Statistical Modelling of Breast Cancer Risk for British Columbian Women

by

Tanja Hoegg

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in

THE COLLEGE OF GRADUATE STUDIES
(Interdisciplinary Studies)

THE UNIVERSITY OF BRITISH COLUMBIA
(Okanagan)
June 2013
© Tanja Hoegg, 2013
Abstract

Although there are many known factors associated with an increased risk of breast cancer, age remains the main eligibility criterion for the current Screening Mammography Program of British Columbia (SMP BC). In the light of recent controversy surrounding regular breast screening, the ability of targeted screening for high-risk women is of current interest as it has the potential to sustain high cancer detection rates, reduce the number of false positive results while controlling the overall costs. Multiple models have been proposed to estimate a woman’s risk of breast cancer given her current risk factor profile, the model introduced by Gail et al. in 1989 being particularly popular.

Using five-year follow-up data of 223,399 SMP BC participants, we investigate whether the probability estimates of the Gail model adequately stratify the British Columbian population into groups of high and low-risk women and, hence, provide a basis for a personalized access criterion into the Screening Mammography Program. Further, we built a breast cancer risk model for British Columbia with the goal to include a stronger set of predictor variables and improve the outcome stratification of the Gail model.

Neither the Gail model nor the new risk prediction model based on SMP BC participants showed adequate stratification properties. Overall, effect sizes of all covariates were too small to clearly separate risk estimates of future breast cancer cases and non-cases. It is questionable whether changes in screening policies should be based on breast cancer risk prediction models.
Preface

This thesis is original, unpublished, independent work by the author, Tanja Hoegg. The research reported in Chapters 2 - 4 was covered by UBC BCCA Ethics Certificate number H11-01617.
# Table of Contents

Abstract ........................................................................................................... ii  
Preface ........................................................................................................... iii  
Table of Contents ............................................................................................ iv  
List of Tables .................................................................................................... vi  
List of Figures ................................................................................................... vii  
Acknowledgements .......................................................................................... ix  

**Chapter 1: Introduction ................................................................. 1**  
1.1 Motivation ......................................................................................... 1  
1.2 The BC Screening Mammography Program ...................................... 2  
1.3 Previous Work .................................................................................. 3  
1.4 Notation and Concepts ................................................................. 5  
  1.4.1 Disease Frequency ................................................................... 5  
  1.4.2 Survival and Hazard ............................................................... 5  
  1.4.3 Survival Models ..................................................................... 6  
  1.4.4 Censoring ............................................................................. 7  

**Chapter 2: Preliminary Analysis ............................................... 9**  
2.1 Study Population and Variables ............................................. 9  
2.2 Data Cleaning and Initial Exclusion Criteria .......................... 10  
2.3 Exploratory Analysis ................................................................. 12  
  2.3.1 Population and SMP BC Comparison .................................. 12  
  2.3.2 Analysis of Survival Times .................................................. 13  
  2.3.3 Risk Factors and Cancer Outcome ...................................... 15  
  2.3.4 Variable Interactions ........................................................... 19
## Table of Contents

2.4 Limitations of Follow-Up ................................................. 21  
2.5 Evaluation of Study Bias .................................................. 23  

Chapter 3: Gail Model Validation .......................................... 26  
3.1 The Gail Model .............................................................. 26  
3.2 Validation Population ...................................................... 29  
3.3 Validation Methods .......................................................... 30  
3.4 Performance Evaluation .................................................... 32  
3.4.1 Risk Distribution ......................................................... 32  
3.4.2 Assessment of Model Calibration ................................. 33  
3.4.3 Discriminatory Ability .................................................. 36  
3.4.4 Comparison to Incidence Model .................................. 37  
3.5 Residual Analysis ........................................................... 38  
3.6 Optimization of Risk Threshold ........................................ 40  

Chapter 4: Model Building ....................................................... 43  
4.1 Survival Models .............................................................. 43  
4.1.1 Cox’s Proportional Hazards Model ................................ 43  
4.1.2 Aalen’s Additive Model ............................................... 45  
4.2 Modelling Population ......................................................... 47  
4.2.1 Estimation of Time to Diagnosis .................................. 47  
4.2.2 Exclusion of Individuals .............................................. 48  
4.3 Analysis of Family History ................................................ 50  
4.4 Modelling of Invasive Cancers .......................................... 52  
4.4.1 Single Covariate Modelling ......................................... 53  
4.4.2 Simultaneous Modelling of Covariates .......................... 56  
4.4.3 Testing Proportional Hazards Assumptions .................... 58  
4.4.4 Functional Form ......................................................... 61  
4.5 Modelling of In Situ Cancers ............................................. 63  
4.5.1 Single Covariate Modelling ......................................... 63  
4.5.2 Simultaneous Covariate Cox Model ......................... 64  
4.5.3 Model Checking .......................................................... 65  
4.6 Predictive Performance ...................................................... 67  

Chapter 5: Conclusion ............................................................. 70  

Bibliography ................................................................. 74
List of Tables

2.1 Variables extracted from the SMP BC database: Coding, description and values. ................................................................. 11
2.2 Contingency table notation of absolute numbers and probabilities (in parenthesis) for exposure groups and cancer outcomes. .................. 16
2.3 Number of missing values (NA), complete observations (n), absolute (nbc) and relative (nbc/n) cancer frequencies and odds ratios θ with 95% confidence intervals for selected risk factor categories. .................. 18

3.1 Gail model logistic regression coefficients for Caucasian, Black and Asian women as provided by the Gail model source code. .................. 27
3.2 Numbers of expected and observed cancers, E/O ratios and 95% confidence intervals for age and ethnic groups for (1) 176,978 women and (2) 78,106 women with complete information. .................. 34
3.3 Person years of 176,978 SMP BC women, B.C. incidence rates, observed cancers (O) and expected cancers predicted by B.C. incidence (E) for five-year age categories. .................. 38

4.1 Parameter estimates, relative hazards, standard errors, Wald statistics and p-values for family history model. ................................. 51
4.2 Coefficients and p-values of univariate Cox and Aalen models for hypothesis tests of significant and time-varying effects. (* refers to all variables levels, ** refers to at least one level) ................................. 54
4.3 Regression coefficients, standard errors, Wald statistics and corresponding p-values under proportional hazards model. .................. 57
4.4 Parameters estimates, standard errors, Wald statistics and p-values for the stratified Cox model. ........................................... 61
4.5 Coefficients, standard errors, Wald statistics and p-values for multivariate Cox model of in situ cancers. ................................. 66
List of Figures

2.1 Age distribution by five-year age groups for 221,545 SMP BC participants and B.C. women aged 40 to 79. ................................. 13
2.2 Ethnic composition of 221,545 SMP BC participants undergoing screening in the year 2000. .................................................. 14
2.3 Kaplan-Meier estimator and 95% confidence interval of breast cancer free survival times for 221,545 SMP BC participants, stratified by invasive (black) and in situ cancers (blue). ........................................ 15
2.4 Relative frequencies of responses (in parenthesis) for variables curr_estr, hysterec, menstr, educ, mammo_dens by five-year age groups. ........ 19
2.5 Age-specific mean breast density of 212,545 SMP BC women stratified by ethnicity. ............................................................. 20
2.6 Histogram of follow-up length in years for 221,545 SMP BC participants screened in the year 2000. ......................................... 22
2.7 Remaining percentage of participants at the end of each follow-up year by five-year age categories. .......................................... 23
2.8 Age-specific breast cancer incidence rates per 100,000 person years for the B.C. population and the data set of 212,545 SMP BC women. .... 24
3.1 Relative cancer frequency by five-year age category for 221,545 individuals when women lost to follow-up are included and removed from the data set. . . . . 30
3.2 Distribution of five-year breast cancer probabilities for 176,978 SMP BC participants as estimated by the Gail model. ......................... 33
3.3 Ratio of expected and observed cancers with 95% confidence interval in five-year age categories and ethnic groups for 78,106 individuals with complete Gail information. .................................................. 36
3.4 Estimated five-year cancer risk densities stratified by cancer outcome for 78,106 women with complete Gail model information (x-axis is truncated at $x = 5$). ................................................................. 37
List of Figures

3.5 Boxplot of Pearson residuals from Gail model stratified by the number of affected relatives. The y-axis is restricted to display the model’s residuals of non-cases only. .......................... 39

3.6 Left: Cancer detection rate above threshold $T$ (black) with lower 95% confidence bound (dashed) and prevalence (blue) for 78,106 women. Right: Proportion of women with risk estimates $p_i$ above risk threshold $T$. ........ 42

4.1 Left: Histogram of time between last recorded SMP BC screening date and cancer diagnosis of 4078 women. Right: Histogram of time between the 2000 screening appointment and diagnosis date for all prevalent cancers. ........ 49

4.2 Relative frequency of affected first degree relatives by type of relative, as a function of five-year age categories. ......................................................... 51

4.3 Cumulative regression coefficients $B(t)$ of univariate Aalen models for variables $age$, $abmam$, $menstr$ and $age_{lastmenstr}$. ............................. 55

4.4 Scaled Schoenfeld residuals for variables $age$ and $abmam$ with smoothing spline as a function of event time $t(k)$. ................................. 60

4.5 Plot of all martingale residuals (left) and martingale residuals of non-cases (right) under multivariate Cox model versus age. .............................. 62

4.6 Cumulative coefficients $B_j(t)$ as a function of time for single covariate Aalen models of $age$ and $abmam$. ........................................... 64

4.7 Scaled Schoenfeld residuals over time for variable $age$ with locally weighted smoothing spline. ............................................. 65

4.8 Scatterplot of five-year breast cancer risk for 97,856 women as estimated under the Gail model and the stratified Cox model. .......................... 68

4.9 Five-year cancer risk densities stratified by cancer outcome for 97,786 women as estimated by the stratified Cox model (x-axis is truncated at $x = 5$). .... 69
Acknowledgements

First and foremost, I would like to thank my supervisors, Dr. Sylvia Esterby and Dr. Paramjit Gill, who guided me through this project from near and far, whose doors were always open, providing advice and encouragement whenever it was needed. This thesis would not have been possible without your support.

I would also like to thank my committee members, Dr. Rasika Rajapakshe and Dr. Cynthia Araujo, from the BC Cancer Agency for inviting me into their team and making this project possible. Your inexhaustible enthusiasm has left a lasting impression on my path.

My sincere thanks go to my committee member Dr. Jason Loeppky for giving his valuable time to review my work and providing me with the latest gadgets like one of his own. Finally, many thanks to Dr. Anatol Sargin, who sent advice and support across the globe in many late night conversations.
Chapter 1

Introduction

1.1 Motivation

In Canada, over 20,000 women are diagnosed with breast cancer each year, making it the most common cancer among Canadian women (Statistics Canada, 2011). Hoping to detect cancer at an early stage and therefore improve chances of a successful treatment, many women consider early detection measures such as regular breast screening.

In British Columbia, the Screening Mammography Program (SMP BC) offers annual mammograms free of charge for women aged 40 to 79. Although many personal and familial factors have been associated with an increased risk of developing breast cancer, age remains the main eligibility criterion for the current Screening Mammography Program. Other established risk factors for breast cancer include early age at menarche (Brinton et al., 1988), nulliparity (Brinton et al., 1983) and first degree relatives affected by breast cancer (Brinton et al., 1982).

The ability to estimate a woman’s probability of developing breast cancer in a future time period given her risk factor profile is of current interest as it provides a basis for targeted screening practice for high-risk women. Based on the projected probability, women at high breast cancer risk could be identified in the population and accepted into the screening program before reaching the age of 40. To further improve the effectiveness of the early detection initiative, more sensitive screening techniques such as Magnetic Resonance Imaging (MRI) could also be implemented for high risk women. Providing more detailed breast imaging, MRI screening has a higher cancer detection rate compared to the currently used mammograms. Overall, targeted screening has the potential to increase cancer detection rates while controlling the overall costs.

Several risk models have been proposed during the last 20 years, the model introduced by Gail et al. (1989) being particularly popular. Although many efforts have been made to
predict a woman’s risk of developing breast cancer (Barlow et al., 2006; Chen et al., 2006), most models showed unsatisfactory predictive performance when evaluated on independent data (Decarli et al., 2006; Rockhill et al., 2001). Therefore, the applicability of individualized risk estimates is still limited and needs improvement.

Using five-year follow-up data of 223,399 SMP BC participants in the year 2000, the following thesis investigates whether the probability estimates of the Gail model sufficiently stratify the British Columbian population into groups of high and low-risk women and, hence, provide a basis for a personalized access criterion into the Screening Mammography Program. Further, we build a breast cancer risk model for British Columbia with the goal to include a stronger set of predictor variables and improve the outcome stratification of the Gail model.

1.2 The BC Screening Mammography Program

Introduced in 1988 as the first breast cancer screening program in Canada, the Screening Mammography Program of B.C. is a population based early detection effort operated by the BC Cancer Agency and is accessible to British Columbian women aged 40 to 79 years (BC Cancer Agency, 2001). Participants are self-referred, invited by promotion activities and advertisement throughout the province which include newspaper ads and information brochures at doctor’s offices. In 1998, the program also introduced an invitation letter sent out to women turning 50 years of age. Age information is based on the B.C. Ministry of Health’s client registry. Women outside the age range of 40 to 79 may also become eligible for screening if referred by a primary healthcare provider and approved by a radiologist.

Once enrolled in the program, participants receive recall letters for re-screening examinations. The frequency of the letters is based on SMP BC recall recommendations, which suggest biannual screening for women aged 50 to 79 and annual screening for women under the age of 50.

At each screening appointment, background information on the participant’s status of established breast cancer risk factors is collected through an eight item questionnaire. Questions cover risk factors related to personal background, reproductive and familial breast cancer history, menopausal status and medical history. The questionnaire is self-administered and completed on a voluntary basis.
1.3 Previous Work

The breast cancer status of participants is ascertained through a linkage with the B.C. Cancer Registry. All breast cancers detected within British Columbia, both within or outside the screening program, are also recorded within the SMP BC database. Once diagnosed with breast cancer, women become ineligible for screening mammograms within the SMP BC.

In the year 2000, there were 38 screening centers throughout the province including four mobile stations serving remote areas of British Columbia. Province wide participation in a 24-month period was estimated at 49% for women between the ages of 50 to 79 (BC Cancer Agency, 2001).

1.3 Previous Work

In 1989, Gail et al. proposed a model to project a woman’s absolute risk of developing invasive and in situ breast cancer in a specified age interval. The model is based on data collected from the Breast Cancer Detection Demonstration Project (BCDDP), a US cohort study following the breast cancer status of 284,780 predominately white women with up to five annual screening examinations from 1973 to 1979 (Baker, 1982). Six explanatory variables are considered in the model and include a woman’s age at menarche, the number of previous breast biopsies, her age at first live birth, the number of affected first degree relatives and a dichotomized age variable separating women below and above fifty years of age.

Based on a matched case-control study nested within the BCDDP cohort, a woman’s breast cancer risk \( r(t|X_i) \) relative to an unexposed woman of the same age is approximated by the odds ratio obtained from a logistic regression model. This relative risk is then combined with an age-specific breast cancer baseline hazard rate, \( h_1(t) \), estimated from the entire BCDDP cohort. Under consideration of risks of competing causes of death, \( h_2(t) \), the probability of breast cancer in the age interval \([a, a + \tau] \) could then be estimated by integration of hazards over the length of the interval \( \tau \). This will be referred to as Gail model 1. The presence of atypical hyperplasia in a previous biopsy was added as an additional explanatory variable in a revised model. All variables were considered ordinal in the modelling process.

The original Gail model 1 was later modified as part of the Breast Cancer Prevention Trial (BCPT) study design to predict the absolute risk of invasive breast cancer only. BCDDP
baseline hazards were replaced with invasive breast cancer rates from the Surveillance, Epidemiology and End Result (SEER) Program and a woman’s race was added as an additional explanatory variable. This version was implemented in an online risk calculator accessible through the website of the National Cancer Institute (www.cancer.gov/bcrisktool/) where women have the ability to calculate their personalized breast cancer risks. In the following chapters, we will refer to this later version as the ‘Gail model’.

The validity of the Gail model has been assessed on independent US and international data. Comparing the number of expected (E) and observed (O) cancers in predefined subgroups such as age categories, risk quintiles and categorical risk factor levels, the overall calibration performance of the Gail model has been found to be adequate in several studies. Rockhill et al. (2001) reported a value of $E/O = 0.94$ (95% CI: 0.89 - 0.99) for the ratio of observed and expected cancers, Costantino et al. (1999) a ratio $E/O = 1.03$ (95% CI: 0.88 - 1.21). However, it has been noted that the model overestimates breast cancer risk for women who do not participate in annual screenings (Spiegelman et al., 1994) and also shows a tendency to overestimate risk for women classified in the higher predicted risk quintiles (Costantino et al., 1999).

Anothaisintawee et al. (2011) offer a systematic review of previous publications assessing the predictive performance of the Gail model. For all validation studies discussed, the Gail model performed rather unsatisfactory in the discrimination of cases and non-cases. The reported concordance statistics range from 0.58 to 0.67, indicating a considerable overlap in the risk distributions for cases and non-cases and therefore a limited applicability in the prediction of individual cancer outcomes.

In an attempt to improve on calibration and discriminatory accuracy, modifications of the Gail model were proposed in several subsequent studies. Adjustments were based around a search for a stronger set of predictor variables in the logistic model. New variables like breast density and weight were added to the model by Chen et al. (2006) while other variables considered in the original model were removed. A change in coding of the explanatory variables from ordinal to categorical was proposed by Decarli et al. (2006). Overall, these modifications only lead to minor improvements over the discriminatory performance of the Gail model. Increased concordance statistic were reported (Chen et al., 2006; Decarli et al., 2006), however, the models’ ability to identify who will and will not develop breast cancer
remained limited.

