Development of a prognostication model for breast cancer in British Columbia

by

Joyce Epp

B.Sc., The University of British Columbia, 2015

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

MASTER OF SCIENCE

in

The College of Graduate Studies

(Computer Science)

THE UNIVERSITY OF BRITISH COLUMBIA

(Okanagan)

July 2020

The following individuals certify that they have read, and recommend to the College of Graduate Studies for acceptance, a thesis/dissertation entitled:

Development of a prognostication model for breast cancer in British Columbia

submitted by Joyce Epp in partial fulfilment of the requirements of the degree of Master of Science

Dr. Jeff Andrews, I. K. Barber School of Arts & Sciences
Supervisor

Dr. Rasika Rajapakshe, I. K. Barber School of Arts & Sciences
Co-supervisor

Dr. John Braun, I. K. Barber School of Arts & Sciences
Supervisory Committee Member

Dr. Ramon Lawrence, I. K. Barber School of Arts & Sciences
Supervisory Committee Member

Dr. Abbas Milani, School of Engineering
University Examiner

Abstract

Detecting breast cancer in its early stages improves patient outcomes, as the disease is often more treatable at early stages. In British Columbia, BC Cancer Breast Screening provides screening mammography as the primary means of early detection. As with any screening test, false positive results do occur. However, false positive results from screening mammography have been shown to be associated with the risk of a future breast cancer. Moreover, the mammographic features that indicate a mammogram is abnormal have been shown to stratify the increase in risk. Models to predict breast cancer risk currently have a modest ability to distinguish between who will be diagnosed and who will not, with concordances in the 0.55 to 0.71 range.

Using risk factor, screening, and diagnosis data from BC Cancer Breast Screening, we build several models that incorporate mammographic features with the aim of creating a risk prediction model that provides a better risk prediction than current models. We fit a Cox proportional hazards model, a time varying Cox model, an accelerated failure time model and a Poisson regression model. We found that mammographic features are associated with an increased risk of breast cancer and increased model concordance by about 0.006. However, while the internal calibration was satisfactory, as it was close to 1, the discrimination of the models was unsatisfactory, with concordances around 0.63. Compared to several models in the literature, our model performance was similar, with concordances near 0.63 and internal calibration close to 1 for both ours and the published models examined. We conclude that breast cancer risk prediction is not substantially improved by the incorporation of mammographic features in a categorical format.

Lay Summary

Sometimes when women get a screening mammogram there are features on the mammogram that indicate an abnormality, but after further testing no breast cancer is found. Though no cancer was found at the time, the features on the mammogram that indicated something was abnormal may increase the woman's risk of a future breast cancer down the road.

Here we use statistical models and mammographic features to predict the risk of a woman developing breast cancer using data from BC Cancer Breast Screening. We confirmed that mammographic features do increase the risk of breast cancer, but not enough to be useful for individual risk prediction.

Preface

This thesis is covered by UBC-BCCA REB Certificate of Ethics Approval number H18-00172.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisors Dr. Rasika Rajapakshe and Dr. Jeff Andrews for their enthusiasm, guidance and advice, and for making this project possible.

I would also like to thank my committee members Dr. Ramon Lawrence and Dr. John Braun and examiner Dr. Abbas Milani for their insights and time taken to review this work and Dr. Mikael Hartman and Dr. Hua Shen for helpful discussions from near and far.

To my family

Chapter 1

Introduction

Breast cancer is the most common cancer in Canadian women, accounting for a quarter of new cancer diagnoses each year [BWD+20]. In 2020, it is estimated that over 27,000 women in Canada will be diagnosed with the disease and 5,100 will die from it [BWD+20]. Fortunately, patient outcomes are improved by detecting a breast cancer in its early stages, while the disease is often more treatable [SBSTL15, KGR+95].

