left(\left(\frac{v_{max}}{v_0}\right)^{\beta} - 1\right)e^{-\beta\rho x}\right]^{-1/\beta},$$

with an additional parameter $\beta > 0$. The parameter β determines the growth deceleration, and the deceleration is greater for smaller values of β. Spratt et al. [115,116] found that the values $\beta = 1/4$ and $v_{max} = 1.1 \times 10^6\text{mm}^3$ gave the best fit to their breast cancer data. The same parameter values were later used by Weedon-Fekjær et al. [99,117] to analyze screening data.

Norton [113] found that the Gompertz model fitted breast cancer data better than the exponential function. However, his analysis was based on data from women who had already displayed symptoms and declined treatment [118]. Fournier et al. [119] analyzed breast cancers in a cohort of women screened at a high frequency. They failed to observe the dampening effect in a Gompertzian growth model, and found the growth rates to be approximately constant, which supports the exponential model. Talkington & Durrett [120] compared the different tumor growth models on data from multiple cancer sites, and also concluded that an exponential growth model is a better fit specifically for breast cancer tumor growth.

To account for the great variation in growth between individual tumors, one can model the tumor growth rate ρ as a random effect—a sample from some parametric distribution.

One distribution that has been used for growth rate random effects is the log-normal distribution. It is the distribution of choice in conjunction with either the Gompertzian [113] or generalized logistic [99,116] tumor growth models. The growth rate ρ then has the probability density function

$$f_{LN}(\rho) = \frac{1}{\rho\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln\rho - \mu)^2}{2\sigma^2}\right).$$

In the exponential growth model, it is more common to model the *inverse* growth rate $r = 1/\rho$, and to use a Gamma distribution as the random effects component [109,110,121]. The probability density function for R is then

$$f_\gamma(r) = \frac{b^a}{\Gamma(a)} r^{a-1} e^{-br}.$$

The inverse growth rate $R = r$ is related to the *tumor volume doubling time* through $DoublingTime = \ln(2)r$.

2.5.2.3 Symptomatic detection

Given enough time, a tumor will progress to a point where it will begin to display symptoms. For continuous tumor growth models the time *U'* from onset to symptomatic detection has been modelled [98,100,109,110] by using a continuous hazard function proportional to the current tumor volume. If *U'* is the time to symptomatic detection, then

$$P(U' \in (u, u + \Delta u] |\, U' > u) = \eta V(u) \Delta u + o(\Delta u),$$

for a hazard rate coefficient $\eta > 0$. Under this model, faster growing tumors surface sooner than slow growing tumors, on average.

Out of the tumor growth functions described above, only the exponential growth gives a closed-form expression for the survival function of *U'*. If we assume an exponential tumor growth, the survival function for symptomatic detection—given the inverse growth rate *r*—is

$$P(U' > u|r) = \exp\left(-\eta r v_0 \left(e^{u/r} - 1\right)\right).$$

Under these assumptions, the tumor volume at symptomatic detection follows a translated exponential distribution, with mean $(\eta r)^{-1}$ and offset v_0 [110]. If we also assume that the inverse growth rate is gamma-distributed, the marginal tumor volume at symptomatic detection follows a Pareto distribution [109].

2.5.2.4 Screening sensitivity

Between the times of breast cancer onset and symptomatic detection, there is an opportunity to detect the tumor early through mammography screening.

Hanin & Yakovlev [109] proposed—for continuous growth models—a model for a tumor being detected at mammography screening. They assumed the probability of a tumor of size *v* being detected at screening to be

$$P(Detection|v) = 1 - e^{-\beta v},$$

for $\beta > 0$, where *v* is the (latent) tumor volume. This is a probability which is similar to the symptomatic detection model above in that it is also proportional to the current tumor volume.

An alternative model was later used by Weedon-Fekjær et al. [99], who instead modelled the screening test sensitivity (STS) at the time of screening as a logistic function dependent on the (latent) tumor diameter. Abrahamsson & Humphreys [100] later extended this model to include the woman's mammographic density.

2.5.3 The CISNET consortium

In 2000, the US National Cancer Institute set up a consortium for studying the observed breast cancer mortality trends from 1975 to 2000 among U.S. women, and for estimating the effects of screening and adjuvant treatment. They named it the Cancer Intervention and Surveillance Network (CISNET). It consists of six research groups from different universities. Each has developed a different natural history model. One is an analytical multi-state model, but the others use microsimulation approaches with different model assumptions. The six research groups, and their respective natural history models, are:

- **Dana-Farber Cancer Institute:** The “DFCI” model is a 6-state Markov model, which differentiates between DCIS and invasive breast cancer, and includes breast cancer death. It is the only model which analytically estimates survival and overdiagnosis. [122]
- **Erasmus MC:** The "MISCAN-Fadia" model is a continuous growth microsimulation model. It assumes exponential tumor growth, and simulates tumour diameter thresholds at which symptomatic detection, screen detection, metastasis, and breast cancer death occur. [4]
- **Georgetown University/Albert Einstein College of Medicine:** The "Spectrum/G-E" model is a multi-state microsimulation model. Starting from the SEER incidence data, a gamma-distributed sojourn time is sampled and subtracted to determine the pre-clinical phase. The screening sensitivity component depends on age, mammographic density and screening round. [123]
- **MD Anderson Cancer Center:** The "MDACC" model is a microsimulation model, which uses approximate Bayesian computation to estimate incidence trends, stage shifts due to screening, and the effects of different treatments. [124]
- **Stanford University:** The "BCOS" model is a continuous growth microsimulation model. It assumes exponential tumour growth with gamma-distributed growth rates. The hazards of symptomatic detection and metastatic spread are proportional to the tumour volume. [125]
- **University of Wisconsin:** The "UWBCS" model is a discrete-event simulation model using 6-month intervals. It uses a Gompertz function for tumour growth, with an individual gamma-distributed growth rate. Screening sensitivity is assumed to be a function of size interval, age, mammographic density, and screening round. [126]

These models have been used, in a joint effort, to estimate the effects of screening and treatment on mortality [127]. More recently, the effects have been estimated based on molecular subtypes (ER and HER2 status) [3].

2.6 PREDICTION MODELS FOR BREAST CANCER

2.6.1 Relative Risk Models

Many statistical models for predicting the future risk of breast cancer exist. The arguably most well-known was introduced in 1989 by Gail et al. [128], and is commonly referred to as the **Gail model**.

The principle behind the Gail model is to first estimate relative risks/odds ratios of the risk factors to be included in the model using methods such as logistic regression, Cox proportional hazards models, or conditional linear regression. In the original model, a logistic regression model was used on the Women's CARE data [129] to estimate odds ratios of family history, age at menarche, age at first birth, and previous breast biopsies [128,130].

The estimated relative risks are then used to infer age-specific baseline hazard rates in external population data (i.e. the SEER data in the original model). Individual risk predictions in the larger population can then be made using the baseline hazard function and the individual risk factors. The US National Cancer Institute has implemented the original Gail model as an online tool called the Breast Cancer Risk Assessment Tool (**BCRAT**) [131].

The Gail model has also been extended to feature additional risk factors, such as HRT use, BMI and lifestyle factors [132], mammographic density [133], and PRS [134].

The Breast Cancer Surveillance Consortium (**BCSC**) has created an alternative risk prediction tool [135] based on the Gail model [136–138]. The major difference to BCRAT is the emphasis on mammographic density, where relative risks of each BI-RADS category has been estimated with a Cox proportional hazards model using the BCSC data. The other incorporated risk factors are family history and benign breast disease [137,139]. Some studies have also incorporated a PRS [138,140].

2.6.2 Pedigree Models

Another type of breast cancer risk prediction model focuses on the genetic risk of breast cancer—specifically the risk surrounding BRCA1 and BRCA2 mutation carriership. Two models, **BRCAPRO** [141,142] and **BOADICEA** (Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm) [143,144] use nearly identical approaches. Each starts by using detailed family histories of both breast cancer and ovarian cancer including the relation, cancer history, and age at diagnosis. The family trees (pedigrees) are then used to estimate the probability of carrying BRCA1, BRCA2, or carrying neither. The probability of having a specific BRCA phenotype given the observed family history is calculated using Bayes' theorem and the population prevalence of each phenotype.

The BOADICEA model also includes a random effect component in the proportional hazard, representing an unmeasured genetic risk, which now can be viewed as a PRS [145].

If the BRCA phenotype is known, the corresponding age-specific incidence rates are used for the risk prediction. Otherwise, the prediction is based on weighing the incidence rates of each phenotype based on the predicted probabilities of belonging to each phenotype.

Arguably the second most well-known breast cancer prediction model is the **Tyrer-Cuzick model** [146]. Starting from the same premise as BRCAPRO and BOADICEA, the Tyrer-Cuzick model uses the same detailed family histories and incidence rates based on BRCA phenotype, but adds an additional hypothetical gene (with lower penetrance) which is also inferred from the family history. The model then uses relative risks for the other risk factors (similarly to the Gail model).

The Tyrer-Cuzick model developed by its original authors currently operates under the name *International Breast Intervention Study* (**IBIS**) *Breast Cancer Risk Evaluation Tool* [147]. The IBIS risk model uses the detailed family history, age at menarche, age at first birth, age at menopause, height, BMI, benign breast disease history, HRT use, and mammographic density [148,149].

2.6.3 The Rosner-Colditz model

A third approach to risk prediction is taken by the **Rosner-Colditz model** [150]. The model is a non-linear Poisson regression model where the log-incidence depends on the reproductive states of menarche, first childbirth, each subsequent childbirth, and menopause. The concept is based on Pike's model for breast cancer incidence [151], where a woman accumulates "tissue age" at a rate depending on which reproductive state she is in. With each reproductive event, the rate decreases [150]. This is an alternative way of incorporating the protective effects of e.g. parity, late menarche, and early menopause. If we know a woman's precise information on when these reproductive events occur, her current "tissue age" can be determined, and her future risk can be predicted.

The original Rosner-Colditz model also included other risk factors in its log-incidence, such as family history, benign breast disease history, HRT use, BMI, and alcohol use [150,152]. Other risk factors have been studied, such as mammographic density and PRS [134]. Since the incidence is time-dependent, these risk factors can be time-dependent as well when such data is available (e.g. precise period of HRT use or pre- and postmenopausal BMI). The model has also been adapted to make subtype-specific risk predictions [153,154].

3 RESEARCH AIMS

This thesis set out to do the following:

- Develop a natural history model for breast cancer that can make use of the wealth of information available in detailed screening cohorts. This thesis was inspired by a previous natural history model developed for modeling tumor size distributions in a cases-only study design. With a cohort, there was an opportunity to study the onset of breast cancer, and to model both the patient age and tumor size at detection.
- Provide better understanding of the underlying processes and events that occur before a breast cancer is detected. In particular, to study—on a population level—the onset, growth, and detectability of breast cancer at mammography. It was also of interest to see how well-established risk factors, known to influence the final incidence of breast cancer, can be separated into factors for the latent processes leading up to the incidence.
- Understand the interplay between the processes involved, and how mammography screening shifts the time of detection and changes the outcome. In doing so, we can study the effect screening attendance has on an individual level, and—before commiting resources to a clinical trial—we can glimpse at the effects of changing the screening patterns.

4 MATERIALS AND METHODS

4.1 DATA SOURCES

4.1.1 KARMA

The data featured in this thesis is the Karolinska Mammography Project for Risk Prediction of Breast Cancer (KARMA) [155]. KARMA is a Swedish prospective breast cancer screening cohort. Women attending their mammography screening at four hospitals (Stockholm South General, Helsingborg, Skåne University, and Landskrona) between January 2011 and March 2013 were invited to participate. 70 877 out of the 210 233 invited women joined.