1.4 Notation and Concepts

1.4.1 Disease Frequency

Occurrence of disease in a population is commonly described by two measures, disease prevalence and incidence. The prevalence of a disease describes the proportion of the population that are diagnosed with the disease,

\[
\frac{\text{number of cases}}{\text{population size}} \tag{1.1}
\]

and measures the likelihood of disease among a population at a specific point in time (Coggon et al., 2003). Disease incidence is defined as

\[
\frac{\text{number of new cases}}{\text{person-years at risk}} \tag{1.2}
\]

and represents the rate at which new cases occur in a population during a specified time interval (Coggon et al., 2003). Person-years at risk is the total length of time (in years) individuals in a population are at risk of the disease. An individual is considered to be at risk only during observation and until diagnosed with the disease. Hence, disease incidence allows for different follow-up periods between individuals.

Throughout this thesis, a woman diagnosed with breast cancer will be referred to as a case while a woman not diagnosed with breast cancer will be denoted as a non-case. At a given point in time, a prevalent case defines an individual whose breast cancer has already developed while an incident case has yet to develop the disease (Krug and McNutt, 2007).

1.4.2 Survival and Hazard

The following notation is consistent with Therneau and Grambsch (2000). Let \( T \) be a non-negative, random variable denoting the time until event occurrence. We assume that \( T \) has a density \( f \) and a cumulative distribution function

\[
F(t) = P(T \leq t).
\]
This probability will also be referred to as the breast cancer risk. The complement defines the time-dependent survival function as

\[ S(t) = P(T > t) = 1 - F(t) \] (1.3)

and represents the probability of no event occurrence before time \( t \). We assume \( S(0) = 1 \). Further, let \( \lambda(t) \) denote the hazard function defined as

\[ \lambda(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t | T \geq t]}{\Delta t} = - \frac{dS(t)/dt}{S(t)}. \] (1.4)

The hazard \( \lambda(t) \) represents the instantaneous probability of event given that the event has not occurred before time \( t \). Although the hazard \( \lambda(t) \) is not a probability, it can be interpreted as a measure of risk as higher values \( \lambda(t) \) indicate higher likelihood of event occurrence. Integrating \( \lambda(t) \) over \( t \) gives

\[ \int_0^t \lambda(u)du = - \int_0^t \frac{S'(u)}{S(u)}du = - \log(S(t)). \] (1.5)

and establishes the following relationship between hazard and survival function

\[ S(t) = \exp(-\Lambda(t)) \] (1.6)

where \( \Lambda(t) = \int_0^t \lambda(u)du \) represents the cumulative hazard function.

1.4.3 Survival Models

For individual \( i \), a binary outcome indicator is defined as

\[ N_i(t) = \begin{cases} 0 & \text{for } T_i > t \\ 1 & \text{for } T_i \leq t \end{cases}, i = 1 \ldots n \] (1.7)

and describes the survival status at time \( t \geq 0 \). \( N_i(t) \) takes a value of 1 if the event has occurred before time \( t \) and a value of 0 if it has not. Further, let \( dN_i(t) \) denote the incremental change in \( N_i \) in the infinitesimal time interval \([t, t + dt]\) and \( \bar{N}(t) \) the total number of events until time \( t \).
1.4. Notation and Concepts

The objective is to estimate the survival function for individual \( i \) based on current covariate information

\[
S_i(t) = P(T_i > t | X_i)
\]

where \( X_i = (X_{i1}, \ldots X_{ip}) \) denotes the covariate vector containing information on \( p \) variables recorded at time \( t = 0 \). These variables are generally thought to affect the survival time \( T_i \). Using the relationship between survival and hazard function given in Equation (1.6), survival probabilities can be estimated through modelling of the hazard \( \lambda_i(t) \) as

\[
\lambda_i(t) = g(X_i, \beta),
\]

where \( g \) is a function of the covariate vector \( X_i \) and a parameter vector \( \beta \). The vector \( \beta \) is estimated in the model fitting process.

1.4.4 Censoring

When following the survival status of a cohort over time, individuals are often able to leave the study before the end of the observational period. Reasons can include death from competing causes, relocation or discontinuation of screening. As a result, the survival status \( N_i(t) \) cannot be observed for the entire study period.

Let \( C_i \) be a non-negative random variable describing the censoring time of individual \( i \). An individual is said to be right censored if the event of interest has not occurred before the individual is lost to follow-up. The survival time \( T_i \) then exceeds the censoring time \( C_i \), leading to \( T_i \) to be unknown. Instead, we only observe

\[
\tilde{T}_i = \min(T_i, C_i).
\]

To indicate whether an individual is still at risk of the event at time \( t \), i.e. \( \tilde{T}_i > t \), we define an indicator variable

\[
Y_i(t) = \begin{cases} 
1 & \text{for } \tilde{T}_i > t \\
0 & \text{for } \tilde{T}_i \leq t 
\end{cases}, \quad i = 1 \ldots n
\]

(1.9)

Right censoring is referred to as non-informative or random if the censoring time \( C_i \) is independent of the survival time \( T_i \). Otherwise, censoring is called informative (Therneau
and Grambsch, 2000).
Chapter 2

Preliminary Analysis

2.1 Study Population and Variables

The data set analyzed in this thesis consists of 223,399 women who underwent a screening mammogram as part of the Screening Mammography Program in the year 2000. The data set was extracted from the SMP BC database after approval from the UBC BCCA Research Ethics Board following a formal data request to the SMP BC. A confidentiality agreement was signed by all individuals involved in the project. All women included in the data set were anonymized through consecutive case ID identifiers by the SMP BC. This research project is covered by UBC BCCA Ethics Certificate number H11-01617.

As a result of the linkage between the B.C. Cancer Registry and the Screening Mammography Program, all breast cancers diagnosed in participants before December 31, 2005, were recorded with the exact diagnosis date. This allows us to follow each woman’s cancer status for a minimum of 5 years from the initial screening date. December 31, 2005, marks the official end of the observational period for the study.

For all participants, we obtained complete information on the women’s birthdate, the 2000 screening date, the 2000 screening result and the women’s breast density estimated from the year 2000 mammographic image. If a woman’s mammogram was found to be abnormal in 2000, the diagnoses and corresponding diagnosis dates of examinations following the abnormal mammogram were also available. Information on the participants’ year of entry into the screening program and the last recorded SMP BC screening date as of July 2012 were complete for all individuals. Based on the participants’ birthdate and the 2000 screening date, we derived an additional variable describing the age at the 2000 screening examination.

Further, we obtained the participants’ personal background and risk factor information collected through the screening appointment questionnaire in the year 2000. As the questionnaire is completed on a voluntary basis and self-administered, the corresponding variables
2.2 Data Cleaning and Initial Exclusion Criteria

The items included in the questionnaire provide a woman’s personal background and the presence of well-established breast cancer risk factors. Beside a woman’s highest level of education and ethnic origin, her reproductive history is identified by information on her age at menarche, the occurrence of previous pregnancies, the number of full term deliveries and, if applicable, the age at first delivery.

Risk factors related to medical history are assessed by questions about the current use of estrogen and the occurrence of a previous breast biopsy. For certain women the number of breast biopsies is also available. If a woman states her menstrual cycles has stopped, her menopausal history is further obtained by information collected on the woman’s age at the last menstruating cycle, the presence of a hysterectomy at the time her menstruation stopped and, if applicable, if both ovaries were removed in the process of the hysterectomy. Familial breast cancer history is ascertained by the number and relationship of first degree relatives diagnosed with breast cancer. If breast cancer has been present among first degree relatives, an additional question whether the cancer occurred before the age of 50 is completed. All participants were anonymized and consecutively numbered with case ID identifiers from 1 to 223,399. Table 2.1 gives a list of relevant variables in the data set, abbreviations used in the following chapters and variable values.

2.2 Data Cleaning and Initial Exclusion Criteria

Several participants were excluded from the data set before the start of the analysis. The data set included 50 women diagnosed with bilateral breast cancer, leading them to be listed twice with identical entries for all variables. These duplicates were removed for each of the 50 women. Participant with case ID 60642 was deleted from the data set because of missing values on all risk factors and a missing birthdate.

Due to the regulations of the SMP BC, we also decided to exclude participants outside the eligible age range of 40 to 79 at the time of the 2000 screening appointment. As these women were only admitted to the screening program as a result of special medical circumstances related to breast disease, they differ from the general SMP BC population and may introduce biased breast cancer risks when included. Therefore, 394 women under the age of 40 and
### Table 2.1: Variables extracted from the SMP BC database: Coding, description and values.

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Values (format)</th>
</tr>
</thead>
<tbody>
<tr>
<td>yearscr1</td>
<td>year of first SMP screen</td>
<td>year (yy)</td>
</tr>
<tr>
<td>lastsmp</td>
<td>last recorded screening date</td>
<td>date (dd/mm/yyyy)</td>
</tr>
<tr>
<td>scrdate</td>
<td>2000 screening date</td>
<td>date (dd/mm/yyyy)</td>
</tr>
<tr>
<td>scr_res</td>
<td>result of 2000 screen</td>
<td>1 (Normal), A (Abnormal low), B (Abnormal moderate), C (Abnormal high)</td>
</tr>
<tr>
<td>scr_diagdate</td>
<td>2000 diagnosis date</td>
<td>date (mm/dd/yyyy)</td>
</tr>
<tr>
<td>scr_diag</td>
<td>2000 diagnosis</td>
<td>B (Benign), M (Invasive), D (DCIS), E (Equivocal disease), L (LCIS)</td>
</tr>
<tr>
<td>mammo_dens</td>
<td>breast density</td>
<td>1 (below 50%), 2 (above 50%)</td>
</tr>
<tr>
<td>bdate</td>
<td>birth date</td>
<td>date (dd/mm/yyyy)</td>
</tr>
<tr>
<td>ethnic</td>
<td>ethnic origin</td>
<td>AP (Aboriginal), BR (British), FR (French), EA (East Asian), SA (South Asian), WE (Western Europe), NE (Northern Europe), SE (Southern Europe), EE (Eastern Europe), BL (Black), OT (other), U (Unknown)</td>
</tr>
<tr>
<td>educ</td>
<td>highest educational level</td>
<td>N (None), S (Some high school), H (High school), C (College/ Some university), G (University graduate), U (Unknown)</td>
</tr>
<tr>
<td>age_men</td>
<td>age at menarche</td>
<td>integer</td>
</tr>
<tr>
<td>ever_preg</td>
<td>ever pregnant</td>
<td>Y (yes), N (no)</td>
</tr>
<tr>
<td>num_del</td>
<td>number of deliveries</td>
<td>integer</td>
</tr>
<tr>
<td>age_1stdel</td>
<td>age at first delivery</td>
<td>integer</td>
</tr>
<tr>
<td>menstr</td>
<td>currently menstruating</td>
<td>Y (yes), N (no)</td>
</tr>
<tr>
<td>age_lastmenstr</td>
<td>age menstruation stopped</td>
<td>integer</td>
</tr>
<tr>
<td>hysterect</td>
<td>hysterectomy</td>
<td>Y (yes), N (no)</td>
</tr>
<tr>
<td>both_ovrem</td>
<td>both ovaries removed</td>
<td>Y (yes), N (no)</td>
</tr>
<tr>
<td>fam_hist</td>
<td>affected first degree relatives</td>
<td>N (None), M (Mother), S (Sister(s)), D (Daughter(s)), F (Father or Brother), 1 (Mother &amp; Sister(s)), 2 (Mother &amp; Daughter(s)), 3 (Mother &amp; Sister(s) &amp; Daughter(s)), A (Adopted), O (Only Child), U (Unknown)</td>
</tr>
<tr>
<td>fam_hist50</td>
<td>relatives affected before 50</td>
<td>Y (yes), N (no)</td>
</tr>
<tr>
<td>curr_estr</td>
<td>currently using estrogen</td>
<td>Y (yes), N (no)</td>
</tr>
<tr>
<td>ever_biop</td>
<td>any breast biopsies</td>
<td>Y (yes), N (no)</td>
</tr>
<tr>
<td>num_biop</td>
<td>number of breast biopsies</td>
<td>integer</td>
</tr>
<tr>
<td>ca_diagdate</td>
<td>breast cancer diagnosis date</td>
<td>date (mm/dd/yyyy)</td>
</tr>
</tbody>
</table>
2.3 Exploratory Analysis

1411 women over the age of 80 were not considered in the following analysis. Overall, 221,545 women remained in the study.

Further, we adjusted the coding scheme of several variables to create consistent codes for missing, refused or non-applicable answers throughout the data set. Blank spaces and categories U (unknown) used to indicate unavailable values were replaced with missing value indicators NA. Categories R (refused) were also considered missing as this category was used in a sporadic pattern with a total of 502 entries in the entire data set. For the variable num_biop, the number of previous breast biopsies, missing values were originally assigned a value of zero, making it not possible to differentiate between women without prior biopsies and women with missing information. We adjusted this variable by replacing all values of zeros with NA for all women with missing or positive responses for the variable ever_biop.

Complete information on all risk factors was available for 54.1% of the 221,545 individuals, while 45.9% had missing information for at least one variable.

2.3 Exploratory Analysis

2.3.1 Population and SMP BC Comparison

After excluding participants outside the eligible SMP BC age range in the initial steps, the remaining women’s ages vary between 40 to 79 at the time of the 2000 screening appointment. Figure 2.1 shows the age distribution by five-year age groups within the SMP BC data set and the population of B.C. women aged 40 to 79. Provincial population estimates for the year 2000 were obtained from Statistics B.C. (Statistics B.C., 2012).

While most British Columbian women eligible for SMP BC screening are between the ages of 40 to 44, the age category of 45 - 49 showed the highest participation in the 2000 screening. This discrepancy is most likely to be explained by invitation letters sent annually to women turning 50 years of age and the resulting increase of awareness. With overall higher frequencies compared to the B.C. population, participants aged 45 to 69 are overrepresented in the data set, while fewer than expected women aged 70 to 79 decided to undergo screening in 2000. Due to the high sample size, a χ² test assessing the global goodness of fit between the observed and expected frequency distribution results in a highly significant p value of < 0.0001, suggesting that the age representation in the data set is not consistent with the
2.3. Exploratory Analysis

Figure 2.1: Age distribution by five-year age groups for 221,545 SMP BC participants and B.C. women aged 40 to 79.

Looking at the ethnic distribution given in Figure 2.2, SMP BC participants in the year 2000 were predominately white, with British and European origins representing the largest groups of 32.1% and 20.7%, respectively. Asian ethnicities rank second with East Asian origins accounting for 11.5% and South Asian origins for 2.8%. First Nations and African heritage were the smallest groups in the data set with 1.0% and 0.2%, respectively.

2.3.2 Analysis of Survival Times

Of 221,545 women, 4078 were diagnosed with breast cancer by the end of 2005, resulting in a relative cancer frequency of 1.85%. In situ cancers accounted for 18.6% of all cases.

The breast cancer free survival time $S(t)$ is estimated with a non-parametric estimator introduced by Kaplan and Meier (1958). For ordered event times $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$, it is defined as

\[
\hat{S}(t) = \prod_{t_{(i)} < t} \frac{n_i - d_i}{n_i}
\]  

(2.1)
2.3. Exploratory Analysis

![Ethnicity Composition](image)

**Figure 2.2:** Ethnic composition of 221,545 SMP BC participants undergoing screening in the year 2000.

where $n_i$ is the number of individuals at risk prior to event time $t_{(i)}$ and $d_i$ the number of events occurring at time $t_{(i)}$. The function `survfit()` available from the R package `survival` was used to calculate the survival function $\hat{S}(t)$.

Figure 2.3 shows the Kaplan-Meier estimator of survival times stratified by in situ and invasive cancers. In comparison to in situ cancers, risks associated with invasive breast cancers are considerably higher as seen from clear separation in survival curves. A noticeable drop in survival takes place shortly after the 2000 screening appointment for both cancer types. It can be assumed that these cancers developed before the start of the observational period and were found in the 2000 screening examination. Especially for invasive cancers, a wavy pattern in the survival curves repeating with a frequency of about one year can be seen. There appears to be a particular increase in cancer detection rates after approximately two years, a result of women following the SMP BC screening recall recommendations. With the exception of the initial drop, the survival proportions for both cancer types follow an overall linear trend with a five-year survival rate of 98.4% and 99.6% for invasive and in situ cancers, respectively.
2.3. Exploratory Analysis

Figure 2.3: Kaplan-Meier estimator and 95% confidence interval of breast cancer free survival times for 221,545 SMP BC participants, stratified by invasive (black) and in situ cancers (blue).

2.3.3 Risk Factors and Cancer Outcome

To investigate whether the development of breast cancer is associated with different levels of risk factor exposure, we calculated breast cancer odds relative to a baseline group defined as

$$\theta = \frac{p_{11}p_{12}}{p_{21}p_{22}}$$  (2.2)

for each risk factor level (Agresti, 2002). Table 2.2 gives the contingency table notation introduced for the following calculations.

Due to low cell counts in the cross classification of two risk factors, odds ratios $\theta$ were estimated using conditional maximum likelihood estimation. We assume a multinomial model for cell counts $n_{ij}$, thus, the distribution of $n_{ij}$ is dependent on the cell probabilities $p_{ij}$, $i, j = 1, 2$. Conditioning on the marginal counts, i.e. the total numbers of exposed ($n_e$) and diagnosed ($n_{bc}$) women, cell count $n_{11}$ determines the value of cell counts $n_{12}$, $n_{21}$ and $n_{22}$. Restating Equation (2.2) in terms of $p_{11}$ and the marginal probabilities, the odds ratio
2.3. Exploratory Analysis

Table 2.2: Contingency table notation of absolute numbers and probabilities (in parenthesis) for exposure groups and cancer outcomes.