The primary method for the early detection of breast cancer in Canada is screening mammography [oCS12]. Canada's first organized screening mammography program was BC Cancer Breast Screening, which started in 1988 and currently provides screening mammograms to women across the province of British Columbia through 36 fixed and 3 mobile screening sites [oC11, Scr18]. Symptom-free women aged 40-74 without a prior breast cancer diagnosis are eligible to self-refer to screening mammography subject to the guidelines of the program. These guidelines stratify the frequency and eligibility with which a woman is recommended to receive screening [Scr18]. In general, women assessed to be in a low risk group are recommended to attend screening less often than a woman deemed to be at a high risk of developing breast cancer. The risk factors currently used by BC Cancer Breast Screening to determine screening eligibility are age, family history and presence of a mutation in the breast cancer genes BRCA1 and BRCA2 [Scr18].

Due to the nature of screening tests, which balance the trade-off between correctly identifying a high number of those who have a disease and recalling an appropriate number for additional tests, an unavoidable byproduct of screening mammography is the phenomenon of false positives [Gor13, GSW+17]. False positives are defined as an abnormality detected by the radiologist that does not result in a cancer diagnosis after further assessments. Since this leads women to sometimes undergo invasive procedures as well as experience the emotional stress of a potentially looming diagnosis, false positives have traditionally been seen as a harm of screening mammography [EBM+98, VdSKDDVR11, MSG02].

However, it has been shown that false positives are associated with the

risk of a future breast cancer [HHS+15, vECRTV12, vECKV14, CRR+13, WvECL+17]. Further, recent work has shown that the mammographic features on a false positive mammogram are a promising risk factor to include in determining screening eligibility [CTRP+16]. Mammographic features are the characteristics on a mammogram that a radiologist uses to determine if a mammogram is abnormal, and thus, whether a woman should go for additional procedures, such as a diagnostic mammogram, ultrasound or biopsy. The features BC Cancer Breast Screening radiologists, and radiologists worldwide, endeavour to detect are masses, calcification, architectural distortion and asymmetry. The presence of these features and the location they appear in on the mammogram are recorded in the Breast Screening registry.

These features stratify the increased risk from a false positive mammogram. For example, of the four features, calcifications have been shown to have the highest risk of a future diagnosis and the lowest increase in risk has been found in asymmetry and distortion [CTRP+16]. Not only that, but women whose mammographic features change from one screen to the next show an even higher risk of developing a future cancer [CTRP+16]. The increased risks that these mammographic features provide are greater than or equal to that attributed to family history, which is currently used to determine screening frequency [BJS+17, Scr18]. It has been suggested that incorporation of mammographic features into risk prediction models will improve the models that screening guidelines are based on by providing better risk stratification, and, in turn, improve the screening guidelines by increasing personalization [CTRP+16, WRB+13].

Models that use risk factors to predict breast cancer risk in the literature currently do not incorporate this mammographic feature information. False positives and previous biopsies have been used, among other well-known risk factors, such as age, family history, mammographic breast density, and various lifestyle and reproductive factors. Further, the ability of these models to distinguish between who will develop breast cancer, and who will not, is fair at best. With the added risk that the false positives contribute, and the stratification from the mammographic features, it is possible that their inclusion could improve the predictive ability.

1.1 Objective and thesis outline

The objective of this thesis is to build a risk prediction model for British Columbian women attending screening mammography which incorporates

the mammographic feature information and predicts risk of breast cancer. The remainder of this thesis is laid out as follows. Chapter 2 will review the literature on models to predict breast cancer risk. In Chapter 3, we detail the methodology for the Cox proportional hazards, time-varying Cox, accelerated failure time, and Poisson models explored. In Chapter 4, we make a comparison of the results to the literature and explore the clinical relevance. Finally, Chapter 5 finishes with a summary and suggestions for future work.

Chapter 2

Background

In this chapter, we define some key concepts and terms and provide a review of breast cancer risk prediction models in the literature.

2.1 Terminology

2.1.1 Risk

Risk is the probability, or how likely it is, that an event will occur. This risk can be described as absolute, which is the number of events divided by the number of people in that group; or relative, which is the absolute risk of one group compared to the absolute risk of another [gro19]. In the context of breast cancer, absolute risk is the probability of developing breast cancer in a given period of time, while relative risk is how much more or less likely one group is to develop breast cancer compared with another.

2.1.2 Risk models, concordance and calibration

Risk models for breast cancer are statistical models that assess the probability that an individual will be diagnosed with breast cancer in a given period of time [AALY+18]. These are known as prognostication models, which are a type of prediction model [SMvdW+13]. The main measures of validation reported in the literature are discrimination and calibration [CGBB+17].