At baseline, the participants answered a detailed web-based questionnaire related to breast cancer risk factors such as reproductive history and various lifestyle factors. The images taken at screening are continuously collected and analyzed as the women continue to attend the screening programme. The women are also continuously matched to national breast cancer registry data to update their disease status. Blood samples were also collected, and as a part of the Breast Cancer Association Consortium, approximately 20 000 women have been genotyped.

4.2 STATISTICAL METHODS

This thesis centers around the development and application of a new statistical model for the natural history of breast cancer. To avoid repetition, the model is not described in detail here– it has already been implicitly introduced in Section 2.5.2, and is a significant part of the results of Study I (Section 5.1). Please refer to these sections or any of the four studies for the model description. Here, a recapitulation is provided of a few important concepts and formulae that are helpful for understanding the model as it is presented in the studies.

4.2.1 Conditional Probabilities

Let X and Y be two random variables with potential outcomes in the respective domains D_X and D_Y. We denote the *joint probability* of $X = x$ and $Y = y$ as $P(X = x, Y = y)$. The conditional probability of Y *given* X, denoted $P(Y|X)$, is defined as

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)},$$

where in the second equality we have used the first equality and the symmetry of the joint probability. This is also known as *Bayes' Theorem*.

We can get the *marginal* probability of $X = x$ from the joint probability by summing over all values of D_Y:

$$P(X = x) = \sum_{y \in D_Y} P(X = x, Y = y) = \sum_{y \in D_Y} P(X = x|Y = y)\, P(Y = y).$$

If Y is a continuous variable and its domain D_Y is an interval (a, b), then the summation turns into an integral over the probability density function of Y, $f_Y(y)$:

$$P(X = x) = \int_a^b P(X = x, Y = y)\mathrm{d}y = \int_a^b P(X = x|Y = y)f_y(y)\mathrm{d}y.$$

In this thesis, the notation $P(Y = y)$ is used for both discrete and continuous variables. The variable type is implied from the context in which the variable is introduced.

These concepts can be collected into the following set of useful equalities:

$$\begin{aligned} P(Y = y|X = x) &= \frac{P(X = x, Y = y)}{P(X = x)} = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)} \\ &= \frac{P(X = x|Y = y)P(Y = y)}{\sum_{y \in D_Y} P(X = x|Y = y)P(Y = y)}. \end{aligned}$$

4.2.2 Left Truncation

Truncation is a type of bias that occurs when sampling, whereby some outcomes of a random variable Y are not included. Left truncation specifically refers to the exclusion of smaller values (from the left side of the random variable's domain). Instead of observing Y with probability function $P(Y = y)$ we observe the conditional probability of Y given that $Y > c$ with probability function $P(Y = y|Y > c)$ for some truncation threshold c.

Left truncation often occurs when a variable relates to time. The example which is relevant to this thesis is that recruitment for the KARMA study was restricted to a limited period of time, and to women attending screening. The data was also restricted to the first breast cancer diagnosis, which had to occur after being recruited. This meant that women under the age of 40 (who were not yet eligible for screening), and women diagnosed with breast cancer before the recruitment period were not included in the study. There is therefore left truncation with respect to the age at detection in the data, since the age at detection is observed conditional on not being detected before the study started.

4.2.3 Maximum Likelihood Estimation

Assume that we have a set of random variables $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$ with a joint probability function which is known except for a set/vector of k parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$. We can represent this by writing $p(\boldsymbol{X}; \boldsymbol{\theta})$. The task is to use the observed outcome $\boldsymbol{X} = \boldsymbol{x}$ to estimate the parameter vector $\boldsymbol{\theta}$.

For each possible $\boldsymbol{\theta}$, the likelihood (joint probability) of the observed outcome $\boldsymbol{X} = \boldsymbol{x}$ will be different. Instead of viewing the joint probability as a function of $\boldsymbol{x}$, we can therefore consider it as a function of $\boldsymbol{\theta}$. We call this the likelihood function of $\boldsymbol{\theta}$, defined as

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}).$$

According to the *maximum likelihood principle* [156], the correct estimate of $\boldsymbol{\theta}$ is the one that maximizes the likelihood function. The intuition for this is that we should assume that the observed outcome $\boldsymbol{X} = \boldsymbol{x}$ is the most common of all the possible outcomes (at least on average).

The maximum likelihood estimate (MLE), $\widehat{\boldsymbol{\theta}}_{MLE}$ is therefore the vector among the possible $\boldsymbol{\theta}$ that maximizes the likelihood function, i.e. formally

$$\widehat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}; \boldsymbol{x}).$$

For computational reasons, the natural logarithm of the likelihood function, $l(\boldsymbol{\theta}; \boldsymbol{x}) = \ln\big(L(\boldsymbol{\theta}; \boldsymbol{x})\big)$, is usually maximized instead. The MLE is the same for both functions.

4.2.3.1 Confidence Intervals

It can be shown that the MLE $\widehat{\boldsymbol{\theta}}_{MLE}$ asymptotically follows a multivariate normal distribution with mean $\boldsymbol{\theta}_{MLE}$ and variance $(-H)^{-1}$, where

$$H = \begin{pmatrix} \frac{\partial^2 l}{\partial \theta_1^2} & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 l}{\partial \theta_2^2} & \cdots & \frac{\partial^2 l}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial \theta_1 \partial \theta_k} & \frac{\partial^2 l}{\partial \theta_2 \partial \theta_k} & \cdots & \frac{\partial^2 l}{\partial \theta_k^2} \end{pmatrix}$$

is the hessian of the log-likelihood function with respect to $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}_{MLE}$. For each $\theta_j \in \boldsymbol{\theta}$, the standard error s_j is given by $\sqrt{(-H)^{-1}_{j,j}}$, i.e. the square root of the *j*:th diagonal element of $(-H)^{-1}$. This can be used to construct an approximate $1 - \alpha$ confidence interval for each θ_j:

$$\widehat{\theta_j} - z(1-\alpha/2)s_j < \theta_j < \widehat{\theta_j} + z(1-\alpha/2)s_j.$$

4.2.4 Likelihood-Ratio Tests

The likelihood ratio test is a goodness-of-fit comparison between two models M_1 and M_2 that are *nested* in the sense that they are both from the same family of model (e.g. linear regression models) and that the set of free/unknown parameters of one model is a proper subset of the free/unknown parameters in the other. If $M_1$is nested in M_2, then M_2 has the parameter vector $\boldsymbol{\theta_2} = (\theta_1, \dots, \theta_{k-d}, \theta_{k-d+1}, \dots, \theta_k)$ with *k* free parameters, and M_1 has the equivalent parameter vector $\boldsymbol{\theta_1} = (\theta_1, \dots, \theta_{k-d}, c_1, \dots, c_d)$ with *k*-*d* free parameters and *d* fixed parameters. Being nested, the two models have the same likelihood function $L(\boldsymbol{\theta}; \boldsymbol{x})$ with a parameter vector $\boldsymbol{\theta}$ of length *k*.

The likelihood-ratio (LR) can be constructed by maximizing L twice—first over $\boldsymbol{\theta_1}$ and then over $\boldsymbol{\theta_2}$—and then taking the ratio of the two maximized likelihood values:

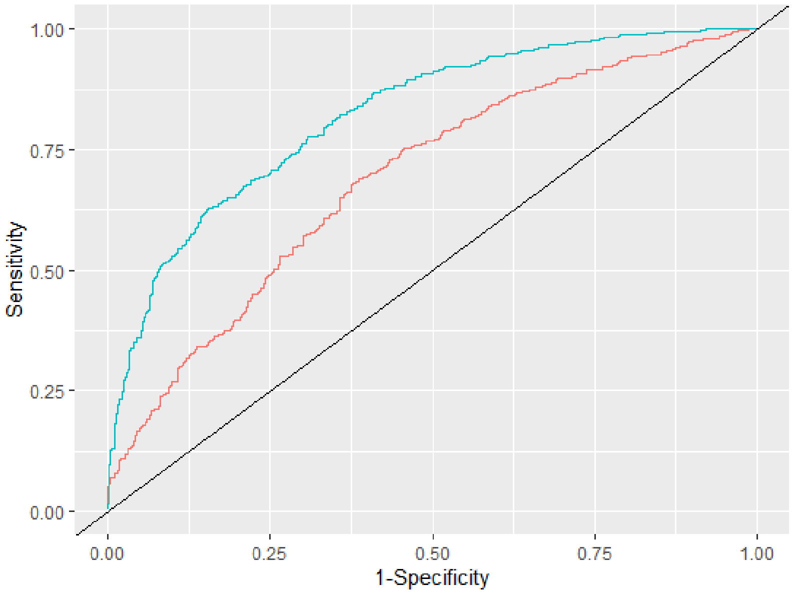
$$LR = \frac{L(\widehat{\boldsymbol{\theta}}_1; \boldsymbol{x})}{L(\widehat{\boldsymbol{\theta}}_2; \boldsymbol{x})}.$$

It can be shown [157] that

$$-2\ln(LR) = -2\left(l(\widehat{\boldsymbol{\theta}}_1; \boldsymbol{x}) - l(\widehat{\boldsymbol{\theta}}_2; \boldsymbol{x})\right)$$

is approximately χ^2-distributed with d degrees of freedom. This statistic can therefore be used to test if M_2 is a statistically better fit to the data than M_1 (i.e. if the maximum likelihood value is significantly greater with M_2 than with M_1). In other words, we can test if adding $\theta_{k-d+1}, \theta_{k-d+2}, \dots, \theta_k$ as free parameters to M_1 provides a better fit than keeping them fixed at $c_1, c_2, \dots, c_d$.

4.2.5 Receiver Operating Characteristics



Let us assume that we have a binary classification problem, where there is some binary outcome (i.e. positive or negative) and a classifier/test attempting to correctly classify/predict the outcome. The positive outcomes that are correctly classified as positive are called *true positives* and the proportion of positive outcomes that are true positives is called the *sensitivity* of the classifier. Similarly, the negative outcomes correctly classified are called *true negatives* and the proportion of negative outcomes that are true negatives is referred to as the *specificity*.

If the outcome of the classifier is continuous—like in a cancer risk prediction model where each person receives a predicted risk between 0-1—a threshold value is chosen, and those scoring above the threshold are classified positive, and those scoring below are classified negative.

The sensitivity and specificity are related in the sense that the more generous the threshold is for a positive classification, the number of true positives surely increases, but the number of true negatives also decreases. To study this behavior, one can plot the sensitivity against one minus the specificity for each possible threshold. This generates an increasing curve called the *receiver operating characteristic* (ROC) curve for the classifier [158]. This allows one to study the classifier's performance at different thresholds, and to also compare different tests by comparing their ROC curves. To get an overall assessment of the performance over all thresholds, one can calculate the area-under-curve (AUC) of the ROC curves.

A diagonal line from (0, 0) to (1, 1) is often included in the ROC curve as a reference. This line represents a test which classifies by pure chance (and has an AUC of 0.5).

5 RESULTS

5.1 STUDY I

Study I focused on defining the natural history model, deriving its likelihood function, and adjusting it for random left truncation.

The model is a combination of four submodels, each representing a latent process in the natural history of breast cancer. Each of these submodels have already been briefly described in Section 2.5.2:

Process	Submodel	Unknown parameters
Onset/Carcinogenesis	Moolgavkar-Venson-Knudson model.	A, B, δ
Tumor growth	Exponential growth function with gamma-distributed inverse growth rates.	a, b
Symptomatic detection	Continuous hazard function proportional to the latent tumor volume.	η
Screening sensitivity	Logistic function of latent tumor diameter.	β_0, β_s

The four submodels are combined into an individual likelihood per woman jointly based on the observed age at detection, observed tumor size at detection, and mode of detection (screening vs. symptomatic). In Study I, the expressions for these individual likelihood contributions were derived separately for screen-detected cases, symptomatically detected cases, and censored (no detected breast cancer). The individual likelihood for being censored could be used to make the adjustment for left truncation.