<table>
<thead>
<tr>
<th></th>
<th>breast cancer</th>
<th>no breast cancer</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>exposed</td>
<td>$n_{11}$ ($p_{11}$)</td>
<td>$n_{12}$ ($p_{12}$)</td>
<td>$n_e$ ($p_e$)</td>
</tr>
<tr>
<td>unexposed</td>
<td>$n_{21}$ ($p_{21}$)</td>
<td>$n_{22}$ ($p_{22}$)</td>
<td>$n_0$ ($p_0$)</td>
</tr>
<tr>
<td>total</td>
<td>$n_{bc}$ ($p_{bc}$)</td>
<td>$n_{bc}$ ($p_{bc}$)</td>
<td>$n$ (1)</td>
</tr>
</tbody>
</table>

$\theta$ is given by

$$\theta = \frac{p_{11}(1 - p_e - p_{bc} + p_{11})}{(p_e - p_{11})(p_{bc} - p_{11})}.$$  

(2.3)

Hence, $p_{11}$ can be expressed as a function of the odds ratio given the marginal probabilities. The conditional distribution of the number of cancers in the exposed group, $n_{11}$, depends only on the parameter $\theta$ and is given by a non-central hypergeometric distribution

$$f(n_{11}|n_{bc}, n_e, \theta) = \frac{\binom{n_{bc}}{n_{11}} \binom{n - n_{bc}}{n_e - n_{11}} \theta^{n_{11}}}{\sum_u \binom{n_{bc}}{u} \binom{n - n_{bc}}{n_e - u} \theta^u}$$  

(2.4)

with summation over the range of all possible integers $u$, $\max(0, n_e - n_{bc}) \leq u \leq \min(n_{bc}, n_e)$ (Agresti, 2002). For an observed number $n_{11}$, the conditional maximum likelihood estimate $\hat{\theta}$ is obtained by maximizing the likelihood function

$$L(\theta|n_{11}, n_{bc}, n_e)$$

with respect to the odds ratio $\theta$.

95% confidence intervals were calculated using Fisher’s exact method. The lower bound of the confidence interval is the value $\theta_0$ such that the hypothesis test of $H_0 : \theta = \theta_0$ against $H_a : \theta > \theta_0$ results in a $p$-value of 0.025. Similarly, the upper bound of the confidence interval is the value $\theta_0$ such that the hypothesis test of $H_0 : \theta = \theta_0$ against $H_a : \theta < \theta_0$ results in a $p$-value of 0.025. The $p$-value can be calculated as

$$p = \sum_{x=i}^{j} f(x|n_{bc}, n_e, \theta)$$  

(2.5)
2.3. Exploratory Analysis

where $i = \max(0, n_e - n_{bc})$, $j = n_{11}$ for the alternative hypothesis of $H_a : \theta < \theta_0$ and $i = n_{11}$, $j = \min(n_{bc}, n_e)$ for $H_a : \theta > \theta_0$ (Agresti, 2002).

Odds ratios and 95% confidence intervals were calculated using the function `oddsratio()` from the R package `epitools` (R Core Team, 2012). The function estimates $\theta$ by conditional maximum likelihood estimation and it’s 95% confidence interval using Fisher’s exact method.

With a breast cancer prevalence of 1.85%, the rare disease assumption appears to be fulfilled and odds ratios can be assumed to closely approximate the relative breast cancer risk defined as (Collett, 2002)

$$ r = \frac{p_{11}}{p_{21}}. \quad (2.6) $$

For each risk factor level, Table 2.3 presents the number of missing values (NA), complete observations ($n$), absolute ($n_{bc}$) and relative ($n_{bc}/n$) numbers of cancers and odds ratios ($\theta$) with 95% confidence intervals. Categories with odds ratios of $\theta = 1$ represent baseline exposure levels. For variable *ethnic*, European and British ethnicities were combined into a new category ‘Caucasian’.

Odds ratios for all levels of age at first delivery, age at menarche, previous pregnancies and the number of full term deliveries were not significantly different from 1, suggesting that breast cancers are equally likely to occur in the exposed and baseline groups. Previous breast biopsies, current estrogen use, a hysterectomy and high breast density, however, showed significantly increased odds relative to the baseline groups and can therefore be associated with an increased probability of developing breast cancer. With a value of $\theta = 1.63$, breast biopsies appear to have a particularly strong effect.

The highest and lowest odds ratios are seen for the numbers of affected relatives and ethnicity groups. In comparison to Caucasian women, breast cancers are significantly less likely to occur in South and East Asian women, resulting in values of $\theta = 0.66$ and $\theta = 0.81$, respectively. Thus, ethnicity appears to be an important risk factor. Conversely, a familial history of breast cancer leads to the overall highest values in odds ratios. One, two and three affected relatives increase breast cancer odds by a factor of 1.64, 2.19 and 7.59, respectively, making it the strongest risk factor in the set of available variables. However, due to the limited number of women with two or more affected relatives, point estimates are subject to higher sampling error as seen from the noticeably wider confidence intervals.
### 2.3. Exploratory Analysis

Table 2.3: Number of missing values (NA), complete observations (n), absolute \( n_{bc} \) and relative \( n_{bc}/n \) cancer frequencies and odds ratios \( \theta \) with 95% confidence intervals for selected risk factor categories.

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>NA</th>
<th>levels</th>
<th>n</th>
<th>( n_{bc} )</th>
<th>( n_{bc}/n )</th>
<th>( \theta )</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>num_rel</td>
<td>12816</td>
<td>0</td>
<td>178948</td>
<td>3023</td>
<td>0.017</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>28019</td>
<td>768</td>
<td>0.027</td>
<td>1.64</td>
<td>1.51-1.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>1736</td>
<td>63</td>
<td>0.036</td>
<td>2.19</td>
<td>1.67-2.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>26</td>
<td>3</td>
<td>0.115</td>
<td>7.59</td>
<td>1.46-25.15</td>
</tr>
<tr>
<td>ethnic</td>
<td>59797</td>
<td>Caucasian</td>
<td>117161</td>
<td>2129</td>
<td>0.018</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>South Asian</td>
<td>6186</td>
<td>75</td>
<td>0.012</td>
<td>0.66</td>
<td>0.52-0.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aboriginal</td>
<td>2225</td>
<td>51</td>
<td>0.023</td>
<td>1.27</td>
<td>0.94-1.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Black</td>
<td>446</td>
<td>7</td>
<td>0.016</td>
<td>0.86</td>
<td>0.34-1.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>East Asian</td>
<td>25483</td>
<td>378</td>
<td>0.015</td>
<td>0.81</td>
<td>0.73-0.91</td>
</tr>
<tr>
<td>ever_bip</td>
<td>1525</td>
<td>no</td>
<td>186526</td>
<td>3144</td>
<td>0.017</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td>33494</td>
<td>908</td>
<td>0.027</td>
<td>1.63</td>
<td>1.51-1.75</td>
</tr>
<tr>
<td>curr_estr</td>
<td>2173</td>
<td>no</td>
<td>159926</td>
<td>2733</td>
<td>0.017</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td>59446</td>
<td>1310</td>
<td>0.022</td>
<td>1.30</td>
<td>1.21-1.39</td>
</tr>
<tr>
<td>breast_dens</td>
<td>163841</td>
<td>0 ≤ 50 %</td>
<td>163841</td>
<td>2918</td>
<td>0.018</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt; 50 %</td>
<td>57704</td>
<td>1160</td>
<td>0.020</td>
<td>1.13</td>
<td>1.06-1.21</td>
</tr>
<tr>
<td>ever_preg</td>
<td>0</td>
<td>no</td>
<td>175659</td>
<td>3230</td>
<td>0.018</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td>45886</td>
<td>848</td>
<td>0.018</td>
<td>1.01</td>
<td>0.93-1.09</td>
</tr>
<tr>
<td>hysterec</td>
<td>14609</td>
<td>no</td>
<td>152997</td>
<td>2741</td>
<td>0.018</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td>53939</td>
<td>1081</td>
<td>0.020</td>
<td>1.12</td>
<td>1.04-1.20</td>
</tr>
<tr>
<td>num_del</td>
<td>732</td>
<td>0</td>
<td>45886</td>
<td>848</td>
<td>0.018</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>26311</td>
<td>484</td>
<td>0.018</td>
<td>0.99</td>
<td>0.89-1.12</td>
</tr>
<tr>
<td>age_men</td>
<td>13971</td>
<td>≥ 14</td>
<td>63973</td>
<td>1140</td>
<td>0.018</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>12 - 13</td>
<td>111986</td>
<td>2107</td>
<td>0.019</td>
<td>1.06</td>
<td>0.98-1.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>≤ 11</td>
<td>31615</td>
<td>616</td>
<td>0.019</td>
<td>1.10</td>
<td>0.99-1.21</td>
</tr>
<tr>
<td>age_1stdel</td>
<td>11371</td>
<td>&lt; 20</td>
<td>25081</td>
<td>453</td>
<td>0.018</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>20-24</td>
<td>64502</td>
<td>1192</td>
<td>0.018</td>
<td>1.02</td>
<td>0.92-1.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25-29</td>
<td>48917</td>
<td>860</td>
<td>0.018</td>
<td>0.97</td>
<td>0.87-1.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt; 30</td>
<td>25840</td>
<td>529</td>
<td>0.020</td>
<td>1.14</td>
<td>0.99-1.29</td>
</tr>
</tbody>
</table>
2.3. Exploratory Analysis

2.3.4 Variable Interactions

Many of the risk factors given in the data set, in particular information about a woman’s menopausal status, are strongly associated with age. To control for potential confounding, it is important to differentiate whether differing cancer probabilities among risk factor levels are due to the risk factor itself or a result of it’s association with age. We found distinct age-related patterns within women’s responses on risk factors \textit{curr\_estr}, \textit{hysterec}, \textit{menstr}, \textit{educ}, \textit{mammo\_dens}. The results are displayed in Figure 2.4.

As to be expected, the relative frequency of women with response “Yes” on variable \textit{menstr} decreases sharply with age and approaches zero by age category 60-64. In contrast, the percentage of menopausal women who underwent a hysterectomy or are currently using estrogen show a non-linear pattern, initially increasing with age and peaking for categories 60-64 and 55-59, respectively. Mammographic density decreases with progressing age as we see a decrease in frequencies for the high density category ‘2’. Changes in educational
2.3. Exploratory Analysis

![Graph showing mean breast density by age and ethnicity.](image)

**Figure 2.5:** Age-specific mean breast density of 212,545 SMP BC women stratified by ethnicity.

level, most notably for women with no education, were also visible over time. Frequencies of this category are highest for age categories 75-80 and are slowly decreasing as women get younger.

Overall, there are clear response patterns as function of age, one of the strongest risk factors for breast cancer. A model which includes the risk factors simultaneously as covariates will show whether a risk factor’s effects on cancer-free survival time can be explained by it’s association with age.

We also found associations between mammographic density and ethnic background. For five-year age categories, Figure 2.5 shows the mean mammographic density stratified by ethnic groups. It can be seen that East Asian (EA) women have consistently higher breast densities for all age categories compared to the remaining ethnicities. Women with African (BL), South Asian (SA) and Caucasian (CA) heritage as well as women classified among the category ‘Other’ (OT) showed similar values with intersecting trends as age categories are increasing. The overall lowest values for all age categories were found for Aboriginal (AP)
2.4 Limitations of Follow-Up

To accurately predict the risk of developing breast cancer from the SMP BC study population, the reliable ascertainment of the population’s cancer status throughout the study is essential. Although participants’ breast cancer diagnoses before the end of the observational period are recorded in the data set, the SMP BC follow-up mechanism has several limitations as it both fails to identify women leaving the study and it also introduces informative censoring.

Based on the available information, it is unclear whether women without a recorded diagnosis date could have left the study through death from competing causes or relocation outside of British Columbia. In both cases, a woman remains classified as undiagnosed by the follow-up mechanism throughout the entire study period. However, as the SMP BC does not account for cancers detected outside of British Columbia, the recorded cancer status may be incorrect. As a result, the cancer status of undiagnosed participants may initially be unreliable and needs further investigation.

For each participant without a recorded diagnosis date, we defined the length of follow-up as the time between the 2000 screening date and the last recorded screening date within the SMP BC program. Figure 2.6 shows the distribution of follow-up times in years. While the majority of women undergo screening for a minimum of five years, 20.6% of all participants drop out of the program before the end of the observational period. The high percentage of participants never returning after their 2000 screening mammogram, a total of 8.5%, is particularly noticeable.

As women become ineligible for SMP BC screening once diagnosed with breast cancer, a negative breast cancer status can be confirmed for all participants with a follow-up period greater than five years. Undiagnosed women with a follow-up period shorter than five years on the other hand may still include women lost to follow-up, making their cancer status subject to error. To adjust for these limitations, the cancer status of undiagnosed individuals will only be considered known until their last recorded SMP BC screening date,
2.4. Limitations of Follow-Up

resulting in right censored survival times after this date.

Further, the follow-up mechanism causes the censoring probabilities to be dependent on the breast cancer status. Consider two individuals deciding to leave the study before the end of the observational period. Also assume that one will be diagnosed with cancer within the five year period. Although both were initially lost to follow-up, the cancer status of the case is updated in the SMP BC data base once diagnosed, causing her to be included in the study again. Therefore, the probability of censoring is higher for a non-case than for a case, as cases are only able to leave the study through relocation or death, but not through discontinuation of screening. As a result, informative censoring is introduced into the data set.

Censoring probabilities were also found to be dependent on risk factor levels. Figure 2.7 shows the remaining percentage of participants at the end of each calendar year. Women between the ages of 40 and 69 and above the age of 70 are split up into ten and five-year age categories, respectively. While women aged 40 to 69 appear to leave the study only in small numbers throughout the entire study period, the participation of women over the age of 70.
declines rapidly. It is especially noticeable for the age category of 75-79 where participation drops below 10 % by the end of 2005. Therefore, older women are more likely to discontinue screening compared to younger women, creating associations between risk groups and follow-up time. This pattern can mostly be explained by SMP BC policies. Women aged 75 to 79 become ineligible for screening examinations in the course of the five-year follow-up period, forcing them to leave the study before the end of the observation.

Overall, the cancer status of certain participants cannot be considered known for the length of the study. Instead, observations are right censored with differing follow-up periods. This will be taken into account for further analyses, where appropriate adjustments will be made to address the presence of informative censoring.

### 2.5 Evaluation of Study Bias

Selection bias enters the study population in several ways. Active participants in cancer screening programs are subject to overdiagnosis (Zahl et al., 2004). Particularly in older
2.5. Evaluation of Study Bias

Figure 2.8: Age-specific breast cancer incidence rates per 100,000 person years for the B.C. population and the data set of 212,545 SMP BC women.

women, screening mammography detects cancers that would never have progressed to cause symptoms or death during the women’s lifetime. In an unscreened population, these cancers are not diagnosed or treated, leading to a lower numbers of new cases overall. Therefore, incidence rates, defined as the number of new cases per 100,000 person years of observation, are higher in screened women than in the general population. A shift in cancer detection to earlier ages as a consequence of screening has also been noted (Gail and Benichou, 1994). Compared against an unscreened population, SMP BC participants receive a lead time in cancer detection and thus contribute more cases to lower age categories. As breast cancer incidence increases rapidly with age for younger women, this produces increased cancer rates in particular for lower age categories.

Figure 2.8 shows age-specific incidence rates for 220,691 SMP BC women without prevalent breast cancers and the population of British Columbia. Provincial incidence rates were calculated as the mean incidence rate of the year 2000 to 2004 in British Columbia available from Statistics Canada (Statistics Canada, 2011). In situ cancers were included in
2.5. Evaluation of Study Bias

provincial incidence estimates. Both populations show a positive, approximately linear relationship between incidence rates and age. For all age categories, rates are consistently higher in the screened population. Differences between rates are smallest for ages 40-44 and are increasing with age, a possible sign of overdiagnosis. A $\chi^2$ test comparing the incidence distributions resulted in a highly significant test statistic of 532.54 with a corresponding p-value of $p < 0.001$ and confirms biased incidence rates in the SMP BC sample. As noted by Gail et al. (1989), these results confirm that the risk estimates derived from the SMP BC population are most appropriate for women considering to undergo regular screening, not the general population of British Columbia.
Chapter 3

Gail Model Validation

In the following chapter, we introduce the Gail model and assess its predictive and discriminatory performance to determine whether it provides a reliable breast cancer prediction tool for the B.C. population. Further, we compare its performance to that of a null model ignoring all covariate information. An analysis of the Gail model’s residuals is conducted to gain insight into potential improvements of the functional form.

3.1 The Gail Model

The Gail model estimates a woman’s probability \( p(\tau|X_i) \) of developing invasive breast cancer within a predefined time interval \([0, \tau]\), given her current constellation of the risk factor vector \( X_i \).

In an initial step, her risk \( r_i \) relative to a woman’s at baseline risk factor exposure and of the same age is approximated by the odds ratio from the logistic regression model

\[
  r_i(t|X_i) = \exp(X_i^T \beta)
\]

(3.1)

Parameters \( \beta \) are estimated from a matched case-control study within the BCDDP cohort. Variables contained in the covariate vector \( X_i \) include an indicator whether a woman is above 50 years of age, her age at menarche, age at first live birth, the number of affected first degree relatives and the number of previous benign breast biopsies. The model also includes an interaction term between the age indicator and the number of breast biopsies as well as an interaction between the age at first delivery and the number of affected relatives.