Discrimination is the ability of a model to separate individuals who will have the disease from those who will not. This is usually reported as the concordance value [HJLM96, HJLC+84]. A concordance of 1 indicates perfect discrimination, or women without a diagnosis had consistently lower risk scores than women who are diagnosed, and a value of 0.5 indicates that the model is no better than randomly guessing at distinguishing between who will get a diagnosis and who will not. It can be calculated as the proportion of concordant pairs, where a concordant pair occurs when the individual with the diagnosis has a higher risk than an individual without the

diagnosis. Graphically, a concordance of 1 would mean we see two distinct, non-overlapping curves on a risk distribution plot, with the higher risk curve for those who had a diagnosis. This would allow us to specify a threshold to divide individuals into diagnosis and no diagnosis groups. A concordance of 0.5 would mean the diagnosis and no diagnosis curves perfectly overlap each other and there is no distinguishing between them.

Calibration is a 'goodness-of-fit' measure which gives the ratio of the observed number of cases to predicted number of cases [vH00, Sch80]. A well-calibrated model will have a ratio close to 1. For a stronger look at calibration, individuals may be grouped into quantiles and the observed to expected ratio examined for each quantile [S$^+$19].

2.1.3 Risk distribution

A risk distribution describes the frequency of absolute risk within the population [S$^+$19]. It is often shown with distinguishing between those with a diagnosis and those without a diagnosis to examine whether those with a diagnosis have a higher risk than those without. It may also be used to decide on risk cut-offs for preventive measures, where individuals above a specified risk threshold are eligible to receive the preventive measure [HJ15].

2.1.4 Hazard function

A hazard function, or the hazard at time t, can be defined as

$$h(t) = \frac{f(t)}{S(t)} \tag{2.1}$$

where $f(t)$ is the probability density function and $S(t)$ is the survival function. $h(t)$ can be interpreted as the probability of experiencing the event in the next small interval of time, divided by the probability of surviving up to that time [TG00].

2.1.5 Surveyimpute procedure in SAS

The surveyimpute procedure in SAS can use hot-deck imputation to impute missing values [SAS19, Inc15]. Observations that do not have missing values are considered donor units (d) and observations with missing values are considered recipient units (m). For each recipient unit, a donor unit is selected at random and the missing values for that recipient unit are filled in with the values from the donor. The selection method chosen for this analysis was simple random sampling with replacement.

2.1.6 Schoenfeld residuals

Schoenfeld residuals are calculated using

$$r_i = z_i - E(Z_i) \tag{2.2}$$

where z_i is the observed value of the variable and $E(Z_i)$ is the expected value [Sch82]. The scaled Schoenfeld residuals can be calculated with

$$R_i = r_i \cdot d \cdot var(\beta) \tag{2.3}$$

where r_i are the Schoenfeld residuals, d is the number of events, and $var(\beta)$ is the covariance matrix.

The expected value of the scaled Schoenfeld residuals can be approximated by adding the estimated coefficient to the coefficient as a function of time [TG00]. As such,

$$E(R_i) \approx \ \beta + \beta(t) \tag{2.4}$$

Therefore, if the proportional hazards assumption is correct, then a plot of $\beta(t)$ vs t should show an approximately linear line at zero.

2.1.7 dfbetas

A model diagnostic that can be examined to asses the influence of each observation on parameter estimation is the difference in an estimated coefficient when all of the data are used and when one observation is deleted from the data [W+86]. The model is fitted with removing one observation at a time and the difference between the coefficient estimation with and without that one observation is computed. A simple plot of the change in coefficient against the observations will reveal if any observations have a large influence on parameter estimation and should be removed.

2.1.8 Martingale residuals

Martingale residuals can be used to assess the functional form of a variable [TGF90]. They are calculated with

$$m_i = \delta_i - H_0(t_i) \exp(z_i^{'}\beta) \tag{2.5}$$

where δ_i is the censoring indicator, $H_0(t_i)$ is the integrated baseline hazard function, and $\exp(z_i^{'}\beta)$ is the linear predictor from the proportional hazards model. Since the Martingale residuals are asymmetric, deviance residuals are an attempt to make the Martingale residuals symmetric around zero.

