

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

# Aggressive Breast Cancer: Epidemiological Studies Addressing Disease Heterogeneity

Johanna Holm



**Karolinska  
Institutet**

Stockholm 2018

.

All previously published papers were reproduced with permission from the publisher.  
If not otherwise stated, illustrations are by the author.

Cover painting by the author, 2017

Published by Karolinska Institutet.

© Johanna Holm

ISBN 978-91-7676-876-1

Printed by Eprint AB 2018.

Institutionen för Medicinsk Epidemiologi och Biostatistik

# Aggressive Breast Cancer: Epidemiological Studies Addressing Disease Heterogeneity

## AKADEMISK AVHANDLING

Som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras  
i hörsal Hillarp, Retzius väg 8, Karolinska Institutet, Solna

**Fredagen den 23 Februari 2018, kl 09:00**

### *Huvudhandledare*

Professor Kamila Czene  
Karolinska Institutet  
Institutionen för Medicinsk  
Epidemiologi och Biostatistik

### *Bihandledare*

Universitetslektor Keith Humphreys  
Karolinska Institutet  
Institutionen för Medicinsk  
Epidemiologi och Biostatistik

Professor Per Hall

Karolinska Institutet  
Institutionen för Medicinsk  
Epidemiologi och Biostatistik

### *Fakultetsopponent*

Professor Lucile Adams-Campbell  
Georgetown University, U.S.A.  
Lombardi Comprehensive Cancer Center

### *Betygsnämnd*

Professor Cecilia Magnusson  
Karolinska institutet  
Institutionen för Folkhälsovetenskap

Docent Johan Ahlgren  
Örebro Universitetssjukhus  
Institutionen för Onkologi

Professor Jesper Lagergren  
Karolinska Institutet  
Institutionen för Molekylär Medicin och Kirurgi

Stockholm, 2018

To Karl and Hedvig

## Abstract

Breast cancer is either the number one or the second most common cause of cancer death in women in the world, depending on region<sup>1</sup>. It is a cancer heterogeneous in many aspects related to the aggressiveness, such as proliferation rate, metastatic capacity and survival. This thesis is seeking to increase our understanding of aggressive breast cancer, and how risk factors may be related to it.

In study I, we compared interval cancers (IC) to cancers detected at screening (SC) and found the group of IC to be characterized by higher frequency of BRCA mutations, family history of breast cancer and use of hormonal therapy (HRT) as compared to SC. IC in non-dense breasts were enriched for aggressive tumour features, and in this group the estimates for family history and BRCA mutations were increased.

In study II, we studied if predictors intended to identify healthy women's risk of breast cancer in future prevention efforts were skewed towards certain tumour characteristics. A 77-SNP breast cancer polygenic risk score appeared to underestimate risk in women with high grade and Oestrogen receptor (ER) negative tumours, as it was on average higher in women with low grade, ER positive tumours. The Tyrer-Cuzic model of breast cancer risk also appeared to underestimate risk in ER negative, high grade tumours but this was restricted to early onset cases. Only mammographic density appeared to be a general risk factor/predictor for all tumours independent of prognosticators.

Study III was a case-control study where we estimated odds ratios for each of four breast cancer molecular subtypes separately in multinomial logistic regression. We found subtype heterogeneity in the odds ratios for genetic risk factors and for breastfeeding. The 77-SNP polygenic risk score was associated with all subtypes except for the basal-like subtype, which showed no association with the score. Although breastfeeding was protective for both luminal and basal-like subtypes, the magnitude and underlying mechanism appeared to differ across subtypes.

In Study IV we assessed the concordance between PAM50 gene expression-based and immunohistochemistry-based molecular subtypes. No proxy showed more than moderate concordance with PAM50, however if luminal A and B subtypes were collapsed into one category, substantial concordance was achieved. Sensitivity for HER2-enriched breast cancer as defined by PAM50 was low, at 0.36 for all proxies investigated, whereas sensitivity and specificity was high for classifying basal-like breast cancer.

---

<sup>1</sup>According to Cancer facts sheets, GLOBOCAN 2012, [IARC]

## List of publications

- I. **Johanna Holm**, Keith Humphreys, Jingmei Li, Alexander Ploner, Abbas Cheddad, Mikael Eriksson, Sven Törnberg, Per Hall, Kamila Czene  
**Risk Factors and Tumor Characteristics of Interval Cancers by mammographic density.**  
*Journal of Clinical Oncology* 2015 Mar 20;33(9):1030-7.
- II. **Johanna Holm**, Jingmei Li, Hatef Darabi, Martin Eklund, Mikael Eriksson, Keith Humphreys, Per Hall, Kamila Czene  
**Associations of Breast Cancer Risk Prediction Tools With Tumor Characteristics and Metastasis.**  
*Journal of Clinical Oncology* 2016 Jan 20;34(3):251-8
- III. **Johanna Holm**, Louise Eriksson, Alexander Ploner, Mikael Eriksson, Mattias Rantalainen, Jingmei Li, Per Hall, Kamila Czene  
**Assessment of Breast Cancer Risk Factors Reveals Subtype Heterogeneity.**  
*Cancer Research* 2017 Jul 1;77(13):3708-3717
- IV. **Johanna Holm**, Nancy Yu, Alexander Ploner, Linda Lindström, Kamila Czene  
**Concordance of immunohistochemistry based and PAM50 molecular subtypes of breast cancer.**  
*Manuscript*

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Aims</b>	<b>2</b>
<b>3</b>	<b>Background</b>	<b>3</b>
3.1	Biology of the mammary glands . . . . .	3
3.2	Prevention . . . . .	6
3.2.1	Primary prevention . . . . .	6
3.2.2	Mammography screening . . . . .	6
3.3	Aggressive breast cancer . . . . .	7
3.3.1	Interval cancer . . . . .	7
3.3.2	Prognosticators . . . . .	8
3.3.3	Intrinsic or molecular subtypes . . . . .	9
3.3.4	Immunohistochemistry subtypes . . . . .	10
3.4	Epidemiology of breast cancer . . . . .	10
3.5	Risk factors by breast cancer subtypes . . . . .	15
3.6	Risk factors by mode of detection . . . . .	18
3.7	Risk prediction . . . . .	19
<b>4</b>	<b>Data material</b>	<b>21</b>
4.1	Breast cancer cohorts (Papers I-IV) . . . . .	21
4.2	Population-based registers (Papers I-IV) . . . . .	23
4.3	Questionnaire data (Papers I-III) . . . . .	26
4.4	Mammography screening attendance and outcomes (Paper I) . . . . .	26
4.5	Mammography images (Papers I-III) . . . . .	26
4.6	Genotype data (Papers II and III) . . . . .	28
4.7	Medical records (Papers III and IV) . . . . .	29
<b>5</b>	<b>Statistical methodology</b>	<b>30</b>
5.1	Statistical models used throughout the thesis . . . . .	30
5.1.1	Logistic regression . . . . .	30
5.1.2	Multinomial logistic regression . . . . .	30
5.1.3	Proportional-Hazards regression . . . . .	31
5.1.4	Random forest . . . . .	31
5.1.5	Measures of concordance and performance of classifiers . . . . .	32
5.1.6	Kaplan-Meier, or Product-limit estimator . . . . .	32

5.2	Study designs . . . . .	33
5.2.1	Paper I . . . . .	33
5.2.2	Paper II . . . . .	34
5.2.3	Paper III . . . . .	35
5.2.4	Paper IV . . . . .	36
6	Main results and interpretations	39
7	Methodological considerations	46
8	Ethical considerations	50
9	Concluding remarks	51
10	Future perspectives	54
11	Acknowledgements	56



## List of abbreviations

The following abbreviations are used throughout the thesis

BC	Breast Cancer
ASR	Age-Standardised incidence Rate
AUC	Area Under the Curve
hCG	Human Chorionic Gonadotropin
TNM	Tumour - Node - Metastasis: TNM staging
TNBC	Triple-Negative Breast Cancer
TCRS	Tyrer-Cuzick Risk Score
SNP	Short Nucleotide Polymorphism
GWAS	Genome-Wide Association Study
PRS	Polygenic Risk Score
MD	Mammographic Density
IC	Interval Cancer (here assumed interval breast cancer)
SC	Screening-detected Cancer (here assumed screening detected breast cancer)
IHC	ImmunoHistoChemistry
HRT	Hormone Replacement Therapy
ER	Estrogen Receptor(s)
PR	Progesterone Receptor(s)
HER2	HER2/ <i>neu</i> oncoprotein, also known as Erb-B2
ErbB2	Gene encoding for HER2/ <i>neu</i> , alternative name for HER2/ <i>neu</i> oncoprotein
Ki-67	Protein marker of cell proliferation. Named after site (Kiel) and well number (in a 96-well plate) of discovery
PCR	Polymerase Chain Reaction
WHO	World Health Organization
SEER	Surveillance, Epidemiology, and End Results program
SNOMED	Systemized NOMenclature of MEDicine
UICC	Union of International Cancer Control

*"However, it is true that statistics cannot provide proofs. They can find out if correlations exist, and in this way point to some possibilities and exclude others."*

Johannes Clemmesen, 1948.

# 1 Introduction

The mammary gland is a common denominator of all animals of the class Mammalia. In fact the name 'Mammal' is taken from the Latin word for breast, 'mamma'. This choice of name and classification has been debated [Schiebinger, 1993, Vaughan et al., 2013], but it certainly highlights the biology of the mammary glands. The mammary glands make up a remarkably plastic organ [Lteif and Javed, 2013, Russo and Russo, 2011]. Sadly, it is also one particularly prone to carcinogenesis [Lakhani and IARC, 2012].

The term 'aggressive cancer' refers to a tumour that "forms, grows, or spreads quickly" [National Cancer Institute, 2017]. It has long been appreciated that breast cancer exhibits substantial heterogeneity in an array of aspects, including proliferation markers and stage at diagnosis [Fisher et al., 2008]. In the same manner, it has long been known that the disease has a high variation in patient survival and a continued risk for lethal recurrence decades after diagnosis, despite the 5-year survival generally being considered high [Adami and Killander, 1984]. Some breast cancers exhibit metastatic spread already at diagnosis, whereas others metastasise decades after diagnosis and treatment, yet others will never establish metastatic clones during the lifetime of the host. It is still an area of active research how to accurately predict which patients could be considered cured after treatment and what patients will experience recurrence or metastasis years after initial diagnosis [Pan et al., 2017]. Predicting short term prognosis, i.e. separating aggressive breast cancers from those primarily at risk of metastasis in the long term, is more readily achievable.

There is considerable variation in tumour biology that can account for the varied level of aggression between cases [Sorlie et al., 2001, Sorlie et al., 2003, Fisher et al., 2008]. Given the complexity of factors explaining breast cancer outcomes and the molecular heterogeneity of the disease, it is of importance that research about risk factors also take into account the subtypes of the disease present [Anderson and Matsuno, 2006]. In recent years, there has been a surge in the field of epidemiological studies assessing characteristics by breast cancer subtypes [Barnard et al., 2015]. There is also a move towards individualising prevention efforts [Howell et al., 2014]. Before embarking on such journeys, it is important to understand more about subtype heterogeneity in predictors. Whether aggressive subtypes of breast cancer share aetiology with the majority of cases with good 5-year survival is essentially unknown. Nor is it fully understood if factors used to predict risk of breast cancer apply equally to aggressive cases, although evidence to suggest there are differences have been presented in recent years [Anderson et al., 2014, Barnard et al., 2015], motivating further work.

## 2 Aims

The overarching aim of this thesis is to increase our understanding of aggressive breast cancer, with a focus on how risk factors identified for the disease in general relate to aggressive subtypes. The specific aims are:

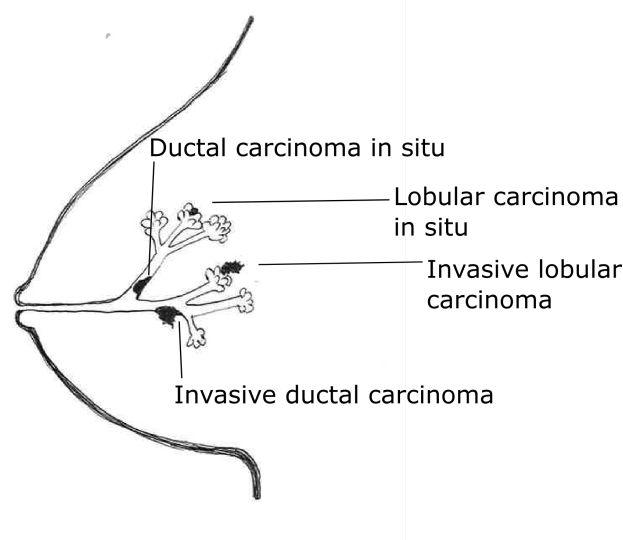
- To assess if aggressive tumour characteristics are enriched among interval cancers diagnosed in women with low mammographic density
- To assess whether standard breast cancer risk factors are equally dispersed between
  - Screening-detected and interval breast cancers
  - Molecular subtypes of breast cancer
- To assess if variables used in risk prediction models are associated with tumour prognosticators and survival
- To assess the concordance between immunohistochemistry marker based classification and gene expression based classification of breast cancer

## 3 Background

### 3.1 Biology of the mammary glands

#### Anatomy of the breast in view of breast cancer

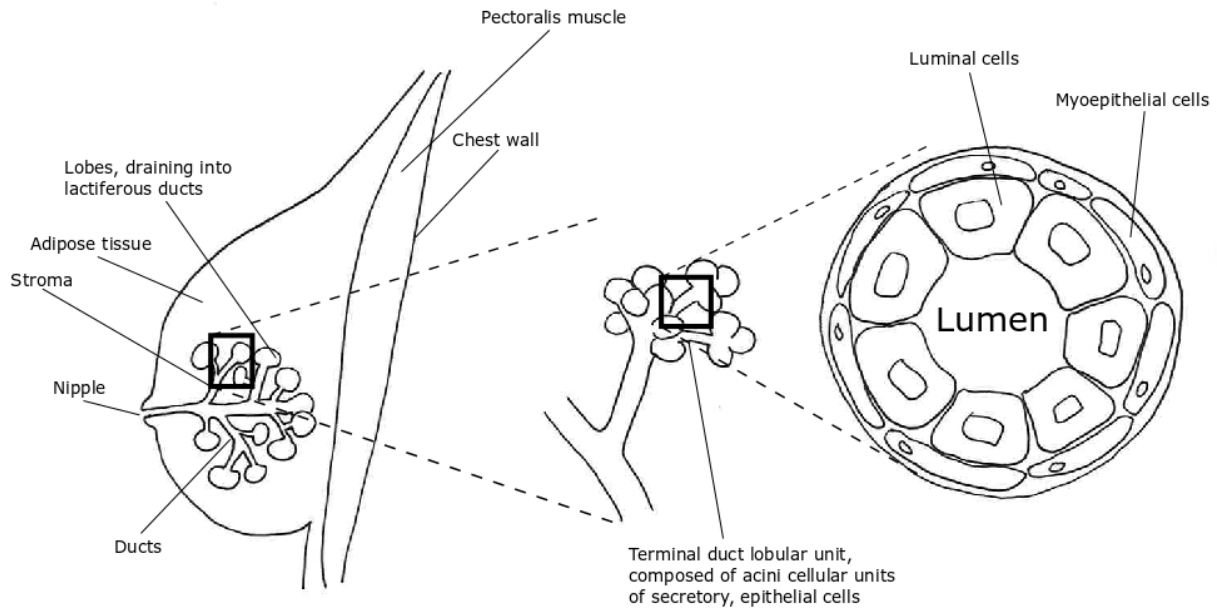
The breast is composed of mainly three types of tissues; the glandular epithelium of the lobes and ducts, the connective tissue that supports the glands, and the adipose tissue that embeds the other two. Breast cancer is an umbrella term for primary tumours diagnosed in the breast. The WHO histological classification of tumours details at least 20 different histological subtypes of breast cancer, with the dominating types being 'invasive carcinoma of no special type' (formerly known as invasive ductal carcinoma, not otherwise specified), varying from 40-70% of diagnoses, and invasive lobular carcinoma, around 5-15% [Lakhani and IARC, 2012]. Thus, with some exceptions, most invasive breast cancers are found in the epithelium of the milk ducts, or in the lobular glands that produce the milk (Figure 1).



**Figure 1** – Cartoon depiction of a human breast in sagittal section, with the location of main histological types of breast cancer outlined

#### Development and plasticity of the human breast

Breast formation is initiated in utero, but the development of the organ is continued throughout the reproductive life, with major structural changes being associated with menarche and puberty, as well as parity and menopause. At birth, a basic mammary gland with about 20 lobes and a certain degree of branching is present. During the first two years of life further branching and nipple maturation occurs, but the tissue remains dormant from the age of two until the onset of puberty and menarche. During puberty a complex growth phase

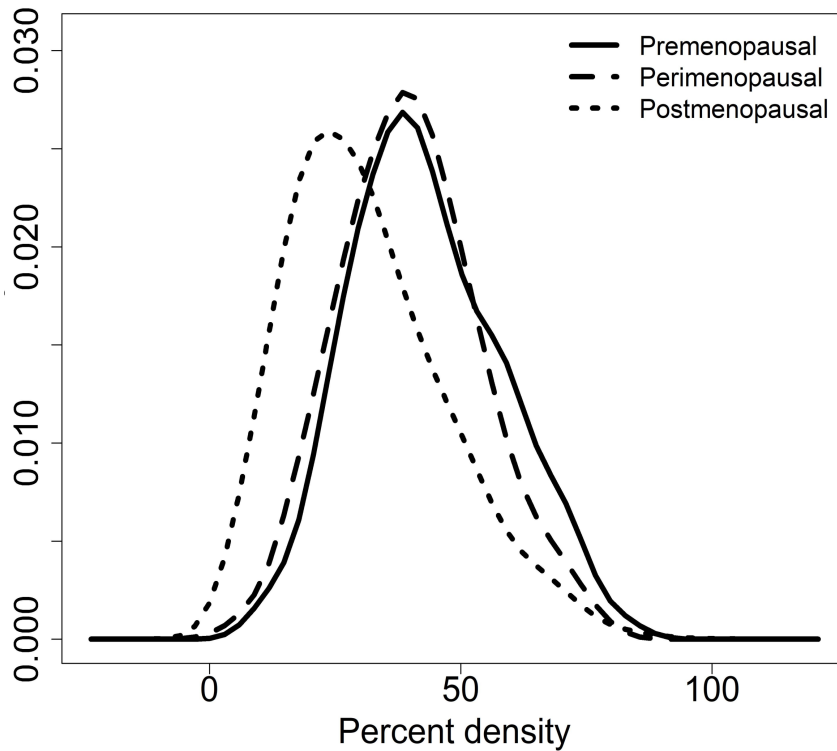
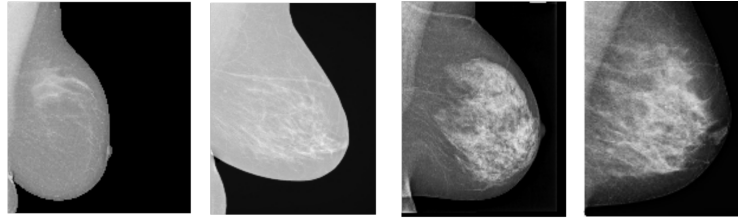


**Figure 2** – Cartoon depiction of (from left to right) a human breast in sagittal section, a terminal duct lobular unit, and a cross-sectional view of a milk duct. Terminal duct lobular units are the functional units of the glands, responsible for milk production. The milk ducts drain into the nipple area and thus transport the milk during lactation. The epithelial cell types consist of myoepithelial cells and luminal cells.

occur in the female breasts under the control of oestrogen, pituitary hormone and IGF-1 [Lteif and Javed, 2013]. Ducts elongate and branch further, and on a macro level the breasts enlarges. It is also during puberty that the glands develop the characteristic bilayered cellular structure, with an outer layer of basal myoepithelial cells close to the basement membrane, and an inner layer of luminal cells towards the lumen (Figure 2, cross-sectional image of milk duct) [Lteif and Javed, 2013]. However, in contrast to other organs it is not fully developed in nulliparous women. The mammary glands reach their highest level of tissue differentiation only under the hormonal influence of the first full-term pregnancy [Russo and Russo, 2004].

The lobules within the breast are classified according to their degree of differentiation, as types 1-4, with type 1 being the least differentiated and type 4 the most [Russo and Russo, 2004, Russo et al., 2005, Lteif and Javed, 2013]. The breast in nulliparous women is mainly composed of lobules of type 1, of the lowest degree of differentiation, and occasionally develop into lobule type 2 during successive menstrual cycles. During pregnancy, the lobules differentiate under the control of a number of hormones. Luteinizing hormone, progesterone and human chorionic gonadotropin (hCG) initiate the substantial tissue growth associated with the first trimester, whereas prolactin stimulates differentiation into the terminally differentiated, lactation-capable type 4 lobules in the very last weeks of pregnancy [Russo and Russo, 2011].

The process of reducing the amount of mammary glands in the breast is termed involution. Two distinct involution processes occur that either inactivates or shrinks the tissue, one at the post-lactational stage and one in connection to the menopausal. During post-lactational involution, tissue remodelling and wound healing processes regress the type 4 lobules into pregnancy-induced type 3 lobules that remain until menopausal, but there is little reduction of glandular tissue at this stage. During perimenopause, the process denoted lobular involution gradually reduces the size and structures of the lobules until they essentially have been replaced by connective tissue and fat by the end of menopause [Radisky and Hartmann, 2009].



**Figure 3** – Kernel density plot of the distribution of percent mammographic density among breast cancer cases, shown by menopausal status. Top panel shows examples of mammography films of breasts with increasing mammographic density. Mammographic density represents the amount of stromal and glandular tissue in the breast. Data and images from the cohorts included in the thesis. Image inspired by [Howell et al., 2014].

## **Breast tissue and mammographic density**

The term 'mammographic density' refers to the radio-density shown on a mammography image. Depending on the amount of stromal and glandular tissue in the breast, the appearance on a roentgen image will vary from largely translucent to largely dense. Due to the changes in breast tissue composition over the lifetime of a woman, premenopausal women tend to have high mammographic density and postmenopausal women have progressively lower mammographic density as they age (See figure 3). Mammographic density is often calculated as percent dense tissue of the total breast area, or as absolute area of dense tissue.

## **3.2 Prevention**

### **3.2.1 Primary prevention**

The scope for primary prevention of breast cancer in Sweden today is mainly restricted to informing the public about modifiable risk factors. Public authorities provide information online on risk factors for the disease, such as overweight, excessive alcohol intake, and physical inactivity. Treatment with hormone replacement therapy (HRT) for extended periods is also stated as a risk factor [Vårdguiden, 2017]. Reproductive and hormonal factors influencing the risk are also mentioned, but stated as non-intervenable risk factors. Actual interventions are limited to women at particularly high risk of breast cancer, by offering mastectomies, hysterectomies and increased screening of women with confirmed inherited mutations of a substantially high penetrance (BRCA1/2 mutations). Pending new evidence from ongoing trials of preventative therapy, primary prevention may however be expanded in the future [Karma, 2017, Howell et al., 2014].