The logistic regression coefficients for Caucasian, Black and Asian American women were extracted from the National Cancer Institute’s Gail model source code, version 3.0 (National Cancer Institute, 2012), and are presented in Table 3.1. For African and Asian
Table 3.1: Gail model logistic regression coefficients for Caucasian, Black and Asian women as provided by the Gail model source code.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Ethnicity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Caucasian</td>
</tr>
<tr>
<td>age ≥ 50</td>
<td>0.011</td>
</tr>
<tr>
<td>age_men</td>
<td>0.094</td>
</tr>
<tr>
<td>num_biop</td>
<td>0.529</td>
</tr>
<tr>
<td>age_1stdel</td>
<td>0.219</td>
</tr>
<tr>
<td>num_rel</td>
<td>0.958</td>
</tr>
<tr>
<td>num_biop : age ≥ 50</td>
<td>-0.288</td>
</tr>
<tr>
<td>age_1stdel : num_rel</td>
<td>-0.191</td>
</tr>
</tbody>
</table>

ethnicities, variables with parameter values of 0 were not significant and are not included in the model. While significant coefficients for Asian and Caucasian women are rather similar, we note that there is a large difference for the model coefficients of Black women. Regression coefficients for Caucasian women are based on the original paper by Gail et al. (1989) while values for Black and Asian women were added as the result of subsequent studies by Gail et al. (2007) and Matsuno et al. (2011), respectively.

A woman’s relative risk \( r_i(t | X_i) \) is then combined with the breast cancer baseline hazard \( h_1(t) \), an age and ethnicity-specific hazard rate for women without exposures to risk factors besides age. This baseline hazard rate \( h_1(t) \) is obtained from the breast cancer incidence rate \( h_1^*(t) \) within the composite SEER population using the following adjustment from Bruzzi et al. (1985).

For \( I \) unique combinations of risk factor levels within the composite population, let \( P_i(t) \) and \( r_i(t) \) be the proportion of women of age \( t \) and relative risk within risk group \( i, i = 1 \ldots I \), respectively. Then

\[
h_1^*(t) = \sum_{i=1}^{I} P_i(t) h_1(t) r_i(t)
\]

For each risk group \( i \), the proportion of cancer cases of age \( t \), \( \rho_i(t) \), is given by

\[
\rho_i(t) = P_i(t) h_1(t) r_i(t) / h_1^*(t)
\]
3.1. The Gail Model

Dividing Equation (3.2) by \( r_i(t) \), multiplying it by \( h_1^*(t) \) and summing on \( i \), we obtain

\[
h_1(t) = h_1^*(t) \sum_{i=1}^{I} \frac{p_i(t)}{r_i(t)} \equiv h_1^*(t) F(t)
\]

where \( p_i(t) \) was estimated by the observed proportion of cases in risk group \( i \) within the SEER cohort and \( r_i(t) \) is obtained from the logistic regression model given in Equation (3.1). The factor \( F(t) \) is equivalent to \( 1 - AR(t) \), where \( AR(t) \) is the population attributable risk fraction for women of age \( t \), the reduction in risk if no risk factors were present in the population.

Under consideration of an age-specific hazard function for competing causes of death, \( h_2(t) \), the probability of developing breast cancer in the age interval \([a, a+\tau]\) is given by

\[
p(a, \tau, r_i) = \int_a^{a+\tau} h_1(t) r_i(t) \exp \left\{ - \int_a^t h_1(u) r_i(u) \, du \right\} \frac{S_2(t)}{S_2(a)} \, dt
\]

where

\[
S_2(t) = \exp \left\{ - \int_0^t h_2(u) \, du \right\}
\]

describes the probability of surviving competing risks up to age \( t \). Dividing the age scale into \( m \) age categories \([\tau_{j-1}, \tau_j]\), \( j = 1 \ldots m \) with widths \( \Delta_j \) and assuming that relative risks \( r_i(t|X_i) \), baseline hazards \( h_1(t) \) and competing risks \( h_2(t) \) are constant within each category \( j \), Equation (3.4) is evaluated as

\[
p(a, \tau, r_i(t)) = \sum_{j} \frac{h_{1j} r_{ij}}{(h_{1j} r_{ij} + h_{2j})} \frac{S_1(\tau_{j-1})}{S_1(a)} \frac{S_2(\tau_j)}{S_2(a)} \frac{S_2(\tau_{j-1})}{S_2(a)}
\]

\[
\times \left[ 1 - \exp \left\{ - \Delta_j (h_{1j} r_{ij} + h_{2j}) \right\} \right]
\]

where \( \Delta_j \) is the length of the age interval \( j \) and \( h_{1j}, h_{2j} \) and \( r_{ij} \) denote the baseline hazard, competing hazard and relative risk in interval \( j \), respectively. The smallest value of \( j \) in the summation is chosen such that \( \tau_{j-1} = a \), the largest such that \( \tau_j = a + \tau \). To simplify calculations, values \( a \) and \( a + \tau \) are assumed to be an element of the set \( \{\tau_0, \ldots \tau_m\} \).
3.2 Validation Population

Several individuals were removed from the data set for the following validation study. As the Gail model predicts the probability of breast cancer occurrence in a predefined time interval in the future, all individuals in the validation set were required to be without breast disease at the beginning of the study. Hence, 855 of 221,545 women with prevalent breast cancers detected at the 2000 screening appointment had to be excluded while all women with a negative screening result remained in the validation set. Further, 600 participants diagnosed with lobular and ductal carcinoma in situ after the 2000 screening appointment were removed from the data set as the Gail model predicts probabilities for invasive cancers only.

The cancer status of all participants was ascertained through December 31, 2005, allowing each woman to be followed for a minimum of five years. Therefore, we calculated five-year breast cancer risks $p(\tau = 5)$ given in Equation (3.6) for all women with a follow-up time of at least five years. As described in Section 2.4, 20.6% of all women left the study before the end of the observational period. Their cancer status cannot be considered known for the full five-year follow-up length, making it inappropriate to include them in a validation study of five-year cancer risks. Thus, these 41,997 women were removed from the validation set. Further adjustments to the remaining observations were required as a mere exclusion of these non-cases led to biased cancer rates in the population and could affect the calibration performance of the assessed model.

Investigations showed that the highest discrepancy between cancer rates before and after exclusion of women lost to follow-up occurred within the variable age. Figure 3.1 shows the change in relative cancer frequencies within five-year age categories for the validation set when women lost to follow-up were removed. While rates remain similar for ages 40 to 69, age categories 70 to 79 show a considerable bias as a result of the exclusion. Cancer rates increase nearly seven fold for the age category 75-79.

Other variables, in particular ethnic background, showed a nearly consistent increase in cancer rates for all variable levels, indicating that the drop out behaviour is mainly associated with age. This assumption is further affirmed as women over the age of 75 are forced to leave the screening program due to SMP BC eligibility policy. Based on these results, we decided to exclude all women over the age of 75 from the validation set in order to minimize
3.3 Validation Methods

For all women in the validation set, five-year breast cancer probabilities were calculated using a program developed by the BC Cancer Agency. The program is based on Version 3.0 (May 2011) of the Gail model source code available from the webpage of the National Cancer Institute (2012) and allowed us to process calculations for all women in batch mode. As an implementation of the Gail model, it returns a woman’s probability of developing invasive breast cancer when standard Gail risk factors age, age at menarche, age at first

![Figure 3.1: Relative cancer frequency by five-year age category for 221,545 individuals when women lost to follow-up are included and removed from the data set.](image-url)

the introduced bias.

Overall, the validation set contained a total of 176,978 women, 2226 of which were diagnosed with invasive breast cancer in the five-year projection interval. Information on the risk factor ‘atypical hyperplasia’ required for the Gail model was not included in the data set and therefore set to NA for all individuals. Only 78,106 had complete information on all other variables.

3.3 Validation Methods

For all women in the validation set, five-year breast cancer probabilities were calculated using a program developed by the BC Cancer Agency. The program is based on Version 3.0 (May 2011) of the Gail model source code available from the webpage of the National Cancer Institute (2012) and allowed us to process calculations for all women in batch mode. As an implementation of the Gail model, it returns a woman’s probability of developing invasive breast cancer when standard Gail risk factors age, age at menarche, age at first
live birth, number of affected first degree relatives, the number of previous benign breast biopsies, atypical hyperplasia and ethnicity are provided. To ensure the reproducibility of our risk assignments, the program’s calculated risks were compared to those of the online calculator (National Cancer Institute, 2012) for 50 randomly selected individuals.

The predictive performance of the Gail model was assessed by comparing the model’s predicted number of breast cancers \( E \) with the observed number \( O \) in a given subgroup of women. Variables considered for the validation analysis are age (by five-year age groups) and ethnicity. The expected value \( E_j \) in category \( j \) is calculated as

\[
E_j = \sum_{i=1}^{n_j} p_{ij} (\tau = 5 | X_i)
\]

(3.7)

where \( p_{ij} \) is the \( i \)th woman’s five-year cancer risk and \( n_j \) the number of women in category \( j \).

To investigate whether the ratio between observed and expected cancers in each category is significantly different from 1, confidence intervals for \( E/O \) ratios were calculated following the steps in Costantino et al. (1999). Assuming expected values \( E \) to be constant and observed values to follow a Poisson distribution \( Pois(O) \) with expected value \( O \), a 95% confidence interval for \( E/O \) is given by

\[
\left( \frac{2E}{\chi^2(0.975, 2(O + 1))}, \frac{2E}{\chi^2(0.025, 2O)} \right)
\]

(3.8)

as

\[
\left( \frac{\chi^2(0.025, 2O)}{2}, \frac{\chi^2(0.975, 2(O + 1))}{2} \right)
\]

(3.9)

is the exact 95% confidence interval for the detected number of breast cancers \( O \) under the assumption of a Poisson distribution.

Further, global goodness of fit over all age and ethnic categories \( j \) was assessed by the \( \chi^2 \) test statistic defined as

\[
X^2 = \sum_{j=1}^{k} \left( \frac{(E_j - O_j)^2}{E_j} \right)
\]

(3.10)

where \( k \) denotes the number of categories. Under the null hypothesis of no difference in frequency distributions, \( X^2 \) follows a \( \chi^2 \) distribution with \( k - 1 \) degrees of freedom.
The discriminatory performance of the model was evaluated by use of the concordance statistic \( c \). The concordance is defined as

\[
c = \Pr(p(\tau | X_i) > p(\tau | X_j) \mid N_i(\tau) = 1, N_j(\tau) = 0)
\]  

(3.11)

and describes the probability that the predicted breast cancer risk \( p(\tau | X_i) \) exceeds \( p(\tau | X_j) \) for two randomly chosen individuals \( i \) and \( j \) with cancer outcomes \( N_i(\tau) = 1 \) and \( N_j(\tau) = 0 \) (Harrell, 2001).

The concordance statistic takes values in the interval \([0, 1]\). A value of \( c = 1 \) indicates that the risk intervals for cases and non-cases can be separated into two mutually exclusive categories by a risk threshold \( T \). Hence, future cancer cases and non-cases can be uniquely identified based on their probability \( p_i \) lying above or below \( T \). A value of \( c = 0.5 \) on the other hand indicates that, based on the model’s prediction, a woman’s assignment into one of the two categories is as accurate as a random guess. The concordance statistic is independent of disease prevalence in the sample (Ash and Shwartz, 1999), hence, the assessment of discriminatory performance is not affected by the selection bias introduced through exclusion of women lost to follow-up. The function \texttt{survConcordance()} available from the \texttt{R} package \texttt{survival} offers calculation of concordance values and corresponding standard errors.

### 3.4 Performance Evaluation

#### 3.4.1 Risk Distribution

For the 176,978 women included in the validation set, the Gail model predicted an average five-year breast cancer risk of 1.11% with minimum and maximum risks of 0.19 % and 12.2 %, respectively. 95% of all individuals had estimated risks below 2.29% while projected risks for the remaining 5% varied between 2.29% and 12.2%, resulting in a skewed distribution of breast cancer probabilities.

Figure 3.2 shows the risk distribution of five-year cancer probabilities for the validation set. The majority of the data appear to be approximately normally distributed. The long, right tail of the histogram containing only a few individuals is particularly noticeable. Cancer
3.4. Performance Evaluation

![Figure 3.2: Distribution of five-year breast cancer probabilities for 176,978 SMP BC participants as estimated by the Gail model.](image)

rates seem to be inflated for these women and require further investigations into covariate combinations responsible for these high risk estimates.

### 3.4.2 Assessment of Model Calibration

We investigated the model’s performance separately for the full validation set as defined in Section 3.2 and for a subset of 78,106 individuals with complete information on all available Gail model variables. Table 3.2 shows the number of individuals $n$, observed and expected breast cancers, $E/O$ ratios and 95% confidence intervals for each validation set, overall and stratified by five-year age categories and ethnic groups.

When all individuals are considered, the Gail model predicted a total number of $E = 2110.25$ breast cancers for the validation set, underestimating the detected number of $O = 2226$ by 5.2%. The difference between expected and observed values is marginally significantly with a 95% confidence interval of (0.91, 0.99) for the $E/O$ ratio. Thus, the Gail model appears to be miscalibrated for the SMP BC population.

Within five-year age categories, the model underestimated the observed number of breast cancers for all age categories with the exception of 40-44 and 60-64. However, $E/O$ ratios
3.4. Performance Evaluation

Table 3.2: Numbers of expected and observed cancers, $E/O$ ratios and 95% confidence intervals for age and ethnic groups for (1) 176,978 women and (2) 78,106 women with complete information.

<table>
<thead>
<tr>
<th>Category</th>
<th>n</th>
<th>O</th>
<th>E</th>
<th>E/O</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40-44</td>
<td>30813</td>
<td>209</td>
<td>218.23</td>
<td>1.04</td>
<td>0.91-1.20</td>
</tr>
<tr>
<td>45-49</td>
<td>40869</td>
<td>383</td>
<td>378.52</td>
<td>0.99</td>
<td>0.89-1.10</td>
</tr>
<tr>
<td>50-54</td>
<td>34447</td>
<td>376</td>
<td>372.3</td>
<td>0.99</td>
<td>0.89-1.10</td>
</tr>
<tr>
<td>55-59</td>
<td>25006</td>
<td>388</td>
<td>338.26</td>
<td>0.87</td>
<td>0.79-0.97</td>
</tr>
<tr>
<td>60-64</td>
<td>19757</td>
<td>286</td>
<td>315.08</td>
<td>1.10</td>
<td>0.98-1.24</td>
</tr>
<tr>
<td>65-69</td>
<td>16159</td>
<td>308</td>
<td>291.51</td>
<td>0.95</td>
<td>0.85-1.06</td>
</tr>
<tr>
<td>70-75</td>
<td>9927</td>
<td>276</td>
<td>196.35</td>
<td>0.71</td>
<td>0.63-0.80</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>94047</td>
<td>1148</td>
<td>1171.25</td>
<td>1.02</td>
<td>0.96-1.08</td>
</tr>
<tr>
<td>South Asian</td>
<td>5048</td>
<td>48</td>
<td>26.98</td>
<td>0.56</td>
<td>0.42-0.76</td>
</tr>
<tr>
<td>Aboriginal</td>
<td>1639</td>
<td>32</td>
<td>18.46</td>
<td>0.58</td>
<td>0.41-0.84</td>
</tr>
<tr>
<td>Black</td>
<td>350</td>
<td>3</td>
<td>3.99</td>
<td>1.33</td>
<td>0.46-6.45</td>
</tr>
<tr>
<td>East Asian</td>
<td>20678</td>
<td>183</td>
<td>119.24</td>
<td>0.65</td>
<td>0.56-0.76</td>
</tr>
<tr>
<td>Total</td>
<td>176978</td>
<td>2226</td>
<td>2110.25</td>
<td>0.95</td>
<td>0.91-0.99</td>
</tr>
</tbody>
</table>

| (2) Age  |     |     |     |      |                 |
| 40-44    | 17415 | 120 | 128.06 | 1.07 | 0.89-1.29       |
| 45-49    | 19694 | 191 | 181.78 | 0.95 | 0.83-1.10       |
| 50-54    | 13970 | 133 | 148.16 | 1.11 | 0.94-1.33       |
| 55-59    | 9864  | 124 | 127.39 | 1.03 | 0.86-1.24       |
| 60-64    | 7698  | 101 | 113.97 | 1.13 | 0.93-1.39       |
| 65-69    | 6010  | 101 | 101.8  | 1.01 | 0.83-1.24       |
| 70-75    | 3455  | 93  | 63.77  | 0.69 | 0.56-0.85       |
| Ethnicity|     |     |     |      |                 |
| Caucasian| 56021 | 654 | 697.23 | 1.07 | 0.99-1.15       |
| South Asian| 3371  | 26  | 18.17  | 0.70 | 0.48-1.07       |
| Aboriginal| 1027  | 19  | 11.82  | 0.62 | 0.40-1.03       |
| Black    | 193   | 2   | 2.13   | 1.07 | 0.29-8.79       |
| East Asian| 12640 | 104 | 76.15  | 0.73 | 0.60-0.90       |
| Total    | 78106 | 863 | 864.94 | 1.00 | 0.94-1.07       |
were only significantly different from 1 for age categories 55-59 and 70-75. The considerable lack of fit for ages 70-75 ($O = 276, E = 196.35$) is most likely caused by elevated cancer rates due to the high proportion of women lost to follow-up in this age group.