2.1.9 Akaike Information Criterion

The Akaike Information Criterion (AIC) is given by

$$AIC = -2l(\beta) + 2k \tag{2.6}$$

where $l(\beta)$ is the value of the log-likelihood of the model and k is the number of parameters in the model [SIK86]. The better the model fits the data, the smaller the value of $-2l(\beta)$. The more complex the model, the larger $2k$ is, which penalizes the model for complexity. Therefore, a better model should have a smaller AIC because it fits the data well with a small number of parameters.

2.1.10 Lexis split

A Lexis split allows for following individuals on multiple time scales [Kei90, Car07, PC$^+$11]. These time scales may be age, calendar time, or time since exposure, to name a few. These times scales are cut into groups, or time bands. During the follow-up of an individual, they pass through different time bands. Each time band they pass through generates a new row in the data for that observation. The hazard is constant in each time band.

2.2 Breast cancer risk prediction models

Models to predict the risk of breast cancer are ubiquitous in the breast cancer literature [LPB$^+$19, EF15, ATW$^+$12, CGBB$^+$17, AALY$^+$18]. Possibly the most well known is the logistic regression Gail model published in 1989, which used white women's age at menarche, number of previous biopsies, age at first live childbirth and number of first-degree female family members who had a breast cancer diagnosis [GBB$^+$89]. The concordance of the model was not reported in the original paper.

Since then, it has been extensively externally validated and updated. For instance, [TCZK05] externally validated the Gail model for mixed ethnicities and updated it using a Cox regression model using the Gail risk factors with the addition of breast density. This resulted in a concordance of 0.68, a promising update. Interestingly, in this same study, they fit a Cox model with age, ethnicity and breast density and this had the same concordance of 0.67 as when they validated the Gail model. [Hoe13] validated the Gail model on BC women attending screening mammography, and found that

while the calibration was good, the concordance was only 0.61. Among others, BMI, alcohol use, diet and exercise have been added to the Gail model to improve the discriminatory performance and modify the model as new risk factors become available. It has since been validated and/or updated on various populations, such as Asian and African, in which it provides a low discrimination for and overestimates risk, respectively [CGBB$^+$17]. Most, if not all, of the modified Gail models used previous breast biopsies as a predictor, however they did not use false positives or the associated mammographic features.

Another well-known model is the Rosner-Colditz model, which was a log-incidence model built from data on white women aged 30-64. This model began with primarily hormonal factors such as age at menarche, menopausal status, age, and age at first and subsequent births. With a concordance value no higher than 0.68, it has been updated to include estrogen use, family history, weight, BMI, alcohol consumption and benign breast disease. Most recently, hormone biomarker measures have been added to this model and the Gail model, such as testosterone and estrogen levels, which show improvement to these models [CGK$^+$19, ZRT$^+$18]. It appears that most, if not all, of the updates to the model have used white women as well, thus this line of models may not be applicable to the multi-ethnic population of women attending screening mammography in BC. While benign breast disease is used in nearly all the updated Rosner-Colditz models, mammographic features and false positives have not been added to this model.

There is also a group of models that include detailed family history and presence of genetically inherited gene mutations. Since the Gail model performs best in women without a family history, [CRT93] built the Claus model for individuals with a family history before the discovery of the genetic mutation on the BRCA1 and BRCA2 genes. The Tyrer-Cuzick model is another genetically based model which uses a host of predictors, such as hormonal factors, biopsies, family history of breast and ovarian cancer, and BMI [CGBB$^+$17]. Though they did not report any performance measures in the original paper, it has been validated on high risk populations and performs better than the Gail and Claus models, although it tends to overestimate risk in certain sub-populations [CGBB$^+$17].

There are also a number of independently derived models, such as the Thai risk prediction model, which is a logistic regression model for Thai women which attained a concordance of 0.65 and calibration of 1.0 using only 4 variables; age, contraceptive use, BMI and menopausal status [ATW$^+$14]. [Hoe13] built a stratified Cox model with BC women, which attained a concordance of 0.65 and 0.63 for *in situ* and invasive, respectively, and among

the typical variables of age, ethnicity, family history, biopsy and estrogen use, incorporated a variable indicating whether the baseline mammogram was abnormal. [UTT+03] built a model for long term risk based on Japanese women using only age, BMI, family history, and age at menarche, but did not report concordance, while [WGB+14] built a model for Chinese women to fit the low-incidence population and achieved a concordance of 0.64.