### **3.2.2 Mammography screening**

Following the results of 31 % reduced mortality in the Two-County trial [Tabár et al., 1985], in 1986 the Swedish National Board of Health and Welfare issued recommendations to all county councils in Sweden to initiate population-based mammography screening programs. Initially, the recommendations were to screen all women aged 40-74, but in 1989, when the Malmö mammography trials published a lack of benefit in the youngest age groups, the recommendations were changed to focus resources primarily on screening women 50-69 years of age [Lind et al., 2010]. The introduction of mammography screening was implemented at different time points across the counties, from 1976 to 1997 [Olsson et al., 2000]. Of relevance for this thesis is the fact that Stockholm introduced screening in 1989, inviting all women 50-69 years of age to mammography every 24 months. Starting in July 2005, women 40-49 years of age were invited at 18 month intervals. A participation rate at around 70 % of



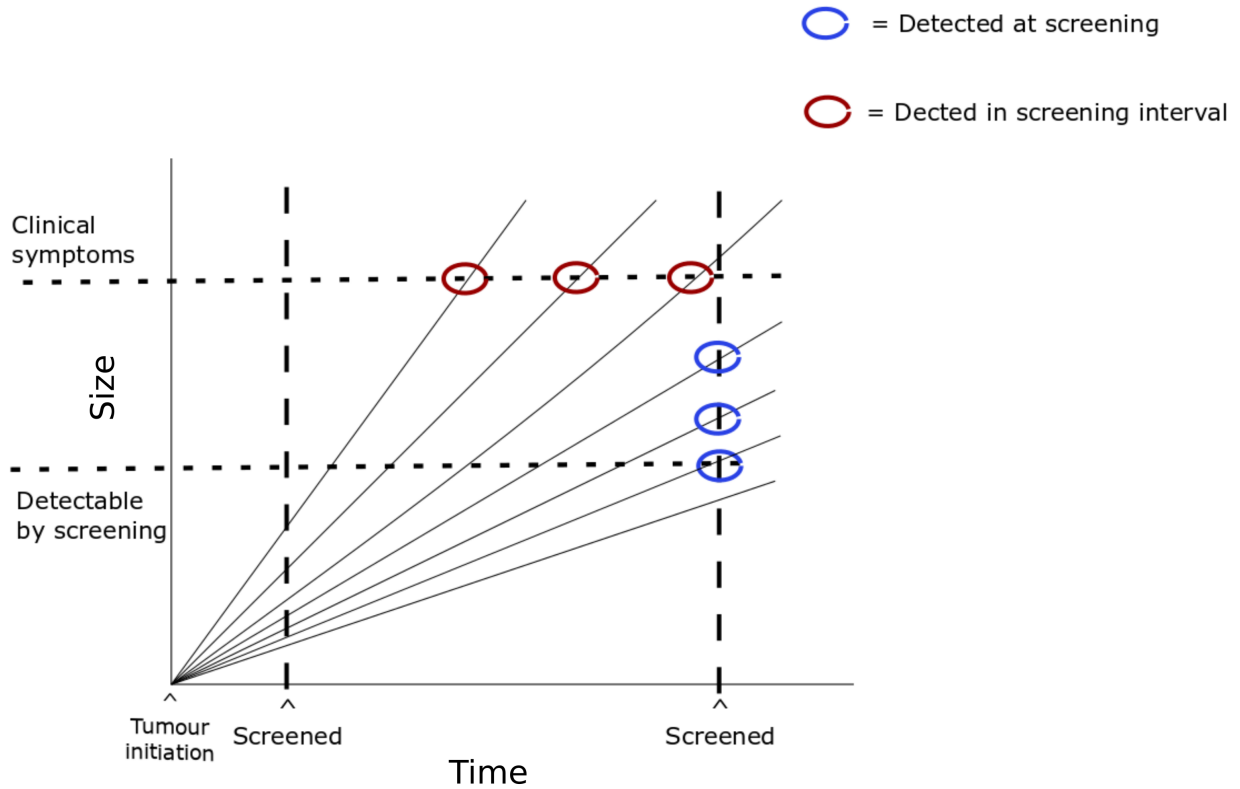
invited women has remained steady throughout the years. Since 2013, invitations are also issued to women aged 70-74 [Lind et al., 2010]. In Southern Sweden, screening intervals as well as age groups called for screening initially varied over time by local screening unit [Olsson et al., 2000]. Following standardisation in 2009, women 40 to 54 were called every 18 months, and women above 55 every 24 months (Personal communication, Boel Hedddson). Participation is generally around 70-80 % [Olsson et al., 2000, Lagerlund et al., 2015].

### **3.3 Aggressive breast cancer**

The term 'aggressive breast cancer' implies breast cancers that metastasise early on in the lifespan of the tumour, and/or grows quickly [National Cancer Institute, 2017]. For the purpose of observational epidemiological studies, fast-growing tumours can only be studied indirectly, as we cannot for ethical reasons passively observe tumours growth over time without intervening in the process. Nor can we be absolutely certain about how early a metastatic process occurs, as we never observe the tumour initiation and tumours progress at different rates. It is therefore necessary to make use of indirect definitions of fast-growing and early metastasising disease.

#### **3.3.1 Interval cancer**

A general feature of any screening program, including mammography screening, is the implicit tendency to preferentially detect slowly progressing disease, a phenomena known as length bias (Figure 4). Rapidly progressing cases are more likely to become detected in the interval between two screening visits, i.e. they become interval cancers. The tendency of interval cancers to contain rapidly progressing disease makes them not only a challenge for screening program designs but also provides means for studying rapidly progressing cases without prior in-depth molecular characterisation. This imbalance in growth rate between interval and screening cancers was used in the first project of this thesis as a proxy for studying characteristics of aggressive breast cancer. Interval cancers are a mix of false-negatives from the last screen and true, fast-growing interval cancers. In either case, they consist a failure of the screening program to detect the cancer and interval cancer rates are part of the evaluation of the quality of a screening program [Törnberg et al., 2010]. In study I, we built on the hypothesis that true interval cancers are enriched among interval cancers diagnosed in women with low mammographic density, i.e. the group with highest screening test sensitivity [Eriksson et al., 2013].



**Figure 4** – Illustration of the principle of length bias. Each line represents a tumour, the slope indicates the growth rate as increase in size over time; the steeper the line, the more aggressive the growth. Inevitable, tumours growing at a faster rate are more likely to be detected in the interval between screening rounds than slow growing tumours, as they take less time to reach the size required for giving symptoms. In this hypothetical example, three out of seven tumours would be interval cancers (red circles) and three would be screening-detected (blue circles), however, in a real setting these proportions would vary by e.g. prevalence of fast growing tumours, screening test sensitivity and vigilance for clinical symptoms in screening-attenders.

### 3.3.2 Prognosticators

Aggressive breast cancer can naturally also be defined by prognosticators of the disease. Among the available prognosticators clinically used today, presence of lymph node metastasis has been shown to have long-term prognostic abilities [Colzani et al., 2011, Fisher et al., 1993]. Additionally, both ER status and tumour size can predict five-year prognosis [Colzani et al., 2011, Thorpe et al., 1986, Fisher et al., 1993]. Tumour grade [Elston and Ellis, 1991] and HER2-status [Paik et al., 1990] are also independent prognostic factors useful in identifying aggressive breast cancer. HER2-positive breast cancer has a much improved prognosis today thanks to targeted therapies, but in the absence of targeted therapy, the nature of the tumour is such that it is prone to early metastasis.

### 3.3.3 Intrinsic or molecular subtypes

Around the turn of the millennium, Charles Perou, Therese Sørli and colleagues established four molecular subtypes of breast cancer based on hierarchical clustering of gene expression of an 'intrinsic' set of 496 genes [Perou et al., 2000]. The main branch clustered tumours according to ER status through differences observed in oestrogen-driven gene expression<sup>2</sup>, denoting them 'luminal' vs. 'non-luminal' clusters. In the non-luminal, ER negative cluster, further subdivision into clusters was seen, denoted 'basal epithelial-like', 'normal breast-like' and 'ErbB2-overexpressing' (In this thesis work, 'ErbB2-overexpressing' will from here on be denoted 'HER2-enriched' and 'basal-epithelial-like' denoted 'basal-like'). These molecular subtypes were corroborated in a larger sample size which also enables further division of the ER+/luminal group into three types, luminal A, B and C [Sorlie et al., 2001].

**Table 1** – Overview of (PAM50) molecular subtypes

Molecular subtype	Characteristics	Recommended treatments
Luminal A	ER+, low grade, luminal epithelial genes.	Anti-oestrogen therapy.
Luminal B	ER+, proliferative, high grade, luminal epithelial genes.	Anti-oestrogen therapy, chemotherapy, anti-HER2 if HER2 +.
HER2-enriched	ER-, most show HER2-amplification proliferative.	Anti-HER2 therapy, chemotherapy.
Basal-like	ER-, HER2-, basal keratines and EGFR, highly proliferative.	Chemotherapy.
Normal-like	Basal and myoepithelial genes, adipose tissue specific genes.	

The division of luminal types into three groups was however not robust enough in sensitivity analysis, with luminal B samples moving into the HER2-enriched group rather than remaining within the main luminal cluster. Luminal B and C tumours were therefore collapsed into one category. These five subtypes (luminal A, luminal B, HER2-enriched, basal epithelial-like and normal breast-like) were robustly confirmed and found to also possess prognostic value in two independent cohorts [Sorlie et al., 2003]. In addition, they have also been independently validated in a Swedish material [Calza et al., 2006]. An overview of these so called molecular subtypes is found in table 1.

In 2009, Joel Parker and colleagues shrank the list of intrinsic genes to 50, with the aim of making it more robust and feasible to classify tumours into molecular subtypes in clinical

<sup>2</sup>The oestrogen receptors are soluble transcription factor regulating the expression of many genes

practice using PCR technology [Parker et al., 2009]. The subtypes determined using their classifier are typically denoted "PAM50-subtypes", after the name of the classifier.

### 3.3.4 Immunohistochemistry subtypes

Parallel to the development of the PAM50-classifier, there was still a need to further simplify the classification of cases according to molecular subtype. With the motivation that not all clinics worldwide have means to assign molecular subtypes to tumours, Maggie Cheang and colleagues developed an immunohistochemistry (IHC) proxy for molecular subtypes based on the staining of ER, PR, HER2 (the protein product encoded by the *ErbB2* gene) and proliferation marker Ki-67 [Cheang et al., 2009]. Their initial proxy was adapted by the St Gallen international expert consensus on the primary therapy of early breast Cancer [Goldhirsch, A. et al., 2013] as a useful way of determining which node-negative, ER positive breast cancers to treat with adjuvant chemotherapy. Prat, Cheang, Parker and colleagues further modified the proxy in 2012 [Prat et al., 2013] to include a criteria of high PR positivity for classifying luminal A tumours. The modified proxy was adapted by the sequential 2015 St Gallen consensus statement [Coates et al., 2015]. The agreement between the PAM50 molecular subtypes and the St Gallen IHC proxy has been evaluated once and found to be moderate ( $\kappa = 0.55$ ) [Romero et al., 2013]. Other IHC proxies for the molecular subtypes of breast cancer exist. A three-marker proxy of ER, PR and HER2 status has been found to provide reasonable accuracy for classifying luminal cases as well as basal-like cases, but low ability to discriminate between luminal A and B, and between the HER2-enriched and luminal B types [Allott et al., 2016]. A six-marker proxy which additionally included IHC stainings for EGFR and for basal-like markers cytokeratin 5/6 has also been used in epidemiological studies of molecular subtypes [Millikan et al., 2008]. To distinguish between gene-expression based and IHC based breast cancer subtypes, 'luminal A', 'luminal B', 'HER2-enriched' and 'basal-like' will be used in the background of this work to denote subtypes defined by gene-expression, and 'luminal A-like', 'luminal B-like', 'HER2-positive' and 'TNBC'<sup>3</sup> will be used to denote IHC based groupings. Distribution of IHC markers for each PAM50 subtype is shown in figure 5 on page 11.

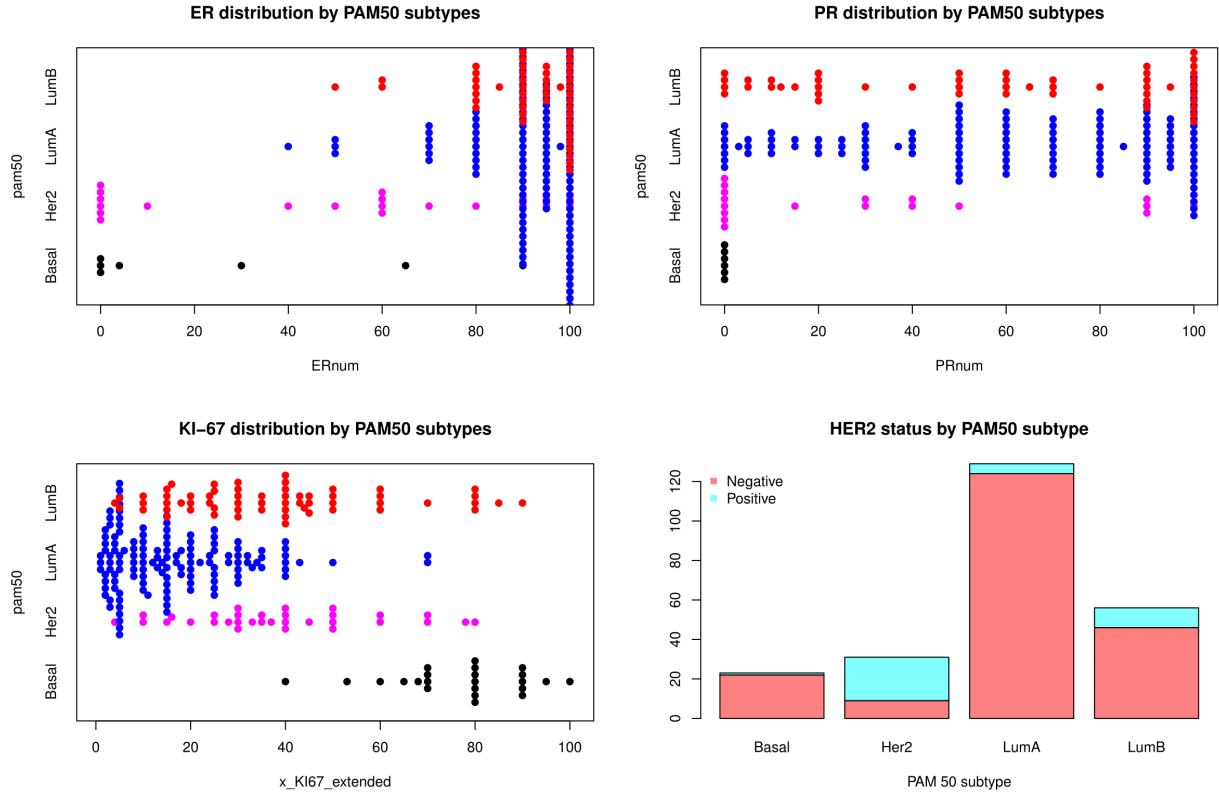
## 3.4 Epidemiology of breast cancer

### Age-related incidence of breast cancer

While the incidence rate of most epithelial cancer increases linearly with age on the log-log-scale [Armitage and Doll, 1954], breast, cervical, uterine and ovarian cancer has a charac-

---

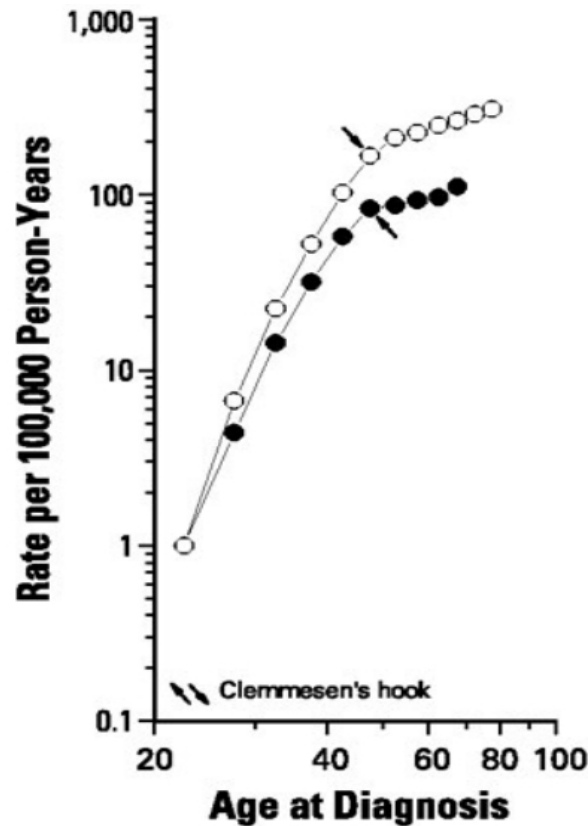
<sup>3</sup>Unless more markers than ER,PR and HER2 were used, in which case the term "basal-like" will be used



**Figure 5** – IHC markers versus PAM50 subtypes in breast tumours. Top panels, and lower left panel: Beeswarm plots of observed values of IHC markers ER, PR and Ki-67 for PAM50 subtypes in Clinseq. Each bee (dot) in the swarm represents one observation. Lower right panel: Bar plots of frequencies of HER2 positive and negative status in Clinseq by PAM50 subtypes.

teristic bend around the (peri)menopausal age, see Figure 6 for the appearance for breast cancer. This bend is denoted "Clemmesens hook"<sup>4</sup>, and distinguishes breast cancer from most other epithelial cancers [Kamangar et al., 2006]. As such, Clemmesen's hook has puzzled epidemiologists for over half a Century [Clemmesen, 1948, Lilienfeld and Johnson, 1955, Anderson and Matsuno, 2006]. Pioneering reports speculated that the change of hormonal milieu in the body during menopause could cause the apparent plateau by pausing cancer progression temporarily [Clemmesen, 1948], or alternatively that it represented a post-menopausal decline in women susceptible to mammary carcinogenesis [Lilienfeld and Johnson, 1955]. During recent years a third interpretation has been added, namely that the pattern corresponds to a bimodal distribution of age at diagnosis, representing two distinct, superimposed distributions of essentially two types of breast cancer; ER positive vs. negative, or premenopausal vs. postmenopausal [Anderson et al., 2002, Anderson and Matsuno, 2006]. ER status of the tumour is correlated with age, with ER positive tumours being more frequent at older ages [Johansson et al., 1984, Mccarty et al., 1983, Pujol et al., 1998]. Consequently, ER status is

<sup>4</sup> After Johannes Clemmesen, Danish epidemiologist and founder of the Danish cancer register

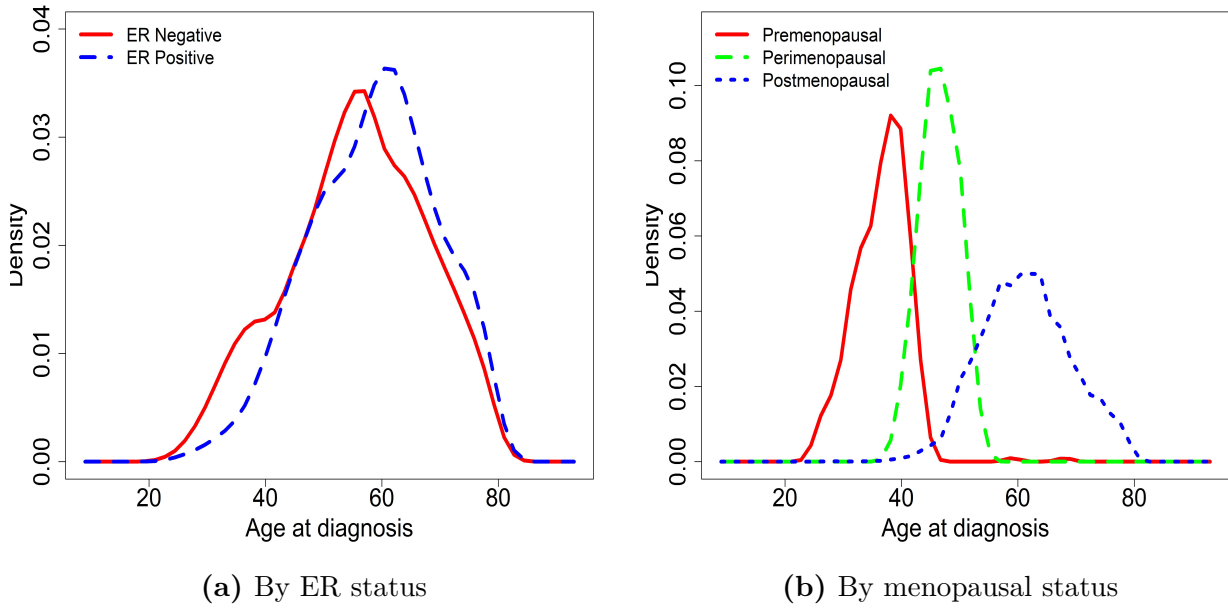


**Figure 6** – Age-specific breast cancer incidence rates in diagnoses 1993-2001 from Cancer Incidence in Five Continents database (CI5VIII), showing Clemmesen's hook. Open circles denote "all regions of Europe, Australia, New Zealand, North America and Japan", filled circles denote "all regions of Africa, Central America, South America, and all regions of Asia except Japan" available in CI5VIII. Image modified from [Kamangar et al., 2006] with permission from the publisher. ©American Society of Clinical Oncology.

also correlated with menopausal status. However, pre- and postmenopausal breast cancer are not synonyms for ER negative and ER positive disease, as shown in figure 7 A and B. In Swedish data for 2001-2008 included in this thesis, the bimodal distribution appears for both ER negative and ER positive cancers (Figure 7). This is similar to results shown from SEER database [Anderson et al., 2002, Anderson et al., 2006] (although our material only includes women below age 80 at diagnosis, cutting our distributions slightly shorter to the right).

### The impact of genes and environment on breast cancer

With the exception of monogenic diseases with complete penetrance, most non-infectious diseases are causally attributed to a combination of both genes and environment. This is certainly true also for breast cancer. Statistical modelling of the level of breast cancer concordance by degree of relatedness in cohorts of twins or other relatives, have provided estimates of 25-27% heritable contribution to the disease at a population level (in Scandi-



**Figure 7** – Distribution of age at diagnosis in (a), all BC diagnoses made in women under 80 years of age 2001-2008 Stockholm-Gotland, by ER status (b), all BC diagnoses made in Libro-1 2001-2008, by menopausal status before diagnosis

navia) [Lichtenstein et al., 2000, Czene et al., 2002]. Indications of the strong environmental contribution to breast cancer came early from descriptive epidemiological studies of migrant populations. Typically, incidence rates for migrant populations alters over generations approaching those of the new country. For instance, in 1960's Okayama in Japan, the BC incidence rates (age-adjusted to the world standard population) for Japanese women were 31.8 and 21.8 for age-groups 35-64, and 65-74 respectively. For the first-generation Japanese immigrants (Issei) in San Francisco, the corresponding incidence rates were 93.6 and 163.4, while for the second-generation (Nisei) the rate was 116.5 (data available for the younger cohort only). By comparison, the rates for San Francisco white Caucasians were 179.4 and 293.4 [Buell, 1973]. This remarkable increase in incidence reproduced in numerous studies [Shimizu et al., 1991, Ziegler et al., 1993], cannot be explained by means other than change of environment as it is present already in the Issei. However the exact causal nature(s) of said environment has remained elusive. For the Japanese migrants, there is also a substantial decrease in the incidence of stomach cancer, with rates approaching that of the U.S. population at large, such that it is unlikely that the changes in breast cancer incidence would be fully explained by potential differences in cancer detection [Maskarinec G, Noh JJ., 2004].