The Gail model shows no significant miscalibration for Caucasian and African ethnicities, but underestimates breast cancers for South Asian, East Asian and Aboriginal ethnic groups significantly by 43.8%, 34.8% and 42.3%, respectively. Testing the overall goodness of fit between expected and observed breast cancers results in a $\chi^2$ test statistic of $X^2 = 43.72$ and $X^2 = 61.11$ for age and ethnic categories. Corresponding p-values were highly significant in both cases, indicating a significant difference in frequency distributions.

With a prediction $E = 864.94$ versus $O = 863$ detected breast cancers, the overall fit of the Gail model improves considerably when all individuals with missing risk factor information are excluded. The model overpredicts the observed number by only 1.94 cancers, resulting in an $E/O$ ratio of 1. As opposed to a ratio of $E/O = 0.95$ observed earlier, the Gail model appears to be well calibrated for the SMP BC population when all necessary risk factor information is known.

Similar improvements can be seen for the assessment within five-year age categories. Figure 3.3 shows the $E/O$ ratios for age and ethnic groups. With the exception of women aged 70-75, the model seems well calibrated for all age groups showing no $E/O$ ratios significantly different from 1. Again, the underestimation of cancers seen in the age group of 70-75 can in part be explained by biased cancer rates and does not necessarily indicate a miscalibration of the model.

When splitting the individuals into ethnic subgroups, the number of breast cancers remains above expectation for Asian and Aboriginal ethnicities, however, differences between observed and expected values were not significant for South Asian and Aboriginal groups. With $E/O = 0.73$ (CI: 0.60-0.90), a significant miscalibration is still present for East Asian ethnicities.

Test statistics for the global $\chi^2$ hypothesis test decreased considerably to $X^2 = 17.5$ and $X^2 = 20.6$ for age and ethnicity, however, remain highly significant with $p < 0.01$. The distribution of observed cancers for both age and ethnicity differs significantly from the expected numbers.
3.4. Performance Evaluation

![Graph showing E/O Ratio vs Age Category and Ethnicity](image)

**Figure 3.3:** Ratio of expected and observed cancers with 95% confidence interval in five-year age categories and ethnic groups for 78,106 individuals with complete Gail information.

3.4.3 Discriminatory Ability

The model’s ability to discriminate between future cases and non-cases using projected breast cancer probabilities $p$ was comparably poor for both validation sets. Concordance values were $c = 0.613$ (95% CI: 0.601-0.625) and $c = 0.611$ (95% CI: 0.591-0.630) for the full and reduced validation set, respectively. With both confidence intervals lying well above the theoretical threshold value of 0.5, the Gail model significantly outperforms a random classification process, however, applicability in the prediction of breast cancer outcomes on an individual level remains limited.

Figure 3.4 shows the substantial overlap between the estimated risk densities of cases and non-cases for 78,106 women with complete Gail model information. Distributions were obtained from the R function `geom_density()` using normal kernel density estimation. Although the mean five-year risk is higher for cases with $\bar{p} = 0.015$ versus $\bar{p} = 0.011$ for non-cases, there is no distinct separation between probability ranges for the two outcomes. If a risk threshold $T$ was to be placed to separate risks into two outcome categories (cancer/no cancer), a considerable number of future cancer patients would be considered as low risk individuals.
Based on these results, the Gail model’s predicted breast cancer risk is not a reliable tool to determine who will and will not develop cancer in a future time period. Overall, the Gail Model appears to be well calibrated to predict the number of cancers in the SMP BC population and various subgroups, but fails to predict cancer outcomes for individual women. Thus, improvements on discriminatory accuracy are needed.

3.4.4 Comparison to Incidence Model

To decide whether the calibration of the Gail model significantly outperforms that of a null model ignoring all covariate information, we compared the observed number of cancers in five-year age categories for 176,978 women with the numbers predicted by British Columbian incidence rates. Provincial incidence rates per 100,000 person-years were calculated as the mean incidence between 2000 and 2004 in British Columbia, available online from Statistics Canada (2011).

For each five-year age category, Table 3.3 shows the person-years contributed by SMP BC participants, provincial incidence rates in British Columbia, the expected and observed...
### Table 3.3: Person years of 176,978 SMP BC women, B.C. incidence rates, observed cancers \((O)\) and expected cancers predicted by B.C. incidence \((E)\) for five-year age categories.

<table>
<thead>
<tr>
<th>age</th>
<th>person years</th>
<th>B.C.</th>
<th>E</th>
<th>O</th>
<th>E/O</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>40-44</td>
<td>67914.1</td>
<td>99.2</td>
<td>67.37</td>
<td>61</td>
<td>1.1</td>
<td>0.86-1.44</td>
</tr>
<tr>
<td>45-49</td>
<td>185256.4</td>
<td>162.8</td>
<td>301.6</td>
<td>321</td>
<td>0.94</td>
<td>0.84-1.05</td>
</tr>
<tr>
<td>50-54</td>
<td>197207.42</td>
<td>204.6</td>
<td>403.49</td>
<td>403</td>
<td>1</td>
<td>0.91-1.11</td>
</tr>
<tr>
<td>55-59</td>
<td>142747.42</td>
<td>247.0</td>
<td>352.59</td>
<td>344</td>
<td>1.02</td>
<td>0.92-1.14</td>
</tr>
<tr>
<td>60-64</td>
<td>110130.34</td>
<td>285.4</td>
<td>314.31</td>
<td>360</td>
<td>0.87</td>
<td>0.79-0.97</td>
</tr>
<tr>
<td>65-69</td>
<td>89044.13</td>
<td>303.0</td>
<td>269.8</td>
<td>285</td>
<td>0.95</td>
<td>0.84-1.07</td>
</tr>
<tr>
<td>70-74</td>
<td>67925.92</td>
<td>328.0</td>
<td>222.8</td>
<td>289</td>
<td>0.77</td>
<td>0.69-0.87</td>
</tr>
</tbody>
</table>

number of cancers in the five-year follow-up period and \(E/O\) ratios with 95\% confidence intervals.

With the exception of age categories 60-64 and 70-74, the ratio of expected and observed cancers is not significantly different from 1. Thus, the incidence model is well calibrated for most age categories without consideration of any risk factor information. The number of breast cancers in category 70-74 was underestimated by 22.9\%. Again, the miscalibration can partly be explained by biased cancer rates in this age category. The Gail model outperformed the null model only for age category 60-64 where the incidence model underestimated the observed number of cancers significantly by 12.7\%. The Gail model, on the other hand, showed no significant miscalibration for this age range.

As a result, the good calibration performance of the Gail model is not surprising. A reasonably accurate prediction of expected cancers can already be obtained by incidence rates alone.

### 3.5 Residual Analysis

Initiated by the highly skewed risk distribution observed in Section 3.4.1, the adequacy of the linear predictor in Equation (3.1) needed investigation. For each individual with complete Gail model information, we calculated the standardized Pearson residuals defined
3.5. Residual Analysis

**Figure 3.5:** Boxplot of Pearson residuals from Gail model stratified by the number of affected relatives. The y-axis is restricted to display the model’s residuals of non-cases only.

\[
\epsilon_i = \frac{N_i - p_i}{\sqrt{p_i(1 - p_i)}}, \quad i = 1 \ldots n
\]  

(3.12)

where \(N_i\) denotes the cancer status at time \(\tau = 5\) for individual \(i\) and \(p_i\) the predicted probability by the model. Systematic patterns in residual plots against the explanatory variables indicate functional misspecifications of the linear predictor (Therneau and Grambsch, 2000).

Figure 3.5 shows the Pearson residuals stratified by the number of affected first degree relatives in a box plot. Variable levels 0, 1, 2 and 3 contained 68292, 9306, 505 and 3 individuals, respectively. To improve visibility of the majority of residuals, the y-axis was restricted and excludes the residuals of all individuals with cancer outcome \(N_i(\tau = 5) = 1\). We see a non-linear, positive pattern in the residuals peaking at two affected relatives, suggesting that the increase in relative risk for women with one or two affected relatives is too high. A very similar pattern was found within the excluded points.

There are two potential causes for the lack of fit in these family history categories. First, this pattern might be caused as the Gail model includes the number of affected relatives as an ordinal rather than a categorical variable. Assuming a log linear increase in relative
risk with increasing numbers of relatives might not be sensible and requires further analysis. This ordinal modelling of family history also appears to be the cause of inflated cancer risks for certain individuals as seen in the highly skewed risk distribution of Figure 3.2. The Gail model’s parameter associated with the variable $\text{num\_rel}$ is the overall highest with a value of 0.958 for Caucasian women, leading to an increase in relative risk by a factor of 2.61 and 6.79 for one and two affected relatives, respectively. In addition, exploratory analysis showed that individuals with the highest five-year Gail risk had a proportionally high number of affected relatives.

Further, stratification of family history with respect to the number of relatives only might not be sufficient. Instead, the type of relative affected should also be taken into account as there is an intuitive difference in risk between a mother and a daughter as a result of the age difference alone. Both alternatives to the current use of family history information in the Gail model will be considered in the following model building process.

### 3.6 Optimization of Risk Threshold

The lack in discriminatory performance of the Gail model is seen in the previous section, what follows is a more detailed analysis in the potential benefits of a population partition based on five-year Gail risks.

An optimal risk threshold $T^*$ defining the partition between a high and low-risk group should simultaneously satisfy two criteria. First, the proportion of women with cancer probabilities greater than threshold $T^*$ cannot exceed a predefined upper bound to ensure affordability of women’s access to enhanced screening procedures. At the same time, this proportion should also be large enough to make a resulting change in screening policies sensible and worthwhile. Second, the difference in relative cancer frequencies $f_j, j = 1, 2$ between the binary partition of the population based on threshold $T^*$ should be maximized.

To find such an optimal value $T^*$, we investigated relative cancer frequencies in the high and low-risk partitions of 78,106 women without missing covariate information as a function of $T$. For each threshold $T$, we evaluated

$$
f(T) = \frac{\sum_i I(N_i(5) = 1 | p_i \geq T)}{\sum_i I(p_i \geq T)}, \quad T = 0, \ldots, \max(p_i) \quad (3.13)
$$
3.6. Optimization of Risk Threshold

where \( N_i \) denotes the cancer status for individual \( i \) and \( p_i \) the five-year risk as predicted by the Gail model. The step size of \( T \) was chosen as 0.1%.

As the number of individuals with risk estimates \( p_i > T \) decreases quickly with \( T \), 95% confidence intervals for relative cancer frequencies \( f(T) \) were calculated using an exact method rather than a normal approximation. Define \( y \) as the absolute number of cancers within \( n \) individuals above threshold \( T \). The confidence bounds are the smallest and largest probabilities for which the occurrence of the observed proportion \( y/n \) has a probability that is at least equal to \( \alpha/2 \) (Collett, 2002). Hence, the lower bound \( p_l(T) \) is the value such that \( p_l(T) \) fulfills the equation

\[
\sum_{j=0}^{y} \binom{n}{j} p_l(T)^j (1 - p_l(T))^{n-j} = \frac{\alpha}{2}
\]

and the upper bound \( p_u(T) \) is the value such that \( p_u(T) \) fulfills

\[
\sum_{j=0}^{y} \binom{n}{j} p_u(T)^j (1 - p_u(T))^{n-j} = \frac{\alpha}{2}.
\]

Exact confidence bounds were calculated using the function \texttt{binom.confint()} available from the \texttt{R} package \texttt{binom}.

The left hand side graph of Figure 3.6 shows the relative cancer frequency in the high-risk group as \( T \) varies between 0 and the maximum observed Gail risk in the sample of 78,106 SMP BC women (black), the 95% lower confidence bound (dashed) and the overall prevalence level (blue). The right hand side graph displays the percentage of women with estimated cancer risks \( p \) exceeding \( T \).

Relative cancer frequencies for women with \( p_i > T \) are consistently greater than the observed prevalence of 0.011 with the exception of the right most bounds. For threshold values greater than 0.082, the relative frequency drops to 0, indicating that the women with the highest estimated risk were free of cancer by the end of the five-year study period. Due to the steep decrease in high risk women, the smoothness of the relative frequency curve decreases with increasing \( T \).

Values of \( f \) are predominately increasing until a local maximum is reached at risk threshold \( T = 0.031 \). A short drop is followed by a steep increase in \( f \) until the overall maximum
3.6. Optimization of Risk Threshold

![Graph](image)

**Figure 3.6:** Left: Cancer detection rate above threshold $T$ (black) with lower 95\% confidence bound (dashed) and prevalence (blue) for 78,106 women. Right: Proportion of women with risk estimates $p_i$ above risk threshold $T$.

of 0.20, however, the lower 95\% confidence bound remains below or in the proximity of the prevalence level for all values $T > 0.037$.

If one was to restrict the smallest possible proportion of high risk women to 1\% of the population, the maximum achievable increase in cancer detection rates relative to prevalence is obtained at $T = 0.031$ where an approximately 2.5 fold increase from 0.011 to 0.026 is observed. For this choice, approximately 1.74\% of the SMP BC population then becomes eligible for advanced screening. The value $T = 0.031$ is indicated on both plots of Figure 3.6 by a vertical dashed line.

Comparing these numbers to the increases in odds ratios seen in Section 2.3.3 through population stratification on a single risk factor alone, a Gail model threshold offers a slim advantage over any of the available risk factor levels. However, increases in relative risk between high and low risk groups still remain too low to consider this stratification as a basis for improved screening policy.
Chapter 4

Model Building

The following chapter introduces two common modelling approaches in the presence of right censored survival data. We investigate the effect of available risk factors on breast cancer free survival times using univariate and multivariate survival modelling and build a breast cancer risk prediction model for the SMP BC population.

4.1 Survival Models

For the modelling of breast cancer free survival times, we consider the non-parametric additive regression model introduced by Aalen (1980) as well as the semi-parametric Proportional Hazards Model (Cox, 1972). The following notation is consistent with Martinussen and Scheike (2006).

4.1.1 Cox’s Proportional Hazards Model

Introduced by D.R. Cox (1972), the proportional hazards model assumes that the hazard function \( \lambda(t) \) can be described as a multiplicative relationship between a time-dependent baseline hazard \( \lambda_0(t) \) and a log-linear relative risk component \( r \). It is given by

\[
\lambda_i(t) = \lambda_0(t) \exp(X_i^T \beta)
\]

where \( X_i = (X_{i1}, \ldots, X_{ip})^T \) denotes a p-dimensional covariate vector and \( \beta \) represents the parameter vector. The baseline hazard \( \lambda_0 \) is assumed to be independent of the covariates \( X_i \) and, hence, identical for all individuals. The model parameter \( \beta_k \) represents the increase in log hazard rates associated with a unit increase of \( X_{ik} \) when all other variables are held constant. The exponent \( \exp(\beta_k) \) can be interpreted as an increase in risk relative to an individual with baseline exposure level of the risk factor \( X_{ik} \).
4.1. Survival Models

A key assumption of the Cox model is that hazard ratios are constant over time \( t \). Consider two individuals \( i \) and \( j \) with covariate vectors \( \mathbf{X}_i \) and \( \mathbf{X}_j \), respectively. The hazard for individual \( i \) relative to \( j \) is

\[
\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) \exp(\mathbf{X}_i^T \beta)}{\lambda_0(t) \exp(\mathbf{X}_j^T \beta)} = \exp(\mathbf{X}_i^T \beta) = \exp(\mathbf{X}_j^T \beta)
\]

and is independent of time \( t \).

The parameter vector \( \beta \) is estimated without specification of \( \lambda_0(t) \) by the method of partial likelihood (Cox, 1972). Using a counting process interpretation of the hazard function \( \lambda_i(t) \), the Cox model can be restated as

\[
\lambda_i(t) = Y_i(t) \lambda_0(t) \exp(\mathbf{X}_i^T \beta), \quad (4.3)
\]

where the risk indicator \( Y_i(t) \) ensures that \( \lambda_i(t) = 0 \) when the individual is not at risk. For untied event times \( t_{(k)} \), the partial likelihood function is defined as

\[
PL(\beta) = \prod_t \prod_{i=1}^n \left( \frac{Y_i(t) \exp(\mathbf{X}_i^T \beta)}{\sum_{j=1}^n Y_j(t) \exp(\mathbf{X}_j^T \beta)} \right)^{dN_i(t)} \quad (4.4)
\]

where \( dN_i(t) \) denotes the \( i \)th individual’s change in event status within the infinitesimal time interval \([t, t + dt]\).

Intuitively, the partial likelihood function can be motivated as a product of conditional probabilities at each event time. Assume individual \( k \) experienced the event of interest at a given event time \( t_{(k)} \). Then, the partial likelihood represents the probability that, knowing an event has occurred at time \( t_k \) among all individuals at risk, it was in fact experienced by individual \( k \). The denominator of Equation (4.4) only considers individuals at risk before time \( t_{(k)} \) and therefore excludes all previously censored observations.

We estimate \( \beta \) through maximization of the partial likelihood function. Differentiating the log partial likelihood with respect to \( \beta \) gives the score vector

\[
U(\beta) = \sum_{i=1}^n \int_0^T (\mathbf{X}_i - \bar{x}(\beta, s)) dN_i(s) \quad (4.5)
\]
4.1. Survival Models

where \( \bar{x}(\beta, s) \) is the weighted mean of \( X \) for individuals still at risk at time \( s \),

\[
\bar{x}(\beta, s) = \frac{\sum_i Y_i(s) \exp(X_i^T \beta)x_i}{\sum_i Y_i(s) \exp(X_i^T \beta)}
\]  \hspace{1cm} (4.6)

The solution \( \hat{\beta} \) of the equation

\[
U(\hat{\beta}) = 0
\]

is a consistent estimator of \( \beta \) and is asymptotically normal with mean \( \beta \) and variance \( E(I(\beta))^{-1} \), the inverse of the expected information matrix (Therneau and Grambsch, 2000). For tests of hypotheses about \( \beta \), \( H_0 : \beta = \beta^{(0)} \), the Wald statistic is given by

\[
T = (\hat{\beta} - \beta^{(0)})^T \hat{I}(\hat{\beta} - \beta^{(0)})
\]  \hspace{1cm} (4.7)

where \( \hat{I} \) is the the observed information matrix at the solution \( \hat{\beta} \). Under the null hypothesis, \( T \) follows a \( \chi^2 \) distribution with \( p \) degrees of freedom.