Study I was published before we had access to the KARMA data. Therefore, parameter values were taken from another study combined with parameter values for the MVK onset model that calibrated well to the cumulative risk according to GLOBOCAN [159]. Since we now have the parameter estimates from KARMA, we can update the results of Study I with these new values.

Figure 6 presents the conditional tumor doubling time distributions for screen-detected tumors (in a biennial screening programme between age 40-74) and interval-detected symptomatic tumors (between age 40-76). It shows the distributions both from simulating a data set (histograms) and from the conditional growth rate formulas derived in the study (lines). The median doubling time is now estimated to be 218 days for screen-detected cases (quartiles: 132 & 339 days), and 134 days for interval cases (quartiles: 71 & 248 days). The purpose of this result is to show that slower tumors are more likely to be screen-detected than faster tumors, simply because they are present for more screening rounds on average.

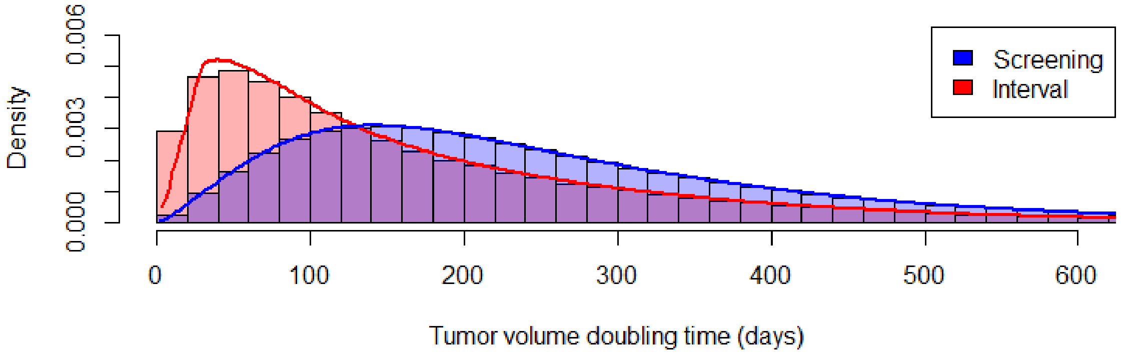


Figure 6: Tumor volume doubling time distributions for screen-detected and interval cancers.

Figure 7 shows the time from onset to symptomatic detection (in the absence of screening), stratified by the age at detection. It shows boxplots from simulations, and the means from both simulations and formula. The mean tumor presence time for cancers detected at age 40 is 7.3 years, and for cancers detected at age 75 the mean time is 10.4 years. The purpose is to show that, even if the tumor growth rate (and thus tumor presence time) is assumed to be independent of the age at onset (as in this study), it is not independent of the age at detection. Faster growing tumors are detected sooner, so with time (age) the slower growing tumors accumulate in the population, and the tumors detected at older ages are found to be slower growing on average than at younger ages.

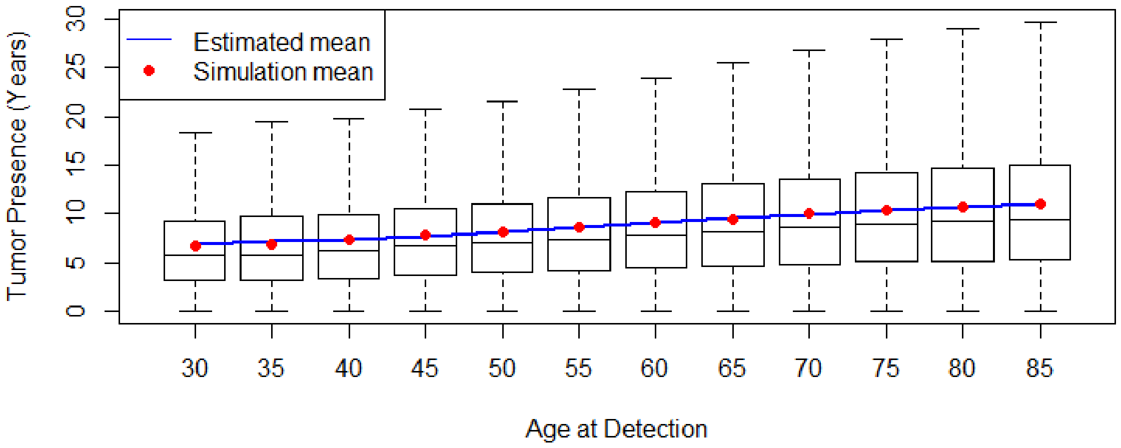


Figure 7: The tumor presence times (from onset to detection in the absence of screening) stratified by age at detection.

5.2 STUDY II

Study II focused on fitting the natural history model developed in Study I to a mammography screening cohort, i.e. KARMA. It also altered the model so that the inverse growth rate depended on the age at onset. From here on, the parameterization of the gamma-distributed inverse growth rates was changed to its mean-value parameterization with the new parameters $\theta = a/b$ and $\phi = 1/a$. This was done so that dependency of the inverse growth rate on the age at onset could be incorporated through the link $\ln\theta = \theta_0 + \theta_1 t$. The parameter estimates are found in Table 1.

Parameter	Estimate	95% CI
A	$-7.22 \cdot 10^{-2}$	(-3.73, -14.00)
B	$1.18 \cdot 10^{-3}$	(0.65, 2.16)
δ	$9.52 \cdot 10^{-2}$	(2.10, 43.19)
ln(η)	-8.82	(-8.98, -8.66)
β_0	-4.99	(-5.39, -4.60)
β_s	0.49	(0.43, 0.55)
β_1 (PD)	-2.09	(-2.93, -1.26)
ϕ	0.56	(0.46, 0.68)
exp(θ_0)	0.52	(0.22, 1.23)
exp(θ_1)	1.011	(0.997, 1.025)

Table 1: The parameter estimates from Study II, including the inverse growth rates dependency on the age at onset through θ_1.

Figure 8 displays some results from estimating the tumor growth submodel. On the left side are the distributions of the tumor volume doubling times (calculated as $\ln 2$ times the inverse growth rate) for tumors with onset at age 40 (top) and onset at age 60 (bottom). The estimated median tumor volume doubling time for tumors with onset at age 40 was 0.46 years or 167 days, and 0.56 years or 207 days for tumors with onset at age 60. The time it takes for a tumor growing at the median tumor doubling time to reach the median tumor size at detection (15mm) from 0.5mm is, according to the parameter estimates, 6.7 years for women with onset at age 40, and 8.4 years for women with onset at age 60.

On the right-hand side of Figure 8 are estimated growth curves for tumors with different inverse growth rates, starting from 10mm at time 0, for tumors with onset at age 40 and 60 respectively for the top and bottom panel. The lines represent the median (solid), 25th & 75th percentiles (dashed) and the 5th & 95th percentiles (dotted) of the inverse growth rate distributions.

In addition to estimating the tumor growth, Study II produced estimates for the age at onset and screening sensitivity. The respective fitted distribution and function is presented in Figure 9. It was estimated that 13.4% of women experience breast cancer onset by age 75. The screening sensitivity for a 13mm tumor (the median size for screen-detected cases) was estimated to be 0.73 for women with a PD of 13%. The sensitivity reduced to 0.58 for women with a PD of 50%, and increased to 0.79 for women with a PD of 2%.

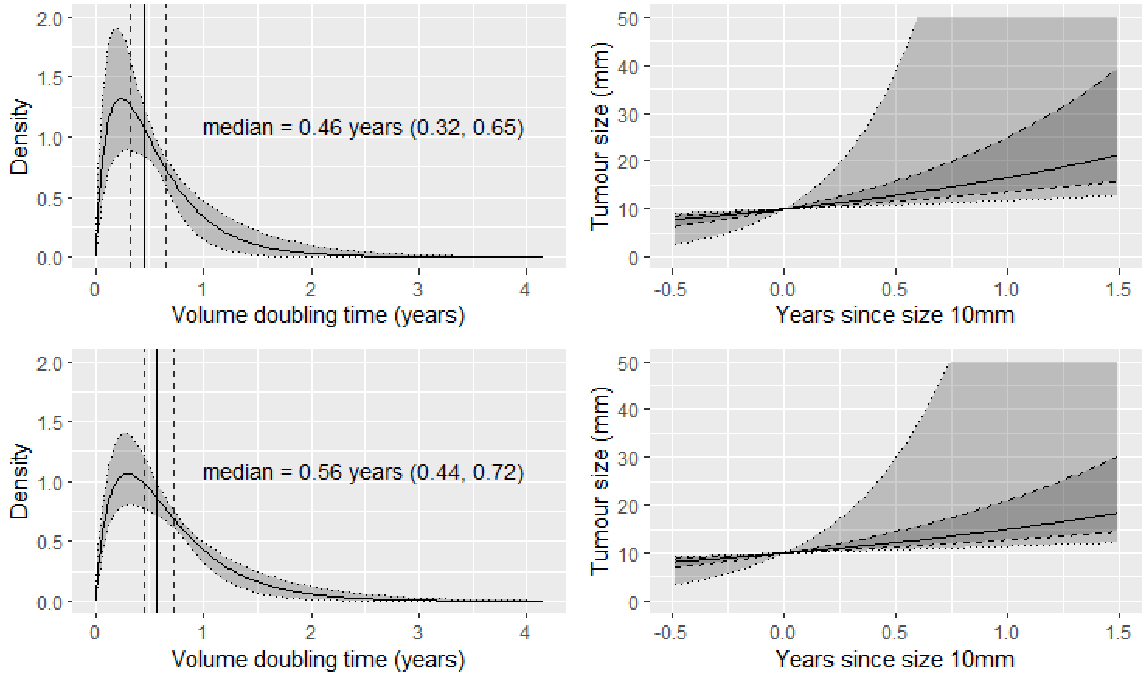


Figure 8: Fitted tumor volume doubling time distributions with 95% confidence regions (left) and tumor growth curves (right), for tumors with onset at age 40 (top) and onset at age 60 (bottom).

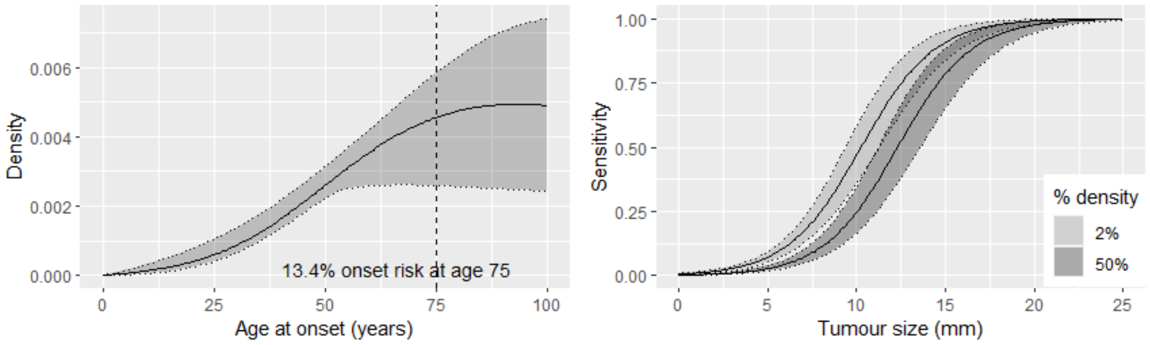


Figure 9: *Left:* The fitted distribution for the age at onset with 95% confidence region. *Right:* Fitted screening sensitivity functions for 2% and 50% PD.

5.3 STUDY III

Study III turned the model into a risk prediction model for breast cancer. It used the data from Study II as a training set and included an additional 26 months of follow-up as a validation set.

The risk factors used in the prediction model—and which submodel they belonged to—are listed in Table 2. To include PD (which greatly changes with age) in the onset submodel based on only one measurement, we defined the age-specific quartile of PD (AQPD). The continuous variables were scaled to have mean 0 and standard deviation 1. In the table are also their estimated coefficients in their respective regressions, and their respective p-values.