**Table 2** – Selected list of established risk factors for breast cancer

Risk factor	Direction	Comment
Female sex	Increases risk	
Age	Increases risk	
Mammographic density	Increases risk	
Family history	Increases risk	Increases with number of affected relatives, stronger effect for early-onset disease
HRT	Increases risk	
Age at menarche	Decreases risk	Early menarche is a risk factor
Age at first birth	Increases risk	Higher risk for aafb >30 compared to 20
Parity	Decreases risk	Long-term protective effect of parity
Breastfeeding	Decreases risk	Especially for premenopausal breast cancer
Age at menopause	Increases risk	Late menopause is a risk factor
Hysterectomy	Decreases risk	
Chest irradiation	Increases risk	Chest irradiation in early ages
Alcohol intake	Increases risk	
Oral contraceptives	Increases risk	Slight risk increase with current use
Physical activity	Decreases risk	
BMI	Depends	Postmenopausal high BMI increases risk, high BMI early in life lowers risk

### Identified risk factors

Increasing age and female sex are the strongest risk factors for breast cancer, together with rare but high-penetrance mutations in the BRCA1 and 2 genes [Clamp et al., 2003]. High mammographic density is also among the stronger risk factors, with a 4 to to 7 times increased risk for women in the top percentiles of density relative to the lowest percentiles. Increasing number of menstrual cycles is also related to risk of breast cancer, as evidenced by risk increases with early menarche and late menopause. A summary of reproductive, hormonal and lifestyle risk factors is given in table 2 on page 14. In addition, a number of genetic variants have been found to associate with risk for breast cancer, identified either through genetic linkage studies of breast cancer families or through large genome wide associations studies (GWAS). These can chiefly be divided into



1. Rare variants with high penetrance, such as mutations in BRCA 1 and 2, PTEN, and TP53
2. Rare variants with medium penetrance, such as mutations in PALB2, ATM, and CHECK2
3. Common variants with low penetrance, i.e. short nucleotide polymorphisms (SNPs)

### 3.5 Risk factors by breast cancer subtypes

In recent years, studies of risk factors for breast cancer as assessed by molecular markers have been performed. Most have typically analysed risk factors by ER status or combined ER, PR and HER2 status [Barnard et al., 2015]. The earliest established findings for heterogeneity were seen for age and menopausal status, BRCA1 mutations, and ethnicity/race [Boyle, 2012]. Young women tend to have more aggressive subtypes, such as ER negative and triple-negative (TNBC) tumours, i.e. tumours negative for tumour markers ER, PR and HER2. Similarly, mutations in the BRCA1, but not BRCA2 gene, tend to yield basal-like and TNBC tumours more than any other type. In the U.S., Non-Hispanic black women have higher age-standardised incidence rates (ASR) of TNBC than Non-Hispanic white women, and Hispanic and Asian/Pacific islander women have the lowest TNBC rates. Non-Hispanic white women have the highest ASR of ER+/HER2- BC, followed by Non-Hispanic black, Hispanic and Asian/Pacific Islander [Kohler et al., 2015]. In countries with high incidence of breast cancer the predominating subtype is postmenopausal breast cancer, meaning low-incidence countries tend to have higher proportion of premenopausal, aggressive breast cancer (although the age-standardized incidence rates for premenopausal breast cancer may not be as dissimilar to those in high-incidence countries) [Ghiasvand et al., 2014]. Here below will follow a review of the literature on subtype heterogeneity, restricted to risk factors assessed in this thesis.

#### Hereditary risk

In 2014 and 2015, two literature reviews of cohort and case-control studies assessing risk factors by IHC subtypes of breast cancer were published [Anderson et al., 2014, Barnard et al., 2015]. Most published studies investigating family history as a risk factor by subtype of breast cancer have found positive associations to all subtypes, with odds ratios in the range of 1.5 to 2 [Barnard et al., 2015]. However, although the magnitude of association is similar for all subtypes, the genetic variants underpinning the associations may still vary. It is well known that the rare but high-risk BRCA1 mutations tend to give rise to triple negative and basal-like breast cancer in carriers [Boyle, 2012, Perou et al., 2000]. Additionally, there are

several indications that individual risk SNPs may confer risk of breast cancer by a certain ER status only. An early study of individual SNPs found heterogeneity by ER status for all five investigated loci [Garcia-Closas, M. et al., 2008]. Two years later, this finding was expanded on when Reeves and colleagues found heterogeneity by ER status for a polygenic risk score comprising of 14 SNPs. Additionally, they saw heterogeneity for three of the individual SNPs [Reeves et al., 2010]. Four years later, a large international genome-wide association study (GWAS) of triple-negative breast cancer (TNBC) could identify 30 SNPs at 25 loci predicting risk for TNBC [Purrington, K. S. et al., 2014].

### **Age at menarche**

Age at menarche appears to be a risk factor for all types of breast cancer. The literature is consistent in reporting early age at menarche as a risk factor for both ER positive breast cancer as well as for TNBC [Anderson et al., 2014, Barnard et al., 2015]. The limited number of studies performed suggest no evidence of heterogeneity for HER2-positive type and luminal B-like. However, the estimates are less precise and there is some uncertainty for the role of age at menarche for both these subtypes at present [Barnard et al., 2015].

### **Parity**

According to the Anderson review, there was evidence of a protective effect of parity for ER positive BC in 19 out of 22 included studies, but for TNBC, a protective effect was seen only in 3 out of 12 studies [Anderson et al., 2014]. These conclusions were corroborated in the Barnard review, in which the authors considered good evidence for a protective effect of parity for luminal A-like cancers. Protective effect of parity on risk for basal-like breast cancer was less evident, with an increased risk or null association in more than half of the included studies. Conflicting results were seen for luminal B-like and HER2-positive subtypes [Barnard et al., 2015].

### **Age at first birth**

The evidence of effects of age at first birth is unclear for all subtypes with great variation of estimates between studies, except for luminal A-like where a protective effect has been consistently reported [Anderson et al., 2014, Barnard et al., 2015].

### **Breastfeeding**

The most consistent observation of heterogeneity in reproductive risk factors for molecular subtypes is a stronger protective effect of breastfeeding as compared to luminal cancers

[Barnard et al., 2015, Anderson et al., 2014]. Breastfeeding was long seen as non-influential or controversial as a risk factor for breast cancer. A literature review of the field up until 2000 concluded that there was little evidence of any effect of lactation on the risk of breast cancer, if any it was confined to premenopausal cases [Lipworth, 2000]. However, in the same year the first systematic review with a meta-analysis was published, which included 23 case-control studies. The meta-analysis concluded that there was evidence of a protective albeit modest effect, not only found to be strongest for risk of premenopausal breast cancer (OR 0.77 (95 % CI 0.72 to 0.84), but also evident in postmenopausal BC (OR 0.92 (95% CI 0.85 to 0.98)). The authors concluded "a slight but significant protective effect of breastfeeding on the risk of breast cancer in premenopausal women" [Bernier, 2000]. Two years later, a re-analysis of data from 47 studies from 30 countries concluded that there was a slight risk reduction conferred by prolonged lactation, observed both in cohort studies as well as case-control studies [The Collaborative Group on Hormonal Factors in Breast Cancer, 2002]. In recent years, studies of breast cancer by ER status have found the largest protective effects of breastfeeding in ER negative disease. A 2015 meta-analysis of studies assessing the associations between breastfeeding and breast cancer by ER status found evidence for a negative association only in triple-negative or ER/PR negative breast cancer (OR in cohort studies 0.84 (0.72, 0.97), OR in case-control studies 0.76 (0.67, 0.86)) [Islami et al., 2015].

## **Hormone replacement therapy**

The risk for breast cancer by molecular subtype from use of hormone replacement therapy (HRT) has been very sparsely reported as compared to reproductive variables. In the prospective cohort Nurse's Health Study, increased risk for use of HRT was seen for luminal A-like and TNBC, but no clear trend was observed for luminal B-like and HER2-positive [Tamimi et al., 2011]. In contrast, a pooling of two case-control studies found increased risk of luminal subtypes for current use of HRT, but no effect on the risk of TNBC [Phipps et al., 2008]. There are further studies examining the risk by ER and PR status. In the European EPIC cohort, another prospective cohort study, the risk of both ER+/PR+ and ER-/PR- breast cancer was found to be increased for current users of HRT, but the magnitude of increase was larger for ER+/PR+ BC [Ritte, R. et al., 2012]. In the prospective California Teachers Study, evidence for increased risk with HRT use was only seen for ER+/PR+ BC, and possibly HER2+, but not for TNBC [Saxena et al., 2010]. In the AMBER consortium, use of HRT was only associated with risk for ER positive BC [Rosenberg et al., 2015]. There is thus disagreement between studies on the presence of an increased risk for TNBC or ER negative BC from HRT use, and evidence by molecular subtypes is scarce.

## **BMI in young ages**

High BMI at a young age has been found to be negatively associated with breast cancer, but limited results are available from analysis of subtype heterogeneity. The Nurse's Health Study found a trend of inverse association of BMI at 18 and risk for basal-like breast cancer [Tamimi et al., 2011]. The AMBER consortium assessed associations between young-adult BMI and risk of breast cancer by ER status and menopausal status at diagnosis. In premenopausal women, they saw null associations for risk of TNBC but a negative association to ER+ breast cancer. Additionally, they found a protective effect for all postmenopausal breast cancer, irrespective of ER, PR and HER2 status [Bandera et al., 2015]. Studies assessing adult premenopausal BMI have generally found negative associations to luminal subtypes, but null associations to TNBC [Barnard et al., 2015]. However, it is unclear at what age BMI was assessed in these studies. One study saw a positive association between premenopausal BMI and risk of TNBC [Gaudet et al., 2011, Barnard et al., 2015].

## **Mammographic density**

A meta-analysis in 2013 of studies analysing mammographic density by receptor status concluded no evidence of heterogeneity by ER status or HER2 status [Antoni et al., 2013]. Our group has previously investigated mammographic density by molecular subtypes in a analysis of 111 cases of breast cancer, without finding any indication of heterogeneity [Eriksson et al., 2012]. In the Carolina Breast Cancer Study, no difference in associations to luminal A-like and basal-like breast cancer was found [Razzaghi et al., 2013]. Similarly, no heterogeneity by IHC subtype was found in the Breast Cancer Surveillance Consortium [Phipps et al., 2012].

## **3.6 Risk factors by mode of detection**

Risk factors for interval cancer is of interest to identify, as women diagnosed with IC have had no benefit from participation in the mammography screening. If individualised screening will be implemented in the future [Shieh et al., 2017], knowledge of factors predisposing to interval cancer could prove useful in screening program designs.

Studies comparing risk factors for breast cancer between interval and screening-detected cancers have generally been low in case numbers. First and foremost, by its influence on mammography screening sensitivity, high mammographic density is overrepresented among interval cancers [Domingo et al., 2014]. Findings are additionally consistent for a higher frequency of current users of HRT among IC [Kirsh et al., 2011, Domingo et al., 2010, Wang et al., 2001, Gilliland et al., 2000]. Studies disagree on whether family history is

more common among IC. Four studies found a trend for over-representation among IC [Mandelson et al., 2000, Kirsh et al., 2011, Lowery et al., 2011, Domingo et al., 2010]. In contrast, three studies found either no difference or the opposite pattern [Gilliland et al., 2000, Musolino et al., 2012, Brekelmans et al., 1994]. The parametrisation of family history varied across these studies, from comparing any history verses none, to analysis by number of affected relatives but no clear patterns related to categorisations are evident.

### 3.7 Risk prediction

A number of statistical models have been developed in the last thirty years, aimed at predicting an individual woman's risk of breast cancer. Such models are or can be employed for selection of high-risk women, based on a predefined cut-off, for chemo-prevention trials, increased screening practices or other means of prevention [Shieh et al., 2017, Howell et al., 2014]. In 1989, Mitchell Gail et al developed one of the earliest and probably most wide-spread models, known as the Breast Cancer Risk Assessment Tool or popularly, the Gail model. The original Gail model was based on age, age at menarche, age at first live birth, number of previous breast biopsies, and the number of first degree relatives with breast cancer [Gail et al., 1989], but has since then been extended. The Gail model is also the model with the most validation reported. The discriminatory capacity has been found to vary from AUC of near 0.50 (i.e. no better than random guesses) to 0.67 in different validation populations [Rockhill et al., 2001, Tice et al., 2005, Adams-Campbell et al., 2007, Mealiffe et al., 2010]. In a cohort study of postmenopausal women, the model was found to be useful only for predicting ER positive BC (AUC = 0.60), as the AUC for ER negative disease was 0.50 [Chlebowski et al., 2007]. Nor has the Gail model shown any value for predicting breast cancer in a cohort of African-American women, with an AUC near 0.50 [Adams-Campbell et al., 2007]. The model has also been found to overestimate breast cancer risk in a Singaporean population [Chay et al., 2012]. Later models have also focused on more detailed information about family history, such as degree of relatedness and age at onset in the relations. Some models are restricted to modelling genetic susceptibility based on information on family history, such as the BOADICEA model [Antoniou, A.C. et. al., 2008]. Evaluated in this thesis, the Tyrer-Cuzick model attempts to estimate the contribution of family history separated into two components of low-penetrance and high-penetrance (BRCA) gene mutations, respectively. It further includes parameters age, parity, age at first birth, height, BMI, age at menopause, use of hormone replacement therapy, presence of lobular carcinoma *in situ* as well as history of atypical hyperplasia [Tyrer et al., 2004]. The Tyrer-Cuzick model has shown comparatively better discriminatory capacity than the Gail model in at least two studies [Powell et al., 2014, Quante et al., 2012].

In a Californian population high in nulliparous women, it was found to have an AUC of 0.65, slightly outperforming the Gail model which showed AUC of 0.62 in this population [Powell et al., 2014]. Similarly, in a study performed in the New York City Breast Cancer Family Registry, the Tyrer-Cuzic model showed an AUC of 0.69, compared to the Gail models AUC of 0.62 in this population [Quante et al., 2012]. However, the Tyrer-Cuzic model has been shown to overestimate BC risk in a population of women with atypical hyperplasia [Boughey et al., 2010]. The added predictive value of information on mammographic density to risk prediction has been evaluated [Tice et al., 2005, Chen et al., 2006], including for the Tyrer-Cuzic model [Warwick et al., 2014]. Synthesised genetic information in the form of polygenic risk scores associated with risk for breast cancer have also been tested, with the aim to predict risk of breast cancer [Mavaddat, N. et al., 2015]. In a Swedish case-control study, the AUC for an extension of the Gail model with added information on BMI, mammographic density and 18 breast cancer risk SNPs was found to be 0.62, as compared to and AUC of 0.56 when using the original Gail model [Darabi et al., 2012].

## 4 Data material

### 4.1 Breast cancer cohorts (Papers I-IV)

This thesis is based on three breast cancer cohorts, two from Stockholm (Libro-1, STO-3), and one from Stockholm and southern Sweden (KARMA)(Figure 8). Study I and II are built solely on cases from Libro-1 whereas, in addition study III and IV includes cases and controls from KARMA, and study IV also includes cases from the STO-3 trial. A graph comparing the recruitment period for each study is given in figure 9 on page 22.



**Figure 8** – Map over origin of cohorts included in the thesis. Libro1 and STO-3 are diagnoses made in Stockholm County. Half of Karma is recruited from Skåne County and half from Stockholm.

#### The Libro-1 study

Libro-1 is a case-only cohort of 5,715 breast cancer cases from the Stockholm-Gotland region of Sweden diagnosed during the years 2001 to 2008. All women diagnosed during these years still alive in 2009 were invited to take part of the study. Starting from the diagnoses listed in the Stockholm-Gotland regional breast cancer quality register (described in a subsection below), all unique breast cancer diagnoses made between 1st of January 2001 and 31st of December 2008 were identified ( $n = 11,707$ ). Cases diagnosed after age 79 ( $n = 1,249$ ) or among male patients ( $n=11$ ) were excluded, and addresses for the remaining 10,447 individuals were accessed from the Swedish Tax Agency, who holds addresses to all Swedish residents. After excluding deceased cases ( $n = 645$ ) as well as those for whom no address could be

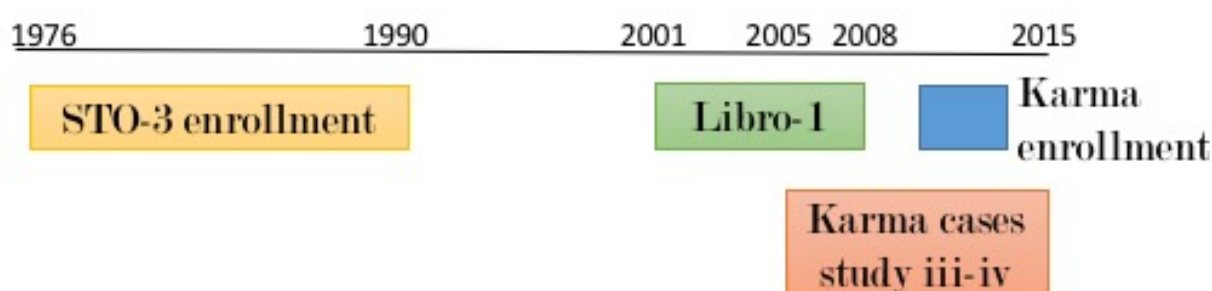
identified ( $n = 454$ ), 9,348 cases were invited by letter to participate in the Libro-1 study. Several attempts were made to contact non-responders, and the final number of consenting cases amounted to 5,715 (consent rate = 61%).

### The KARMA study

KARMA, or KARolinska MAMmography project for risk prediction of breast cancer, is a large, prospective cohort study that invited all women who underwent population-based mammography screening or clinical radiology examinations at Stockholm South General Hospital, Helsingborg Hospital, Skåne University Hospital, Lund, and Landskrona Hospital between January 2011 and March 2013. In total 210,233 women were invited and 70,877 consented (consent rate = 34 %). The cohort profile of KARMA has recently been published [Gabrielson et al., 2017].

### The STO-3 trial

The STO-3 is a subgroup of the Stockholm trial on adjuvant tamoxifen in early breast cancer, a randomized trial comparing the survival benefit of adjuvant tamoxifen to no tamoxifen [Rutqvist and and, 2007]. The trial enrolled postmenopausal women < age 71 with histologically verified invasive breast cancer diagnosed in Stockholm. Recruitment began in 1976 and the trial remained open until 1990, when evidence of a survival benefit from adjuvant tamoxifen emerged, rendering continued randomization unethical. STO-3, the material included in this thesis, consisted of the subgroup of 891 trial participants (63% of the full trial) classified as "low-risk", defined as lymph node negative (N0), with tumour size (pT) < 30 mm.



**Figure 9** – Overview of cohorts included in the thesis. STO-3, recruitment 1976-1990. Libro-1, diagnosis 2001-2008 in Stockholm county. Karma, enrollment 2011-2013. Prevalent and incident Karma cases diagnosed 2005-2015 used in studies iii-iv.



## 4.2 Population-based registers (Papers I-IV)

In addition to study-specific data collected in each cohort, the studies have been merged to official registers for information on cancer diagnosis, follow-up and detailed information regarding the tumours.

### A short note on population registers in Sweden

The practise of registering the population has a long tradition in Sweden, originating from the church books kept by each local parish, enforced nation wide by law in 1686. The first attempt at centralized record keeping was made in 1749 when "Office of Tables"<sup>5</sup>, Tabellverket, was launched to provide population-based statistics for knowledge-based governance and research [Wannerdt, 1982]. Tabellverket was eventually re-launched as Statistics Sweden, the government authority still responsible for official statistics. Today, anyone born in or intending to live in Sweden more than a year is required to be included in the Swedish Population Register<sup>6</sup> [Ludvigsson et al., 2009]. For this purpose, a unique ten-digit Swedish personal identity number (PNR) is assigned to each individual by the Swedish Tax agency<sup>7</sup>. Having a PNR ensures both the individuals right to voting and social benefits, as well as their duties, such as paying taxes. The Tax Agency's population register is updated and transferred daily to the government agency Statistics Sweden, which keeps the Total Population Register [Ludvigsson et al., 2016]. Through the PNR, record keeping and linkages are made possible for several beneficial purposes, including tracking individual medical records across different health care institutions, keeping health care registers for evaluation of health care as well as enabling population-based epidemiological research [Ludvigsson et al., 2009]. Of relevance for this thesis, the National Board of Health and Welfare<sup>8</sup> is the governmental agency responsible for official statistics concerning health and disease including health care and causes of death.

### The Swedish Cancer Register

All Swedish physicians irrespective of health care provider are obliged to report their diagnoses of malignant tumours, as well as certain benign tumours, to the Swedish Cancer Register. In addition, pathologists and cytologists are also obliged to report all tumours diagnosed from excised specimens, aspirates and autopsies. The register was initiated in 1958. Since mid-1980's, cancer registration is made to the six Regional Cancer Centres<sup>9</sup> (RCC),

---

<sup>5</sup>Fantastic name

<sup>6</sup>Swedish: Folkbokföringsregistret

<sup>7</sup>Swedish: Skatteverket

<sup>8</sup>Swedish: Socialstyrelsen

<sup>9</sup>Swedish: Regionala cancercentrum

which are locally responsible for coordinating, coding and ensuring quality of the reporting. The RCCs in turn report cancer statistics annually to the National Board of Health and Welfare, holder of the Swedish Cancer Register [Talbäck, 2011, Barlow et al., 2009]. The completeness of reporting to the Swedish Cancer Register has been evaluated twice, showing to be 95.5 % for all cancer-sites in 1978 [Mattsson and Wallgren, 1984] and 96.4 % in 1998 [Barlow et al., 2009]. Breast cancer was reported at 98.1 % in 1978, and evidenced the highest level of completeness of all cancer sites reported in 1998. Most under-reporting was seen in patients older than 70 at diagnosis and in cases lacking histological confirmation of the cancer [Mattsson and Wallgren, 1984, Barlow et al., 2009]. Information available for each tumour diagnosed includes;

- Date of diagnosis (defined as earliest health care contact made, most often the date when the cytology/biopsy referral was written to pathology)
- Invasive, yes or no (malign/benign)
- TNM stage variables from clinical examination (Available from 2003-09-15), according to UICC version 6
- SNOMED coding of histological type

### **The Breast Cancer Quality Register**

The Cancer Register includes all types of tumours diagnosed, and as such, the scope for detailed information for specific cancers is limited. Instead, detailed information for breast cancer tumours and their treatment is available through the Breast Cancer Quality Register (BCQR). From 1976 to 2006, a regional quality register was kept at each of the six RCC units, but in 2007 a national breast cancer quality register named INCA was launched. Our study participants have been diagnosed both before and after the launch of INCA, and thus both versions of the BCQR have been used in this thesis. Variables derived from the BCQR for use in the thesis include:

- Date of diagnosis (defined as earliest health care contact made, most often the date when the cytology/biopsy referral was written to pathology)
- Invasive, yes/no
- Tumour size in mm, pT and clinical T
- Lymph node involvement, pN and clinical N
- Distant metastasis at diagnosis, pM and clinical M

- ER status, positive or negative (Available from 2001)
- ER percent staining (Available from 2007)
- PR status, positive or negative (Available from 2001)
- PR percent staining (Available from 2007)
- Nottingham Histological Grade, 1 to 3 (Available from 2004)
- HER2/neu status (Available from 2007)

The reporting to the register has not been validated overall in any independent publication, but the registry-reported treatment has been validated against medical records and found to have a high degree of accuracy [Wennman-Larsen et al., 2016]. Additionally, comparisons of data collected for this thesis on ER and PR staining from medical records was compared to information reported to the register and found to be of high concordance (Lin's concordance correlation coefficient 0.92 (0.91 - 0.93), unpublished data). As mentioned above, the reporting of a cancer diagnoses is made to the RCC quality register and through RCC further reported to the National Cancer Register. In order to harmonize the data between the BCQR and the Cancer Register, the staff at RCC performs quality checks of the reporting and curates any discrepancies in reporting of diagnosis date etc. Thus, if the date of diagnosis differs between the BCQR and the cancer register, the earliest date is entered for both registers <sup>10</sup>.