The \texttt{R} package \texttt{survival} (R Core Team, 2012) offers the Cox model fitting algorithm \texttt{coxph()} and is used to estimate model parameters \( \beta \) and conduct tests of hypotheses.

4.1.2 Aalen’s Additive Model

A linear regression approach in modelling of the hazard function \( \lambda(t) \) is taken by Aalen (1980). Covariates are assumed to act additively on the baseline hazard \( \lambda_0(t) \), giving the functional form

\[
\lambda_i(t) = \lambda_0(t) + X_i^T \beta(t)
\]  \hspace{1cm} (4.8)

where \( \beta(t) \) is a \( p \)-dimensional, locally integrable parameter vector. Unlike in the previously introduced Cox model, the model coefficients are no longer assumed to remain constant over time but may vary with \( t \). Hence, no functional form is imposed, making Aalen’s model a non-parametric method.

We define a time-dependent \( n \times p \) design matrix,

\[
X(t) = (Y_1(t)X_1, \ldots, Y_n(t)X_n)^T.
\]

The rows of \( X(t) \) contain the covariate vectors \( X_i, i = 1 \ldots n \) if individual \( i \) is at risk at time
4.1. Survival Models

$t$ and a null vector $(0,\ldots,0)^T$ otherwise. For reasons of simplicity and faster convergence rates, model parameters $\beta$ are estimated in cumulative form

$$B(t) = \int_0^t \beta(s)ds$$

(4.9)

through linear regression models for each increment $dB(t)$, leading to estimating equations

$$d\hat{B}(t) = X^-(t)dN(t)$$

(4.10)

where $X^-$ is a generalized inverse of $X$

$$X^- = (X^TX)^{-1}X^T.$$  

(4.11)

In integral form, we obtain the estimator of cumulative effects

$$\hat{B}(t) = \int_0^t X^-(s)dN(s).$$

(4.12)

Martinussen and Scheike (2006) show that $\sqrt{n}(\hat{B}(t) - B(t))$ converges to a mean zero Gaussian process.

To test the null hypothesis of no effect on survival times for covariate $j$, $H_0: B_j(t) = 0, \forall t \in [0,\tau]$,

we consider the maximum deviation from zero in the time interval $[0,\tau]$,

$$T_1 = \sup_{t \in [0,\tau]} |\hat{B}_j(t)|.$$  

(4.13)

The percentiles of the test statistic $T_1$ are estimated as follows. If model (4.8) holds, the asymptotic distribution of $\sqrt{n}(\hat{B}(t) - B(t))$ may be approximated by

$$n^{1/2} \sum_{i=1}^n \int_0^t (X^TX)^{-1}X_i dM_i(s)$$

(4.14)

where $M_i(s)$ is a mean zero Gaussian martingale, i.e. a stochastic process with $E[dM_i(s)] =$
4.2 Modelling Population

0. The percentiles of the process in Equation (4.14) can then be estimated by replacing \( dM_i(s) \) with \( G_i dN_i(s) \) and simulating \( N \) realizations of \( G_i \) where \( G_i \sim N(0,1) \).

Further, effect constancy of \( \beta_j(t) \) over time given by the null hypothesis

\[
H_0 : B_j(t) = B_j t
\]

is assessed using the maximum observed deviation between the estimated parameter \( \hat{B}_j(t) \) and the expected, linear trend under the null hypothesis, \( \hat{B}_j(\tau) \tau / \tau \),

\[
T_2 = \sup_{t \in [0, \tau]} |\hat{B}_j(t) - \hat{B}_j(\tau) \frac{t}{\tau}| \tag{4.15}
\]

Again, the simulation approach described above is used to estimate the percentiles of the test statistic (Martinussen and Scheike, 2006).

Aalen’s additive model provides a useful supplement to the Cox model in applications where effect constancy is questionable or needs initial investigation. The model’s parameters can be visualized as a function of time and provide a useful tool in assessment of a covariate’s effect over time. However, the additional flexibility of time-varying effects comes at the cost of statistical power, potentially allowing smaller variable effects to go undetected by tests of hypotheses. If the proportional hazards assumption holds for a particular set of covariates, the Cox model is also preferable in terms of computational time.

In the following analysis, the function `aalen()` available from the R package `timereg` (R Core Team, 2012) was used to estimate model coefficients and conduct tests of hypothesis.

4.2 Modelling Population

4.2.1 Estimation of Time to Diagnosis

As described in Section 2.4, the follow-up mechanism of the SMP BC database introduces informative censoring into the sample. A woman’s censoring probability cannot be considered independent from her disease status as women initially lost to follow-up find their way back into the database once diagnosed with breast cancer. It follows that

\[
P(C_i < 5 | T_i < 5) < P(C_i < 5 | T_i > 5), \tag{4.16}
\]
implying a dependency between censoring time $C_i$ and event time $T_i$. However, justification
of the conditional probability argument of Cox’s partial likelihood Equation (4.4) requires
random censoring.

To obtain valid results, the modelling population is adjusted by exclusion of all women
with cancers diagnosed outside the SMP BC screening program. It will be assumed that
cancer is detected through SMP BC screening if the cancer diagnosis follows within a pre-
defined time period after the last recorded screening date. What is needed is an estimate of
this time difference.

For 855 women with prevalent cancer found at the 2000 screening appointment, the data
set provided both screening and diagnosis dates, allowing us to investigate the time between
cancer detection and final diagnosis. The results are presented in the right hand side graph
of Figure 4.1. It can be seen that most cancers are diagnosed within 50 days of the screening
mammogram, however, the histogram is skewed to the right with some outlying observations
diagnosed after more than 200 days.

The left hand side graph shows the time differences between the last recorded screening
date and cancer diagnosis for all non-prevalent cases in the data set of 221,545 women. Again,
the histogram is highly skewed with a long right tail. The local minimum seen around 200
days appears to be a good threshold value, especially as it is consistent with the maximum
diagnosis time observed in 2000. However, since only women with a negative result at the
last screening (i.e. no cancer was detected) are subject to exclusion, this interval seems too
short. Time for cancer development, detection and diagnosis needs to be taken into account.

Finally, we decided to exclude all cancers found after one year of the last recorded
screening date. This allowed us to reduce the presence of informative censoring, while
assuring that enough cases remain in the data set to maintain power of hypothesis testing
(Therneau and Grambsch, 2000).

4.2.2 Exclusion of Individuals

The sample used for the model building process was restricted to individuals with com-
plete information on all the risk factors, resulting in 101,790 of 221,545 women to be excluded
from the analysis. As cancer rates between women with complete and incomplete variable
information differed by 0.2%, missing information was assumed to be independent of the
survival status. Further, 8757 individuals without follow-up, i.e. whose 2000 screening date is also the last recorded SMP BC screening date, were excluded. To fit the survival framework, women entering the sample are also required to be without breast cancer at the beginning of the follow-up period. Therefore, 855 women with prevalent cancers detected at the 2000 screening appointment were removed from the data set. Based on the time intervals described above, we estimated that 847 invasive and 31 in situ cancers were detected outside the SMP BC program. These cases were excluded as well.

Detected cancers were stratified by type to allow for separate risk modelling and comparisons of risk factor effects on survival status. Overall, the two resulting data sets obtained for the model building process include 108,921 undiagnosed individuals and 942 and 311 women diagnosed with invasive and in situ cancers, respectively.

For non-cases with recorded SMP BC screening dates later than 2005, the censoring time \( \tilde{T}_i \) was defined as the time between the 2000 screening mammogram and December 31, 2005. For non-cases not returning for SMP BC screening, the censoring time is marked as the last recorded SMP BC screening appointment. The time to event for all women diagnosed with

\[ \text{Figure 4.1:} \text{ Left: Histogram of time between last recorded SMP BC screening date and cancer diagnosis of 4078 women. Right: Histogram of time between the 2000 screening appointment and diagnosis date for all prevalent cancers.} \]
breast cancer is defined as the time between the 2000 screening date and the cancer diagnosis date. There were no tied event times.

4.3 Analysis of Family History

Residual analysis in Section 3.5 showed the inadequacy of a woman’s family history use in the current Gail model. Ordinal modelling of family history and stratification based on the number of affected relatives alone resulted in systematic patterns within the model’s residuals and led to inflated cancer risks. To improve on these issues, associations between the cancer outcomes of first degree relatives are examined in the following.

The analysis is based on 108,921 non-cases and 942 cases diagnosed with invasive cancer as defined in Section 4.2.2. The variable $fam_{hist}$ was converted into four binary variables indicating the presence of an affected mother ($mother$), sister ($sister$), daughter ($daughter$) and father or brother ($father$). Overall, 7.99% of women had a mother diagnosed with breast cancer, 4.61% a sister and only 0.24% an affected daughter. None of the 24 women with an affected father or brother were diagnosed with breast cancer in the five-year follow-up period. Thus, variable $father$ will not be considered in the following analysis.

Figure 4.2 shows the age-specific relative frequency of affected relatives stratified by type. With the overall lowest values, relative frequencies of affected daughters within the SMP BC sample show a gradual increase with age. A similar but steeper incline can be seen for frequencies of affected sisters. Both trends are easily explained by the rising incidence associated with increasing age. However, unlike the trends for daughters and sisters, the presence of an affected mother becomes less likely with increasing age. In this case, the effect of age appears to be overwhelmed by introduction of early detection programs, leading to the increased detection of cancer in recent years.

Estimates of the effect of relatives diagnosed with breast cancer on survival times were obtained by fitting a Cox proportional hazards model containing three indicator variables derived from the variable $fam_{hist}$. Due to the strong, age related patterns seen in Figure 4.2, the woman’s age was also included. The linear predictor of the model is given by

$$r_i = \exp(\beta_1 mother + \beta_2 sister + \beta_3 daughter + \beta_4 age)$$

(4.17)
4.3. Analysis of Family History

where the parameters $\beta_i, i = 1, 2, 3, 4$, represent the increase in hazard ratios associated with the four explanatory variables.

Due to the limited number of cases with daughters diagnosed with breast cancer, five overall, the corresponding model effect was not significant and the variable daughter was excluded from the model. For the remaining variables, Table 4.1 shows the parameter estimates $\hat{\beta}_k$, relative risks, standard errors and p-values of the hypothesis test $H_0 : \beta_k = 0$. Both mothers and sisters diagnosed with breast cancer have a significant effect on the survival time with corresponding p-values $< 0.0001$. An affected sister leads to the overall highest

Table 4.1: Parameter estimates, relative hazards, standard errors, Wald statistics and p-values for family history model.

<table>
<thead>
<tr>
<th>variable</th>
<th>coef</th>
<th>exp(coef)</th>
<th>s.e.(coef)</th>
<th>z</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0.034</td>
<td>1.035</td>
<td>0.003</td>
<td>10.347</td>
<td>0.000</td>
</tr>
<tr>
<td>mother:yes</td>
<td>0.509</td>
<td>1.664</td>
<td>0.098</td>
<td>5.170</td>
<td>0.000</td>
</tr>
<tr>
<td>sister:yes</td>
<td>0.680</td>
<td>1.973</td>
<td>0.110</td>
<td>6.193</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Figure 4.2: Relative frequency of affected first degree relatives by type of relative, as a function of five-year age categories.
increase in hazard rates by a factor of 1.973, while an affected mother increases the hazard rate by a factor of 1.664.

Based on these results, we see that different types of relatives have different effects on survival times, making it important to differentiate not only by absolute numbers but also by the relationship. Ordinal representation also seems inappropriate. If the same model is fitted with stratification by numbers only,

\[ \lambda(t) = \lambda_0(t) \exp(\beta \text{num}_\text{rel}) \] (4.18)

the increase in hazard ratios associated with the variable \text{num}_\text{rel} is estimated as \( \exp(\beta) = 1.73 \). Comparing models (4.17) and (4.18), the relative risk for a woman with an affected mother, daughter and sister increases by a factor 1.664 \cdot 1.973 = 3.28 and 1.73^3 = 5.18, respectively, a considerable difference given the overall magnitude of the parameters.

A particular problem with the collected family history information is that we have no information whether the participants have sisters or daughters in their family. The category ‘O’ indicating that the woman is the only child in the family seems to have been used in an inconsistent fashion with 51 entries in the original data set and therefore, cannot be considered a reliable indicator. While information on the breast cancer status of the women’s mother is reliable in this respect, sisters and daughters are always assumed to be free of cancer without knowing about their existence. Without adjustment, this leads to biased probability estimates in the risk model, as baseline exposure levels are always assumed for each of the unknown cases. For proper use of family history information, models need to take this into consideration, and more detailed data should to be collected by the SMP BC.

4.4 Modelling of Invasive Cancers

We consider 17 risk factors given in Table 2.1 as the candidate covariates in the modelling of invasive breast cancer for 109,863 SMP BC participants. As a result of Section 4.3, the variable \text{fam_hist} is converted into three binary variables indicating the presence of an affected mother, sister and daughter. To reduce categories and hence, parameters to be estimated for variable \text{ethnic}, six European ethnicities were combined into a new category ‘Caucasian’. An additional explanatory variable is derived converting the year 2000 screening
result into a binary indicator describing the presence of an abnormal result. As prevalent cancers were removed from the data set, the detected abnormalities for the remaining women were found to be non-cancerous in further examinations following the abnormal result. For individuals whose response to variables *hysterec*, *both_ovrem*, *fam_hist50* is missing due to non-applicability, the corresponding covariates were recoded to baseline exposure levels.

### 4.4.1 Single Covariate Modelling

In an initial univariate analysis, we fitted proportional hazards and Aalen models for each of the available explanatory variables and investigated their effects on breast cancer free survival time. For the Cox model, the significance of variable effects was established using likelihood ratio tests, for Aalen’s model using the supremum test described earlier. The model coefficients $\beta$ and p-values of the hypothesis test $H_0 : \beta = 0$ are given in Table 4.2. Parameter values $\hat{\beta}$ for categorical variables with more than two levels are not listed.

For a single covariate Cox model, significant effects on survival times with $p < 0.01$ were found for the presence of an abnormal mammogram, age, age at last menstrual cycle, current estrogen use, ethnicity, previous breast biopsies, a hysterectomy, menopausal status and an affected mother and sister. Mammographic density was marginally significant at $p = 0.041$. For education indicator *educ* ($p = 0.787$) and reproductive history variables *age_men* ($p = 0.732$), *age_1stdel* ($p = 0.374$), *num_del* ($p = 0.357$) and *ever_preg* ($p = 0.343$) the null hypothesis of $H_0 : \beta = 0$ could not be rejected.

As seen in Section 2.3.4, a woman’s response on variables *curr estr*, *hysterec*, *menstr* is highly dependent on her age. As a result, the significant effects corresponding with these covariates in the Cox model could not only been caused by the variables themselves but their association with age, one of the strongest risk factors for breast cancer. A simultaneous covariate model containing these variables as well as age will show whether this is the case.

Although only partly comparable due to different null hypotheses, the univariate results for Aalen’s additive model were overall very similar. Again, a woman’s reproductive history described by the variables *age_men* ($p = 0.792$), *ever_preg* ($p = 0.754$), *age_1stdel* ($p = 0.719$) and *num_del* ($p = 0.34$) did not show a significant effect over time. Unlike in the Cox model, variables *mammo_dens* ($p = 0.138$) and *hysterec* ($p = 0.065$) were non-significant. The category “No education” of variable *educ* was significant at $p = 0.02$. Recalling results
Table 4.2: Coefficients and p-values of univariate Cox and Aalen models for hypothesis tests of significant and time-varying effects. (* refers to all variables levels, ** refers to at least one level)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cox Model</th>
<th></th>
<th>Aalen Model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}$</td>
<td>$p (\beta = 0)$</td>
<td>$p (\beta(t) = 0)$</td>
<td>$p (\beta(t) = \beta)$</td>
</tr>
<tr>
<td>abmam</td>
<td>0.342</td>
<td>&lt; 0.01</td>
<td>&lt; 0.001</td>
<td>0.014</td>
</tr>
<tr>
<td>age</td>
<td>0.036</td>
<td>&lt; 0.01</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>age_lastmenst</td>
<td>0.010</td>
<td>&lt; 0.01</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>age_men</td>
<td>-0.007</td>
<td>0.732</td>
<td>0.792</td>
<td>0.924</td>
</tr>
<tr>
<td>age_1stdel</td>
<td>0.003</td>
<td>0.372</td>
<td>0.719</td>
<td>0.450</td>
</tr>
<tr>
<td>curr_estr</td>
<td>0.226</td>
<td>&lt; 0.01</td>
<td>&lt; 0.001</td>
<td>0.408</td>
</tr>
<tr>
<td>educ</td>
<td>-</td>
<td>0.787</td>
<td>&lt; 0.025**</td>
<td>&gt; 0.141*</td>
</tr>
<tr>
<td>ethnic</td>
<td>-</td>
<td>&lt; 0.01</td>
<td>&lt; 0.001**</td>
<td>&gt; 0.198*</td>
</tr>
<tr>
<td>ever_biop</td>
<td>0.469</td>
<td>&lt; 0.01</td>
<td>&lt; 0.001</td>
<td>0.887</td>
</tr>
<tr>
<td>ever_preg</td>
<td>0.081</td>
<td>0.343</td>
<td>0.754</td>
<td>0.497</td>
</tr>
<tr>
<td>hysterec</td>
<td>0.050</td>
<td>&lt; 0.01</td>
<td>0.065</td>
<td>0.768</td>
</tr>
<tr>
<td>mammo_dens</td>
<td>0.046</td>
<td>0.041</td>
<td>0.139</td>
<td>0.376</td>
</tr>
<tr>
<td>menstr</td>
<td>-0.436</td>
<td>&lt; 0.01</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>mother</td>
<td>0.504</td>
<td>&lt; 0.01</td>
<td>&lt; 0.001</td>
<td>0.233</td>
</tr>
<tr>
<td>num_del</td>
<td>0.021</td>
<td>0.357</td>
<td>0.340</td>
<td>0.746</td>
</tr>
<tr>
<td>sister</td>
<td>0.876</td>
<td>&lt; 0.01</td>
<td>&lt; 0.001</td>
<td>0.344</td>
</tr>
</tbody>
</table>

from Section 2.3.4, it is most likely caused by the strong association between a woman’s age and educational level. The remaining variables were again important predictors of survival with highly significant p-values < 0.001.