Parameter	**Estimate**	**95% CI**	**exp(Est)**	**p-value**
Age at onset				
AQPD = 1	0	Reference	1	
AQPD = 2	0.625	(0.426, 0.825)	1.869	<0.001
AQPD = 3	0.764	(0.556, 0.972)	2.147	<0.001
AQPD = 4	0.917	(0.696, 1.137)	2.501	<0.001
Family history (1st deg.)	0.537	(0.382, 0.692)	1.711	<0.001
Family history (2nd deg.)	0.381	(0.171, 0.590)	1.463	<0.001
Benign breast disease	0.249	(0.110, 0.388)	1.283	<0.001
HRT (current use)	0.504	(0.260, 0.748)	1.656	<0.001
Parity (y/n)	0.012	(-0.178, 0.203)	1.012	0.899
Age at 1st childbirth	0.097	(0.027, 0.167)	1.102	0.006
Age at menarche	-0.045	(-0.110, 0.020)	0.956	0.179
BMI	0.181	(0.107, 0.254)	1.198	<0.001
Tumor growth				
BMI	-0.057	(-0.166, 0.052)	0.945	0.304
Symptomatic detection				
Total breast area	-0.336	(-0.469, -0.202)	0.715	<0.001
Screening sensitivity				
Percent density	-0.497	(-0.667, -0.328)		<0.001

Table 2: List of the risk factors used in the prediction model, with parameter estimates.

The overall risk prediction performance is summarized in Figure 10. The AUC for the ROC was 0.64. The mean predicted risk 0.056 while the observed incidence was 0.060 over the 26-month validation follow-up. When separating the predicted risks by observed outcome, the cases had an average predicted risk of 0.0075 and non-cases had 0.0056. Cases had on average 1.33 times the predicted risk of the non-cases.

In addition to predicting the overall risk of breast cancer detection, the model can make more specific predictions. This study derived formulas for making predictions for experiencing onset (including currently having an undetected tumor), detecting a tumor at the next screening or in the interval between, and the probability of having a tumor detected below a certain size depending on screening attendance.

Figure 11 displays the predicted probabilities of a detected tumor within the 26-month validation follow-up being less than 10mm in diameter, comparing the probabilities if the women attend their next screening or not. The women were separated into groups according to their AQPD. Overall, the average probability went up from 0.09 if not attending, to 0.22 if attending the screening round. For the lowest AQPD, the probabilities were 0.08 versus 0.26, and for the highest AQPD it was 0.10 versus 0.17.

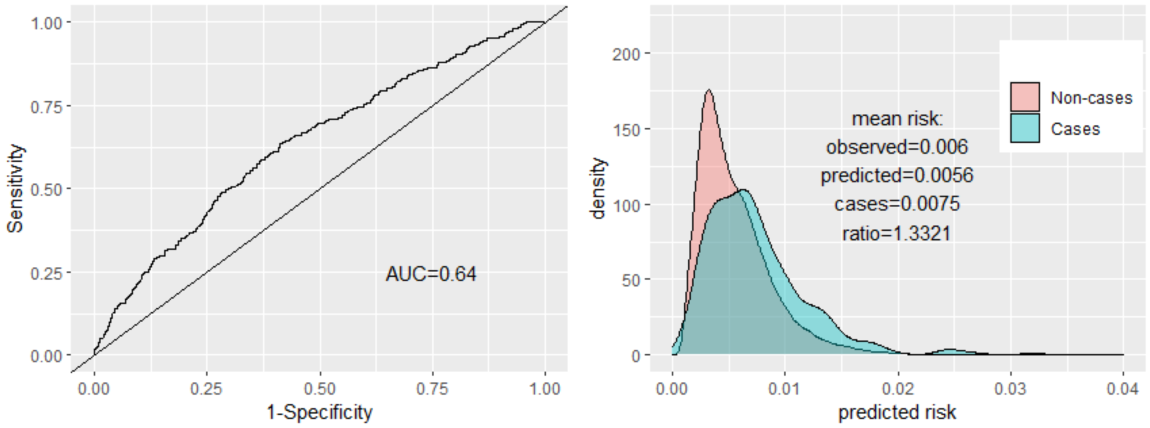


Figure 10: Performance of the risk prediction. Left: ROC curve. Right: Distributions of the predicted risks, separated by the observed case status.

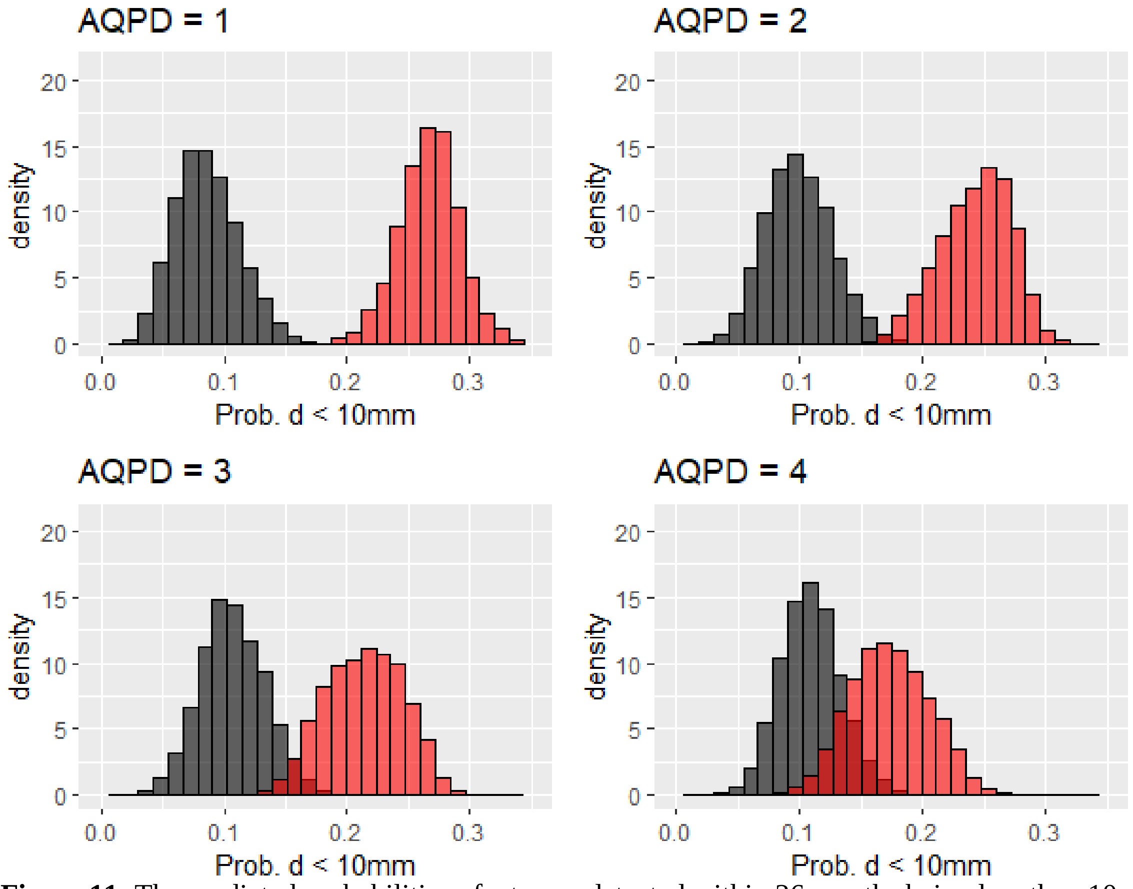


Figure 11: The predicted probabilities of a tumor detected within 26 months being less than 10mm in diameter, if attending the next screening (red), or not attending (grey). One panel for each category of AQPD.

5.4 STUDY IV

This study delved deeper into the mammography screening sensitivity. Several factors at the mammography determine the final image result. The study investigated the effects that some of these acquisition parameters—namely the compressed breast thickness (CBT), breast compression pressure (CP) and total x-ray exposure (EXP)—could have on the screening sensitivity. The resulting parameter estimates of the screening sensitivity submodel are found in Table 3. It was found that, in addition to percent density (PD), CBT had a statistically significant effect, with greater thickness being associated with a reduced screening sensitivity.

Parameter	Estimate	95% CI	Scaled	p-value
β_0 (intercept)	-2.948	(-4.184, -1.711)		
β_s	0.522	(0.461, 0.582)		
PD	-3.522	(-4.620, -2.425)	-0.660	<0.001
CBT	-0.272	(-0.423, -0.121)	-0.340	<0.001
CP	-0.135	(-0.760, 0.490)	-0.052	0.672
EXP	-0.004	(-0.010, 0.003)	-0.093	0.240

Table 3: The estimated parameters for mammography screening sensitivity.

Figure 12 compares the estimated screening sensitivity functions when the model includes both PD and CBT versus only including PD (as in Study II and Study III). We can see that the addition of CBT helps separate women with low and high screening sensitivity.

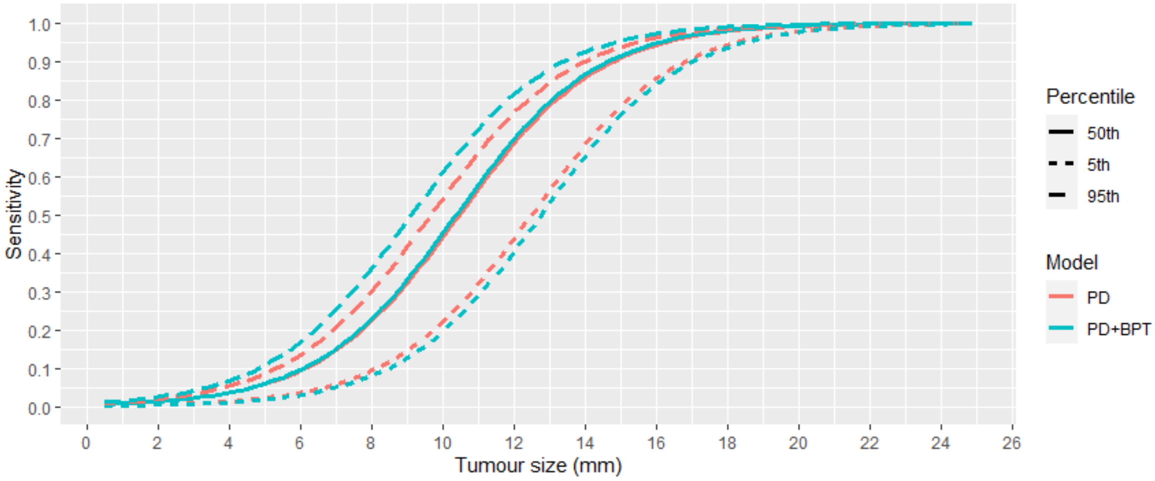


Figure 12: The estimated screening sensitivity functions for different percentiles of PD, comparing models with and without CBT.

The study also shows examples of mammograms with estimated sensitivities and how they change when factoring in CBT.

6 DISCUSSION AND FUTURE PERSPECTIVES

The work presented in this thesis should be thought of as a beginning. The foundations of a natural history model have been laid out, and examples of how it could be useful have been presented. What follows are some thoughts on what can be improved, and where to go from here.

We recognize that the model in the thesis relies on strong parametric assumptions. These strong assumptions are necessary in order to piece together the underlying natural history based on only data from diagnosis. Our confidence in the assumptions is strengthened by the biological motivations behind them. The MVK model for onset has a direct cancer-based origin. For the tumor growth, the exponential function has been shown *in vivo* to hold reasonably [119,120], and the median tumor doubling times estimated in this thesis are close to those estimated in *in vivo* studies [119,160]. Multi-state Markov models, for example, also rely on strong assumptions, but do not have the same biological justifications behind them.