## **The Cause of Death Register**

The Cause of Death register was used to gain information on follow-up of survival outcomes in study II and IV. Since 1952, cause of death for anyone with a Swedish PNR is available for research electronically through the Swedish Cause of Death Register. The register has recently been described fully in an international publication by Brooke and colleagues [Brooke et al., 2017]. Cause of death is ascertained by physicians in accordance with WHO standards and coded using the International Statistical Classification of Diseases and Related Health Problems (ICD) codes. Before 1991, the register is essentially complete as the issuing of a death certificate was legally required for a burial up until 1991. Since 1991, *the number of deaths* is essentially complete, but 0.9 % of deaths are registers as missing death certificate. 96 % of all entries have a specific cause of death noted [Brooke et al., 2017]. In two studies, the quality of the reporting to the register was found to be high for malignancies, with 90% accuracy for malignancies overall [Johansson et al., 2009], and 93 % accuracy found

---

<sup>10</sup>Personal communication, Annette Asterkvist, administrator of the National Cancer Register and BCQR at RCC Stockholm-Gotland

specifically for breast cancer as compared to reviewed case summaries from patient records [Nyström et al., 1995].

### **4.3 Questionnaire data (Papers I-III)**

In studies I, II and IV we made use of questionnaire data filled out by the study participants in Libro-1 and KARMA at time of enrolment. Web-based questionnaires were used as the main format. However, in Libro-1 women could also opt to have a paper questionnaire mailed to them at their home address and 30 % of women chose this option. The questionnaires share a large proportion of the questions posed, but the KARMA questionnaire, which builds on the Libro-1 questionnaire, is more extensive and detailed on questions regarding family history, parity and breastfeeding, and also covers aspects such as diet which was not included in Libro-1. However for this thesis, only questions regarding areas covered in both questionnaires have been included. Whenever depth of information varied across cohorts for a variable, the information was truncated to the lowest common level of information.

### **4.4 Mammography screening attendance and outcomes (Paper I)**

#### **Data collection**

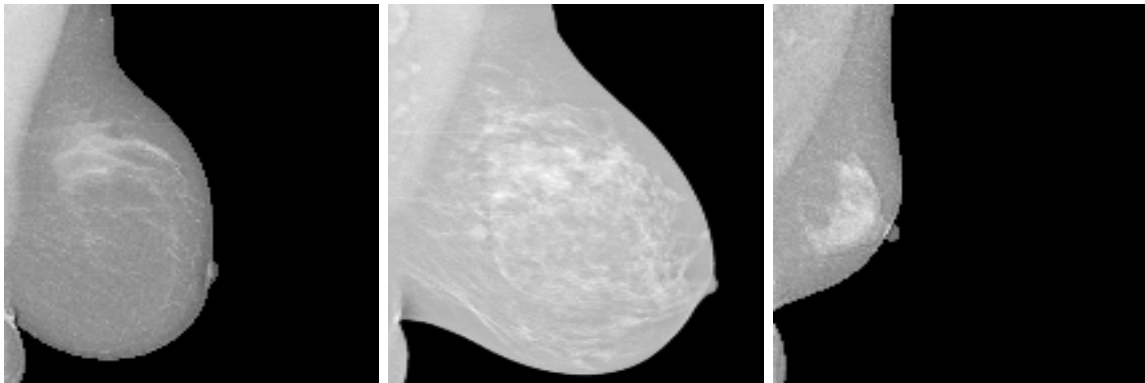
Individual-level data on mammography screening attendance underlined the study outcome definition in study I. In the Stockholm-Gotland region, data on participation and outcome of the population-based mammography screening program is transferred from each mammography clinic to a database kept at the regional cancer centre (RCC). Stockholm-Gotland was the only RCC unit that kept centralized records of the mammography screening, for all other regions raw data must be retrieved from each mammography clinic. By linkage through the PNR, for each individual the date of mammography visit, initial outcome ('selected for follow-up' vs. 'healthy') and follow up of abnormal findings all the way to a potential confirmation of a cancer diagnosis is retrievable. We obtained data from the RCC mammography screening database of all visits undertaken by women diagnosed in 2001-2008 and alive in 2009 (n= 10,447), including all Libro-1 participants, and the associated screening outcomes.

### **4.5 Mammography images (Papers I-III)**

#### **Data collection**

In studies I, II and IV, measurements on mammographic density were included in analysis. These measurement were derived from mammography radiology images collected in the Libro-1 and KARMA studies. For Libro-1, analogue mammograms were collected in 2011-2012

from hospital archives at the four major mammography clinics in the Stockholm region, and all digital mammograms were retrieved electronically and stored in house. Historically, clinics performed analogue imaging in the radiology department, but during the period 2006-2008, all four Stockholm clinics converted to digital imaging. South Hospital was first to go digital in April 2006, followed by St Göran hospital in August 2007, Karolinska hospital in May 2008 and Danderyd hospital in August 2008. The proportion of women with radiology films identified and collected from treating clinic was 68 %, range from 63 - 95 % across clinics, after the initial collection. To increase completeness of the data a second phase of collection of analogue films was carried out in 2013-2014. The image archives at all four clinics were revisited, this time seeking images for all study participants at all clinics, and films which had been sent for long-term archiving outside of the clinics were retrieved. The final proportion of Libro-1 women with at least one analogue film collected was 84 %. All analogue films were digitized using an Array 2905HD Laser Film Digitizer (Array Corporation, Tokyo, Japan) before being returned to the clinics.



**Figure 10** – Example of MLO view mammography images from one of the study populations

### Measurement of mammographic density

Mammographic density was measured using a fully automated approach previously developed by our group in 2011 [Li et al., 2012]. Briefly, the approach was as follows: Image pixel features ( $n = 772$ ) were extracted using ImageJ software, from analogue mammography images from an independent case-control cohort. The dimensionality of the features was reduced in two steps to avoid over-fitting: Firstly, principal components analysis was applied reducing the information into 93 principal components (PCs). Secondly, by using penalized linear regression with the lasso method, variable selection was performed under cross-validation. Specifically, the 93 PCs were entered as covariates into two separate lasso regression models to predict total area of the breast and total dense breast area respectively, as annotated by Cumulus. Using the resulting regression models, total breast area and total dense breast area

could then be predicted on unseen images, given their image features. Percent mammographic density (PD) could then simply be calculated as dense area/total area. The method showed a Pearson correlation coefficient of 0.88 with Cumulus, and the mean difference in PD between the two methods was 0.019 (95% CI, -1.66 to 1.69)[Li et al., 2012]. This model was applied to the analogue images for the Libro-1 and KARMA study participants, predicting percent and absolute area of mammographic density. In paper IV, an updated version of the above method was used in order to also predict accurately PD on digital images in KARMA. This was achieved by repeating the algorithm outlined above with a new training dataset that also including digital images. With the amount of mammography images available in Libro-1 and KARMA (> 1,000,000), a high-throughput approach of analysing images is preferred over manual or semi-automated approaches such as Cumulus, for reasons of both time and cost, as well as precision of the measurement.

## 4.6 Genotype data (Papers II and III)

### Data collection

In studies II and IV, data on individual genetic markers (SNPs) analysed in DNA extracted from blood was included in the analysis. Libro-1 participants had been provided 2x10ml EDTA-buffered blood-sampling tubes with transportation kits at study invitation and were instructed to donate blood through phlebotomy by staff at their local clinic. Instructions on blood sampling protocols for the study were also provided as a separate letter addressed to the medical professionals. Through phlebotomy performed by trained research nurses belonging to the study, KARMA participants donated 24 ml of blood at the KARMA study centres. Samples collected from LIBRO-1 and KARMA participants were sent to the UK and genotyped using the Illumina iSelect SNP Array (iCOGS) of 211,155 SNPs. Missing information on genotypes were imputed using 1000 Genomes [specifically phase I integrated variant set release (v3) in National Center for Biotechnology Information build 37 (hg19) coordinates].

### Calculation of polygenic risk scores

Polygenic risk scores used in this thesis were calculated as described in [Li et al., 2015]. Briefly, each score was based on the same 77 SNPs, namely all SNPs discovered to date of construction of the PRS, in either the COGS GWAS consortia or previous discovery GWAS [Mavaddat, N. et al., 2015]. A complete list of SNPs included is given in [Li et al., 2015]. The polygenic risk scores were calculated by summing the number of risk alleles for each of the 77 SNPs (0, 1 or 2 alleles), weighted by the per-allele log odds ratio for the minor alleles

reported by Mavaddat and colleagues [Mavaddat, N. et al., 2015]. The formula for the PRS was thus calculated as follows for each individual

$$PRS = \beta_1 x_1 + \beta_2 x_2 \dots \beta_n x_n$$

where 1- $n$  is the respective SNP,  $\beta_n$  is the per-allele log odds ratio of SNP $n$ , and  $x$  is the number of alleles (0, 1 or 2) for SNP $n$ . Three PRS were calculated, one for log-odds ratios for associations to breast cancer in general, and one each for associations to ER positive and ER negative breast cancer respectively.

## 4.7 Medical records (Papers III and IV)

### Data collection

In studies III and IV, data collected from medical records and pathology reports on immunohistochemical analysis of tumour specimens was included in the analysis. Although the BCQR nowadays includes variables on percentage staining of the estrogen and progesterone receptors (ER, PR), this was not the case before year 2007. In addition, information on HER2 status was not registered prior to 2007, and information on Ki-67 staining was only included in 2013, reported as "low" or "high". To obtain information on percentage staining and status of the IHC variables, the thesis author or a research nurse instructed by the author, extracted raw information from digitized and/or digital pathology reports and medical journals. After having performed a pilot data extraction on 100 individuals, data was entered into a study-specific data sheet. Quality control of the collection was assessed by the author by random re-assessment of data entries made by the research nurse, as well as of 5 % of the data collected each day by the author. Additional checks were made by comparing the cleaned data to the data reported to the Breast Cancer Quality Register for Libro-1 cases with overlapping information. Completeness of continuous data on ER and PR staining was higher in the study-collected data, but the agreement of staining according to both sources was high (Lin's CCC 0.92 (0.91 - 0.93)).

## 5 Statistical methodology

### 5.1 Statistical models used throughout the thesis

#### 5.1.1 Logistic regression

In regression analysis, it is possible to control for several variables at once when the regressing an exposures on the outcome. For this reason regression models are some of the most used tools in analytical epidemiology. For many of our study questions, we have been dealing with categorical study outcomes of a binary nature, and hence logistic regression was used in studies I-II. The logistic regression model is a linear predictor function given by

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

The logistic transformation of  $(p)$  is necessary because the right hand side of the equation may take on values greater than 1 or less than 0, which are not defined for a probability. By taking the logarithm we arrive at  $\beta$  expressed as log-odds. By exponentiating the  $\beta$ 's we get the odds ratio (OR) for the exposures in question.

Logistic regression assumes that the observations are independent of each other, and that collinearity between the exposure variables is low. Further, it assumes that continuous variables are linear in the logit, and that there are no strong outliers.

#### 5.1.2 Multinomial logistic regression

Multinomial logistic regression (MLR) was used in studies I-III to estimate adjusted odds ratios. MLR is an extension to the logistic regression, and is for unknown reasons further known by many English names, such as polytomous logistic regression, polychotomous logistic regression and multiclass logistic regression. It is useful when we have a  $n > 2$  categorical outcome variable which in addition is nominal, such as subtypes of breast cancer. Multinomial logistic regression estimates OR's separately for each outcome category by fitting  $k-1$  models, where the  $k^{th}$  equation is relative to the referent outcome group and  $k$  is the number of outcome categories <sup>11</sup>.

---

<sup>11</sup>Technically, the estimates from a MLR are not odds ratios, but a ratio of two relative risks, denoted relative risk ratios (RRR's). When there are only two categories in an MLR, the RRR is equivalent to the corresponding OR from a binary logistic regression. The RRR is commonly interpreted as odds ratios and have been denoted as OR's throughout all studies included in this thesis. For more information see <https://www.stata.com/statalist/archive/2005-04/msg00678.html>



### 5.1.3 Proportional-Hazards regression

Proportional-Hazards regression, popularly known as the Cox model after its initiator, was used to model survival in study II. With it, the hazard  $h(t)$  of an event at time  $t$  is modelled as a function of a baseline hazard and the independent variables, given by

$$h(t) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)$$

where  $h_0(t)$  is the baseline hazard at time  $t$ , for a person with the value 0 for all the independent variables. The advantage of this model is that no assumption about the shape of the underlying hazard is needed. A key assumption of the Proportional-hazards model is that censoring is non-informative, i.e. the mechanism behind an observation being censored should not be related to the probability of the event of interest occurring. Furthermore, the hazards for different strata must be proportional over time. This can in some settings be a strong assumption, and it must always be checked by i.e. inspecting the plot of Schoenfeldt residuals against time.

### 5.1.4 Random forest

Random forest classifiers were used in studies III and IV. A random forest belongs to the category of supervised machine learning algorithms, i.e. statistical classifiers or predictors that learn to classify unseen observations by teaching themselves decision rules for classification in a training data set. The term 'supervised' refers to the use of an annotated training data with "correct" answers, which serves as a guide or supervisor to the algorithm. This can be contrasted against unsupervised learning, for example the clustering algorithms used to investigate the patterns of different gene expression clusters between breast cancer tumours [Perou et al., 2000].

The name '**Random forest**' describes in two words the essential components of the algorithm. 'Forest' hints that the algorithm builds multiple decision trees, which together forms a forest of trees. Rather than relying on a single tree, increased robustness is achieved by consulting an ensemble of trees<sup>12</sup>. 'Random' refers to the statistical process of constructing the trees in the forest: For each tree construction, the algorithm is handed a randomly drawn-with-replacement sample (a bootstrapped sample) of the same size  $n$  as the training data. In addition, the algorithm is for every tree handed a random subset of the features (also known as predictors, or variables) available for classification. Thus, the resulting forest of decision trees consist of trees built from random representations from the underlying training data on both the observation and the feature level. This diversity of trees generally intends

---

<sup>12</sup>Random forest is thus an example of ensemble learning methods

to ensure robustness of the resulting forest, whereas isolated decision trees tend to over-fit to the training data resulting in low external validity. The decision on how to classify an unseen observation is taken by a majority vote of the decisions reached by each individual tree in the forest. For example: An illustration of three decision trees is provided in Figure 11 on page 38. Together, they make up a small forest, albeit not a random forest. Given an observation classified as ER positive, HER2 negative, PR positive, Ki-67 high, each tree would classify this as luminal B (tree (a) and tree (b)) or luminal A (tree (c)). The majority vote by this forest would classify that observation as luminal B.

### 5.1.5 Measures of concordance and performance of classifiers

In study IV, we evaluated several IHC proxies in their ability to classify tumours by PAM50 status. Several metric for evaluating classifiers exist [Hossin and Sulaiman, 2015], however we settled on Cohen’s kappa metric and accuracy. Additionally, we calculated class-wise sensitivity and specificity. The metrics are described in table 3 below. Accuracy is essentially the percentage agreement, i.e. the fraction of observations that are classified correctly out of the total number of observations. Kappa estimates the excess in agreement observed above that expected from random guesses. It is usually interpreted as values  $\leq 0$  indicating no agreement beyond random guessing, 0.01 to 0.20 as slight, 0.21 to 0.40 as fair, 0.41 to 0.60 as moderate, 0.61 to 0.80 as substantial, and 0.81 to 1.00 as near perfect agreement [McHugh, 2012].

**Table 3** – Metrics used for evaluations of IHC proxies in study IV. TP = true positive. TN = true negative. FP = false positive. FN = false negative. Pr(o) = observed agreement (defined the same as Accuracy). Pr(e) = expected agreement if random guessing.

Metric	Formula	Interpretation
Accuracy	$\frac{tp+tn}{tp+tn+fp+fn}$	The percentage agreement, or fraction of correct decision of all decision made
Kappa	$\frac{Pr(o)-Pr(e)}{1-Pr(e)}$	The excess agreement observed, above that from random guessing
Sensitivity	$\frac{tp}{tp+fn}$	The fraction of true positives correctly identified
Specificity	$\frac{tn}{tn+fp}$	The fraction of true negatives correctly identified

### 5.1.6 Kaplan-Meier, or Product-limit estimator

Kaplan-Meier plots were used in study IV. With large enough sample sizes, the Kaplan-Meier estimator attempts to estimate the survival function over time for a population. If the status

at the last observations (either right-censored or having the event) and the time from start of follow-up to the last observation is known for all subjects, the curves of said estimates can be plotted overall or by groups of interest. The curve will be displayed as a step-wise declining function, which declines as events occur. Censored observations are denoted by a small vertical tick on the line at time of censoring. A limitation of the Kaplan-Meier method is that it cannot take into account the influence of other covariates; for adjusting, other methods such as the Cox model has to be used. Moreover, due to the declining number of observations over time, the estimated probability of survival is more precise immediately after start of follow-up than towards the end.

## 5.2 Study designs

### 5.2.1 Paper I

Paper I was a case-only study with the aim of comparing interval breast cancers (IC) to screening detected breast cancers (SC) with respect to tumour characteristics and risk factors, overall and stratified by mammographic density quartiles. All invasive breast cancer diagnoses made 2001-2008 in Stockholm-Gotland region in females aged 40-71 at diagnosis were eligible for analysis regarding tumour characteristics. Diagnoses where mode of detection was not IC or SC were excluded. In analysis stratified by mammographic density, and in all assessments of risk factors, analysis was by necessity restricted to participants in the Libro-1 study, which had been recruited from the aforementioned population.

#### **Classification of outcome: Mode of detection**

Starting from the diagnosis date as entered in the BCQR, the closest preceding date of mammography screening was identified for each individual (denoted "index screen"). In the absence of any record of a pre-diagnostic screening visit, the individual was classified as a symptomatically detected cancer in a non-attender, and thus not included in study I. For the remaining cases, the time between index screen and diagnosis was calculated and compared to a normal screening interval, which for the region and time period corresponded to 18 months for women ages 40-49 at diagnosis, and 24 months for women ages 50 or above. If the time in days between the index screen and cancer diagnosis exceeded the normal screening interval, the individual was excluded from the study and classified as symptomatically detected cancer in a former screening-attendant. By definition, for all remaining cases, the mode of detection must either be screening-detected (SC) cancer or interval cancer (IC).

Screening-detected cancer was defined as having a diagnosis of cancer within the normal screening interval after the index screen, where the individual had been selected for follow-up

of a suspicious mammography finding and had a PAD code confirming a diagnosis of breast cancer following the index screen follow-up. Interval cancer was defined as having a diagnosis of cancer within the normal screening interval, for individuals declared healthy and not selected for further follow-up at index screen. A special category were individuals who had been selected for follow-up at the index screen but lacked a PAD code confirming the diagnosis ( $n = 173$ , whereof invasive  $n=143$ , whereof Libro-1 participants  $n = 99$ ). Approximately 50 % of these cases (48 Libro-1 cases) were diagnosed within 3 months of the index screen and therefore likely true SC's despite missing PAD code. Moreover, approximately 50 % (51 Libro-1 cases) were diagnosed  $\geq 180$  days after the index screen, thus likely interval cancers which were false negatives from the index screen. For the sake of having a clean study definition, all 173 cases selected for follow-up at index screen that lacked PAD-code were excluded from analysis.

### **5.2.2 Paper II**

Study II was designed as a case-only study. All Libro-1 women with a primary breast cancer diagnosis were eligible for inclusion. For this study question, non-invasive tumours were included in analysis. The aim of the study was to assess whether scores from the Tyrer-Cuzick model, the PRS for breast cancer and percent mammographic density were equally dispersed between ER negative and ER positive cases, between tumours of high and low grade, between lymph node positive and negative tumours, as well as tumours of different size. This was achieved by regression analysis of the scores against outcome variables tumour prognosticators. Additionally, proportional hazards regression was performed to compare hazard rates of distant metastasis for cases with high and low scores of the risk prediction tools respectively. Cases were followed from date of diagnosis until either censoring (at date of death from any cause or end of follow-up in September 2014) or the event of interest occurred.

#### **Classification of outcome variables: prognosticators and date of distant metastasis**

Information on tumour prognosticators was obtained from the BCQR. ER, PR and HER2 status was used exactly as coded (positive or negative) in the BCQR. Survival analysis was performed with distant metastasis-free survival as the outcome. Information on date of distant metastasis was obtained from the BCQR, and date of death was obtained from the Cause of Death Register.

## Calculation of Tyrer-Cuzic Risk Score

For study II, the 10-year individual risk of breast cancer was calculated for each participant, using the Tyrer-Cuzic risk prediction model. The resulting risk was denoted "Tyrer-Cuzic Risk Score" (TCRS). Information on variables included in the TCRS calculation was obtained from the study questionnaire and the TCRS was calculated using the online tool IBIS from the developers of the Tyrer-Cuzic model [Tyrer et al., 2004].

### 5.2.3 Paper III

The aim of study III was to assess heterogeneity in risk factors for breast cancer across molecular subtypes of the disease. The material was analysed with a case-control design, as well as case-only analysis restricted to the cases. Cases were combined from Libro-1 and KARMA studies, and controls were drawn from breast cancer-free women in KARMA. Inclusion criteria for cases were;

1. Invasive cancers only
2. Primary breast cancer diagnosis made 2005-2015
3. Full information on ER, PR, HER2 and Ki-67 staining of the tumour available in medical records.

The motivation behind the second criteria was two-fold; Firstly, the practice of staining for HER2 and Ki-67 was not widespread in clinical pathology practice before those years. Secondly, as we aimed to assess risk factors by molecular subtypes, we wanted to minimize the survivor bias as certain subtypes have a higher risk of metastasis the first years after diagnosis.