Tests about effect constancy over time revealed that the hypothesis of proportionality could be rejected for age ($p < 0.001$), the presence of an abnormal mammogram ($p = 0.014$), menopausal status ($p < 0.001$) and the age at last menarche ($p < 0.001$). For these variables, plots of the estimated cumulative regression parameter $\hat{B}_j(t)$ and their 95% confidence bands are presented in Figure 4.3. Constancy of the effect $\beta_j(t)$ over time corresponds to a linear trend of the function $B_j(t)$.

For the cumulative effect of age, it appears as if both time periods before and after about two years after the begin of the study follow an approximately linear pattern with the same slope, however, there is a steep increase in effect size shortly after the two years mark.
4.4. Modelling of Invasive Cancers

This short-lived pattern change is doubtfully caused by age alone, but by screening recall recommendations leading to an increased detection of cancer in this time period. A similar abrupt increase or decrease shortly after year two can be seen in the cumulative effects of \textit{menstr} and \textit{age\_lastmenstr}. This leads to the conclusion that their corresponding effects may not necessarily violate proportionality, but are affected by cancer detection clustered in time after the first screening recall. Overall, the importance of exact timelines until disease occurrence is clearly shown, a property that is not fulfilled due to the nature of the cancer detection process.

For an abnormal mammogram, however, the time-varying effect on survival times appears to be justified. A steep increase in effect size for approximately one and a half years is followed by a nearly constant progression and further but less steep increases. Hence, the
effect of an abnormal mammogram at the first screening decreases with time, indicating that the risk of abnormalities progressing into invasive cancers is highest in the 1.5 years following their detection, after this period this chance decreases. This difference could be caused by the type of cancer and whether it is slow growing or aggressive.

4.4.2 Simultaneous Modelling of Covariates

Due to the particularly high p-values seen for univariate models, the variables $age_{\text{men}}$, $ever_{\text{preg}}$, and $age_{\text{1stdel}}$ will not be considered for a multivariate model of survival times. Starting with an initial proportional hazards model including all remaining covariates, the model is reduced in a backward elimination fashion by exclusion of covariates with the highest p-values until only significant variables remain in the model. Variables were excluded in the following order: $age_{\text{lastmenstr}}$ ($p = 0.97$), $menstr$ ($p = 0.62$), $educ$ ($p = 0.59$), $num_{\text{del}}$ ($p = 0.18$), $curr_{\text{estr}}$ ($p = 0.11$) and $hysterec$ ($p = 0.07$). Once these variables were removed and a final model determined, all excluded variables were added individually once again, however, a significant effect could not be found.

To confirm the exclusions, the backward selection function `selectCox()` from the R package `pec` was applied to the initial model. The selection method is based on Akaike’s Information Criterion defined as

$$AIC = 2k - 2\log(L),$$

where $k$ represents the number of model parameters and $L$ the maximized value of the likelihood. Models with smaller values of AIC are preferred. At each step, the function compares the AIC of the proposed model with the AIC of all models resulting from the exclusion of one covariate. The model with the smallest AIC value is selected as the new candidate model until no reduction in AIC can be achieved through exclusion of covariates. The selection algorithm resulted in the same subset of predictor variables as described above.

When added to the final model, interactions between $ever_{\text{biop}}$ and $age$ ($p = 0.47$) and $num_{\text{rel}}$ and $age_{\text{1stdel}}$ ($p = 0.99$) as included in the Gail model did not show a significant effect. An interaction between $mammo_{\text{dens}}$ and $ethnic$ could also be declined with $p = 0.56$. 

56
4.4. Modelling of Invasive Cancers

Table 4.3: Regression coefficients, standard errors, Wald statistics and corresponding p-values under proportional hazards model.

<table>
<thead>
<tr>
<th>variable</th>
<th>coef</th>
<th>exp(coef)</th>
<th>s.e.(coef)</th>
<th>z</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0.036</td>
<td>1.036</td>
<td>0.003</td>
<td>10.336</td>
<td>0.000</td>
</tr>
<tr>
<td>ethnic:Aboriginal</td>
<td>0.401</td>
<td>1.493</td>
<td>0.261</td>
<td>1.535</td>
<td>0.125</td>
</tr>
<tr>
<td>ethnic:Black</td>
<td>-0.218</td>
<td>0.804</td>
<td>0.708</td>
<td>-0.308</td>
<td>0.758</td>
</tr>
<tr>
<td>ethnic:EastAsian</td>
<td>-0.297</td>
<td>0.743</td>
<td>0.111</td>
<td>-2.681</td>
<td>0.007</td>
</tr>
<tr>
<td>ethnic:Other</td>
<td>0.065</td>
<td>1.067</td>
<td>0.129</td>
<td>0.503</td>
<td>0.615</td>
</tr>
<tr>
<td>ethnic:SouthAsian</td>
<td>-0.211</td>
<td>0.810</td>
<td>0.212</td>
<td>-0.994</td>
<td>0.320</td>
</tr>
<tr>
<td>abmam:abnormal</td>
<td>0.371</td>
<td>1.449</td>
<td>0.115</td>
<td>3.221</td>
<td>0.001</td>
</tr>
<tr>
<td>mammo_dens:2</td>
<td>0.297</td>
<td>1.346</td>
<td>0.076</td>
<td>3.924</td>
<td>0.000</td>
</tr>
<tr>
<td>mother:yes</td>
<td>0.474</td>
<td>1.607</td>
<td>0.099</td>
<td>4.790</td>
<td>0.000</td>
</tr>
<tr>
<td>sister:yes</td>
<td>0.653</td>
<td>1.922</td>
<td>0.110</td>
<td>5.944</td>
<td>0.000</td>
</tr>
<tr>
<td>ever_biop=yes</td>
<td>0.283</td>
<td>1.327</td>
<td>0.082</td>
<td>3.448</td>
<td>0.001</td>
</tr>
</tbody>
</table>

For all variables found to have a significant effect on survival times, the coefficients, standard errors and p-values for the test of $H_0 : \beta = 0$ are displayed in Table 4.3. As previously suspected, the associations between survival status and the variables `curr_estr`, `hysterec`, `menstr` and `educ` were introduced by a strong linkage with age. In a multivariate model including age as a main effect, these variables no longer significantly affect survival times.

The variable age resulted in a highly significant model parameter of $\beta = 0.036$, suggesting that breast cancer hazards increase by a factor of 1.036 for each additional year of life. We notice non-significant model parameters for all ethnic categories with the exception of $\beta = -0.297$ for variable level “East Asian”. Breast cancer hazards for East Asian women are therefore significantly reduced by a factor of 0.743, while there is no significant difference in hazards between the baseline of Caucasian women and the remaining ethnicities. A parameter estimate of $\beta = 0.06$ for the variable level “Other” is not surprising as a result of the ethnic composition of this category. Ethnicities include various Caucasian, Asian and African backgrounds not available as answers on the SMP BC questionnaire and show no consistency. The category “Black” consists of only 287 women and thus shows comparably large standard errors for it’s effect on survival times.

Medical history also plays an important role in the prediction of survival times. An
abnormal mammogram and previous breast biopsies show highly significant coefficients of 0.371 and 0.283, respectively, leading to a rise in hazard by factors 1.449 and 1.327 relative to baseline exposure levels.

Mammographic density, a variable not included in the current Gail model, is found to be a significant predictor in survival times. Women with predominately dense breasts are subject to increased breast cancer hazard by a factor of 1.346. Even with the limitation of the binary density classification provided in the data set, breast density appears to be an important risk factor and could help in the stratification of cancer outcomes.

The highest increases in hazard ratios are found for women with a family history of breast cancer. For individuals whose mother is affected by breast cancer, hazards increase by a factor 1.607 relative to baseline levels. An even stronger effect can be seen for an affected sister where cancer hazards are 1.922 times higher compared to women without an affected sister. The difference in effects, a factor of 1.607 versus 1.922, confirms the choice of categorical inclusion of the variable `fam_hist`, stratified by relative type.

It seems reasonable to combine ethnic categories without significant model parameters into category “Other” to reduce on the number of estimated parameters. Although statistical testing of this reduced model’s adequacy through the method of likelihood ratio is not possible, the maximized log likelihood decreased only slightly by a value of 0.44 from $L = -15079.3$ to $L = -15079.74$ when all women with African heritage were assigned to the category “OT”.

### 4.4.3 Testing Proportional Hazards Assumptions

Hypothesis tests of effect constancy are conducted using a residual definition proposed by Schoenfeld (1980). Let $0 < t_{(1)} < t_{(2)} < \ldots$ denote ordered event times and

$$
V(\beta, t) = \frac{\sum_i Y_i(t) r_i [X_i - \bar{x}(\beta, t)]^T [X_i - \bar{x}(\beta, t)]}{\sum_i Y_i(t) r_i} \quad (4.20)
$$

a weighted variance of the covariate matrix $X$ at time $t$. The Schoenfeld residual at event time $t_{(k)}$ is defined as

$$
s_k = X_{(k)} - \bar{x}(\hat{\beta}, t_{(k)}) \quad (4.21)
$$
4.4. Modelling of Invasive Cancers

where $X_{(k)}$ is the covariate vector of the individual experiencing the $k$th event and

$$\bar{x}(\beta, t_{(k)}) = \frac{\sum_i Y_i(t_{(k)})r_i X_i}{\sum_i Y_i(t_{(k)})r_i}$$

(4.22)

is the weighted mean of $X$ over those individuals still at risk at time $t_{(k)}$.

Grambsch and Therneau (1994) showed that the time-varying effect $\beta_j(t)$ for variable $j$ can be approximated by the sum of the expected scaled Schoenfeld residual $s_{kj}^* = V^{-1}(\hat{\beta}, t_{(k)}) s_{kj}$ and the estimated parameter value under the Cox model

$$\beta_j(t) \approx E(s_{kj}^*) + \hat{\beta}_j.$$  (4.23)

The hypothesis test of a constant model parameter, $H_0 : \beta_j(t) = \beta_j$, is motivated by the regression approach

$$\beta_j(t) = \beta_j + \theta_j(t - \bar{t})$$

and translated into an equivalent hypothesis $H_0 : \theta_j = 0$ about the regression parameter $\theta_j$. Using the approximation given in Equation (4.23) and the fact that $\hat{\beta}_j$ is an estimator of $\beta_j$, it follows that

$$E(s_{kj}^*) \approx \theta_j(t - \bar{t})$$

(4.24)

Estimation of $\theta$ and it’s variance $Q$ result in the test statistic

$$T = \hat{\theta}^T Q^{-1} \hat{\theta},$$

(4.25)

following a $\chi^2$ distribution with $p$ degrees of freedom under the null hypothesis $\beta(t) = \beta$. Insights into the nature of deviations from the proportional hazards assumption can also be gained from a scatterplot of the scaled Schoenfeld residuals as a function of event times $t_{(k)}$.

For the Cox model as given in Table 4.3, tests of proportional hazards showed a significant p-value of 0.002 associated with the presence of an abnormal mammogram. The null hypothesis of constant effects over time, $H_0 : \beta(t) = \beta$, can be rejected. For the remaining variables, the null hypothesis could not be rejected with p-values $>0.299$. In addition, the global test was not significant with a p-value of 0.211.

Plots of the scaled Schoenfeld residuals for variables abmam and age, the latter sus-
4.4. Modelling of Invasive Cancers

Expected to be in violation of constant effects in the univariate Aalen model, are presented in Figure 4.4. To ease visualization of the time-varying progression of the expected Schoenfeld residuals $E(s_{kj}^*)$, a locally weighted smoothing spline was added to the plot and can be seen by the blue line. If the proportional hazards assumption holds, it is an approximately horizontal line.

The effect for age shows a wavy pattern without any strong increase or decreases in slope. Interestingly, small increases in effect size can be found shortly before two and four years, the screening recall recommendations of the SMP BC. As seen from the vertical clustering of points within these time periods, cancer detection rates are elevated, leading to the conclusion that the variations are not necessarily caused by a time-varying effect of age, but rather by screening patterns leading to comparably higher cancer detection. However, a slight, underlying trend over time can also be detected, a potential indicator of a wrongly specified functional form of age.

Unlike for variable age, the effect associated with an abnormal mammogram at the first screening does show a dominant downward trend, but again briefly interrupted around the two year time point. It is clear that proportionality cannot be assumed in this case, as
Table 4.4: Parameters estimates, standard errors, Wald statistics and p-values for the stratified Cox model.

<table>
<thead>
<tr>
<th>variable</th>
<th>coef</th>
<th>exp(coef)</th>
<th>s.e.(coef)</th>
<th>z</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0.036</td>
<td>1.036</td>
<td>0.003</td>
<td>10.377</td>
<td>0.000</td>
</tr>
<tr>
<td>ethnic:EastAsian</td>
<td>-0.252</td>
<td>0.777</td>
<td>0.105</td>
<td>-2.393</td>
<td>0.017</td>
</tr>
<tr>
<td>ethnic:Other</td>
<td>-0.044</td>
<td>0.957</td>
<td>0.123</td>
<td>-0.354</td>
<td>0.723</td>
</tr>
<tr>
<td>ethnic:SouthAsian</td>
<td>0.052</td>
<td>1.053</td>
<td>0.181</td>
<td>0.288</td>
<td>0.773</td>
</tr>
<tr>
<td>mammom_dens:2</td>
<td>0.262</td>
<td>1.300</td>
<td>0.074</td>
<td>3.523</td>
<td>0.000</td>
</tr>
<tr>
<td>mother:yes</td>
<td>0.491</td>
<td>1.633</td>
<td>0.099</td>
<td>4.974</td>
<td>0.000</td>
</tr>
<tr>
<td>sister:yes</td>
<td>0.655</td>
<td>1.926</td>
<td>0.110</td>
<td>5.966</td>
<td>0.000</td>
</tr>
<tr>
<td>ever_biop:yes</td>
<td>0.293</td>
<td>1.341</td>
<td>0.082</td>
<td>3.584</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Hazard ratios decrease with increasing time $t$. Adjustments to the original Cox model to allow for a time-varying effect need to be made.

As $abmam$ is the only variable found to violate proportional hazards assumptions, we decided to replace the initial model with a stratified Cox model. The categorical nature of $abmam$ provides a natural stratification by variable levels ‘$\geq 50\%$’ and ‘$< 50\%$’. As a result, an abnormal mammogram is not a part of the linear predictor $\exp(X_i^T\beta)$ but will be represented by a second baseline hazard $\lambda_0(t)$ applicable to individuals with an abnormal mammogram only (Therneau and Grambsch, 2000).

The coefficients, standard errors and p-values of the stratified Cox model are presented in Table 4.4. To reduce non-significant variable levels of $ethnic$, categories ‘Black’ and ‘Aboriginal’ were combined with the category ‘Other’. Only minor changes at the second decimal place of the model’s coefficients could be seen after these adjustments were made. The coefficient for South Asian ethnicities showed the highest change from -0.211 to 0.052 and is now closer to zero, an acceptable change considering that the effect was not significant in both models.

4.4.4 Functional Form

The adequacy of the linear predictor $r$ for the continuous variable $age$ needs to be confirmed. As described in Therneau and Grambsch (2000), a plot of the martingale residuals against the covariate gives insight whether the functional form is specified in the correct
The Martingale residual for individual $i$ is defined as

$$M_i(t) = N_i(t) - E_i(t)$$

where $E_i(t) = Y_i(t) \Lambda_0(t) r(X_i, \beta)$ is the cumulative hazard at follow-up time $t$. A plot of the martingale residuals from the fitted Cox model versus age is shown in Figure 4.5. As dense horizontal bands complicate visualization, a locally weighted smoothing spline is added to assist in the pattern detection. The left hand side graph includes the residuals for all individuals while the right hand side graph focused on the main concentration of points in the lower half of the first graph, the residuals for all non-cases. Due to the large number of individuals displayed, the points in the scatterplot were displayed transparent rather than in solid black.

The smoothing spline is approximately horizontal with a slight wavy pattern around age 50, but there are no major deviations suggesting a misspecified model. The variable age appears to be included in the correct functional form.
4.5 Modelling of In Situ Cancers

In the following paragraph, we investigate risk factor effects on event free survival times for in situ breast cancers only and build a multivariate proportional hazards model with the objective to compare between variable effects for in situ and invasive cancers. With only 311 in situ cases remaining in the modelling set, model parameters will be subject to considerably more variability than in the previous analysis of invasive cancers, impacting the ability to detect effect sizes as small as in the initial model. For risk factor effect comparisons between cancer types, this needs to be kept in mind.