The model featured in this thesis uses the Moolgavkar-Venson-Knudson model for carcinogenesis as its onset component. It is based on the Knudson two-hit hypothesis [107] which is in turn based on the apparent heritability of retinoblastoma. While it is now known that the two-event hypothesis is an old and outdated concept, what we were primarily after was a flexible function that could resemble a time-shifted breast cancer incidence curve. The MVK model gave us just that, while also having a connection to cancer biology. It also has closed expressions for its hazard, survival, and density functions.

One could conceivably develop an alternative onset submodel which requires more than two events. The Armitage-Doll model of carcinogenesis [99] can handle any number of events, but has been shown to fit poorly to breast and ovarian cancer incidence data [102]. Currently, ten major ‘hallmarks’ of cancer have been identified [161], and thus a model with up to ten events using the same construction as the MVK model could be motivated. But its utility is questionable when the focus is on studying the post-onset natural history or when studying breast cancer screening.

The submodel that we identify as needing improvement the most is the one for symptomatic detection. It has been borrowed from previous natural history models, and has a few convenient properties when combined with the exponential growth and gamma random effects [110]. For example, if we view the volume as being proportional to the third power of the diameter, one could experiment with different powers. A more general relationship between the hazard rate and tumor size could involve e.g. polynomials or cubic splines. Further developments to this submodel could lead to better calibration to the observed tumor sizes among symptomatically detected cases (Study II), and to better mode-specific risk predictions (Study III).

To model cancer detection through mammography screening, we assumed that the screening sensitivity followed a logistic function of tumor diameter. In a study without parametric assumptions for this kind of screening sensitivity, Wang et al. [162] found that the logistic function both underestimates the sensitivity for small tumor sizes and overestimates the sensitivity for large tumors—even suggesting that the sensitivity does not reach 1 until after tumors reach 70mm. An alternative function with these features could possibly replace the logistic function. We can anecdotally note that replacing the diameter in the logistic function with the logarithm of the diameter would behave in this way. However, this appears to lead to optimization issues in the model likelihood (when it is combined with the other submodels). More investigation is necessary.

In its current state, the model only considers the primary tumor, i.e. the T-stage of the cancer. While the tumor size at diagnosis is associated with the prognosis of breast cancer [163,164], it is metastasis which predominantly determines the lethality of the disease. Isheden et al. [165] has developed a model for the number of affected lymph nodes at diagnosis, compatible with a continuous growth model for the primary tumor, which calibrated well to breast cancer case data. Recently, Gasparini & Humphreys [166] developed a similar model for the seeding and presence of distant metastasis. If these models for the spread of breast cancer can be successfully incorporated into the model presented in this thesis, stage-specific breast cancer predictions can be made. In Study III, we made predictions relating a specific stage-shift for the T-stage, namely about the tumor being less or greater than 10mm at detection, depending on whether or not a woman attended her next screening round. With these proposed extensions for regional and distant metastasis, other highly relevant types of stage-shifts of the N- and M-stages can be studied (e.g. shifting from M1 to M0 or from N1+ to N0). These types of stage-shifts are important when evaluating screening, and thus also important when considering risk-based screening.

Another important prognostic factor that is absent from the model is the molecular subtype of the tumor. Currently, the model simply assumes that some of the heterogeneity in the (inverse) tumor growth rates is due to the tumors belonging to different subtypes, and that the subtypes can be considered to occupy different sections of the gamma distribution. The main reason why this has not yet been incorporated into the model is that the subtype is unknown for the latent tumors, which makes the implementation less straight-forward. One could conceivably treat each subtype as competing risks, where the first to lead to onset determines the subtype of the final tumor. If one is to use the MVK model for onset it is unclear if, for example, the subtypes should be incorporated as separate states within one MVK-like model, or if they should have separate MVK models. Also, breast cancers in younger women are associated with more aggressive subtypes at diagnosis [167,168], suggesting that the relative hazard rates for each subtype should differ with age.

Once the onset for each subtype has been handled, the differences in aggressiveness between the subtypes can then be formalized in the model with different growth rate distributions for each (e.g. gamma distributions with different means). This could be done either for the four intrinsic subtypes (i.e. luminal A/B, HER2-enriched, and basal-like) or for specific receptor statuses (e.g. ER, PR, and HER2 status).

Menopausal status and the age when menopause occurs is a possible extension to the onset model which is ready to be implemented. In this case, the hazard function of the MVK model (specifically the parameter δ) could be extended to be piece-wise constant—either with a discontinuity point at 45 or 50 as a proxy for the menopausal transition, or at the specific age at menopause when it is known. The same idea can be extended to age at menarche and age at first birth (which are currently time-constant effects in the onset model). The resulting hazard function would then be reminiscent of Pike's model [151] (which is also the basis for the Rosner-Colditz risk prediction model [150,169]), except for modeling onset rather than incidence.

This thesis heavily featured mammographic percent density—both as a masking factor in the screening sensitivity, and as a risk factor for onset. It is well-known that PD greatly changes with age [52,53]. In Study II, we had only a single measurement of PD per woman at baseline. In Study III, where we wanted to use PD as a risk factor for breast cancer onset, we used the baseline measurement of PD and the age at baseline to define the age-specific quartile of PD (AQPD). Since we could not assume that the PD would be constant with age, we instead assumed that the assigned AQPD was constant with age (i.e. if a woman was in the top quartile at age 50 among the women aged 50, she was also in the top quartile at age 30 among the women aged 30, etc.). In Study IV, we had access to multiple longitudinal measurements of PD (and the other acquisition parameters).

To directly and properly solve the above issue with using PD in the onset submodel, we need information about PD from before the screening starts. With multiple longitudinal measurements of PD per woman, it might be viable to jointly model the individual PD with the rest of the model, for example by assuming it follows a Gompertz or generalized logistic function with individual random effects [53].

Eriksson et al. [170] developed a breast cancer risk prediction model which included the presence of microcalcifications and masses in the breast as novel features. With these additions their prediction model managed to achieve an AUC of 0.71—a considerable improvement compared to alternatives without. This provides a possibility for more dynamic risk prediction, whereby findings at a negative mammogram can be used to inform the next screening interval or modality. It could therefore be of interest to jointly model the formation of these image features together with the natural history model in this thesis. This could for example be done through a Poisson process where the rate depends on the presence of and/or the size of the tumor.

While we know of the reduction in breast cancer mortality due to screening, it is also known that the screen-detected breast cancers have less aggressive properties (e.g. lower grade, ER+, HER2-) on average than interval cancers detected between screening rounds [171,172]. This phenomenon relates to a type of bias called *length bias*, which refers to the length of time that cancer can be detected by screening (before being detected e.g. symptomatically). This time is, on average, longer for the less aggressive cancer subtypes, which allows for more opportunities to detect those cancer early, compared to more aggressive subtypes. Thus, screen-detection is biased towards slow-growing and less aggressive tumors. Study I (see Figure 6 above) presented conditional tumor volume doubling time distributions for screen-detected and interval cancers. We could see a clear selection for slower growing tumors among the screen-detected cancers—merely by the mechanics from the assumptions made in the model. This shows that the model can be a useful tool when studying length bias and interval cancers.

The main concern surrounding mammography screening programmes is *overdiagnosis*—defined as the proportion of breast cancer cases which are screen-detected but would not have been detected without screening (symptomatically) in the woman's lifetime, i.e. where the order of events is screen-detection, death from other causes, and then hypothetical symptomatic detection. These breast cancers are therefore not clinically relevant to the health of the woman, and the unnecessary treatment in these cases is only harmful.

The model presented in this thesis jointly models both processes of symptomatic detection and mammography screen-detection. With it, a formula can be derived for the conditional distribution of the time to (would-be) symptomatic detection given age and tumor size at screen-detection (similarly to what has been done for a similar natural history model [173]). If this is combined with competing risks of death from other causes, this can lead to a formulation for the probability of overdiagnosis by considering the conditional probability that death occurs before would-be symptomatic detection, given age and tumor size at screen-detection.

7 CONCLUSIONS

This thesis has introduced a new biologically motivated natural history model for the onset, growth and detection of breast cancer. The model can incorporate a wide range of risk factors for breast cancer and can study their effects on specific processes in the natural history.

The model is also capable of making risk predictions, with similar performance as the current roster of breast cancer risk prediction models. The model stands out in that it can also make more detailed predictions regarding the mode of detection and tumor size at detection. Since it models the underlying processes, it can even predict counterfactual events, such as the outcome depending on if a woman attends her next screening or not.

By combining the natural history modelling framework with the risk prediction framework, and by further studying which factors determine the detectability of breast cancer at mammography, we might be able to make better informed assessments about mammography screening and decisions about individualized screening.

8 ACKNOWLEDGEMENTS

Thank you to **Keith Humphreys**, my principal supervisor. Thank you for giving me this opportunity to grow as a statistician and as an academic researcher. Thank you for letting me share my own ideas, and for reining them in when necessary. Thank you for all the discussion, the ups and downs, the guidance, and the encouragement from start to finish.

Thank you to my co-supervisors **Kamila Czene** and **Per Hall**. Thank you for letting me tap into your vast knowledge of breast cancer and for bringing a fresh and practical perspective to my otherwise very methodological thesis.

Thank you to my mentor **Therese Andersson.** Thank you for the discussions and advice on becoming a researcher in biostatistics.

Thank you to **Mark Clements**, for your advice and suggestions on how to improve the computation times of my programs. Maybe someday they will find their way onto GitHub.

Thank you to **Marie Reilly** for taking me in as a master's student and for providing the introduction to the biostatistics group at MEB.

Thank you to the members of my research group, past and present: **Linda Abrahamsson**, **Maya Alsheh Ali**, **Benjamin Christoffersen**, **Alessandro Gasparini**, **Gabriel Isheden**, and **Letizia Orsini**. Thank you for welcoming me and for sharing ideas. Thank you for letting me work next to such great minds.

Thank you to my office mates, past and present: **Elizabeth Dahlqwist**, **Pablo Gonzales Ginestet**, **Shuang Hao**, **Andreas Karlsson**, **Þorgerður Pálsdóttir**, and **Nikolaos Skourlis**. Thank you for creating healthy and kind work atmosphere and thank you for good thermostat management.

Thank you to **Marie Jansson**, **Alessandra Nanni**, and **Camilla Ahlqvist**, for making being a PhD student at MEB so seamless by expertly helping with the administrative aspects.

Thank you to the other PhD students I got to know during my time at MEB, including **Shadi Azam, Nurgul Batyrbekova, Hanna Bower**, **Adam Brand, Enoch Yi-Tung Chen**, **Wenjiang Dang**, **Hilda Daníelsdóttir, Bénédicte Delcoigne**, **Mikael Eriksson**, **Johanna Holm, Tyra Lagerberg, Jiayao Lei**, **Yulia Leontyeva**, **Yuying Li**, **Xingrong Liu, Frida Lundberg**, **Xinhe Mao**, **Emilio Ugalde Morales**, **Zheng Ning**, **Henrik Olsson, Lu Pan**, **Peter Ström, Jet Termorshuizen**, **Caroline Weibull, Philippe Weitz, Jingru Yu,** and **Erwei Zeng**. Thank you for the shared experiences and thank you for sharing the pain.

Thank you to the rest of the biostatistics group at MEB, past and present, including **Julien Bryois, Iuliana Ciocanea-Teodorescu, Sophie Debonneville, Paul Dickman, Martin Eklund, Erin Gabriel, Anna Johansson, Robert Karlsson, Cecilia Lundholm, Juni Palmgren, Yudi Pawitan, Alexander Ploner, Michael Sachs, Sven Sandin, Arvid Sjölander, Elisavet Syriopoulou, Agnieszka Szwajda, Valentin Vancak, Nghia Trung**

Vu, Henric Winell, and **Li Yin.** Thank you for creating a good research environment and for providing a sense of belonging for a statistician.