### Classification of outcome variables: Molecular subtypes

Starting from information on IHC markers, we used a random forest classifier to assign molecular subtypes to cases. The classifier was trained on a subset of the data (n=258) which was annotated with PAM50 subtypes. Input variables to the classifier were ER, PR, HER2 (all binary), Ki-67 and age (continuous). After training, the random forest predicted subtype for the remained of observations.

PAM50 subtypes had been assigned previously based on RNA-sequencing of gene expression of the tumours as described in detail in [Wang et al., 2016], and was used as such for the present study. Briefly, RNA was extracted from fresh-frozen bio banked tumours and sequenced on the Illumina HiSeq 2500 platform. After quality control and data pre-processing detailed in [Wang et al., 2016], molecular subtype was assigned using the Nearest

Shrunken Centroid Classifier for the PAM50 gene set, with parameters estimated from the publicly available dataset The Cancer Genome Atlas (TCGA) breast cancer study (<http://cancergenome.nih.gov/>).

Data regarding ER, PR, HER2 and Ki-67 was obtained directly from medical records and cleaned and coded by the author.

- ER and PR were coded as positive or negative, by (1), applying a cutoff at 10% percent immunohistochemistry staining and (2), if percent staining was unavailable, as noted "positive" or "negative" in the records.
- HER2 was coded from two variables, IHC staining of 0+ to 3+, and FISH analysis of gene amplification of the ErbB2 (HER2) gene. HER2 was coded as negative if IHC was below 2+, or 2+ to 3+ with no FISH amplification. Confirmed amplification through FISH was required for coding positive HER2 status.
- Ki-67 was used as percentage cells stained. According to Swedish guidelines for pathology of breast cancer, Ki-67 should be measured in hot spot regions of the slides and entered as percentage cells positive [Grabau, 2014].

As sensitivity analysis, the St Gallen IHC proxy was used an alternative approach for assigning molecular subtypes. The proxy is described in figure 11 A, page 38.

#### 5.2.4 Paper IV

Study IV was performed in two separate cohorts and was largely cross-sectional, with the exception for the survival analysis. The aim of the study was to assess the concordance of three distinct IHC proxies to the PAM50 method of classifying breast cancer into molecular subtypes. The first cohort, denoted "Clinseq", used a subset of the cases included in study III, namely those 258 with RNA-sequencing and PAM50 subtyping available through the Clinseq project. The second cohort was STO-3, tumours diagnosed 1976-1990. The STO3 included only node negative, <30 mm tumours.

Analysis was performed separately for each cohort. All cases were classified according to each IHC proxy, and classifications were compared to their previously ascertained PAM50 status. For each proxy, Cohen's kappa statistic and the overall accuracy were calculated with 95 % confidence intervals. Sensitivity and specificity were calculated for each subtype against all other subtypes. Additionally, the datasets were restricted to PAM50 luminal subtypes, calculating sensitivity and specificity for classifying luminal A vs B.

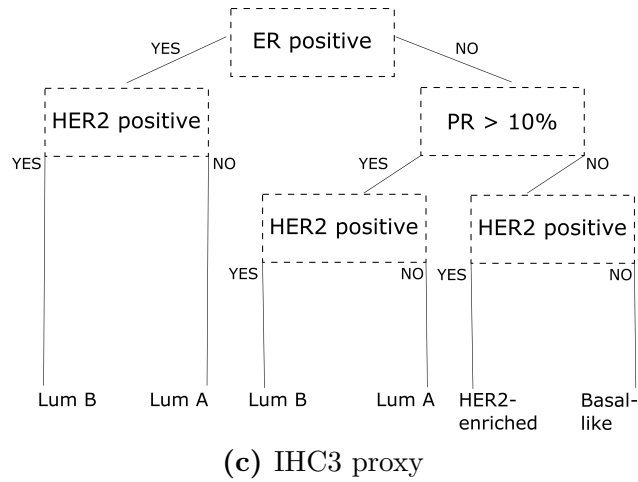
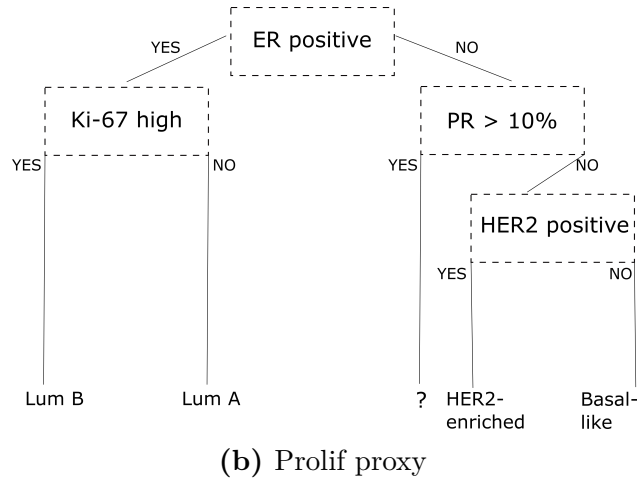
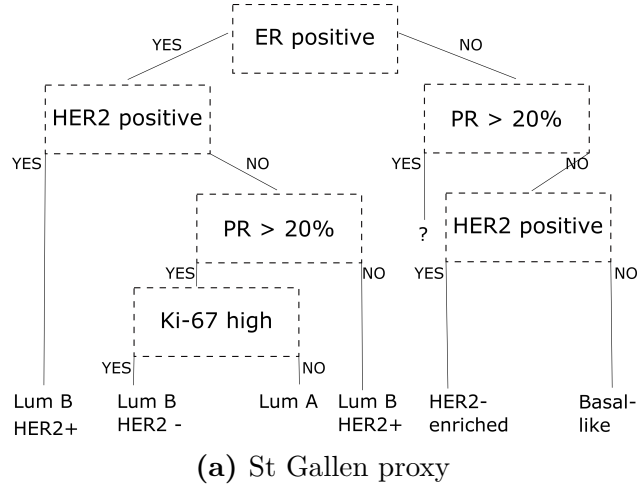
## **Classification of outcome variables: IHC proxies**

Cases were classified according to three different IHC based proxies. One proxy (denoted 'IHC3') was based on ER, PR and HER2, and two additionally included Ki-67 (denoted 'St Gallen' and 'Prolif'). In figure 11 on page 38, each IHC proxy is depicted in the form of a decision tree. The IHC proxies were based on information on ER, PR, HER2 and Ki-67 staining for the tumours.

- For the Libro-1 and KARMA materials, ICH markers were based on the material collected and coded as described for study III above.
- For STO-3, tumours were retrospectively stained for immunohistochemistry of ER, PR, HER2 and Ki-67 status centrally. FFPE tissue blocks were sliced, randomly annotated with an ID and shipped to the University of California Davis Medical Centre for pathology.
  - ER and PR were coded as positive or negative by applying a cutoff at 10% percent immunohistochemistry staining for ER, and 10 or 20 % for PR.
  - HER2 was coded as positive if IHC was 3+, and negative for 0 to 2+.
  - Ki-67 was measured as percentage cells stained in whole-slide sections, not hot-spots.

## **Assignment of PAM50 subtypes**

In Clinseq, PAM50 subtypes were assigned as described for study III above. In STO-3, PAM50 subtypes were assigned from RNA gene expression data hybridized onto Agilent microarray platforms as described earlier [Esserman et al., 2017], and used for the study. As described earlier [Esserman et al., 2017], assignment of PAM50 subtypes using the PAM50 classifier was done after centering the gene expression values to the median expression level for each gene by ER status calculated in a sub-sample of all ER negative tumours and an equal number of randomly selected ER positive tumours.



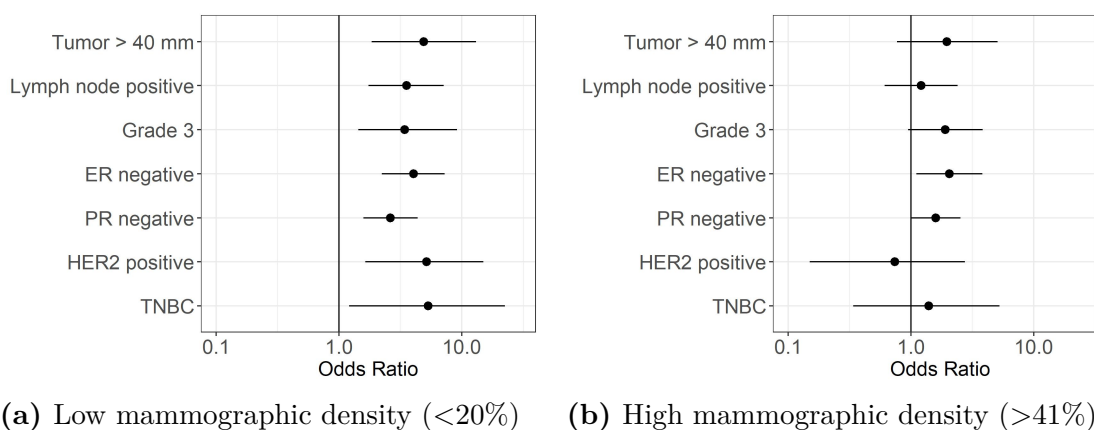
**Figure 11** – Decision trees for the IHC proxies. The use of '?' for St Gallen and Prolif proxies denotes unclassifiable combinations of markers.



## 6 Main results and interpretations

### Interval cancers are more aggressive than screening-detected cancers, but only in women with low mammographic density

In study I, we found support for our hypothesis that interval cancers are particularly enriched for true, fast-growing breast cancer in women with non-dense breasts. In the strata of low mammographic density, interval cancer were more commonly lymph node positive, high grade, ER and PR negative, HER2 positive, and triple negative as compared to screening-detected cancers. In dense breasts, only tumour size and ER status differed between groups (Figure 12).



**Figure 12** – Forestplot of odds ratios for assessed tumour characteristics in study I, comparing interval detected versus screening-detected breast cancer

These observations are consistent with the idea that true interval cancers (i.e. aggressive, fast-progressing disease) are more common in the group of women with high screening test sensitivity. If a tumour of a detectable size had been present at the previous screen, it is much more likely to have been found in a woman whose breast had low mammographic density. Likewise, the interval cancers among women with high mammographic density are more likely to have been masked at the previous screen (false negatives), due to the lower sensitivity of the screening test in this group [Domingo et al., 2014].

### Current users of HRT were overrepresented among interval cancers

The investigation into the distribution of risk factors across IC and SC revealed differences primarily in family history, current use of hormone replacement therapy (HRT) and BMI. Higher frequency of BRCA mutation carriers and lower frequency of high parity was also observed in IC [Holm et al., 2015]. The finding of HRT has consistently been observed in previous works [Kirsh et al., 2011, Domingo et al., 2010, Wang et al., 2001, Gilliland et al., 2000], but the

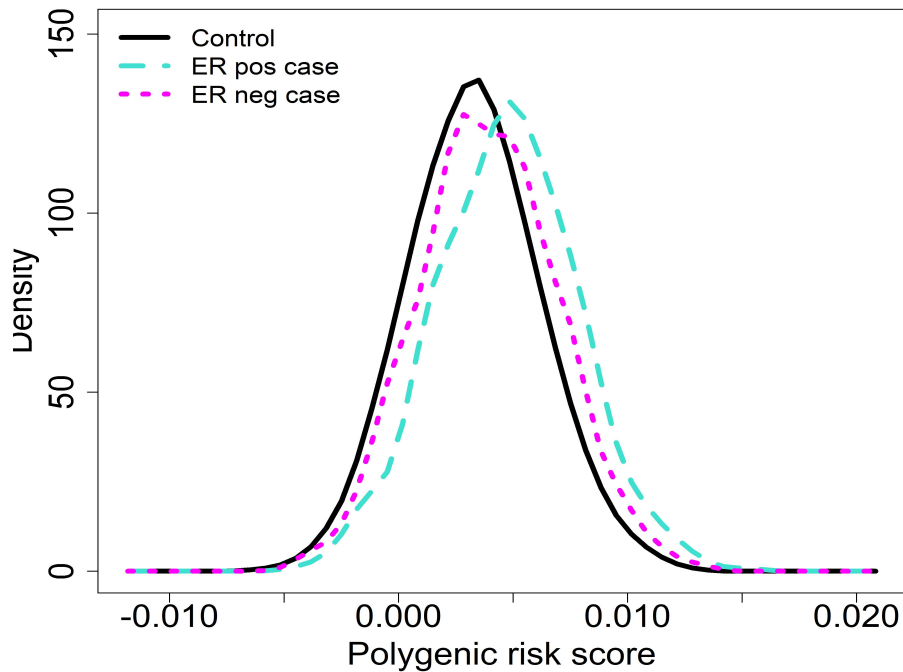
impact of mammographic density in explaining this observation was not known. It has been argued that the association between IC and HRT is mediated through increasing the mammographic density which leads to lowered screening sensitivity and higher probability of IC [Laya et al., 1996, Kavanagh et al., 2000, Wang et al., 2001]. If this was the only explanation, we would expect the association to be present mainly among dense breasts. However we saw similar associations in both strata. This is not necessarily indicative of HRT giving a more aggressive phenotype. Likely explanations would also include a surveillance-bias, as women undergoing current HRT treatment during the period were advised to be vigilant for symptoms of breast cancer and to attend sporadic screening outside of the screening program. In study III we evaluated HRT use as a risk factor by molecular subtypes and saw no indication of HRT use being a risk factor for the more aggressive subtypes. This would indicate that surveillance bias, rather than aggressive biology, is a more probable explanation for the observed association with current HRT use and IC diagnosis.

### **Family history was more common among interval cancers - especially in the low mammographic density strata**

We saw a higher frequency of family history among diagnoses of IC, especially in the non-dense strata with aggressive IC. Family history has been found to be more common in IC than SC in some [Mandelson et al., 2000, Kirsh et al., 2011, Lowery et al., 2011, Domingo et al., 2010], but not all [Gilliland et al., 2000, Musolino et al., 2012, Brekelmans et al., 1994] previous studies. Definition of family history has varied across studies, and most of these works were of low sample sizes (100 to 700 IC and SC total) additionally with few hypothesis test performed, making conclusions hard to reach. Our study was the largest to date to investigate family history and IC. The association with family history and IC has been confirmed in a Spanish study (HR for true IC, 2.11, 95% CI 1.60 to 2.78) [Domingo et al., 2014]. It could be that the observed association for family history is due to a higher vigilance for symptoms and checking up symptoms among relatives to breast cancer cases, however the larger estimate in the non-dense strata could speak against that as sole explanation. Associations for BRCA mutations could possibly also be the effect of a surveillance bias, if the mutation status or a strong heredity in the family was known before the diagnosis and caused the woman to be under increased clinical surveillance. We did not have information on the timing of BRCA testing to evaluate this. Part of the association for BRCA could be due to the known enrichment of basal-like breast cancer for BRCA1 mutation carriers [Boyle, 2012], as we found that TNBC, a proxy for basal-like BC, were approx. 5-fold more common in true IC. Recent work in our group has also found a higher prevalence of rare, high penetrance mutations beyond BRCA among interval cancers in low dense breasts [Li 2018, unpublished

work]. These genes were typically not tested for clinically during the study period, and it is thus unlikely that those associations are due to surveillance bias. Instead, they could be explained by association between these genes and an aggressive phenotype.

### The breast cancer polygenic risk score of 77 SNPs is primarily a score for identifying risk of ER positive tumours



**Figure 13** – Distribution of the polygenic risk score for controls and cases in study III, scores separated by ER status among cases.

Two of the included works investigated heterogeneity of polygenic risk score between subtypes of breast cancer. In both studies, we observed a pattern where subtypes with favourable prognosis tended to have the highest values in the score. In study II, the score was found to be on average lower in ER negative tumours (20 % decrease in odds of ER negative tumour per standard deviation increase in the score), tumours above 40 mm (14% decrease) and grade 3 tumours (14% decrease vs grade 1) [Holm et al., 2016]. In study III, the score was associated with a linear increase in odds per standard deviation of the score for luminal A (74% increase), luminal B (43%) and HER2-enriched subtypes (36%), but not for the basal-like subtype (15%, n.s.) as compared to controls [Holm et al., 2017]. The dose-response appeared strongest for the luminal subtypes (Figure 14, page 43). The above findings are understandable if one considers the distribution of the score among ER negative cases, which was hardly distinguishable from the distribution among controls, see

figure 13 on page 41. Better separation for ER negative cases versus controls is unlikely to be achieved unless more SNPs specific for predicting ER negative BC can be included. The majority of cases included in large GWAS discovery datasets have been ER positive [Mavaddat, N. et al., 2015, Pharoah et al., 2002]. Individual SNPs have recently been found to differ in associations to molecular subtypes, in both the Carolina Breast Cancer Study and among Chinese Han women [O'Brien et al., 2013, Xu et al., 2017]. Encouragingly, recent GWAS efforts focusing on associations to ER negative or TNBC specifically have advanced knowledge on risk variants for this aggressive subtype [Purrington, K. S. et al., 2014, Li, 2014, Ruiz-Narváez et al., 2016].

### **The Tyrer-Cuzic risk model may under-perform for identifying early-onset, aggressive breast cancer**

In study II, the TCRS was higher in cases with ER positive and grade 1 tumours as compared to cases with aggressive features (ER negative and grade 3). This finding was restricted to early-onset, pre-menopausal cases, below age 40 at diagnosis. These results may indicate that young women at risk for aggressive tumours do not share the set of reproductive risk factors underpinning the TCRS. The estimates of association increased when analysis was restricted to women without a family history of breast cancer. Moreover, depending on the degree of contrast in TCRS between young women with aggressive breast cancers and healthy controls, the TCRS may not prospectively identify these women as being at risk of breast cancer. However, we did not compare the scores to a distribution of scores in controls, as this was a case only design. It should be noted that we did not include information on history of atypical hyperplasia in calculating the score, as we did not have access to such data at the time. It would probably had little impact on our estimates, as the number of women with atypical hyperplasia was found to be too low for analysis in study III where we had information on previous benign breast diseases available, but should be kept in mind.

### **Reproductive risk factors appear differently associated with risk of basal-like and non-basal like breast cancer**

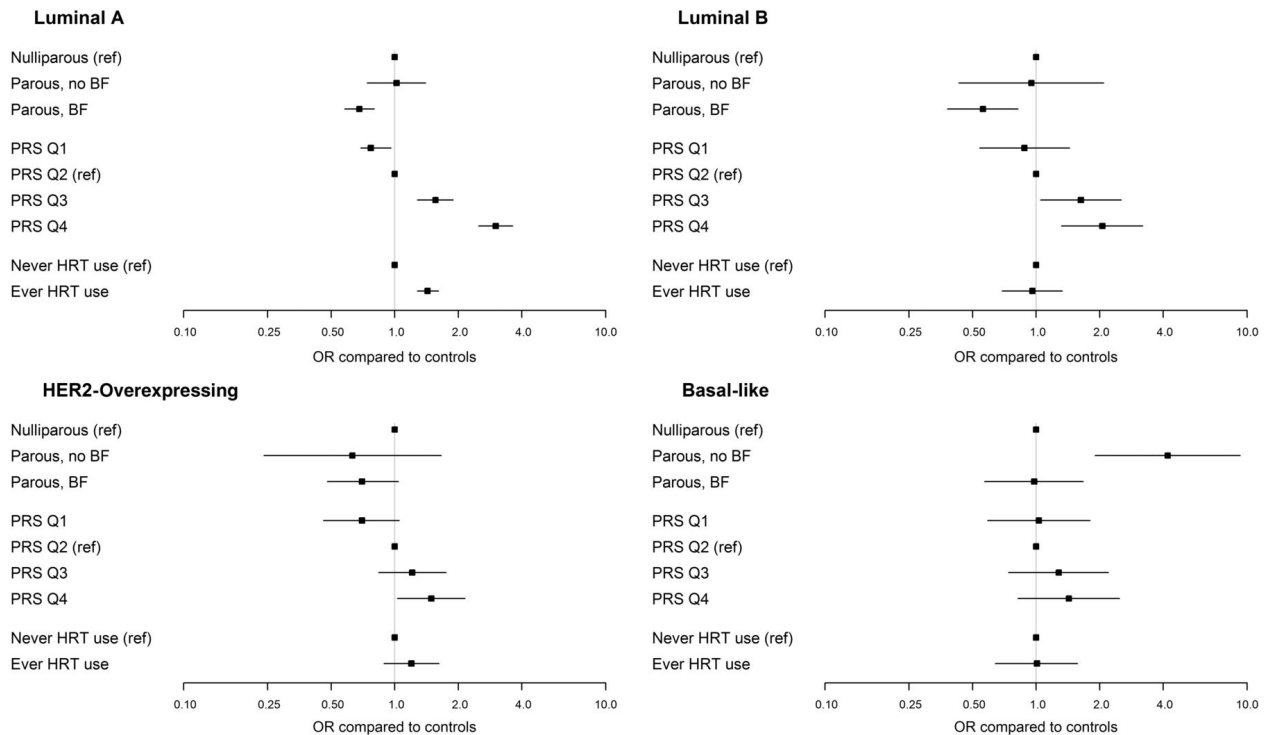
We saw in study III, that the pattern of associations to parity, breastfeeding and age at first birth appeared to differ for the molecular subtypes:

- Parity was protective for all subtypes except basal-like. ORs for having more than two children vs. nulliparity were 0.61 - 0.63 for luminal and HER2 subtypes but no difference was seen for basal-like (Table 3, [Holm et al., 2017]).
- Having the first child after age 30, as compared to before 30, was associated with an

increase in odds of luminal A, B and HER2 subtypes (ORs 1.32 and 1.42 respectively), but a lower OR of 1.16 for basal-like. These differences did not yield statistically significant heterogeneity, but agreed with the overall pattern.

- Breastfeeding did show heterogeneity, with odds ratios of luminal A and B subtypes of 1.49 and 1.74 respectively, null effect for HER2-enriched, and 4-fold increased odds for basal-like disease (OR 4.2, 95 % CI 2.20 to 7.99).

A composite variable of parity and breastfeeding revealed that, relative to nulliparous women, the added effect of breastfeeding to parity varied for basal-like and non-basal like subtypes, with no increased risk for never breastfeeding for non-basal like types, but an increased risk for the basal-like subtype in women never breastfeeding (OR 4.17; 95% CI 1.89 to 9.21, see forestplot in figure 14). Corresponding estimate for triple negative breast cancer was OR 2.95; 95% CI 1.47 to 5.90) (Supplementary table 6, [Holm et al., 2017]).



**Figure 14** – Summary of findings from Study III: Forest plot of associations to risk factors where heterogeneity was observed, as estimated by odds ratios, for each subtype of breast cancer examined.