4.5.1 Single Covariate Modelling

Following the steps in the variable selection process of Section 4.4, we built univariate Cox and Aalen models for each available explanatory variable. For the proportional hazards model, non-significant effects in the time interval were found for age at menarche \((p = 0.95)\), ethnicity \((p = 0.6)\), previous pregnancies \((p = 0.08)\), the number of deliveries \((p = 0.17)\), age at first delivery \((p = 0.42)\), hysterectomy \((p = 0.40)\), current estrogen use \((p = 0.16)\) and the presence of an affected sister \((p = 0.38)\) or daughter \((p = 0.65)\). Age \((p = 0.001)\), an abnormal mammogram at the 2000 screening \((p < 0.001)\), mammography density \((p < 0.001)\) and an affected mother \((p < 0.001)\) were all highly significant. Education \((p = 0.04)\), menopausal status \((p = 0.03)\), age at last menarche \((p = 0.01)\) and a family member diagnosed before the age of 50 \((p = 0.004)\) were also significant, however, most likely for reasons of confounding as noted earlier.

For Aalen’s model, highly significant time varying effects with \(p < 0.001\) in the time interval \([0, 5]\) were found for age, but not for an abnormal mammogram as observed in the modelling of invasive cancers. The corresponding parameter estimates are displayed as functions over time in Figure 4.6. As seen from the first graph, the time-varying effect for age was again artificially created by biannual screening recalls. Excess risks increase steeply around the two and four year time points. Interestingly, there is virtually no increase in cumulative effect size between these points as lines remain nearly horizontal or are even slightly decreasing. In comparison to invasive cancers where a continuous increase in effect could be seen, this gives an indication about the non-aggressive nature of in situ cancers. These cancers
4.5. Modelling of In Situ Cancers

are predominately found through screening examinations and remain undetected in the time intervals between recall examinations.

The progression of effect size for an abnormal mammogram leads to similar conclusions. After a longer first period with no diagnoses and a short but steep incline, cumulative effects follow an approximately linear trend until year five. A decrease in slope cannot be seen, suggesting that the abnormalities found in the 2000 screen are detected as cancers at the same rate as in the initial years after screening. For invasive cancers, comparably more cancers were detected in this first one and a half years, indicating a faster and more aggressive growth.

4.5.2 Simultaneous Covariate Cox Model

Based on the high p-values found in the univariate analysis, variables \textit{age\_men}, \textit{ethnic}, \textit{age\_1stdel}, \textit{hysterec} and \textit{daughter} will not be considered as candidate variables in the modelling of in situ cancers. Although an affected sister showed an equally high p-value, \textit{sister} will be initially included with the remaining variables as it showed the highest risk increase for invasive cancers.

Subsequent elimination of the variables with the highest p-values for the hypothesis of non significant effects resulted in the following order of exclusion: number of deliveries
4.5. Modelling of In Situ Cancers

(p = 0.697), current estrogen use (p = 0.686), an affected sister (p = 0.62), education level (p = 0.601), a family member diagnosed before the age of 50 (p = 0.249), menopausal status (p = 0.104), age at last menstrual cycle (p = 0.324) and previous pregnancies (p = 0.082). The remaining variables age, an abnormal mammogram, mammographic density, an affected mother and previous breast biopsies were all highly significant with p-values < 0.001. All excluded variables were again added individually to the final model, however, no significant variables were found. The number of deliveries showed the lowest p-value with 0.08. The backward selection algorithm of `selectCox()` arrived at the same variables with the exception for previous biopsies.

4.5.3 Model Checking

Tests of time-varying effects using scaled Schoenfeld residuals resulted in non-significant p-values for all variables except age, where we found a marginally significant p-value of 0.046. A plot of the residuals over time can be seen in Figure 4.7.

For the first one and a half years, the effect of age is below 0 and shows a slightly
4.5. Modelling of In Situ Cancers

Table 4.5: Coefficients, standard errors, Wald statistics and p-values for multivariate Cox model of in situ cancers.

<table>
<thead>
<tr>
<th>variable</th>
<th>coef</th>
<th>exp(coef)</th>
<th>s.e.(coef)</th>
<th>z</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0.025</td>
<td>1.026</td>
<td>0.006</td>
<td>4.107</td>
<td>0.000</td>
</tr>
<tr>
<td>abmam:abnormal</td>
<td>1.048</td>
<td>2.852</td>
<td>0.153</td>
<td>6.828</td>
<td>0.000</td>
</tr>
<tr>
<td>mammo_dens:2</td>
<td>0.535</td>
<td>1.707</td>
<td>0.122</td>
<td>4.390</td>
<td>0.000</td>
</tr>
<tr>
<td>mother:yes</td>
<td>0.694</td>
<td>2.002</td>
<td>0.160</td>
<td>4.336</td>
<td>0.000</td>
</tr>
<tr>
<td>ever_biop:yes</td>
<td>0.497</td>
<td>1.644</td>
<td>0.137</td>
<td>3.643</td>
<td>0.000</td>
</tr>
</tbody>
</table>

increasing trend over time. This suggests that in the initial years after screening, older women could be less likely to be diagnosed with in situ breast cancer, most likely explained by slower tumour growth. Due to the low number of events seen in the first year, the trend of the smoothing spline is subject to higher variability.

Table 4.5 displays the coefficients for the final model. Due to the small number of cases, standard errors are considerably higher for all parameter estimates. A two sample z-test showed that, with the exception of the effect for an abnormal mammogram ($p < 0.001$), the parameters for variables included in both invasive and in situ models are not significantly different. As the effect of an abnormal mammogram was found to be time-varying for invasive cancers, parameter estimates are thus only comparable as averages over the time interval $[0, 5]$. Some observations will be made based on the change in point estimates.

The coefficient for mammography density, an affected mother and previous biopsies, each variable with a approximately constant effect for both models, increased and appears to lead to higher relative risks for in situ than for invasive cancers. The model coefficient for an abnormal mammogram increased from 0.371 to $\hat{\beta} = 1.048$ for in situ cancers. Most notably, an affected sister, the most important risk factor for invasive cancers with a nearly two fold increase in relative risk, did not have a significant effect on event-free survival time from in situ breast cancer. When sister is added to the final model as given in Table 4.5, the corresponding model parameter was estimated as $\hat{\beta} = 0.028$ and was not significantly different from 0 with $p = 0.91$.

The internal concordance statistic is higher for in situ cancers with $c = 0.649$ versus $c = 0.633$ for invasive cancers while containing two explanatory variables less.
4.6 Predictive Performance

The predictive performance of the stratified Cox model of invasive cancers needs to be assessed and compared to the performance of the Gail model. We will restrict the validation analysis to the discriminatory accuracy of the new model as it is the most important property in the present context.

Due to the already limited number of cases available for the model building process, we were unable to retain a subsample for external validation of the fitted model. Instead, a ‘leave-one-out’ cross validation will be performed on the modelling data set to evaluate the discriminatory performance on independent data.

The steps are as follows. For each individual \( k \), \( k = 1 \ldots n \), fit a proportional hazards model

\[
\lambda_i^{(k)}(t) = \lambda_0(t) \exp(X_i \beta^{(k)})
\]  

(4.27)

and estimate model parameters \( \hat{\beta}^{(k)} \) using all individuals with the exception of the \( k \)th. The hazard \( \hat{\lambda}_k \) for individual \( k \) is then predicted using the fitted model

\[
\hat{\lambda}_k(t|X_k) = \lambda_0(t) \exp(X_k \hat{\beta}^{(k)}).
\]  

(4.28)

This method ensures that each training subsample contains at least \( n_{bc} - 1 \) cases and standard errors remain comparably low. For all individuals \( i \), five-year breast cancer probabilities

\[
P(T_i \leq 5) = 1 - \exp(-\Lambda_i(t = 5))
\]

were projected following the 2000 screening appointment. Cumulative baseline hazards \( \Lambda_0(t = 5) \) were estimated by the Nelson-Aalen estimator

\[
\hat{\Lambda}_0(t) = \int_0^t \frac{d\bar{N}(s)}{\sum_{i=1}^n Y_i(s) r_i},
\]  

(4.29)

where \( \bar{N}(s) \) is the total number of failures up to time \( s \) and \( r_i \) the relative risk estimate under the stratified Cox model (Therneau and Grambsch, 2000). Although an evaluation of the Cox model’s discriminatory performance can be performed using relative risk estimates \( r_i \) only, this is not possible for a stratified model as one explanatory variable is represented
4.6. Predictive Performance

Figure 4.8: Scatterplot of five-year breast cancer risk for 97,856 women as estimated under the Gail model and the stratified Cox model.

by an additional baseline hazard. Thus, additional error associated with an estimation of the baseline hazards needs to accepted.

For reasons explained in Section 3.2 as part of the Gail model validation, the following performance evaluation is restricted to individuals with at least five years of follow-up time. Thus, predicted probabilities of 12,077 individuals were excluded from the validation set, leaving a total of 97,856 individuals.

The fitted model showed a concordance of 0.635 (95% CI: 0.617-0.652), only a slight improvement over the Gail model’s value of 0.611. However, including seven instead of nine explanatory variables, the new model’s set of predictors appears to have more explanatory power while containing two fewer variables. The overall mean risk decreased to 0.79% versus 1.1% predicted by the Gail model. Five-year mean risks for non-cases and cases were 0.79 and 1.02%, respectively. Due to categorical modelling of family history information, the maximum observed risk decreased from 12.2% to 8.3%, suggesting that cancer risks for individuals with multiple affected relatives are less inflated.

Figure 4.8 shows a scatterplot of estimated risks for 97,856 women under the stratified
4.6. Predictive Performance

Figure 4.9: Five-year cancer risk densities stratified by cancer outcome for 97,786 women as estimated by the stratified Cox model (x-axis is truncated at $x = 5$).

Cox model and the Gail model. Risks are positively correlated with the majority of points below the diagonal. Probability values for individuals with the highest Gail risk were clearly reduced by the new model while remaining above average for the stratified Cox model.

Probability separation between cases and non-cases remains poor as seen from Figure 4.9 where five-year risk densities estimated under the Cox model were stratified by cancer outcome. There is still a considerable overlap between the risk distributions of cases and non-cases, making the new model an inadequate tool for predictions about an individual’s future cancer status.

Although the stratified Cox model did not meet the performance requirements to be considered a reliable tool for the SMP BC, a stronger set of predictors could be found. Two new variables $abmam$ and $mammo\_dens$ were able to replace reproductive history variables $age\_men$ and $age\_1stdel$ and interaction terms while slightly improving the predictive performance.
Chapter 5

Conclusion

The data set provided by the Screening Mammography Program of British Columbia presented many initial data cleaning challenges, mostly caused by misleading variable coding and improper survival ascertainment.

The cancer status indicator did not differentiate between women not diagnosed with breast cancer and women lost to follow-up. In order to apply survival analysis methods, a censoring variable needed to be created using the last recorded SMP BC screening date. In turn, this definition of follow-up introduced informative censoring into the data set and lead to further adjustments.

Variable coding for missing information was very inconsistent throughout the data set. Some numerical variables indicated missing covariate information with values of 0 and 99. For variable *num_biop* this was a particular problem as ‘0’ is also a plausible value for the number of breast biopsies. Cross-classification tables with other variables usually provided insights into the true meaning. For categorical variables, missing variables information was represented by several different indicators and required recoding to reduce variable levels.

The first objective of this project was a validation study of the widely used Gail et al. model and in particular the question whether it can be used as a basis for targeted screening policies. Although the Gail model showed good calibration properties for the B.C. population, it’s ability to predict an individual’s five-year breast cancer status was inadequate. Probability estimates for cancer cases were found to overlap substantially with those of non-cases, a conclusion also drawn in the validation studies by Costantino et al. (1999) and Decarli et al. (2006). If a Gail risk criterion was to separate the population into high and low-risk groups, cancer detection rates would show a maximum increase of only twice the prevalence level. Once converted into absolute terms, cancer detection rates increase from approximately 1% to 2%, a difference that is clearly too small to provide a basis for a change in screening policies. Given the large percentage of East Asian women in British
Columbia, 11.5% of the SMP BC population, the Gail model’s significant underestimation of breast cancer risk for this ethnic group is particularly noteworthy.

In the light of these results, the practice of presenting patients with individualized breast cancer risks also seems rather questionable. Providing good results for the ‘average’ patient is irrelevant when individualized probabilities do not reflect information about a woman’s future cancer status. This is a particular problem if online risk calculators are not accompanied by confidence intervals or statements about accuracy of risk estimates.

The main objective of this research, a risk model with improved outcome stratification over the Gail model, was not met. The majority of available risk factors did not have enough explanatory power to predict who will and will not develop cancer in a five-year time interval. Although a number of predictor variables were found to have a significant effect on event-free survival time, effect sizes were generally too small to set probability values of future cases apart from those of non-cases. Significance was achieved not through effect size but as a result of the very large sample size. The overall strongest risk factor, a sister affected by breast cancer, increased a woman’s five-year probability significantly by a factor of 1.926. However, once the estimated probabilities are put into an absolute context, increases associated with the available risk factors are nearly irrelevant. With the majority of estimated risks no larger than 1%, even the strongest risk factor increases these five-year risks only to less than 2%. Generally, estimated probabilities for all cancer cases were too low and did not reflect the five-year outcome. If longer follow-up data was available, one could investigate whether the projection interval of five years is too short for cancers of presumed high-risk women to develop.

Although the analysis of in situ cancers was certainly limited due to the low number of available cases, comparisons between the two cancer types gave initial insights into potential differences. Model parameters showed higher relative risks associated with mammographic density and an abnormal mammogram. The question whether the non-significant effect for a sister diagnosed with breast cancer was only the result of low case numbers could be investigated when more data are available.

There were several shortcomings in the validation analysis and modelling fitting. Trying to avoid false assumptions about the cancer status of women lost to follow-up, many individuals and more importantly, cases, were excluded during the analysis. This led to biased cancer
rates in the validation of the Gail model. Although we attempted to minimize impacts on the predictive performance, observed breast cancers in the highest five-year age group were consistently underestimated. For model fitting, we tried to minimize the presence of informative censoring by identifying cases diagnosed outside the program. As a result, only 978 of the 4078 initially available cancer cases were eligible, making parameter estimates subject to considerably more variability than anticipated. For in situ cancer, the number was even lower with 311 eligible cases. In retrospect, there is a tradeoff between a potentially false cancer status of what might be a small percentage of women and higher standard errors for model parameters due to smaller sample size. One could investigate the differences between the two alternatives by comparing the results of Section 4.4.2 with a proportional hazards model assuming that all women were followed for the full five-year period.

As seen from the poor model performances, breast cancer is too complex to be described by the available risk factors. Especially all information related to a woman’s reproductive history seemed to be of little importance. Although new, stronger risk factors are crucial to improve model performances, medical history and an accumulation of breast cancer in the family showed the most promising predictive abilities and should be the focus point for further risk modelling.

This approach is also taken in a recent addition to the class of breast cancer prediction models proposed by Tyrer et al. (2004). At the basis of the Tyrer-Cuzick Model stands a genetic submodel predicting the presence of a hypothetical risk gene from breast cancer occurrences in first and second degree relatives. A recent study by Quante et al. (2012) suggests that it outperforms the Gail model in terms of population stratification and also shows a superior calibration performance. As family history information provided by the SMP BC data set was limited to first degree relatives only and gave no information about their existence, this approach could not be taken.

As seen in a study by Cecchini et al. (2012), the importance of mammographic density as a breast cancer risk factor could also be confirmed in the analysis. Even with the binary density classification available from data set, a significant effect comparable that of a previous benign biopsy was found and should encourage the collection of breast density data in a four category format.

If the Screening Mammography Program questionnaire were to be redesigned, it is there-
fore important to collect more detailed information about the family’s breast cancer history. Questions about the existence of sisters and daughters need to be added such that we are in a position to differentiate between women having no affected relatives because there are no cancer cases in the family and women with no family members beside their mothers. Women with large families are more likely to have a family member diagnosed with cancer based on chance alone than women with small families. Relative instead of absolute numbers for sisters and daughters could be collected. It seems that an answer ‘Only child’ was once included on the SMP BC questionnaire, however, the appearance of this category in the data set was too low to be consistent with population values of families with one child only. Currently, the collected information does not contain enough explanatory power to build a breast cancer risk prediction model.
Bibliography


population attributable risk for multiple risk factors using case-control data. American
Journal of Epidemiology 122(5), 904–914.

Cecchini, R. S., J. P. Costantino, J. A. Cauley, W. M. Cronin, D. L. Wickerham, H. Bandos,
J. L. Weissfeld, and N. Wolmark (2012). Baseline mammographic breast density and the
risk of invasive breast cancer in postmenopausal women participating in the nsabp study
of tamoxifen and raloxifene (star). Cancer Prevention Research 5(11), 1321–1329.

Chen, J., D. Pee, R. Ayyagari, B. Graubard, C. Schairer, C. Byrne, J. Benichou, and M. H.
Gail (2006). Projecting absolute invasive breast cancer risk in white women with a model
that includes mammographic density. Journal of the National Cancer Institute 98(17),
1215–1226.

Coggon, D., G. Rose, and D. J. P. Barker (2003). Epidemiology for the Uninitiated (5th

Taylor & Francis.

Wieand (1999). Validation studies for models projecting the risk of invasive and total

Series B (Methodological) 34(2), 187–220.

for prediction of absolute risk of invasive breast cancer: independent evaluation in the
florence–european prospective investigation into cancer and nutrition cohort. Journal of
the National Cancer Institute 98(23), 1686–1693.

Journal of the National Cancer Institute 86(8), 573.


Bibliography