Thank you to the rest of the talented researchers and personnel at MEB, where the list of names gets completely out of hand. You all played a part in making my time at the department what it was, though you may not have known it (and also didn’t have a choice).

Tack till mina föräldrar, **Joakim Strandberg** och **Madeleine Strandberg**. Tack för allt ert stöd och för att jag fick ta chansen alla dessa år sedan till att göra något så säreget och så annorlunda.

A big thank you to my dear wife, **Viktoria Strandberg**. You have changed my life, and I could not have done this without you. I love you.

9 REFERENCES

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021 May 4;71(3):209–49.

2. Altobelli E, Rapacchietta L, Angeletti P, Barbante L, Profeta F, Fagnano R. Breast Cancer Screening Programmes across the WHO European Region: Differences among Countries Based on National Income Level. Int J Environ Res Public Health. 2017 Apr 23;14(4):452.

3. Plevritis SK, Munoz D, Kurian AW, Stout NK, Alagoz O, Near AM, et al. Association of Screening and Treatment With Breast Cancer Mortality by Molecular Subtype in US Women, 2000-2012. JAMA. 2018 Jan 9;319(2):154.

4. van den Broek JJ, van Ravesteyn NT, Heijnsdijk EA, de Koning HJ. Simulating the Impact of Risk-Based Screening and Treatment on Breast Cancer Outcomes with MISCAN-Fadia. Med Decis Mak. 2018 Apr 19;38(1_suppl):54S-65S.

5. 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 Jul 5;155(1):10.

6. Louro J, Posso M, Hilton Boon M, Román M, Domingo L, Castells X, et al. A systematic review and quality assessment of individualised breast cancer risk prediction models. Br J Cancer. 2019 Jul 22;121(1):76–85.

7. Shieh Y, Eklund M, Madlensky L, Sawyer SD, Thompson CK, Stover Fiscalini A, et al. Breast Cancer Screening in the Precision Medicine Era: Risk-Based Screening in a Population-Based Trial. J Natl Cancer Inst. 2017;109(5):1–8.

8. French DP, Astley S, Astley S, Brentnall AR, Cuzick J, Dobrashian R, et al. What are the benefits and harms of risk stratified screening as part of the NHS breast screening Programme? Study protocol for a multi-site non-randomised comparison of BC-predict versus usual screening (NCT04359420). BMC Cancer. 2020;20(1):1–14.

9. Javed A, Lteif A. Development of the Human Breast. Semin Plast Surg. 2013 May 23;27(01):005–12.

10. Russo J, Russo IH. Development of the human breast. Maturitas. 2004 Sep 24;49(1):2–15.

11. Atashgaran V, Wrin J, Barry SC, Dasari P, Ingman W V. Dissecting the Biology of Menstrual Cycle-Associated Breast Cancer Risk. Front Oncol. 2016 Dec 26;6.

12. Russo IH, Russo J. Pregnancy-Induced Changes in Breast Cancer Risk. J Mammary Gland Biol Neoplasia. 2011 Sep;16(3):221–33.

13. Radisky DC, Hartmann LC. Mammary Involution and Breast Cancer Risk: Transgenic Models and Clinical Studies. J Mammary Gland Biol Neoplasia. 2009 Jun 30;14(2):181–91.

14. Russo J, Rivera R, Russo IH. Influence of age and parity on the development of the human breast. Breast Cancer Res Treat. 1992;23(3):211–8.

15. Russo J, Moral R, Balogh GA, Mailo D, Russo IH. The protective role of pregnancy in breast cancer. Breast Cancer Res. 2005 Jun 7;7(3):131.

16. Larønningen S, Ferlay J, Bray F, Engholm G, Ervik M, Gulbrandsen J, et al. NORDCAN: Cancer Incidence, Mortality, Prevalence and Survival in the Nordic Countries, Version 9.1 (27.09.2021) [Internet]. Association of the Nordic Cancer Registries. 2021 [cited 2022 Feb 1]. Available from: https://nordcan.iarc.fr/

17. Eheman CR, Shaw KM, Ryerson AB, Miller JW, Ajani UA, White MC. The Changing Incidence of In situ and Invasive Ductal and Lobular Breast Carcinomas: United States, 1999-2004. Cancer Epidemiol Biomarkers Prev. 2009 Jun;18(6):1763–9.

18. Harris J, Lippman M, Morrow M, Osborne C. Diseases of the Breast. 5th ed. Lippincott Williams & Wilkins; 2012.

19. Ellis IO, Galea M, Broughton N, Locker A, Blamey RW, Elston CW. Pathological prognostic factors in breast cancer. II. Histological type. Relationship with survival in a large study with long-term follow-up. Histopathology. 2007 Apr 3;20(6):479–89.

20. Louwman MWJ, Vriezen M, Van Beek MWPM, Tutein Noltlienius-Puylaert MCBJE, Van Der Sangen MJC, Roumen RM, et al. Uncommon breast tumors in perspective: Incidence, treatment and survival in the Netherlands. Int J Cancer. 2007;121(1):127–35.

21. Erbas B, Provenzano E, Armes J, Gertig D. The natural history of ductal carcinoma in situ of the breast: A review. Breast Cancer Res Treat. 2006;97(2):135–44.

22. Wu Q, Li J, Zhu S, Wu J, Chen C, Liu Q, et al. Breast cancer subtypes predict the preferential site of distant metastases: A SEER based study. Oncotarget. 2017;8(17):27990–6.

23. Saadatmand S, Bretveld R, Siesling S, Tilanus-Linthorst MMA. Influence of tumour stage at breast cancer detection on survival in modern times: Population based study in 173 797 patients. BMJ. 2015;351.

24. Walters S, Maringe C, Butler J, Rachet B, Barrett-Lee P, Bergh J, et al. Breast cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK, 2000-2007: A population-based study. Br J Cancer. 2013;108(5):1195–208.

25. Young JJ, Roffers S, Ries L, Fritz A, Hurlbut A. SEER Summary Staging Manual - 2000 Codes and Coding Instructions. Natl Cancer Inst. 2001;NIH Pub. N(01):118–59.

26. Hortobagyi GN, Connolly JL, D'Orsi CJ, Edge SB, Mittendorf EA, Rugo HS, et al. AJCC Cancer Staging Manual. 8th ed. Cham: Springer International Publishing; 2017. 589–628 p.

27. Hagemeister FB, Buzdar AU, Luna MA, Blumenschein GR. Causes of death in breast cancer a clinicopathologic study. Cancer. 1980;46(1):162–7.

28. Wills L, Pearson C. 10-year cancer survival by stage for patients diagnosed in the East of England , 2007 to 2017. 2017.

29. Narod SA, Iqbal J, Giannakeas V, Sopik V, Sun P. Breast cancer mortality after a diagnosis of ductal carcinoma in situ. JAMA Oncol. 2015;1(7):888–96.

30. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. C. W. Elston & I. O. Ellis. Histopathology 1991; 19; 403-410. Histopathology. 2002 Sep;41(3a):151–151.

31. Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T, Eusebi V, et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Res. 2010 Aug 30;12(4):207.

32. Schwartz AM, Henson DE, Chen D, Rajamarthandan S. Histologic Grade Remains a Prognostic Factor for Breast Cancer Regardless of the Number of Positive Lymph Nodes and Tumor Size: A Study of 161 708 Cases of Breast Cancer From the SEER Program. Arch Pathol Lab Med. 2014 Aug 1;138(8):1048–52.

33. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001;98(19):10869–74.

34. Brinton LA, Gaudet MM, Gierach GL. Breast Cancer. In: Thun MJ, Linet MS, Cerhan JR, Haiman CA, Schottenfeld D, editors. Cancer Epidemiology and Prevention. 4th ed. New York: Oxford University Press; 2018. p. 861–88.

35. Howlader N, Cronin KA, Kurian AW, Andridge R. Differences in breast cancer survival by molecular subtypes in the United States. Cancer Epidemiol Biomarkers Prev. 2018;27(6):619–26.

36. Johansson ALV, Trewin CB, Fredriksson I, Reinertsen K V., Russnes H, Ursin G. In modern times, how important are breast cancer stage, grade and receptor subtype for survival: a population-based cohort study. Breast Cancer Res. 2021;23(1):1–10.

37. Koo MM, von Wagner C, Abel GA, McPhail S, Rubin GP, Lyratzopoulos G. Typical and atypical presenting symptoms of breast cancer and their associations with diagnostic intervals: Evidence from a national audit of cancer diagnosis. Cancer Epidemiol. 2017;48:140–6.

38. Swedish Breast Cancer Group. Bröstcancer - Nationellt vårdprogram. Regionala cancercentrum i samvekan; 2020.

39. Tan AR, Swain SM. Ongoing Adjuvant Trials with Trastuzumab in Breast Cancer. Semin Oncol. 2003;30(5 SUPPL. 16):54–64.

40. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: An independent review. Br J Cancer. 2013;108(11):2205–40.

41. Beau AB, Andersen PK, Vejborg I, Lynge E. Limitations in the effect of screening on breast cancer mortality. J Clin Oncol. 2018;36(30):2988–94.

42. Olsson S, Andersson I, Karlberg I, Bjurstam N, Frodis E, Håkansson S. Implementation of service screening with mammography in Sweden: from pilot study to nationwide programme. J Med Screen. 2000 Mar 1;7(1):14–8.

43. Swedish Breast Cancer Group. Nationellt vårdprogram bröstcancer. Regionala cancercentrum i samvekan; 2019.

44. Lind H, Svane G, Kemetli L, Törnberg S. Breast Cancer Screening Program in

Stockholm County, Sweden – Aspects of Organization and Quality Assurance. Breast Care. 2010;5(5):353–7.

45. Li T, Sun L, Miller N, Nicklee T, Woo J, Hulse-Smith L, et al. The Association of Measured Breast Tissue Characteristics with Mammographic Density and Other Risk Factors for Breast Cancer. Cancer Epidemiol Biomarkers Prev. 2005 Feb 1;14(2):343–9.

46. D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA, others. ACR BI-RADS Atlas, Breast Imaging Reporting and Data System. Reston, VA, American College of Radiology; 2013.

47. Boyd NF, Byng JW, Jong RA, Fishell EK, Little LE, Miller AB, et al. Quantitative Classification of Mammographic Densities and Breast Cancer Risk: Results From the Canadian National Breast Screening Study. JNCI J Natl Cancer Inst. 1995 May 3;87(9):670–5.

48. Li J, Szekely L, Eriksson L, Heddson B, Sundbom A, Czene K, et al. High-throughput mammographic-density measurement: a tool for risk prediction of breast cancer. Breast Cancer Res. 2012 Aug 30;14(4):R114.

49. Sovio U, Li J, Aitken Z, Humphreys K, Czene K, Moss S, et al. Comparison of fully and semi-automated area-based methods for measuring mammographic density and predicting breast cancer risk. Br J Cancer. 2014 Apr 20;110(7):1908–16.

50. Checka CM, Chun JE, Schnabel FR, Lee J, Toth H. The Relationship of Mammographic Density and Age: Implications for Breast Cancer Screening. Am J Roentgenol. 2012 Mar 1;198(3):W292–5.

51. 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 Oct 20;16(5):439.

52. Lokate M, Stellato RK, Veldhuis WB, Peeters PHM, van Gils CH. Age-related Changes in Mammographic Density and Breast Cancer Risk. Am J Epidemiol. 2013 Jul 1;178(1):101–9.

53. Krishnan K, Baglietto L, Stone J, Simpson JA, Severi G, Evans CF, et al. Longitudinal Study of Mammographic Density Measures That Predict Breast Cancer Risk. Cancer Epidemiol Biomarkers Prev. 2017 Apr;26(4):651–60.

54. 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 Jan 18;356(3):227–36.

55. Gómez-Raposo C, Zambrana Tévar F, Sereno Moyano M, López Gómez M, Casado E. Male breast cancer. Cancer Treat Rev. 2010 Oct 1;36(6):451–7.