## Results for breastfeeding in light of previous works

The numbers of parous women who never breastfed were low (367 controls, 88 cases), and therefore our reported estimates are imprecise and need further replication. Our findings of

an increased risk of basal-like breast cancer among parous women who never breastfed, is corroborated by the one previous report in the literature assessing a similar parametrisation [Millikan et al., 2008]. In the Carolina Breast Cancer Study, a population-based case-control study, an odds ratio for basal-like breast cancer of 1.9 (95% CI 1.1 to 3.3) was seen for parous women who never breastfed, relative to nulliparous women [Millikan et al., 2008]. They defined basal-like breast cancer as ER/PR/HER negative, Cytokeratin5/6 and EGFR positive, and thus differed to our definition. Recently a population-based, prospective study of Norwegian women born between 1886 and 1928 investigated breastfeeding, using the same definition of basal-like BC as Millikan and colleagues [Horn et al., 2014]. They reported an odds ratio for never vs. ever breastfeeding (n.b. restricted to parous women) to be 1.06 for basal-like cancer, thus not confirming the results in Millikan et al. Intriguingly, the same study also assessed cases that were five-marker negative (ER/PR/HER2/Cyt5/EGFR negative) and for this group the OR for never breastfeeding was 3.85 (95% CI 1.10 to 13.56) [Horn et al., 2014]. This estimate is similar to the results for basal-like breast cancer reported in our study. Different results using varying definitions of basal-like disease, and the relatively low number of cases who never breastfed in these studies, merits caution in the interpretations at this stage. It appears more studies are required before we may know if the association between never breastfeeding and TNBC truly differs by cytokeratin and EGFR markers. Although for TNBC as a whole, there is robust evidence of a negative association to breastfeeding [Islami et al., 2015, Anderson et al., 2014].

### **Mammographic density increases the risk of interval cancer, but not by increasing the risk of aggressive phenotypes**

In study I, we saw that mammographic density is higher on average among interval cancers compared to screening-detected cancers. This is explained by breast density's influence on the screening test [Domingo et al., 2014]. Consistently in both study II and III, mammographic density was associated with all types of the disease, irrespective of aggressiveness in phenotype. Our results thus agree with the consensus understanding from the literature that mammographic density is a general risk factor for all breast cancers [Antoni et al., 2013]. They are also consistent with the hypothesis that the risk conferred through high breast density is due to increased mammary gland mass at risk for malignant transformation, i.e., mammographic density is a surrogacy measure for underlying number of cells at risk [Trichopoulos et al., 2007].

## **IHC proxies show moderate agreement with PAM50 gene expression based subtypes**

In study IV, we compared several methods used in the literature to determine molecular subtype. We found that IHC based subtypes had fair to moderate agreement with PAM50 gene expression based subtyping. The kappa values were higher in the Clinseq material for all comparisons, likely due to a wider spread of the Ki-67 distribution in this cohort, achieving better separation. Nevertheless, there was not a large difference between the kappa values in STO-3 and Clinseq, and the best agreement to PAM50 was seen with the simplistic Prolif IHC proxy in both cohorts.

## **IHC proxies and PAM50 classification often disagree on classification of luminal A and B subtypes**

One reason that the agreement to PAM50 was only moderate, was the challenge of distinguishing between luminal A and B. When these classes were collapsed into one luminal category, the kappa increased to 0.71 (95% CI 0.65 to 0.78) in STO-3 and 0.69 (95% CI 0.58 to 0.80) in Clinseq, indicating substantial agreement, and accuracy rose to at around 90 % in both cohorts. Important information may also lie in the direction of re-classification. Inspection of cross-tabulations of proxy vs PAM50 revealed a pattern of the St Gallen proxy over declaring luminal A as luminal B, and the reverse of over declaring luminal B as luminal A using the Prolif proxy. The IHC3 proxy declared almost all of the luminal B as luminal A in our material.

## **Sensitivity was low for the HER2-enriched subtype**

When collapsing luminal A and B into one luminal category, sensitivity and specificity for luminal breast cancer was very high, at 0.98-0.99 and 0.77-0.66 respectively in STO-3 as well as Clinseq. Similarly, sensitivity and specificity was high for the basal-like subtype, at 0.83-0.85 and 0.96-0.98, respectively. However, although HER2-enriched subtype had a very high specificity at 0.99, sensitivity was only around 0.36 in both cohorts. HER2-enriched subtype as defined by PAM50 were misclassified as basal-like, luminal B and even luminal A by the IHC proxies, but the most common confusion was with luminal B.

## 7 Methodological considerations

Our studies have all been observational in nature, and therefore suffer from the curse common to all observational designs: Claims of causality can never be made. Nevertheless, observational studies are sometimes the only means to address scientific questions in human populations, as not all exposures are permissible to randomization for either ethical or practical reasons. Just as caution must be made in extrapolating findings from experiments in cell lines or a specific animal model to other settings, so must caution be exercised in interpreting observational data.

### Selection bias

#### Study I

The analysis of tumour characteristics between IC and SC has little selection bias, as it essentially consists of the entire source population for Libro-1 that were active screening participants. All registered breast tumour diagnoses in females below age 80 during 2001-2008 that were found to be either SC or IC were included in our analysis, and cancer registration in Sweden has high completeness [Barlow et al., 2009]. However, the analysis stratified by mammographic density and analysis which included information on risk factors had two levels of potential selection biases:

1. Cases had to be participants in the Libro-1 study in order to consent to collection of images and questionnaire data, so non-responder bias applies by definition.
2. There is a survivor bias build into Libro-1. As the study was initiated in 2009, only cases surviving until 2009 could be included.

To address these concerns, sensitivity analysis was performed. For concern (1), we assessed whether the associations between mode of detection and tumour characteristics were similar when restricting analysis to Libro-1 participants. For concern (2), we assessed the findings in analysis restricted to the most recent diagnoses made, in 2005-2009. The same conclusions were reached in both cases.

#### Study II

In study II, the same general concerns related to selection bias for participating in Libro-1 apply as for study I. To address the issue of survivor bias, sensitivity analysis restricted to diagnoses 2005-2008 was performed, yielded same conclusions regarding analysis of TCRS, PRS and MD with near identical estimates for PRS and MD, and slightly attenuated estimates



for TCRS with ER status as outcome. Similarly, in survival analysis restricted to diagnoses made 2005-2008, same conclusions were reached (HR above median TCRS/PRS 0.60, 95 % CI 0.37 to 0.97 all women, HR 0.17 (95 % CI 0.02 to 1.34) in young women).

### **Study III**

In study III, the survivor bias in Libro-1 was less of an issue, as we only included diagnoses from 2005. However, participation bias still applied. The main concern for study III lies in the unusual case-control design. Although the participants in Libro-1 were recruited from the same regions as half of the Karma study and during approximately the same calendar period, there are likely differences between Libro-1 and Karma participants on group level due to differences in recruitment approaches. This is a potential issue as control women were exclusively from Karma. Libro-1 can be considered a population-based cohort in that all diagnoses made for a certain period as well as region were invited. Karma on the other hand recruited participants mainly through the population-based mammography screening program, thus technically limiting inclusion to women willing to partake in screening. Non-attendance has been linked to levels of physical activity, adhering to a vegetarian- or vegan diet, low self-rated health, stress, alcohol abstinence and smoking [Lagerlund et al., 2015]. Social factors such as employment, income, marital status, immigration from a non-Nordic country and having five or more children have also been linked to non-attendance of screening in Sweden [Zidar et al., 2015, Lagerlund et al., 2002]. Some of the potential bias is likely lessened by the fact that both studies rely on women willing to partake in scientific studies. To address the potential differences between studies, all estimates were adjusted for education level and being born in Sweden or not. However, unmeasured confounding was likely still an issue. For this reason, we assessed also OR's for breast cancer in general, to ascertain whether estimates obtained with our design mirror expected magnitude and direction of associations to risk factors. Additionally, we restricted analysis to a case-only design where imbalance between Karma and Libro-1 was not a potential issue, and could draw the same conclusions about heterogeneity from this analysis.

### **Study IV**

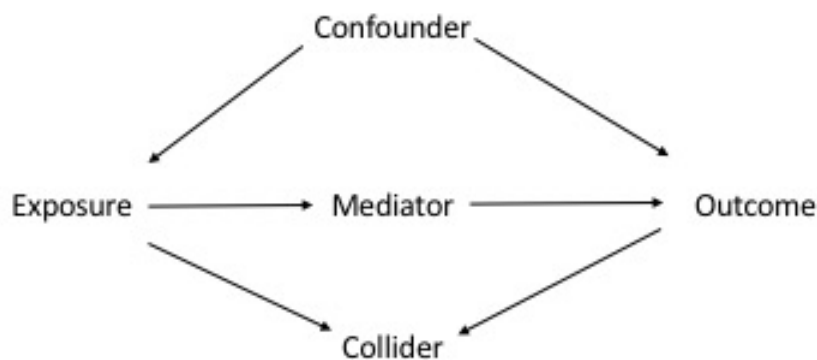
In study IV, the tumours included were on average larger than the underlying population of tumours. This was inevitable as availability of tumour materials for analysis was a requirement for performing the study, and very small tumours had to be excluded. We do however believe that the inference regarding concordance between IHC and PAM50 is valid for breast cancers in general.

## Handling of missing data

Throughout the thesis work, in final analysis, missing data was handled by complete case analysis. However, in Study I, complementary data collection was made to ensure as high as possible completeness of the data on mammographic images. In study III and IV, we opted for using data on IHC markers collected by us from medical records rather than using data from the BCQR. This was motivated by the fact that data on Ki-67 was not available the BCQR, but also since data on HER2 and percent ER/PR staining was incomplete or absent in the BCQR for the first years of the study period.

## Confounding

One of the greatest challenges in making inference from observational data (and in some circumstances, randomised trials [Rothman, 2014]) is the presence of confounding. This was mainly a challenge for studies I and III. For study II, we did not attempt to draw any causal conclusions, instead the inference was concerned with prediction. Study IV was a descriptive study in nature, and inference was limited to concordance measures. For study I and III, we chose to include potential confounders in our statistical models based on the concept of directed acyclic graphs (Figure 15) [Schisterman et al., 2009]. By blindly including every variable available to the researcher, over-fitting is achieved. However it is also possible to create biased models already when adjusting for a limited number of variables, for instance by adjusting for a mediator when assessing an overall effect, or adjusting for a collider, or by introducing collinearity [Schisterman et al., 2009].



**Figure 15** – Directed acyclic graph of the relationship between an exposure, outcome, mediator, confounder and collider

There is thus never an easy answer to the question of what variables to include in a model; we have tried to make these judgements based on subject matter knowledge on known or

suspected confounders. We have however been limited to the variables available through registers and questionnaires, and there is the ever present risk of unmeasured confounding affecting the reported associations, as subject matter knowledge is limited to our current understanding of the subject of study.

Additionally, when adjusting for BMI in study III we should note that BMI was only available for post-diagnosis measurements in Libro-1. The estimates of interest did not change after adjusting for BMI. However, it is possible adjustment for pre-diagnostic BMI would have resulted in altered estimates. For some women, weight changes as a result of breast cancer diagnosis and/or treatment, and this is likely not random [Playdon et al., 2015]. For this reason, we chose not to evaluate BMI as a risk factor in study III. BMI can be investigated as a risk factor by subtype in these cohorts in the future, when the KARMA study has gathered more follow-up time and cases. Post-diagnostic BMI was however evaluated in study I, and found to be negatively associated with interval cancer. The results for BMI in study I thus merit some caution in interpretation.

## 8 Ethical considerations

Sadly, medical history is not short of examples of abuse of other humans and unethical study protocols [Abbott, 2008, Nature, 2012, Nature, 2017]. It should be with great responsibility and deep sense of respect for the other human being that we continue to practice medical research to this day. History has taught us that we are always running the risk of putting our own goals and needs above those of others - especially when we are considering other humans far removed from us, such as study subjects. Observational studies are certainly no exception. Throughout the analysis, we have handled highly sensitive and personal data about our study participants. This would not have been legal without their consent, obtained after informing each participants of the study scope, goal, data types to obtain, personal risks and rights. Furthermore, before the recruitment into each study could begin, the study protocols and study questions, weighing the potential benefits of the study against potential risks of harm for the participants, had to be approved by an independent ethical committee.

When analysing collected data or data from national registers, a researcher never has access to the raw personal information such as name and personal registration number. However, it was necessary for the data collection from medical records that the author, in role of data collector, had access to such information during the actual collection. As soon as data was entered into the study, this information was removed and not kept for any part of the analysis. All study participants had given their consent to withdraw data from journals and the responsible at each clinic approved the extraction which was carried out with a temporary licence for access to each electronic journal system, logging the collectors every move.

The main issue for our studies using sensitive personal data lies in the handling of data in a secure, and pseudomised manner. 'Pseudomised' refers to de-identification of the personal registration numbers into an assigned study-ID. Although this may seem identical to 'anonymised', it is at least theoretically possible to identify a certain individual based on the combination matrix of data at hand, especially in small studies with low frequencies within strata - although this is strictly prohibited by law -, hence the term 'pseudomisation'. Every PhD student at our department must undergo a mandatory education in good data management, which includes information on the legislation relating to handling sensitive data. This could be compared to the clinical practice, where caregivers are legally obliged to maintain patient confidentiality [Vårdguiden, 2018]. Perhaps there is scope for a similar, code of conduct more formally sworn in by each data analyst, too?

## 9 Concluding remarks

The final verdict on whether we have come any closer to answering the questions posed in this thesis can only come with time and in retrospect. In the scientific method lies the willingness and aptitude to correct hypotheses which does not hold up in light of new data. However, at this point in time, and with the data at hand from this thesis, a few conclusions may be noted.

In study I, we could establish that there are clear differences in aggressiveness between IC diagnosed in dense and non-dense breasts. This is an important difference to account for when re-thinking screening program designs, as neither of these IC groups currently benefit from their participation in screening but for separate reasons. In dense breasts, the issue is mainly the low sensitivity of the screening modality being used. Among women at risk for aggressive IC, it would be more important to consider modifying screening intervals, if this group of women can be prospectively identified.

Our approach of stratifying IC by mammographic density has been used in later studies not included in this thesis, to disentangle molecular differences between IC and SC. We have shown that relative to SC, IC in non-dense breast have lower polygenic risk score values [Li et al., 2015], in line with what we found for aggressive cancer as defined by receptor status in studies II and III. We have also seen an over-representation of the aggressive basal-like and HER2-enriched molecular subtypes in this category of women [Li et al., 2017]. Ongoing work in our group has begun to evaluate the impact of high risk, rare germline mutations beyond BRCA1/2 in the risk for IC, with preliminary findings of higher prevalence of such variants in non-dense IC. Future studies of risk factors for interval cancer may make use of mammographic density to separate false negative SC from true IC.

As far as prediction of breast cancer risk is concerned, we have found reasons to believe that polygenic risk scores based on the SNPs known to date of study II, are less likely to identify women at risk of ER negative breast cancer, particularly breast cancer of the basal-like type. If SNPs are to be used in predicting risk of TNBC and basal-like BC, they should probably be the SNPs identified in GWAS with a high proportion of cases with ER negative disease. Concerns regarding the validity of the Tyrer-Cuzick model for predicting risk of early onset ER-negative breast cancer are also raised from our results, although we have only performed case-only analysis for TCRS. Ideally these analysis should be performed in independent cohorts, and contrasting the effects to controls, ideally in a prospective setting. Continued caution in generalizing the validity of a prediction model to other populations than the intended target population in which it was developed is warranted. It is however reassuring that mammographic density genuinely appears to be associated with risk of breast

cancer independent of subtype, as observed both in studies II and III. If risk-prediction models are to be used to identify women at medium- or high risk of breast cancer for targeted screening, such models should take into account the different profile in risk factor of TNBC and basal-like breast cancer. These cancers are otherwise potentially double-cursed to evade screening detection: First at the invitation level due to the low prediction level for basal-like breast cancer in the PRS and other potentially other risk prediction models, and secondly by their tendency to surface in screening intervals, because of their aggressive nature.

The challenge of the IHC proxies to agree with PAM50 subtyping on what constitutes an HER2-enriched, a luminal A and a luminal B tumour introduces some question marks as to what these gene-expression clusters represent. When Calza and colleagues used two separate statistical approaches to cluster tumours into molecular subtypes from their gene-expression, they too found the highest rate of discordance to be between luminal A and B, and between luminal B and HER2 [Calza et al., 2006]. Given that the original papers on molecular subtypes also struggled to settle on one, two or three luminal subtypes and that tumours typically reclassified between luminal and HER2-enriched groups [Perou et al., 2000, Sorlie et al., 2001], it is not that surprising that IHC proxies struggle to mimic these classifications.

If no distinction between luminal subtype is needed, IHC proxies do well compared to the PAM50. However, the St Gallen IHC proxy was developed for the very purpose of distinguishing between luminal A and B. For this purpose, it should be noted that we saw a high degree of misclassification for all proxies investigated. Moreover, the confusion between luminal A and B tumours by classifiers may hamper discovery of differences in risk factors, if such differences do exist.

The high specificity for identification of HER2-enriched and basal-like subtypes is reassuring when drawing conclusions regarding their respective risk factors, in our studies as well as others. The low sensitivity for the HER2-enriched type means IHC proxy definitions of luminal cancers are diluted by HER2-enriched cancer to various extents. If HER2-enriched and luminal tumours truly have different risk factors, this would attenuate estimates for luminal cases.

One question which has lingered in my mind (and many other's) for some time, is whether there exists separate aetiology for breast cancer by subtypes - be it by pre and post-menopausal, by ER status, or by PAM50 subtypes. The finding that PRS is not associated with risk for basal-like BC, and the strong preference for BRCA1 mutations to result in basal-like BC seems to indicate that there may be some merit to such a hypothesis. Associations seen between germ line mutations and phenotype are less likely to be due to confounding, compared to exposures encountered later on in a lifetime. Additionally, with the exception of breastfeeding, there appears to be weak if any relationships with reproductive risk factors and

risk of basal-like breast cancer. Observations from stem cell research suggest that basal-like breast cancer most closely resembles luminal progenitor cells, whereas all other subtypes are more similar to mature, differentiated mammary epithelial cells [Visvader and Stingl, 2014]. Whether this is explained by distinct cell-of-origins for these breast cancers, or by other means of differences in aetiology is unclear. Until experiments such as careful lineage-tracing can be done in breast cancer (stem) cells, this question will remain to be answered.

## 10 Future perspectives

As is often the case, when trying to answer one question we end up generating many new ones to answer. A few such questions which spring to my mind for the future are,

- How would the identified differences in associations between reproductive risk factors, genetic risk factors and BC subtypes hold up in a large, prospective study with gene-expression based definitions of subtypes?
- What would the conclusion of another methodological study comparing several statistical methods to assign molecular subtype of breast cancer be?
- Why do we see a preference for BRCA1 mutations to yield basal-like tumours?
- How would the observed associations between TCRS and prognosticators translate into differences in absolute risk in a prospective setting? Are they clinically relevant?
- Is breastfeeding (or lack thereof) truly causal in its association to basal-like breast cancer, or are the observations we see due to unmeasured confounding? Alternatively, are they in some cases the result of early signs of the disease manifesting itself?
- It would be valuable to see larger studies investigate reasons for cessation of breastfeeding in the context of risk for TNBC, to see if voluntary and involuntary early cessation of breastfeeding affects risk of basal-like breast cancer equally. This was after all the case in the Carolina Breast Cancer Study [Millikan et al., 2008], and it would be interesting to see it expanded upon.
- Should we develop separate risk prediction models aimed at predicting risk of basal-like and non-basal-like breast cancer?
- Will models such as the Tyrer-Cuzic be useful to predict women at risk of breast cancers that are diagnosed as interval cancers? Or will it be skewed towards predicting screening-detectable disease, through its associations to grade and ER status?
- What will the conclusions of the recently started ATHENA trial [Shieh et al., 2017] of risk-prediction based mammography screening verses age-based screening be? Will it be the same for all subtypes? This trial will be very interesting to follow up on in the future.

Some of the questions pertaining to causality and aetiology must clearly be answered by a combination of laboratory experimental studies in cell lines, animal model and patient



derived tissue samples as well as epidemiology. However, there is a lot to do still in the field of breast cancer epidemiology, specifically with regards to analysing more data on tumour and patient biological markers and coupling this to national and regional health registers. This is especially so beyond Scandinavia and other high-incidence countries in general. Gathering information on reasons for not breastfeeding would be informative in future cohorts, as would information on PAM50 available in larger sample sizes be. Additionally, to achieve less selected, more population based studies on interval cancer and molecular subtypes, information collected in the quality registers on e.g. mammographic density and PAM50 subtype would also enable better scope for future research. This must however be contrasted to administrative burdens for clinicians entering the data at clinic level, and the security and integrity for the patients and participants.

## 11 Acknowledgements

*None of these works would have been possible without the availability of data. For this I wish to acknowledge:*

**The principal investigators, original data collectors and data base ppl** Thank you for your efforts to create STO-3, KARMA and Libro-1.

**All the participants in the included study cohorts** As a scientist working with these projects I hope we will however slowly, piece by piece, turn the trust you have put in us into contributions to improved practice of medicine.

**Funding** Libro-1 and KARMA projects were funded by the Swedish research council, as part of the translational research umbrella environment CRisP. KARMA was mainly funded by a substantial donation from the Märta and Hans Rausing initiative against breast cancer. The STO-3 trial was funded originally by King Gustav V jubilee fund and the Swedish Cancer Society. Further funding for this PhD has come from the Swedish research council and the Swedish Cancer Society.

*Writing my PhD thesis would not have been possible without the support and insight shared with and provided for me by a large number of people. I am indebted to, among others:*

**Kamila Czene, Main supervisor** For always being there to support every project when you sit down and methodologically devote your time and presence to the question at hand. Thank you also for the opportunity and trust to be involved in planning and carrying out data collection, which has been truly educational. Your sharp intellect and problem-solving skills are truly inspirational. I have learnt so much during these years and under your guidance, and I am forever grateful to you for taking me on as a student.

**Per Hall, Co-supervisor** For providing a glimpse of the real world outside of researcher corridors when giving clinical perspectives on our studies, which is invaluable, but also for being a pillar of combined optimism and realism. Thank you also for initiating the study cohorts we used. Without data, what are we?

**Keith Humphreys, Co-supervisor** For being a source of thoughtful scepticism to learn from, but also a friendly face in the forest of PhD studies, and for many an interesting

discussions about life in general. I have always enjoyed every conversation we have had, both over a coffee as well as over data. If co-supervisors were decades, please feel free to call yourself the seventies.

In addition to my thesis advisors, two people have had more or less formal roles as well as contributed to the education of the author.

**Mats Hamberg** External mentor, chemistry professor. You have been a constant inspiration for many of your students, your enthusiasm for discovery and mechanisms combined with your stringent logic and way of science sets the standard against which I tend to evaluate my work. Thank you for believing in me as a student and recognizing life's various struggles. Thank you also for pushing me out of my comfort zone when needed and for convincing me I could belong in research.

**Alexander Ploner** Statistician, teacher, self-acknowledged professional pessimist (I beg to differ) and, I may add, language diversifier. For statistical teaching since Anno 2009, study guidance, discussions, endless moral support and humour along the way, and for always sharing your truly genuine enthusiasm for data, statistics, science, programming and curiosity with your colleagues - thank you. 'Magic numbers' still annoys me. And hopefully one day I will understand this whole 'degrees of freedom' concept... Any day now. A day with a meeting with Alex is a happy day at work, I know so many colleagues that can agree to that.