While it has been found that false positives are associated with the risk of breast cancer, so far, only one published model has explicitly used them. False positives were added by [BWBB+06] which gave a concordance of 0.63 when fitting a logistic regression model with 11,638 diagnoses and 2 million screens.

Chapter 3

Methods

3.1 Data

The data used in this study were from the BC Cancer Breast Screening registry. It contained risk factor, screening, and diagnosis information collected during BC Cancer Breast Screening's routine operations from 6,455,264 screening mammograms for 1,099,298 women who attended screening mammography from 1988 to March 31, 2019.

Ethics approval was obtained from the UBC-BCCA Research Ethics Board with certificate number H18-00172. Access to the data was granted after a formal BC Cancer Data Access Request was approved by BC Cancer on July 22, 2018. Data were provided to us de-identified by BC Cancer Breast Screening with a randomly assigned numeric subject identifier.

Risk factors that are assumed to be unchanging when a woman first attends screening at BC Cancer Breast Screening are collected with a voluntary background survey at the first screening mammography appointment. These factors include ethnicity, highest education level achieved, number of full-term pregnancies, the age at the first full-term pregnancy, if applicable, and the age at menarche [LZW$^+$12, DQ20, RCW94, MCL$^+$70, oHFiBC$^+$12].

Risk factors that can change over time are voluntarily surveyed at each screening visit. These include the woman's family history of breast cancer, current estrogen use, whether she has had a breast biopsy, menopausal status and if she has had a hysterectomy and ovary removal [PDD$^+$97, KKLL18, CG01, DYFC15, oHFiBC$^+$12, SPF$^+$97]. Since the risk factor information in this dataset is obtained from voluntary surveys, it is subject to errors and missingness. The risk factors collected by BC Cancer Breast Screening have been shown to affect breast cancer risk and have been used in previous risk prediction studies.

Once screening mammograms are completed, they are assessed by credentialed and trained radiologists [CP12]. The radiologists ascertain and record the mammographic breast density, the abnormal status, and if abnormal, which mammographic features were observed along with their location. As such, the screening mammogram information available in the Breast

Screening registry relies completely on the radiologist's interpretation of the mammogram.

Screen-detected and non-screen detected cancers were also available in the data. Screen-detected cancers are those which are diagnosed as a result of abnormal findings from a screening mammogram. Non-screen detected cancers include: interval cancers, which are cancers discovered within the recommended screening interval from the last normal screen; false negatives, which are cancers discovered after additional assessments from an abnormal screen did not find cancer; and cancers that occur when a woman does not screen according to the recommended guidelines.

Each woman can have at most one diagnosis. Once a woman is diagnosed with ductal carcinoma *in situ* (DCIS) or invasive breast cancer then she is no longer eligible to attend screening mammography. Recurrences and treatment are not recorded in the BC Cancer Breast Screening Registry, but are in the BC Cancer Registry.

3.2 Participants

The individuals in the study can be divided into two groups; those who did not have a false positive and those who had at least one false positive. These two groups are important because normal screens for those who did not have a false positive serve as the reference category for which risks from false positive screens with mammographic features can be compared. As in [CTRP$^+$16], for women who did not have a false positive, the start of follow-up was defined to begin from the first screening mammogram, but for women who did have a false positive, their follow-up began from the first false positive (Figure 3.1). This adjustment to the index screen was done to better understand the risk following a false positive screening mammogram. In the following analysis, the methods require choosing a time origin, which will be the index screening mammogram. From this index screen, women were followed for 5 years, until diagnosis, until their last screen if there was no diagnosis, or until study end, whichever came first. The official study end was Dec 31, 2016 as the data were considered complete (as of August 2019) by BC Cancer until this date, thus an additional censoring of all screens and diagnoses was done at this date.