56. Schneider R. Comparison of age, sex, and incidence rates in human and canine breast cancer. Cancer. 1970 Aug;26(2):419–26.

57. Madigan MP, Ziegler RG, Benichou J, Byrne C, Hoover RN. Proportion of Breast Cancer Cases in the United States Explained by Well-Established Risk Factors. JNCI J Natl Cancer Inst. 1995 Nov 15;87(22):1681–5.

58. Johns PC, Yaffe MJ. X-ray characterisation of normal and neoplastic breast tissues.

Phys Med Biol. 1987 Jun 1;32(6):675–95.

59. Clemons M, Goss P. Estrogen and the Risk of Breast Cancer. Epstein FH, editor. N Engl J Med. 2001 Jan 25;344(4):276–85.

60. Magnusson C, Colditz G, Rosner B, Bergström R, Persson I. Association of family history and other risk factors with breast cancer risk (Sweden). Cancer Causes Control. 1998;9(3):259–67.

61. Anderson KN, Schwab RB, Martinez ME. Reproductive risk factors and breast cancer subtypes: A review of the literature. Breast Cancer Res Treat. 2014;144(1):1–10.

62. Barnard ME, Boeke CE, Tamimi RM. Established breast cancer risk factors and risk of intrinsic tumor subtypes. Biochim Biophys Acta - Rev Cancer. 2015 Aug;1856(1):73–85.

63. Byrne C, Ursin G, Martin CF, Peck JD, Cole EB, Zeng D, et al. Mammographic Density Change With Estrogen and Progestin Therapy and Breast Cancer Risk. JNCI J Natl Cancer Inst. 2017 Sep 1;109(9).

64. Rice MS, Tamimi RM, Bertrand KA, Scott CG, Jensen MR, Norman AD, et al. Does mammographic density mediate risk factor associations with breast cancer? An analysis by tumor characteristics. Breast Cancer Res Treat. 2018 Jul 3;170(1):129–41.

65. Woods KL, Smith SR, Morrison JM. Parity and breast cancer: evidence of a dual effect. BMJ. 1980 Aug 9;281(6237):419–21.

66. Lambe M, Hsieh C, Trichopoulos D, Ekbom A, Pavia M, Adami H-O. Transient Increase in the Risk of Breast Cancer after Giving Birth. N Engl J Med. 1994 Jul 7;331(1):5–9.

67. Hsieh C, Pavia M, Lambe M, Lan S-J, Colditz GA, Ekbom A, et al. Dual effect of parity on breast cancer risk. Eur J Cancer. 1994 Jan;30(7):969–73.

68. Albrektsen G, Heuch I, Kvåle G. The short-term and long-term effect of a pregnancy on breast cancer risk: a prospective study of 802,457 parous Norwegian women. Br J Cancer. 1995 Aug;72(2):480–4.

69. MacMahon B. Reproduction and Cancer of the Breast. Cancer. 1993;71(10):3185–8.

70. Russo J, Tay LK, Russo IH. Differentiation of the mammary gland and susceptibility to carcinogenesis. Breast Cancer Res Treat. 1982 Mar;2(1):5–73.

71. Lambe M, Hsieh C, Chan H, Ekbom A, Trichopoulos D, Adami H-O. Parity, age at first and last birth, and risk of breast cancer: A population-based study in Sweden. Breast Cancer Res Treat. 1996 Oct;38(3):305–11.

72. Pharoah PDP, Day NE, Duffy S, Easton DF, Ponder BAJ. Family history and the risk of breast cancer: A systematic review and meta-analysis. Int J Cancer. 1997 May 29;71(5):800–9.

73. Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. Nat Genet. 2008 Jan 27;40(1):17–22.

74. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, et al. Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families. Am J Hum Genet. 1998 Mar;62(3):676–89.

75. Antoniou AC, Pharoah PDP, McMullan G, Day NE, Stratton MR, Peto J, et al. A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. Br J Cancer. 2002;86(1):76–83.

76. Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet. 2015 Apr 9;47(4):373–80.

77. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017 Nov 23;551(7678):92–4.

78. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature. 2007;447(7148):1087–93.

79. Mavaddat N, Antoniou AC, Easton DF, Garcia-Closas M. Genetic susceptibility to breast cancer. Mol Oncol. 2010;4(3):174–91.

80. Mavaddat N, Pharoah PDP, Michailidou K, Tyrer J, Brook MN, Bolla MK, et al. Prediction of breast cancer risk based on profiling with common genetic variants. J Natl Cancer Inst. 2015;107(5):1–15.

81. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet. 2018/12/13. 2019 Jan 3;104(1):21–34.

82. Neuhouser ML, Aragaki AK, Prentice RL, Manson JE, Chlebowski R, Carty CL, et al. Overweight, Obesity, and Postmenopausal Invasive Breast Cancer Risk. JAMA Oncol. 2015 Aug 1;1(5):611.

83. The Premenopausal Breast Cancer Collaborative Group. Association of Body Mass Index and Age With Subsequent Breast Cancer Risk in Premenopausal Women. JAMA Oncol. 2018 Nov 8;4(11):e181771.

84. Pizot C, Boniol M, Mullie P, Koechlin A, Boniol M, Boyle P, et al. Physical activity, hormone replacement therapy and breast cancer risk: A meta-analysis of prospective studies. Eur J Cancer. 2016 Jan;52:138–54.

85. Gaudet MM, Gapstur SM, Sun J, Diver WR, Hannan LM, Thun MJ. Active Smoking and Breast Cancer Risk: Original Cohort Data and Meta-Analysis. JNCI J Natl Cancer Inst. 2013 Apr 17;105(8):515–25.

86. Chen WY, Rosner B, Hankinson SE, Colditz GA, Willett WC. Moderate Alcohol Consumption During Adult Life, Drinking Patterns, and Breast Cancer Risk. JAMA. 2011 Nov 2;306(17):1884.

87. Jung S, Wang M, Anderson K, Baglietto L, Bergkvist L, Bernstein L, et al. Alcohol consumption and breast cancer risk by estrogen receptor status: in a pooled analysis of 20 studies. Int J Epidemiol. 2016 Jun;45(3):916–28.

88. Duffy SW, Chen H-H, Tábar L, Day NE. Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase. Stat Med. 1995;14(14):1531–43.

89. Prevost T, Launoy G. Estimating sensitivity and sojourn time in screening for

colorectal cancer a comparison of statistical approaches. Am J Epidemiol. 1998;148(6):609–19.

90. Uhry Z, Hedelin G, Colonna M, Asselain B, Arveux P, Rogel A, et al. Multi-state Markov models in cancer screening evaluation: a brief review and case study. Stat Methods Med Res. 2010;19(5):463–86.

91. Wu JC-Y, Hakama M, Anttila A, Yen AM-F, Malila N, Sarkeala T, et al. Estimation of natural history parameters of breast cancer based on non-randomized organized screening data: subsidiary analysis of effects of inter-screening interval, sensitivity, and attendance rate on reduction of advanced cancer. Breast Cancer Res Treat. 2010 Jul 7;122(2):553–66.

92. Duffy SW, Day NE, Tabár L, Chen H-H, Smith TC. Markov Models of Breast Tumor Progression: Some Age-Specific Results. JNCI Monogr. 1997 Jan;1997(22):93–7.

93. Chen H-H, Yen AM-F, Tabár L. A Stochastic Model for Calibrating the Survival Benefit of Screen-Detected Cancers. J Am Stat Assoc. 2012;107(500):1339–59.

94. Tan KHX, Simonella L, Wee HL, Roellin A, Lim Y-W, Lim W, et al. Quantifying the natural history of breast cancer. Br J Cancer. 2013 Oct 1;109(8):2035–43.

95. Chiu SYH, Duffy SW, Yen AM-F, Tabár L, Smith RA, Chen H-H. Effect of baseline breast density on breast cancer incidence, stage, mortality, and screening parameters: 25-Year follow-up of a Swedish mammographic screening. Cancer Epidemiol Biomarkers Prev. 2010;19(5):1219–28.

96. Wu Y-Y, Yen M-F, Yu C-P, Chen H-H. Risk Assessment of Multistate Progression of Breast Tumor with State-Dependent Genetic and Environmental Covariates. Risk Anal. 2014 Feb;34(2):367–79.

97. Taghipour S, Banjevic D, Miller AB, Montgomery N, Jardine AKS, Harvey BJ. Parameter estimates for invasive breast cancer progression in the Canadian National Breast Screening Study. Br J Cancer. 2013 Feb 15;108(3):542–8.

98. Bartoszyński R, Edler L, Hanin LG, Kopp-Schneider A, Pavlova L, Tsodikov A, et al. Modeling cancer detection: Tumor size as a source of information on unobservable stages of carcinogenesis. Math Biosci. 2001;171(2):113–42.

99. Weedon-Fekjær H, Lindqvist BH, Vatten LJ, Aalen OO, Tretli S. Breast cancer tumor growth estimated through mammography screening data. Breast Cancer Res. 2008 Jun 8;10(3):R41.

100. Abrahamsson L, Humphreys K. A statistical model of breast cancer tumour growth with estimation of screening sensitivity as a function of mammographic density. Stat Methods Med Res. 2016;25(4):1620–37.

101. Isheden G, Humphreys K. Modelling breast cancer tumour growth for a stable disease population. Stat Methods Med Res. 2019 Mar 6;28(3):681–702.

102. Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. Br J Cancer. 1954;8(1):1–12.

103. Armitage P. Multistage Models of Carcinogenesis. Environ Health Perspect. 1985 Nov;63:195–201.

104. Moolgavkar SH, Venzon DJ. Two-event models for carcinogenesis: incidence curves for childhood and adult tumors. Math Biosci. 1979;47(1–2):55–77.

105. Moolgavkar SH, Knudson AG. Mutation and Cancer: A Model for Human Carcinogenesis2. JNCI J Natl Cancer Inst. 1981 Jun 1;66(6):1037–52.

106. Moolgavkar SH, Luebeck G. Two-Event Model for Carcinogenesis: Biological, Mathematical, and Statistical Considerations. Risk Anal. 1990 Jun;10(2):323–41.

107. Knudson AG. Mutation and Cancer: Statistical Study of Retinoblastoma. Proc Natl Acad Sci. 1971 Apr 1;68(4):820–3.

108. Heidenreich WF, Luebeck EG, Moolgavkar SH. Some Properties of the Hazard Function of the Two-Mutation Clonal Expansion Model. Risk Anal. 1997 Jun;17(3):391–9.

109. Hanin LG, Yakovlev AY. Multivariate distributions of clinical covariates at the time of cancer detection. Stat Methods Med Res. 2004;13(6):457–89.

110. Plevritis SK, Salzman P, Sigal BM, Glynn PW. A natural history model of stage progression applied to breast cancer. Stat Med. 2007 Feb 10;26(3):581–95.

111. Laird AK. Dynamics of Tumor Growth. Br J Cancer. 1964 Sep;18(3):490–502.

112. Norton L, Simon R, Brereton HD, Bogden AE. Predicting the course of Gompertzian growth. Nature. 1976 Dec 1;264(5586):542–5.

113. Norton L. A Gompertzian Model of Human Breast Cancer Growth. Cancer Res. 1988;48(24 Part 1):7067–71.

114. Norton L. Conceptual and Practical Implications of Breast Tissue Geometry: Toward a More Effective, Less Toxic Therapy. Oncologist. 2005 Jun 1;10(6):370–81.

115. Spratt JA, von Fournier D, Spratt JS, Weber EE. Decelerating growth and human breast cancer. Cancer. 1993 Mar 15;71(6):2013–9.

116. Spratt JA, von Fournier D, Spratt JS, Weber EE. Mammographic assessment of human breast cancer growth and duration. Cancer. 1993;71(6):2020–6.