**My co-authors not previously mentioned, in alphabetical order: Abbas Cheddad, Hatef Darabi, Martin Eklund, Louise Eriksson, Mikael Eriksson, Jingmei Li, Linda Lindström, Mattias Rantalainen, Sven Törnberg, Nancy Yu.** For making each study better than it was before. Special thanks to **Sven Törnberg**, head of screening at RCC, for facilitating my visits to the regional cancer centre, always making me feel so welcome. **Hatef**, for warning me about the perils of trusting data whatever the source, "Even if it is from me", and for being so reliable and generous with sharing your knowledge. **Abbas**, for your understanding and warm spirit and for all your help in my chasing after understanding the structure of our imaging data. **Nancy**, for helping me put together the analysis for our manuscript so effortlessly, despite working in separate campi. **Linda**, I still remember when we first met and discussed Ki-67! Thank you for enabling study IV to happen, and for all the interesting discussions we have had. I have so much to learn from you. **Mikael**, You have been instrumental as Keeper of The Data and all things related to Karma and Libro1. Thank

you for the endless help I have received upon every email or spontaneous knock on your door (and there were MANY visits - sorry!). Thank you for being the professional problem-solver never making me feel out of place.

*Additionally, a number of people have contributed to my education and understanding of the thesis subject. I am indebted to:*

**Dr Eduard Azavedo** Dedicated head of the Karolinska mammography unit, for his generous time and help in tracing lost mammographic images, always taking time to explain the details of radiological work stations and daily proceedings of the mammography clinic so that the quality of the science on interval cancer and screening is ensured. Dr Azavedo also played an important role in solving some issues with my Study 1. In this context, I also wish to thank **Fredrik Strand** for putting me in contact with Dr Azavedo.

**The staff at the KS mammography unit** Who showed me the workings of an examination and answered my questions regarding the process of image acquisition.

**Sini Kilpeläinen, Agneta Lissmats, Ulla Johansson and Anette Asterkvist** at the Regional Cancer Center Stockholm-Gotland, who helped me understand the breast cancer quality registers and whose collaboration was essential in disentangling the mammography database.

**Hanna Fredholm** for letting me to shadow you during your clinical work. These experiences added much value to my understanding of breast cancer, and it was incredibly kind of you to invite me. Thank you for being so dedicated to the patient care and quality of science.

**Martin Malmberg, Cecilia Arnesson & Boel Heddson** For all your help with accessing and understanding pathology and mammography data in Southern Sweden.

**Agneta Lönn & Erik Olason** Thank you for great efforts made in the complementary data collection in Libro-1, and for being so kind to me throughout the process.

**Aki Tuuliainen & Michael Broms** For helping me dig deeper into the origins of Libro-1.

**Ola Steinberg** For helpful suggestions on the handling of collaborations.

**Marika Gabrielsson, Fredrik Strand, Judith Brandt** For stimulating discussions on the topic of breast cancer biology. Marika, also a thank you (swe-english) for letting me assist with teaching in the medical program.

*Of course, I owe a lot to the Department of MEB. In particular, I wish to thank:*

**The MEB Biostat group**, especially (in approximate order of encounter) **Arvid Sjölander, Alex Ploner, Therese Andersson, Yudi Pawitan, Cecilia Lundholm, Henrik Olsson, Mark Clements, Paul Dickman, Anna Johansson, Caroline Weibull, Sandra Eloranta, Marie Reilly**. Thank you for sharing your knowledge with the department during many interesting meetings, seminars and teachings. You are a big reason for why MEB is a great department to work at. I would also like to make yet another acknowledgment to **Paul Dickman, Anna Johansson and Gustaf Edgren** for sharing your teaching of SAS programming with the world. Without your web page, I would never have achieved my masters thesis within the time allowed, and the doctoral thesis would have been slow from the beginning.

**Jonas Ludvigsson and Fang Fang** Thank you for frequently sharing your knowledge and experiences with the doctoral students. It has been inspirational and humbling and inspirational again!

**Camilla Ahlqvist, Erika Nordenhagen, Frank Pettersson, Marie Jansson** And the rest of MEB's TA staff: For always welcoming questions, always being one step ahead and always figuring out the solution, with a kind, reassuring attitude: Thank you.

**My growing list of office mates, past and present: Robert Karlsson, Robert Szulkin, Adina Feldman, Dounya Schoormans, Camilla Sjörs, Emilio Ugalde Morales, Andreas Jangmo** Thank you for all the company and fun throughout the years! Extra thanks to Robert K and Ndegwa Gichora for generously sharing with me (and others) the L<sup>A</sup>T<sub>E</sub>Xtemplate for this thesis.

**Vilhelmina Ullemar** - For your constant peership and support on and off for now more than 10 years at KI. Thank you not the least for the constant intellectual challenge and creative exchanging of ideas, high or low, in academia or related to life as we know it. To me, it is and has been, invaluable. I am so blessed to have been at MEB during your years here. I agree, let us not end there.

**The entire open cohort of PhD students at MEB** For offering invaluable moral support, teaching and fun at work. I am afraid of forgetting names, so I want to thank everyone in one go! Thank you for all the lunches and coffee breaks, especially the sunny days at the roof terrace I will never forget. Plus, extra thanks to everyone engaged in the PhD group and arranging of seminars. Good stuff.

*Family and friends*

**"Glitter, sparkles, and rainbows doctors"** Haris, Teresa, Luisa, Milind, Joanna, Sus and everything granny cooked and baked for us - Even the food deserves an acknowledgement of its own. A special thank you to **Susann** for making my Lund data collection trips whilst pregnant a bit more dignifient and fed.

**Morfar** I hope I don't run in the opposite direction when I see numbers - I have you and the many afternoons spent with games and marble bags to thank for that. I wish we had had more time with you, but only because you were so dear to us. We miss you.

**My family** Your love means everything, and I am so, so grateful for having you. For long-distance, long term, and close-up support as well as fresh perspectives: Tack, Asante, Kiitos, Thank you & *ευχαριστω*. Särskilt tack till **Mamma, Pappa och Ndegwa**, for caring a little extra for Hedvig during these last weeks of thesis writing during Christmas and New Years. And to **Eva and Ndegwa**, for proof-reading my thesis despite writing your own. Jag hade aldrig blivit klar i tid om det inte vore för er. **Mume wangu, Karl and Hedvig**. *You are my everything*. I cannot believe how lucky I am to have you. *Niwakupenda sana* ♡.

## References

- [Abbott, 2008] Abbott, A. (2008). Lessons from the dark side. Nature, 451(7180):755–755.
- [Adami and Killander, 1984] Adami, H. O. and Killander, D. (1984). Prediction of survival in breast cancer. principles and current status of hormone receptors and dna content as prognostic factors. Acta Chirurgica Scandinavica Suppl, 519:25–34.
- [Adams-Campbell et al., 2007] Adams-Campbell, L. L., Makambi, K. H., Palmer, J. R., and Rosenberg, L. (2007). Diagnostic Accuracy of the Gail Model in the Black Women’s Health Study. The Breast Journal, 13(4):332–336.
- [Allott et al., 2016] Allott, E. H., Cohen, S. M., Geradts, J., Sun, X., Khoury, T., Bshara, W., Zirpoli, G. R., Miller, C. R., Hwang, H., Thorne, L. B., O’Connor, S., Tse, C.-K., Bell, M. B., Hu, Z., Li, Y., Kirk, E. L., Bethea, T. N., Perou, C. M., Palmer, J. R., Ambrosone, C. B., Olshan, A. F., and Troester, M. A. (2016). Performance of Three-Biomarker Immunohistochemistry for Intrinsic Breast Cancer Subtyping in the AMBER Consortium. Cancer Epidemiol. Biomarkers Prev., 25(3):470–478.
- [Anderson et al., 2014] Anderson, K. N., Schwab, R. B., and Martinez, M. E. (2014). Reproductive risk factors and breast cancer subtypes: A review of the literature.
- [Anderson et al., 2002] Anderson, W. F., Chatterjee, N., Ershler, W. B., and Brawley, O. W. (2002). Estrogen receptor breast cancer phenotypes in the surveillance, epidemiology, and end results database. Breast Cancer Research and Treatment, 76(1):27–36.
- [Anderson and Matsuno, 2006] Anderson, W. F. and Matsuno, R. (2006). Breast cancer heterogeneity: A mixture of at least two main types? JNCI: Journal of the National Cancer Institute, 98(14):948–951.
- [Anderson et al., 2006] Anderson, W. F., Pfeiffer, R. M., Dores, G. M., and Sherman, M. E. (2006). Comparison of age distribution patterns for different histopathologic types of breast carcinoma. Cancer Epidemiology Biomarkers & Prevention, 15(10):1899–1905.
- [Antoni et al., 2013] Antoni, S., Sasco, A. J., dos Santos Silva, I., and McCormack, V. (2013). Is mammographic density differentially associated with breast cancer according to receptor status? A meta-analysis. Breast Cancer Res. Treat., 137(2):337–47.
- [Antoniou, A.C. et. al., 2008] Antoniou, A.C. et. al. (2008). The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. British Journal of Cancer, 98(8):1457–1466.

- [Armitage and Doll, 1954] Armitage, P. and Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. British Journal of Cancer, 8(1):1–12.
- [Bandera et al., 2015] Bandera, E. V., Chandran, U., Hong, C.-C., Troester, M. A., Bethea, T. N., Adams-Campbell, L. L., Haiman, C. a., Park, S.-Y., Olshan, A. F., Ambrosone, C. B., Palmer, J. R., and Rosenberg, L. (2015). Obesity, body fat distribution, and risk of breast cancer subtypes in African American women participating in the AMBER Consortium. Breast Cancer Res. Treat., 150(3):655–666.
- [Barlow et al., 2009] Barlow, L., Westergren, K., Holmberg, L., and Talbäck, M. (2009). The completeness of the Swedish Cancer Register: a sample survey for year 1998. Acta Oncol., 48(1):27–33.
- [Barnard et al., 2015] Barnard, M. E., Boeke, C. E., and Tamimi, R. M. (2015). Established breast cancer risk factors and risk of intrinsic tumor subtypes.
- [Bernier, 2000] Bernier, M. O. (2000). Breastfeeding and risk of breast cancer: a meta-analysis of published studies. Human Reproduction Update, 6(4):374–386.
- [Boughey et al., 2010] Boughey, J. C., Hartmann, L. C., Anderson, S. S., Degnim, A. C., Vierkant, R. A., Reynolds, C. A., Frost, M. H., and Pankratz, V. S. (2010). Evaluation of the tyrer-cuzick (international breast cancer intervention study) model for breast cancer risk prediction in women with atypical hyperplasia. Journal of Clinical Oncology, 28(22):3591–3596.
- [Boyle, 2012] Boyle, P. (2012). Triple-negative breast cancer: Epidemiological considerations and recommendations. Ann. Oncol., 23(SUPPL. 6).
- [Brekelmans et al., 1994] Brekelmans, C. T., Peeters, P. H., Faber, J. A., Deurenberg, J. J., and Collette, H. J. (1994). The epidemiological profile of women with an interval cancer in the DOM screening programme. Breast Cancer Res. Treat., 30(3):223–32.
- [Brooke et al., 2017] Brooke, H. L., Talbäck, M., Hörnblad, J., Johansson, L. A., Ludvigsson, J. F., Druid, H., Feychting, M., and Ljung, R. (2017). The swedish cause of death register. European Journal of Epidemiology, 32(9):765–773.
- [Buell, 1973] Buell, P. (1973). Changing incidence of breast cancer in japanese-american women. Journal of the National Cancer Institute, 51(5):1479–1483.
- [Calza et al., 2006] Calza, S., Hall, P., Auer, G., Bjöhle, J., Klaar, S., Kronenwett, U., Liu, E. T., Miller, L., Ploner, A., Smeds, J., Bergh, J., and Pawitan, Y. (2006). Intrinsic



molecular signature of breast cancer in a population-based cohort of 412 patients. Breast Cancer Research, 8(4).

- [Chay et al., 2012] Chay, W. Y., Ong, W. S., Tan, P. H., Leo, N. Q. J., Ho, G. H., Wong, C. S., Chia, K. S., Chow, K. Y., Tan, M., and Ang, P. (2012). Validation of the gail model for predicting individual breast cancer risk in a prospective nationwide study of 28, 104 singapore women. Breast Cancer Research, 14(1).
- [Cheang et al., 2009] Cheang, M. C. U., Chia, S. K., Voduc, D., Gao, D., Leung, S., Snider, J., Watson, M., Davies, S., Bernard, P. S., Parker, J. S., Perou, C. M., Ellis, M. J., and Nielsen, T. O. (2009). Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. J. Natl. Cancer Inst., 101(10):736–750.
- [Chen et al., 2006] Chen, J., Pee, D., Ayyagari, R., Graubard, B., Schairer, C., Byrne, C., Benichou, J., and Gail, M. H. (2006). Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. JNCI: Journal of the National Cancer Institute, 98(17):1215–1226.
- [Chlebowski et al., 2007] Chlebowski, R. T., Anderson, G. L., Lane, D. S., Aragaki, A. K., Rohan, T., Yasmeen, S., Sarto, G., Rosenberg, C. A., and and, F. A. H. (2007). Predicting risk of breast cancer in postmenopausal women by hormone receptor status. JNCI Journal of the National Cancer Institute, 99(22):1695–1705.
- [Clamp et al., 2003] Clamp, A., Danson, S., and Clemons, M. (2003). Hormonal and genetic risk factors for breast cancer. The Surgeon, 1(1):23–31.
- [Clemmesen, 1948] Clemmesen, J. (1948). I. results from statistical research. The British Journal of Radiology, 21(252):583–590.
- [Coates et al., 2015] Coates, A. S., Winer, E. P., Goldhirsch, A., Gelber, R. D., Gnant, M., Piccart-Gebhart, M., Thürlimann, B., and Senn, H.-J. (2015). Tailoring therapies—improving the management of early breast cancer: St gallen international expert consensus on the primary therapy of early breast cancer 2015. Annals of Oncology, 26(8):1533–1546.
- [Colzani et al., 2011] Colzani, E., Liljegren, A., Johansson, A. L., Adolfsson, J., Hellborg, H., Hall, P. F., and Czene, K. (2011). Prognosis of patients with breast cancer: Causes of death and effects of time since diagnosis, age, and tumor characteristics. Journal of Clinical Oncology, 29(30):4014–4021.

- [Czene et al., 2002] Czene, K., Lichtenstein, P., and Hemminki, K. (2002). Environmental and heritable causes of cancer among 9.6 million individuals in the swedish family-cancer database. International Journal of Cancer, 99(2):260–266.
- [Darabi et al., 2012] Darabi, H., Czene, K., Zhao, W., Liu, J., Hall, P., and Humphreys, K. (2012). Breast cancer risk prediction and individualised screening based on common genetic variation and breast density measurement. Breast Cancer Res., 14(1):R25.
- [Domingo et al., 2010] Domingo, L., Sala, M., Servitja, S., Corominas, J. M., Ferrer, F., Martínez, J., Macià, F., Quintana, M. J., Albanell, J., and Castells, X. (2010). Phenotypic characterization and risk factors for interval breast cancers in a population-based breast cancer screening program in Barcelona, Spain. Cancer Causes Control, 21(8):1155–64.
- [Domingo et al., 2014] Domingo, L., Salas, D., Zubizarreta, R., Baré, M., Sarriugarte, G. n., Barata, T., Ibáñez, J., Blanch, J., Puig-Vives, M., Fernández, A. B., Castells, X., and Sala, M. (2014). Tumor phenotype and breast density in distinct categories of interval cancer: results of population-based mammography screening in Spain. Breast Cancer Res., 16(1):R3.
- [Elston and Ellis, 1991] Elston, C. and Ellis, I. (1991). pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology, 19(5):403–410.
- [Eriksson et al., 2013] Eriksson, L., Czene, K., Rosenberg, L. U., Törnberg, S., Humphreys, K., and Hall, P. (2013). Mammographic density and survival in interval breast cancers. Breast Cancer Res., 15(3):R48.
- [Eriksson et al., 2012] Eriksson, L., Hall, P., Czene, K., dos Santos Silva, I., McCormack, V., Bergh, J., Bjohle, J., and Ploner, A. (2012). Mammographic density and molecular subtypes of breast cancer. British Journal of Cancer, 107(1):18–23.
- [Esserman et al., 2017] Esserman, L. J., Yau, C., Thompson, C. K., van ’t Veer, L. J., Borowsky, A. D., Hoadley, K. A., Tobin, N. P., Nordenskjöld, B., Fornander, T., Stå l, O., Benz, C. C., and Lindström, L. S. (2017). Use of Molecular Tools to Identify Patients With Indolent Breast Cancers With Ultralow Risk Over 2 Decades. JAMA Oncol.
- [Fisher et al., 2008] Fisher, B., Redmond, C. K., and Fisher, E. R. (2008). Evolution of knowledge related to breast cancer heterogeneity: A 25-year retrospective. Journal of Clinical Oncology, 26(13):2068–2071.

- [Fisher et al., 1993] Fisher, E., Constantino, J., Fisher, B., and Redmond, C. (1993). Pathological findings from the national surgical adjuvant breast project (Protocol-4) - Discriminants for 15-year survival. Cancer, 71(6, S):2141–2150. Workshop on prognostic indicators in breast cancer, San Francisco, CA, Jan 17-19, 1992.
- [Gabrielsson et al., 2017] Gabrielsson, M., Eriksson, M., Hammarström, M., Borgquist, S., Leifland, K., Czene, K., and Hall, P. (2017). Cohort profile: The Karolinska Mammography Project for Risk Prediction of Breast Cancer (KARMA). Int. J. Epidemiol.
- [Gail et al., 1989] Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., and Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. JNCI Journal of the National Cancer Institute, 81(24):1879–1886.
- [Garcia-Closas, M. et al., 2008] Garcia-Closas, M. et al. (2008). Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. PLoS Genetics, 4(4):e1000054.
- [Gaudet et al., 2011] Gaudet, M. M., Press, M. F., Haile, R. W., Lynch, C. F., Glaser, S. L., Schildkraut, J., Gammon, M. D., Douglas Thompson, W., and Bernstein, J. L. (2011). Risk factors by molecular subtypes of breast cancer across a population-based study of women 56 years or younger. Breast Cancer Res. Treat., 130(2):587–97.
- [Ghiasvand et al., 2014] Ghiasvand, R., Adami, H.-O., Harirchi, I., Akrami, R., and Zendejdel, K. (2014). Higher incidence of premenopausal breast cancer in less developed countries; myth or truth? BMC Cancer, 14(1).
- [Gilliland et al., 2000] Gilliland, F. D., Joste, N., Stauber, P. M., Hunt, W. C., Rosenberg, R., Redlich, G., and Key, C. R. (2000). Biologic characteristics of interval and screen-detected breast cancers. J. Natl. Cancer Inst., 92(9):743–9.
- [Goldhirsch, A. et al., 2013] Goldhirsch, A. et al. (2013). Personalizing the treatment of women with early breast cancer: Highlights of the st gallen international expert consensus on the primary therapy of early breast Cancer 2013. Ann. Oncol., 24(9):2206–2223.
- [Grabau, 2014] Grabau, D. (2014). KVAST dokument brösttumörer, volume 3.1.3. Svensk förening f  r patologi - Svensk f  rening f  r klinisk cytologi.
- [Holm et al., 2017] Holm, J., Eriksson, L., Ploner, A., Eriksson, M., Rantalainen, M., Li, J., Hall, P., and Czene, K. (2017). Assessment of breast cancer risk factors reveals subtype heterogeneity. Cancer Research, 77(13):3708–3717.

- [Holm et al., 2015] Holm, J., Humphreys, K., Li, J., Ploner, A., Cheddad, A., Eriksson, M., Hall, P., Czene, K., and Törnberg, S. (2015). Risk factors and tumor characteristics of interval cancers by mammographic density. J. Clin. Oncol., 33(9):1030–1037.
- [Holm et al., 2016] Holm, J., Li, J., Darabi, H., Eklund, M., Eriksson, M., Humphreys, K., Hall, P., and Czene, K. (2016). Associations of Breast Cancer Risk Prediction Tools With Tumor Characteristics and Metastasis. J. Clin. Oncol., 34(3):251–258.
- [Horn et al., 2014] Horn, J., Opdahl, S., Engstrøm, M. J., Romundstad, P. R., Tretli, S., Haugen, O. A., Bofin, A. M., Vatten, L. J., and Åsvold, B. O. (2014). Reproductive history and the risk of molecular breast cancer subtypes in a prospective study of norwegian women. Cancer Causes & Control, 25(7):881–889.
- [Hossin and Sulaiman, 2015] Hossin, M. and Sulaiman, M. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2):01–11.
- [Howell et al., 2014] Howell, A., Anderson, A. S., Clarke, R. B., Duffy, S. W., Evans, D. G., Garcia-Closas, M., Gescher, A. J., Key, T. J., Saxton, J. M., and Harvie, M. N. (2014). Risk determination and prevention of breast cancer. Breast Cancer Res., 16(5):446.
- [Islami et al., 2015] Islami, F., Liu, Y., Jemal, A., Zhou, J., Weiderpass, E., Colditz, G., Boffetta, P., and Weiss, M. (2015). Breastfeeding and breast cancer risk by receptor status—a systematic review and meta-analysis. Annals of Oncology, page mdv379.
- [Johansson et al., 2009] Johansson, L. A., Björkenstam, C., and Westerling, R. (2009). Unexplained differences between hospital and mortality data indicated mistakes in death certification: an investigation of 1, 094 deaths in sweden during 1995. Journal of Clinical Epidemiology, 62(11):1202–1209.
- [Johansson et al., 1984] Johansson, R., Vanharanta, R., and Soderholm, J. (1984). Estrogen receptors in mammary cancer: Correlation with age, menopausal status, and response to therapy. Journal of Cancer Research and Clinical Oncology, 107(3):221–224.
- [Kamangar et al., 2006] Kamangar, F., Dores, G. M., and Anderson, W. F. (2006). Patterns of cancer incidence, mortality, and prevalence across five continents: Defining priorities to reduce cancer disparities in different geographic regions of the world. Journal of Clinical Oncology, 24(14):2137–2150.
- [Karma, 2017] Karma (2017). Karma Intervention Study - Karisma. Available from: <https://karmastudy.org/ongoing-research/>.