Although all women who underwent screening at BC Cancer Breast Screening until March 31, 2019 were included in the data we received, we had to exclude some individuals prior to analysis (Table 3.1). First, 680 women were excluded due to missing diagnosis information or birth date.

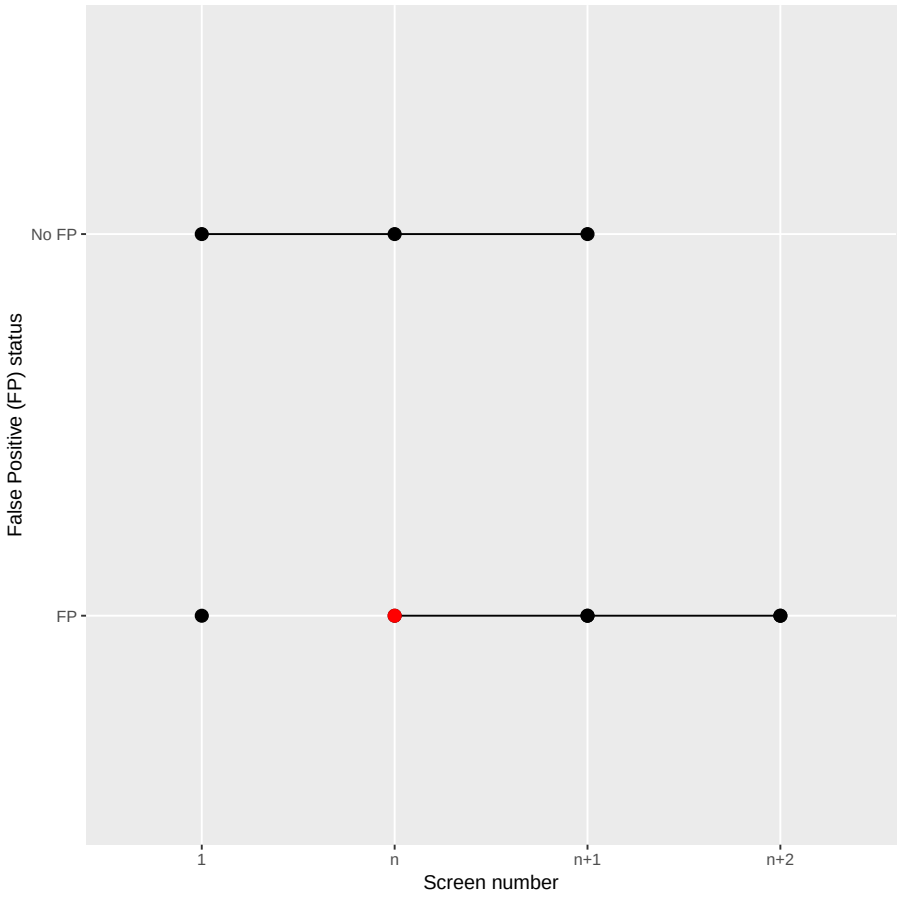

Figure 3.1: Example follow-up of two women undergoing screening. Normal screens are black and abnormal screens are red.

Table 3.1: Exclusions made to get to the final cohort

Reason removed	Women remaining	Women removed
all women	1,099,298	
incorrect diagnosis date	1,099,288	10
missing birth date	1,099,286	2
missing diagnosis details	1,098,618	668
age outside 40-74	1,054,660	43,958
index screen after Dec 30, 2016	962,318	92,342
true positive at index screen	957,276	5,042
false negative at index screen	954,760	2,516
one screen	813,280	141,480
Total	**813, 280**	**287, 736**

As the screening program regularly screens women in the age range of 40-74, it is not helpful to predict risk for women outside these ages. Therefore, 43,958 women younger than 40 or older than 74 at the time of the index screen were excluded. 92,342 women who had their index screen after the study end date were also excluded.

We further excluded 5,042 women with a true positive at the index screen and 2,516 women with a false negative at the index screen. True positives were excluded because the cancer had developed prior to the start of follow-up and were only detected at the index screen because it was already present. False negatives were excluded because these ideally should have been true positives and would have been excluded for the same reasons as true positives. Lastly, 141,480 women who only had one screen were excluded, since no useful information could be learned from them.