117. Weedon-Fekjær H, Tretli S, Aalen OO. Estimating screening test sensitivity and tumour progression using tumour size and time since previous screening. Stat Methods Med Res. 2010 Oct 31;19(5):507–27.

118. Bloom HJG, Richardson WW, Harries EJ. Natural History of Untreated Breast Cancer (1805-1933). BMJ. 1962 Jul 28;2(5299):213–21.

119. Fournier D v, Weber E, Hoeffken W, Bauer M, Kubli F, Barth V. Growth rate of 147 mammary carcinomas. Cancer. 1980 Apr 15;45(8):2198–207.

120. Talkington A, Durrett R. Estimating Tumor Growth Rates In Vivo. Bull Math Biol. 2015 Oct 19;77(10):1934–54.

121. Abrahamsson L, Czene K, Hall P, Humphreys K. Breast cancer tumour growth modelling for studying the association of body size with tumour growth rate and symptomatic detection using case-control data. Breast Cancer Res. 2015;17(1):116.

122. Lee SJ, Li X, Huang H, Zelen M. The Dana-Farber CISNET Model for Breast Cancer

Screening Strategies: An Update. Med Decis Mak. 2018 Apr 19;38(1_suppl):44S-53S.

123. Schechter CB, Near AM, Jayasekera J, Chandler Y, Mandelblatt JS. Structure, Function, and Applications of the Georgetown–Einstein (GE) Breast Cancer Simulation Model. Med Decis Mak. 2018 Apr 19;38(1_suppl):66S-77S.

124. Huang X, Li Y, Song J, Berry DA. A Bayesian Simulation Model for Breast Cancer Screening, Incidence, Treatment, and Mortality. Med Decis Mak. 2018 Apr 19;38(1_suppl):78S-88S.

125. Munoz DF, Xu C, Plevritis SK. A Molecular Subtype–Specific Stochastic Simulation Model of US Breast Cancer Incidence, Survival, and Mortality Trends from 1975 to 2010. Med Decis Mak. 2018 Apr 19;38(1_suppl):89S-98S.

126. Alagoz O, Ergun MA, Cevik M, Sprague BL, Fryback DG, Gangnon RE, et al. The University of Wisconsin Breast Cancer Epidemiology Simulation Model: An Update. Med Decis Mak. 2018;38(1_suppl):99S-111S.

127. Berry DA, Cronin KA, Plevritis SK, Fryback DG, Clarke L, Zelen M, et al. Effect of Screening and Adjuvant Therapy on Mortality from Breast Cancer. N Engl J Med. 2005 Oct 27;353(17):1784–92.

128. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. JNCI J Natl Cancer Inst. 1989 Dec 20;81(24):1879–86.

129. Marchbanks PA, Mcdonald JA, Wilson HG, Burnett NM, Daling JR, Bernstein L, et al. The NICHD Women’s Contraceptive and Reproductive Experiences Study: Methods and operational results. Ann Epidemiol. 2002;12(4):213–21.

130. Gail MH, Costantino JP, Pee D, Bondy M, Newman L, Selvan M, et al. Projecting Individualized Absolute Invasive Breast Cancer Risk in African American Women. JNCI J Natl Cancer Inst. 2007 Dec 5;99(23):1782–92.

131. NIH National Cancer Institute. Breast Cancer Risk Assessment Tool [Internet]. [cited 2022 Feb 14]. Available from: https://bcrisktool.cancer.gov/about.html

132. Boyle P, Mezzetti M, La Vecchia C, Franceschi S, Decarli A, Robertson C. Contribution of three components to individual cancer risk predicting breast cancer risk in Italy. Eur J Cancer Prev. 2004;13(3):183–91.

133. Chen J, Pee D, Ayyagari R, Graubard B, Schairer C, Byrne C, et al. Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. J Natl Cancer Inst. 2006;98(17):1215–26.

134. Zhang X, Rice M, Tworoger SS, Rosner BA, Eliassen AH, Tamimi RM, et al. Addition of a polygenic risk score, mammographic density, and endogenous hormones to existing breast cancer risk prediction models: A nested case–control study. Zheng W, editor. PLOS Med. 2018 Sep 4;15(9):e1002644.

135. Breast Cancer Surveillance Consortium. Breast Cancer Surveillance Consortium Risk Calculator [Internet]. 2015 [cited 2022 Feb 14]. Available from: https://tools.bcsc-scc.org/BC5yearRisk/

136. Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K. Using Clinical Factors and Mammographic Breast Density to Estimate Breast Cancer

Risk: Development and Validation of a New Predictive Model. Ann Intern Med. 2008 Mar 4;148(5):337.

137. Tice JA, Bissell MCS, Miglioretti DL, Gard CC, Rauscher GH, Dabbous FM, et al. Validation of the breast cancer surveillance consortium model of breast cancer risk. Breast Cancer Res Treat. 2019;175(2):519–23.

138. Shieh Y, Hu D, Ma L, Huntsman S, Gard CC, Leung JWT, et al. Breast cancer risk prediction using a clinical risk model and polygenic risk score. Breast Cancer Res Treat. 2016;159(3):513–25.

139. Tice JA, Miglioretti DL, Li C, Vachon CM, Gard CC, Kerlikowske K. Breast Density and Benign Breast Disease: Risk Assessment to Identify Women at High Risk of Breast Cancer. J Clin Oncol. 2015 Oct 1;33(28):3137–43.

140. Vachon CM, Pankratz VS, Scott CG, Haeberle L, Ziv E, Jensen MR, et al. The contributions of breast density and common genetic variation to breast cancer risk. J Natl Cancer Inst. 2015;107(5):1–4.

141. Parmigiani G, Berry DA, Aguilar O. Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. Am J Hum Genet. 1998;62(1):145–58.

142. Mazzola E, Blackford A, Parmigiani G, Biswas S. Recent enhancements to the genetic risk prediction model BRCAPRO. Cancer Inform. 2015;14:147–57.

143. Antoniou AC, Gayther SA, Stratton JF, Ponder BAJ, Easton DF. Risk models for familial ovarian and breast cancer. Genet Epidemiol. 2000;18(2):173–90.

144. Antoniou AC, Pharoah PPD, Smith P, Easton DF. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. Br J Cancer. 2004;91(8):1580–90.

145. Antoniou AC, Cunningham AP, Peto J, Evans DG, Lalloo F, Narod SA, et al. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: Updates and extensions. Br J Cancer. 2008;98(8):1457–66.

146. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med. 2004 Apr 15;23(7):1111–30.

147. Cuzick J. IBIS Breast Cancer Risk Evaluation Tool [Internet]. [cited 2022 Feb 7]. Available from: https://ems-trials.org/riskevaluator/

148. Brentnall AR, Cuzick J, Buist DSM, Bowles EJA. Long-Term accuracy of breast cancer risk assessment combining classic risk factors and breast density. JAMA Oncol. 2018;4(9).

149. Cuzick J, Brentnall A. Models for Assessment of Breast Cancer Risk. DI Eur. 2016;(October):54–5.

150. Rosner B, Colditz GA. Nurses' Health Study: Log-Incidence Mathematical Model of Breast Cancer Incidence. JNCI J Natl Cancer Inst. 1996 Mar 20;88(6):359–64.

151. Pike MC, Krailo MD, Henderson BE, Casagrande JT, Hoel DG. "Hormonal" risk factors, "Breast tissue age" and the age-incidence of breast cancer. Nature. 1983;303(5920):767–70.

152. Colditz GA, Rosner B. Cumulative Risk of Breast Cancer to Age 70 Years According

to Risk Factor Status: Data from the Nurses’ Health Study. Am J Epidemiol. 2000 Nov 15;152(10):950–64.

153. Glynn RJ, Colditz GA, Tamimi RM, Chen WY, Hankinson SE, Willett WW, et al. Extensions of the Rosner-Colditz breast cancer prediction model to include older women and type-specific predicted risk. Breast Cancer Res Treat. 2017 Aug 6;165(1):215–23.

154. Colditz GA, Rosner BA, Chen WY, Holmes MD, Hankinson SE. Risk factors for breast cancer according to estrogen and progesterone receptor status. J Natl Cancer Inst. 2004;96(3):218–28.

155. Gabrielson M, Eriksson M, Hammarström M, Borgquist S, Leifland K, Czene K, et al. Cohort Profile: The Karolinska Mammography Project for Risk Prediction of Breast Cancer (KARMA). Int J Epidemiol. 2017 Dec 1;46(6):1740-1741g.

156. Berger JO, Wolpert RL. The likelihood principle: A review, generalizations, and statistical implications. Institute of Mathematical Statistics Lecture Notes - Monograph Series, Volume 6; 1988.

157. Wilks SS. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. Ann Math Stat. 1938 Mar 1;9(1):60–2.

158. Brown CD, Davis HT. Receiver operating characteristics curves and related decision measures: A tutorial. Chemom Intell Lab Syst. 2006;80(1):24–38.

159. 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 Nov;68(6):394–424.

160. Zhang S, Ding Y, Zhu Q, Wang C, Wu P, Dong J. Correlation factors analysis of breast cancer tumor volume doubling time measured by 3D-ultrasound. Med Sci Monit. 2017;23:3147–53.

161. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell. 2011;144(5):646–74.

162. Wang J, Gottschal P, Ding L, Veldhuizen D. van, Lu W, Houssami N, et al. Mammographic sensitivity as a function of tumor size: A novel estimation based on population-based screening data. The Breast. 2021 Feb;55:69–74.

163. Zheng YZ, Wang L, Hu X, Shao ZM. Effect of tumor size on breast cancer-specific survival stratified by joint hormone receptor status in a SEER population-based study. Oncotarget. 2015;6(26):22985–95.

164. Narod SA. Tumour size predicts long-term survival among women with lymph node-positive breast cancer. Curr Oncol. 2012;19(5):249–53.

165. Isheden G, Abrahamsson L, Andersson T, Czene K, Humphreys K. Joint models of tumour size and lymph node spread for incident breast cancer cases in the presence of screening. Stat Methods Med Res. 2019 Dec 3;28(12):3822–42.

166. Gasparini A, Humphreys K. Estimating latent, dynamic processes of breast cancer tumour growth and distant metastatic spread from mammography screening data. Stat Methods Med Res. 2022;

167. Arvold ND, Taghian AG, Niemierko A, Abi Raad RF, Sreedhara M, Nguyen PL, et al. Age, breast cancer subtype approximation, and local recurrence after breast-conserving therapy. J Clin Oncol. 2011;29(29):3885–91.

168. Acheampong T, Kehm RD, Terry MB, Argov EL, Tehranifar P. Incidence Trends of Breast Cancer Molecular Subtypes by Age and Race/Ethnicity in the US From 2010 to 2016. JAMA Netw open. 2020;3(8):e2013226.

169. Colditz GA, Rosner B. Cumulative risk of breast cancer to age 70 years according to risk factor status: Data from the nurses' health study. Am J Epidemiol. 2000;152(10):950–64.

170. Eriksson M, Czene K, Pawitan Y, Leifland K, Darabi H, Hall P. A clinical model for identifying the short-term risk of breast cancer. Breast Cancer Res. 2017 Dec 14;19(1):29.

171. Meshkat B, Prichard RS, Al-Hilli Z, Bass GA, Quinn C, O'Doherty A, et al. A comparison of clinical-pathological characteristics between symptomatic and interval breast cancer. Breast. 2015;24(3):278–82.

172. Weber RJP, van Bommel RMG, Louwman MW, Nederend J, Voogd AC, Jansen FH, et al. Characteristics and prognosis of interval cancers after biennial screen-film or full-field digital screening mammography. Breast Cancer Res Treat. 2016;158(3):471–83.

173. Abrahamsson L, Isheden G, Czene K, Humphreys K. Continuous tumour growth models, lead time estimation and length bias in breast cancer screening studies. Stat Methods Med Res. 2020;29(2):374–95.