- [Kavanagh et al., 2000] Kavanagh, A. M., Mitchell, H., and Giles, G. G. (2000). Hormone replacement therapy and accuracy of mammographic screening. The Lancet, 355(9200):270–274.
- [Kirsh et al., 2011] Kirsh, V. A., Chiarelli, A. M., Edwards, S. A., O’Malley, F. P., Shumak, R. S., Yaffe, M. J., and Boyd, N. F. (2011). Tumor characteristics associated with mammographic detection of breast cancer in the ontario breast screening program. J. Natl. Cancer Inst., 103:942–950.
- [Kohler et al., 2015] Kohler, B. A., Sherman, R. L., Howlader, N., Jemal, A., Ryerson, A. B., Henry, K. A., Boscoe, F. P., Cronin, K. A., Lake, A., Noone, A.-M., Henley, S. J., Ehemann, C. R., Anderson, R. N., and Penberthy, L. (2015). Annual report to the nation on the status of cancer, 1975-2011, featuring incidence of breast cancer subtypes by race/ethnicity, poverty, and state. JNCI: Journal of the National Cancer Institute, 107(6).
- [Lagerlund et al., 2015] Lagerlund, M., Drake, I., Wirfält, E., Sontrop, J. M., and Zackrisson, S. (2015). Health-related lifestyle factors and mammography screening attendance in a swedish cohort study. European Journal of Cancer Prevention, 24(1):44–50.
- [Lagerlund et al., 2002] Lagerlund, M., Maxwell, A. E., Bastani, R., Thurffjell, E., Ekblom, A., and Lambe, M. (2002). Cancer Causes and Control, 13(1):73–82.
- [Lakhani and IARC, 2012] Lakhani, S. R. and IARC (2012). IARC WHO Classification of Tumours of the Breast (IARC WHO Classification of Tumours). World Health Organization.
- [Laya et al., 1996] Laya, M. B., Larson, E. B., Taplin, S. H., and White, E. (1996). Effect of estrogen replacement therapy on the specificity and sensitivity of screening mammography. JNCI Journal of the National Cancer Institute, 88(10):643–649.
- [Li et al., 2015] Li, J., Holm, J., Bergh, J., Eriksson, M., Darabi, H., Lindström, L. S., Törnberg, S., Hall, P., and Czene, K. (2015). Breast cancer genetic risk profile is differentially associated with interval and screen-detected breast cancers. Ann. Oncol., 26(3):517–22.
- [Li et al., 2017] Li, J., Ivansson, E., Klevebring, D., Tobin, N. P., Lindström, L. S., Holm, J., Prochazka, G., Cristando, C., Palmgren, J., Törnberg, S., Humphreys, K., Hartman, J., Frisell, J., Rantalainen, M., Lindberg, J., Hall, P., Bergh, J., Grönberg, H., and Czene, K. (2017). Molecular differences between screen-detected and interval breast cancers are largely explained by pam50 subtypes. Clinical Cancer Research, 23(10):2584–2592.

- [Li et al., 2012] Li, J., Szekely, L., Eriksson, L., Hedddson, B., Sundbom, A., Czene, K., Hall, P., and Humphreys, K. (2012). High-throughput mammographic density measurement: a tool for risk prediction of breast cancer. Breast Cancer Res., 14(4):R114.
- [Li, 2014] Li, J. e. a. (2014). 2Q36.3 Is Associated With Prognosis for Oestrogen Receptor-Negative Breast Cancer Patients Treated With Chemotherapy. Nat. Commun., 5(May):4051.
- [Lichtenstein et al., 2000] Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., and Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer — analyses of cohorts of twins from sweden, denmark, and finland. New England Journal of Medicine, 343(2):78–85.
- [Lilienfeld and Johnson, 1955] Lilienfeld, A. M. and Johnson, E. A. (1955). The age distribution in female breast and genital cancers. Cancer, 8(5):875–882.
- [Lind et al., 2010] Lind, H., Svane, G., Kemetli, L., and Törnberg, S. (2010). Breast Cancer Screening Program in Stockholm County, Sweden - Aspects of Organization and Quality Assurance. Breast Care (Basel)., 5(5):353–357.
- [Lipworth, 2000] Lipworth, L. (2000). History of Breast-Feeding in Relation to Breast Cancer Risk: a Review of the Epidemiologic Literature. J. Natl. Cancer Inst., 92(4):302–312.
- [Lowery et al., 2011] Lowery, J. T., Byers, T., Hokanson, J. E., Kittelson, J., Lewin, J., Risendal, B., Singh, M., and Mouchawar, J. (2011). Complementary approaches to assessing risk factors for interval breast cancer. Cancer Causes Control, 22(1):23–31.
- [Lteif and Javed, 2013] Lteif, A. and Javed, A. (2013). Development of the human breast. Seminars in Plastic Surgery, 27(01):005–012.
- [Ludvigsson et al., 2016] Ludvigsson, J. F., Almqvist, C., Bonamy, A.-K. E., Ljung, R., Michaëlsson, K., Neovius, M., Stephansson, O., and Ye, W. (2016). Registers of the swedish total population and their use in medical research. European Journal of Epidemiology, 31(2):125–136.
- [Ludvigsson et al., 2009] Ludvigsson, J. F., Otterblad-Olausson, P., Pettersson, B. U., and Ekblom, A. (2009). The swedish personal identity number: possibilities and pitfalls in healthcare and medical research. European Journal of Epidemiology, 24(11):659–667.
- [Mandelson et al., 2000] Mandelson, M. T., Oestreicher, N., Porter, P. L., White, D., Finder, C. A., Taplin, S. H., and White, E. (2000). Breast density as a predictor of mammographic

- detection: comparison of interval- and screen-detected cancers. J. Natl. Cancer Inst., 92(13):1081–7.
- [Maskarinec G, Noh JJ., 2004] Maskarinec G, Noh JJ. (2004). The effect of migration on cancer incidence among japanese in hawaii. Ethnicity & Disease, 4:431–439.
- [Mattsson and Wallgren, 1984] Mattsson, B. and Wallgren, A. (1984). Completeness of the Swedish Cancer Register. Non-notified cancer cases recorded on death certificates in 1978. Acta Radiol Oncol, 23(5):305–313.
- [Mavaddat, N. et al., 2015] Mavaddat, N. et al. (2015). Prediction of breast cancer risk based on profiling with common genetic variants. J. Natl. Cancer Inst., 107(5).
- [McCarty et al., 1983] McCarty, K. S., Silva, J. S., Cox, E. B., Leigh, G. S., Wells, S. A., and McCarty, K. S. (1983). Relationship of age and menopausal status to estrogen receptor content in primary carcinoma of the breast. Annals of Surgery, 197(2):123–127.
- [McHugh, 2012] McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia Medica, pages 276–282.
- [Mealiffe et al., 2010] Mealiffe, M. E., Stokowski, R. P., Rhees, B. K., Prentice, R. L., Pettinger, M., and Hinds, D. A. (2010). Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. JNCI Journal of the National Cancer Institute, 102(21):1618–1627.
- [Millikan et al., 2008] Millikan, R. C., Newman, B., Tse, C.-K., Moorman, P. G., Conway, K., Dressler, L. G., Smith, L. V., Labbok, M. H., Geradts, J., Bensen, J. T., Jackson, S., Nyante, S., Livasy, C., Carey, L., Earp, H. S., and Perou, C. M. (2008). Epidemiology of basal-like breast cancer. Breast Cancer Res. Treat., 109(1):123–39.
- [Musolino et al., 2012] Musolino, A., Michiara, M., Conti, G. M., Boggiani, D., Zatelli, M., Palleschi, D., Bella, M. A., Sgargi, P., Di Blasio, B., and Ardizzoni, A. (2012). Human epidermal growth factor receptor 2 status and interval breast cancer in a population-based cancer registry study. J. Clin. Oncol., 30(19):2362–8.
- [National Cancer Institute, 2017] National Cancer Institute (2017). NCI Dictionary of Cancer Terms. Available from: <https://www.cancer.gov/publications/dictionaries/cancer-terms>.
- [Nature, 2012] Nature (2012). Hypocritical oaths. Nature, 482(7384):132–132.
- [Nature, 2017] Nature (2017). Science must acknowledge its past mistakes and crimes. Nature, 549(7670):5–6.

- [Nyström et al., 1995] Nyström, L., Larsson, L.-G., Rutqvist, L. E., Lindgren, A., Lindqvist, M., Rydén, S., Andersson, I., Bjurstam, N., Fagerberg, G., Frisell, J., and Tabár, L. (1995). Determination of cause of death among breast cancer cases in the swedish randomized mammography screening trials: A comparison between official statistics and validation by an endpoint committee. Acta Oncologica, 34(2):145–152.
- [O’Brien et al., 2013] O’Brien, K. M., Cole, S. R., Engel, L. S., Bensen, J. T., Poole, C., Herring, A. H., and Millikan, R. C. (2013). Breast cancer subtypes and previously established genetic risk factors: A bayesian approach. Cancer Epidemiology Biomarkers & Prevention, 23(1):84–97.
- [Olsson et al., 2000] Olsson, S., Andersson, I., Karlberg, I., Bjurstam, N., Frodis, E., and Håkansson, S. (2000). Implementation of service screening with mammography in sweden: from pilot study to nationwide programme. Journal of Medical Screening, 7(1):14–18.
- [Paik et al., 1990] Paik, S., Hazan, R., Fisher, E. R., Sass, R. E., Fisher, B., Redmond, C., Schlessinger, J., Lippman, M. E., and King, C. R. (1990). Pathologic findings from the national surgical adjuvant breast and bowel project: prognostic significance of erbB-2 protein overexpression in primary breast cancer. Journal of Clinical Oncology, 8(1):103–112.
- [Pan et al., 2017] Pan, H., Gray, R., Braybrooke, J., Davies, C., Taylor, C., McGale, P., Peto, R., Pritchard, K. I., Bergh, J., Dowsett, M., and Hayes, D. F. (2017). 20-year risks of breast-cancer recurrence after stopping endocrine therapy at 5 years. New England Journal of Medicine, 377(19):1836–1846.
- [Parker et al., 2009] Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., and Bernard, P. S. (2009). Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. J. Clin. Oncol., 27(8):1160–1167.
- [Perou et al., 2000] Perou, C. M., Sørli, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O. y., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen Dale, A.-L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. Nature, 406(6797):747–752.



- [Pharoah et al., 2002] Pharoah, P. D., Antoniou, A., Bobrow, M., Zimmern, R. L., Easton, D. F., and Ponder, B. A. (2002). Polygenic susceptibility to breast cancer and implications for prevention. Nature Genetics, 31(1):33–36.
- [Phipps et al., 2012] Phipps, A. I., Buist, D. S. M., Malone, K. E., Barlow, W. E., Porter, P. L., Kerlikowske, K., O’Meara, E. S., and Li, C. I. (2012). Breast density, body mass index, and risk of tumor marker-defined subtypes of breast cancer. Ann. Epidemiol., 22(5):340–8.
- [Phipps et al., 2008] Phipps, A. I., Malone, K. E., Porter, P. L., Daling, J. R., and Li, C. I. (2008). Reproductive and hormonal risk factors for postmenopausal luminal, HER-2-overexpressing, and triple-negative breast cancer. Cancer, 113(7):1521–6.
- [Playdon et al., 2015] Playdon, M. C., Bracken, M. B., Sanft, T. B., Ligibel, J. A., Harrigan, M., and Irwin, M. L. (2015). Weight gain after breast cancer diagnosis and all-cause mortality: Systematic review and meta-analysis. Journal of the National Cancer Institute, 107(12):djv275.
- [Powell et al., 2014] Powell, M., Jamshidian, F., Cheyne, K., Nititham, J., Prebil, L. A., and Ereman, R. (2014). Assessing breast cancer risk models in marin county, a population with high rates of delayed childbirth. Clinical Breast Cancer, 14(3):212–220.e1.
- [Prat et al., 2013] Prat, A., Cheang, M. C. U., Martín, M., Parker, J. S., Carrasco, E., Caballero, R., Tyldesley, S., Gelmon, K., Bernard, P. S., Nielsen, T. O., and Perou, C. M. (2013). Prognostic significance of progesterone receptor-positive tumor cells within immunohistochemically defined luminal A breast cancer. J. Clin. Oncol., 31(2):203–9.
- [Pujol et al., 1998] Pujol, P., Daures, J.-P., Thezenas, S., Guilleux, F., Rouanet, P., and Grenier, J. (1998). Changing estrogen and progesterone receptor patterns in breast carcinoma during the menstrual cycle and menopause. Cancer, 83(4):698–705.
- [Purrington, K. S. et al., 2014] Purrington, K. S. et al. (2014). Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple-negative breast cancer. Carcinogenesis, 35(5):1012–9.
- [Quante et al., 2012] Quante, A. S., Whittemore, A. S., Shriver, T., Strauch, K., and Terry, M. B. (2012). Breast cancer risk assessment across the risk continuum: genetic and nongenetic risk factors contributing to differential model performance. Breast Cancer Research, 14(6).

- [Radisky and Hartmann, 2009] Radisky, D. C. and Hartmann, L. C. (2009). Mammary involution and breast cancer risk: Transgenic models and clinical studies. Journal of Mammary Gland Biology and Neoplasia, 14(2):181–191.
- [Razzaghi et al., 2013] Razzaghi, H., Troester, M. A., Gierach, G. L., Olshan, A. F., Yankaskas, B. C., and Millikan, R. C. (2013). Association between mammographic density and basal-like and luminal a breast cancer subtypes. Breast Cancer Research, 15(5).
- [Reeves et al., 2010] Reeves, G. K., Travis, R. C., Green, J., Bull, D., Tipper, S., Baker, K., Beral, V., Peto, R., Bell, J., Zelenika, D., and Lathrop, M. (2010). Incidence of breast cancer and its subtypes in relation to individual and multiple low-penetrance genetic susceptibility loci. JAMA, 304(4):426–34.
- [Ritte, R. et al., 2012] Ritte, R. et al., title = Adiposity, hormone replacement therapy use and breast cancer risk by age and hormone receptor status: a large prospective cohort study, j. . B. (2012). 14(3).
- [Rockhill et al., 2001] Rockhill, B., Spiegelman, D., Byrne, C., Hunter, D. J., and Colditz, G. A. (2001). Validation of the gail et al. model of breast cancer risk prediction and implications for chemoprevention. JNCI Journal of the National Cancer Institute, 93(5):358–366.
- [Romero et al., 2013] Romero, A., Prat, A., García-Sáenz, J. Á., del Prado, N., Pelayo, A., Furió, V., Román, J. M., de la Hoya, M., Díaz-Rubio, E., Perou, C. M., Cladés, T., and Martín, M. (2013). Assignment of tumor subtype by genomic testing and pathologic-based approximations: implications on patient’s management and therapy selection. Clinical and Translational Oncology, 16(4):386–394.
- [Rosenberg et al., 2015] Rosenberg, L., Bethea, T. N., Viscidi, E., Hong, C.-C., Troester, M. A., Bandera, E. V., Haiman, C. A., Kolonel, L. N., Olshan, A. F., Ambrosone, C. B., and Palmer, J. R. (2015). Postmenopausal female hormone use and estrogen receptor–positive and –negative breast cancer in african american women. Journal of the National Cancer Institute, 108(4):djv361.
- [Rothman, 2014] Rothman, K. J. (2014). Six persistent research misconceptions. Journal of General Internal Medicine, 29(7):1060–1064.
- [Ruiz-Narváez et al., 2016] Ruiz-Narváez, E. A., Lunetta, K. L., Hong, C.-C., Haddad, S., Yao, S., Cheng, T.-Y. D., Bensen, J. T., Bandera, E. V., Haiman, C. A., Troester, M. A.,

- Ambrosone, C. B., Rosenberg, L., and Palmer, J. R. (2016). Genetic variation in the insulin, insulin-like growth factor, growth hormone, and leptin pathways in relation to breast cancer in african-american women: the AMBER consortium. npj Breast Cancer, 2(1).
- [Russo and Russo, 2011] Russo, I. H. and Russo, J. (2011). Pregnancy-induced changes in breast cancer risk. J. Mammary Gland Biol. Neoplasia, 16(3):221–233.
- [Russo et al., 2005] Russo, J., Moral, R., Balogh, G. A., Mailo, D., and Russo, I. H. (2005). The protective role of pregnancy in breast cancer. Breast Cancer Res., 7(3):131–42.
- [Russo and Russo, 2004] Russo, J. and Russo, I. H. (2004). Development of the human breast. Maturitas, 49(1):2–15.
- [Rutqvist and and, 2007] Rutqvist, L. E. and and, H. J. (2007). Long-term follow-up of the randomized stockholm trial on adjuvant tamoxifen among postmenopausal patients with early stage breast cancer. Acta Oncologica, 46(2):133–145.
- [Saxena et al., 2010] Saxena, T., Lee, E., Henderson, K. D., Clarke, C. A., West, D., Marshall, S. F., Deapen, D., Bernstein, L., and Ursin, G. (2010). Menopausal hormone therapy and subsequent risk of specific invasive breast cancer subtypes in the california teachers study. Cancer Epidemiology Biomarkers & Prevention, 19(9):2366–2378.
- [Schiebinger, 1993] Schiebinger, L. (1993). Why mammals are called mammals: Gender politics in eighteenth-century natural history. The American Historical Review.
- [Schisterman et al., 2009] Schisterman, E. F., Cole, S. R., and Platt, R. W. (2009). Over-adjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology, 20(4):488–495.
- [Shieh et al., 2017] Shieh, Y., Eklund, M., Madlensky, L., Sawyer, S. D., Thompson, C. K., Fiscalini, A. S., Ziv, E., van’t Veer, L. J., Esserman, L. J., and Tice, J. A. (2017). Breast cancer screening in the precision medicine era: Risk-based screening in a population-based trial. Journal of the National Cancer Institute, 109(5):djw290.
- [Shimizu et al., 1991] Shimizu, H., Ross, R. K., Bernstein, L., Yatani, R., Henderson, B. E., and Back, T. M. (1991). Changing incidence of breast cancer in japanese-american women. Journal of the National Cancer Institute, 63(6):1479–1483.
- [Sorlie et al., 2001] Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese,

- J. C., Brown, P. O., Botstein, D., Lønning, P. E., and Borresen Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. U. S. A., 98(19):10869–74.
- [Sorlie et al., 2003] Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lønning, P. E., Brown, P. O., Borresen-Dale, A.-L., and Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc. Natl. Acad. Sci., 100(14):8418–8423.
- [Tabár et al., 1985] Tabár, L., Gad, A., Holmberg, L., Ljungquist, U., Fagerberg, C., Baldetorp, L., Grøntoft, O., Lundström, B., Månson, J., Eklund, G., Day, N., and Pettersson, F. (1985). Reduction in mortality from breast cancer after mass screening with mammography. The Lancet, 325(8433):829–832.
- [Talbäck, 2011] Talbäck, M. (2011). Cancer Patient survival in Sweden - theory and application. PhD thesis, Karolinska Institutet.
- [Tamimi et al., 2011] Tamimi, R. M., Colditz, G. A., Hazra, A., Baer, H. J., Hankinson, S. E., Rosner, B., Marotti, J., Connolly, J. L., Schnitt, S. J., and Collins, L. C. (2011). Traditional breast cancer risk factors in relation to molecular subtypes of breast cancer. Breast Cancer Res. Treat., 131(1):159–167.
- [The Collaborative Group on Hormonal Factors in Breast Cancer, 2002] The Collaborative Group on Hormonal Factors in Breast Cancer (2002). Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50 302 women with breast cancer and 96 973 women without the disease. The Lancet, 360(9328):187–195.
- [Thorpe et al., 1986] Thorpe, S., Rose, C., Rasmussen, B., King, W., Desombre, E., Blough, R., Mouridsen, H., Rossing, N., and Andersen, K. (1986). Steroid-hormone receptors as prognostic indicators in primary breast-cancer. Breast Cancer Research and Treatment, 7(S):91–98.
- [Tice et al., 2005] Tice, J. A., Cummings, S. R., Ziv, E., and Kerlikowske, K. (2005). Mammographic breast density and the gail model for breast cancer risk prediction in a screening population. Breast Cancer Research and Treatment, 94(2):115–122.
- [Törnberg et al., 2010] Törnberg, S., Kemetli, L., Asuncion, N., Hofvind, S., Anttila, A., Sèradour, B., Paci, E., Guldenfels, C., Azavedo, E., Frigerio, A., Rodrigues, V., and Ponti,

- A. (2010). A pooled analysis of interval cancer rates in six European countries. Eur. J. Cancer Prev., 19:87–93.
- [Trichopoulos et al., 2007] Trichopoulos, D., Adami, H.-O., Ekbom, A., Hsieh, C.-C., and Laggiou, P. (2007). Early life events and conditions and breast cancer risk: From epidemiology to etiology. International Journal of Cancer, 122(3):481–485.
- [Tyrer et al., 2004] Tyrer, J., Duffy, S. W., and Cuzick, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. Stat. Med., 23(7):1111–30.
- [Vaughan et al., 2013] Vaughan, T. A., Ryan, J. M., and Czaplewski, N. J. (2013). Mammalogy (Jones & Bartlett Learning Titles in Biological Science). Jones & Bartlett Learning.
- [Visvader and Stingl, 2014] Visvader, J. E. and Stingl, J. (2014). Mammary stem cells and the differentiation hierarchy: current status and perspectives. Genes & Development, 28(11):1143–1158.
- [Vårdguiden, 2017] Vårdguiden (2017). Vårdguiden. Available from: <https://www.1177.se/Stockholm/Tema/Cancer/Cancerformer-och-fakta/Cancerformer/Brostdcancer>.
- [Vårdguiden, 2018] Vårdguiden (2018). Vårdguiden. Available from: <https://www.1177.se/Stockholm/Regler-och-rattigheter/Tystnadsplikt-och-sekretess/>.
- [Wang et al., 2001] Wang, H., Bjurstam, N., Bjørndal, H., Braaten, A., Eriksen, L., Skaane, P., Vitak, B., Hofvind, S., and Thoresen, S. O. (2001). Interval cancers in the Norwegian breast cancer screening program: frequency, characteristics and use of HRT. Int. J. Cancer, 94(4):594–8.
- [Wang et al., 2016] Wang, M., Klevebring, D., Lindberg, J., Czene, K., Grönberg, H., and Rantalainen, M. (2016). Determining breast cancer histological grade from RNA-sequencing data. Breast Cancer Res., 18(1):48.
- [Wannerdt, 1982] Wannerdt, A. (1982). Den svenska folkbokföringens historia under tre sekler. Skatteförvaltningen.
- [Warwick et al., 2014] Warwick, J., Birke, H., Stone, J., Warren, R. M. L., Pinney, E., Brentnall, A. R., Duffy, S. W., Howell, A., and Cuzick, J. (2014). Mammographic breast density refines Tyrer-Cuzick estimates of breast cancer risk in high-risk women: findings from the placebo arm of the International Breast Cancer Intervention Study I. Breast Cancer Res., 16(5):451.

- [Wennman-Larsen et al., 2016] Wennman-Larsen, A., Nilsson, M. I., Saboonchi, F., Olsson, M., Alexanderson, K., Fornander, T., Sandelin, K., and Petersson, L.-M. (2016). Can breast cancer register data on recommended adjuvant treatment be used as a proxy for actually given treatment? European Journal of Oncology Nursing, 22:1–7.
- [Xu et al., 2017] Xu, Y., Chen, M., Liu, C., Zhang, X., Li, W., Cheng, H., Zhu, J., Zhang, M., Chen, Z., and Zhang, B. (2017). Association study confirmed three breast cancer-specific molecular subtype-associated susceptibility loci in chinese han women. The Oncologist, 22(8):890–894.
- [Zidar et al., 2015] Zidar, M. N., Larm, P., Tillgren, P., and Akhavan, S. (2015). Non-attendance of mammographic screening: the roles of age and municipality in a population-based swedish sample. International Journal for Equity in Health, 14(1).
- [Ziegler et al., 1993] Ziegler, R., Hoover, Pike, M., Hildesheim, A., Nomura, A., West, D., Wu-Williams, A., Kolonel, L., Horn-Ross, P., JF, J. F. R., and Hyer, M. (1993). Migration patterns and breast cancer risk in asian-american women. Journal of the National Cancer Institute, 85(22):1819–1827.

*Wisdom is like a baobab tree. No one individual can embrace it.*<sup>13</sup>

---

<sup>13</sup>Akan proverb