This left us with a cohort of 813,280 women, 11,166 diagnoses and 2,347,901 screening mammograms, an average of 2.9 ± 1.3 screens per person within the 5 years of follow-up (Figure 3.2). Of the 2,347,901 screening mammograms, 296,774 (12.6%) were false positives and 259,794 (87.5%) of these false positives were the woman's first false positive (which occurred at the index screen as defined earlier for women with a false positive). The remaining 12.5% of the false positives were additional false positives, e.g. second or third, that occurred on a subsequent screen at some point during follow-up. The median follow-up for women with and without a false positive was 4.97 and 5 years, respectively.

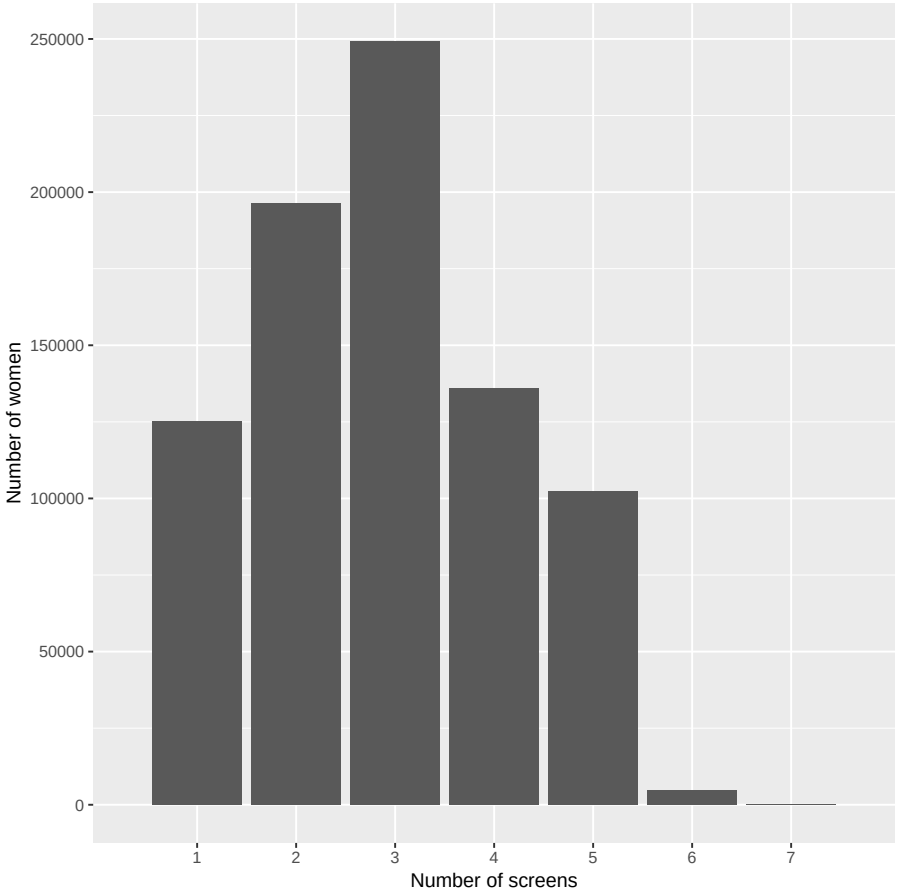

Figure 3.2: Total number of screens in the cohort.

3.3 Variable characteristics

Data cleaning was undertaken prior to model fitting and the results are summarized in Table 3.2. First, there were several different ways missing entries were recorded in the data, such as 'Refuse to Answer', 'Unknown', 'NA' , 'N', and sometimes a special numeric value such as 0 or 99. These were recoded to a single value, 'NA', for consistency.

Second, at the start of the screening program, breast density was recorded as a binary variable, representing $< 50\%$ and $\geq 50\%$ density. However, in 2005, density started being recorded as a quaternary variable, representing 0-24%, 25-49%, 50-74% and $\geq 75\%$ density, but the changeover was gradual and only 24% of the mammograms in the dataset are recorded in the new way. While less informative, to reduce the number of missing values, all breast density values were recoded as a binary factor.

Education and ethnicity were also recoded to contain fewer categories in an effort to make the categories more meaningful. European and British ethnicities were combined into one 'Caucasian' ethnicity. Education levels were grouped into more coarse groups defining no high school education ('No HS'), some or graduated high school ('Some/HS'), and some or graduated post-secondary ('Some/Co').

Family history was initially coded as a 19 level categorical variable representing combinations of family members ever diagnosed with breast cancer, such as 'mother and sister' or 'sister and father'. Family history was converted into a binary variable to indicate 'Yes' or 'No'. In light of the findings from [Hoe13], who found that the type of relationship is important and not just the presence of family history, two additional binary variables were created from the original family history variable indicating whether their mother or sister had a history of breast cancer. Women who indicated that their family history was unknown were given the baseline value of 'No' to both these variables.

The mammographic features also underwent some modifications. Originally, each of the four features was in its own data field; these were combined into one field. If a feature was seen at least once and was the only feature observed on the mammogram, it was labeled with '*feature*_only'. For example, a mammogram with only calcification would be coded as 'calc_only'. If only two features were seen on the mammogram at least once each, this combination received its own category, but 3 or more distinct features were grouped into one category due to the low number of occurrences. While grouping the features into one variable and creating a category for each combination makes for many categories, it is important clinically to see the

Table 3.2: Table describing the final risk factors in the data

Variable	Description	Values
age	age at index screen	numeric
ethnic	ethnic/cultural heritage of client	First Nations
		African
		East Asian
		South Asian
		Caucasian
		Other
education	highest level of education achieved	No HS
		Some/Co
		Some/HS
num_full_term _deliveries	number of full-term pregnancies	integer
age_at_1st _delivery	age at first full-term pregnancy	integer
age_at_menarche	age menstruation began	integer
ever_mom	mother had breast cancer	0 (No), 1 (Yes)
ever_sis	sister had breast cancer	0 (No), 1 (Yes)
Ever_Estrogen	currently using estrogen	0 (No), 1 (Yes)
Ever_Biopsy	ever had a breast biopsy	0 (No), 1 (Yes)
menstr	menopausal status	0 (No), 1 (Yes)
hyster	had hysterectomy	0 (No), 1 (Yes)
ovary	both ovaries removed	0 (No), 1 (Yes)
result	abnormal status	0 (No), 1 (Yes)
density	breast density	1 ($< 50\%$)
		2 ($\geq 50\%$)
feat	mammographic features	3_or_more
		arch_asym
		arch_calc
		arch_only
		asym_calc
		asym_only
		calc_only
		mass_arch
		mass_calc
		mass_only
		missing
		no_feat

influence of the solo features and combinations. To handle missing features, an additional category 'missing' was created to represent screens that the data indicated were abnormal, but were missing the corresponding mammographic feature(s) present. This occurred for 107,525 screens, which is roughly a third of all the false positives (Table 3.3). Prior to 2003, only the presence and location of an abnormality was recorded, but not the type of mammographic feature, resulting in the missing features.

Table 3.3: Table showing the breakdown of false positives and other screens on the index screens and all screens

	Count	%
Index screen		
normal screen	553,486	68
false positive	259,794	32
false positive with features recorded	164,519	20
false positive without features recorded	95,275	12
Total	813,280	100
All screens		
non-false positive screen	2,051,127	87
false positive	296,774	13
false positive with features recorded	189,249	8
false positive without features recorded	107,525	5
Total	2,347,901	100

After data cleaning, some exploratory analysis was performed to understand the data. We found that of the 813,280 index screens, 32% (259,794) were false positives (Table 3.3). Of these false positive screens, 37% (95,275) did not have mammographic features recorded (Figure 3.3). Therefore, of the 813,280 index screens, 20% (164,519) were false positives with features recorded and 12% (95,275) were false positives without any mammographic features recorded.

From Table 3.3, we note that roughly one third of the women who attended screening mammography had a false positive at some point during their screening journey. This is a similar finding to [EBM+98], who found that for women attending screening with a median of 4 screening mammograms, one third of the women screened had a false positive.

The distribution in Figure 3.3 of the index screens with mammographic features present reveals that asymmetry ('asym_only') is recorded most often (32%) and the combination of architectural distortion with calcifications

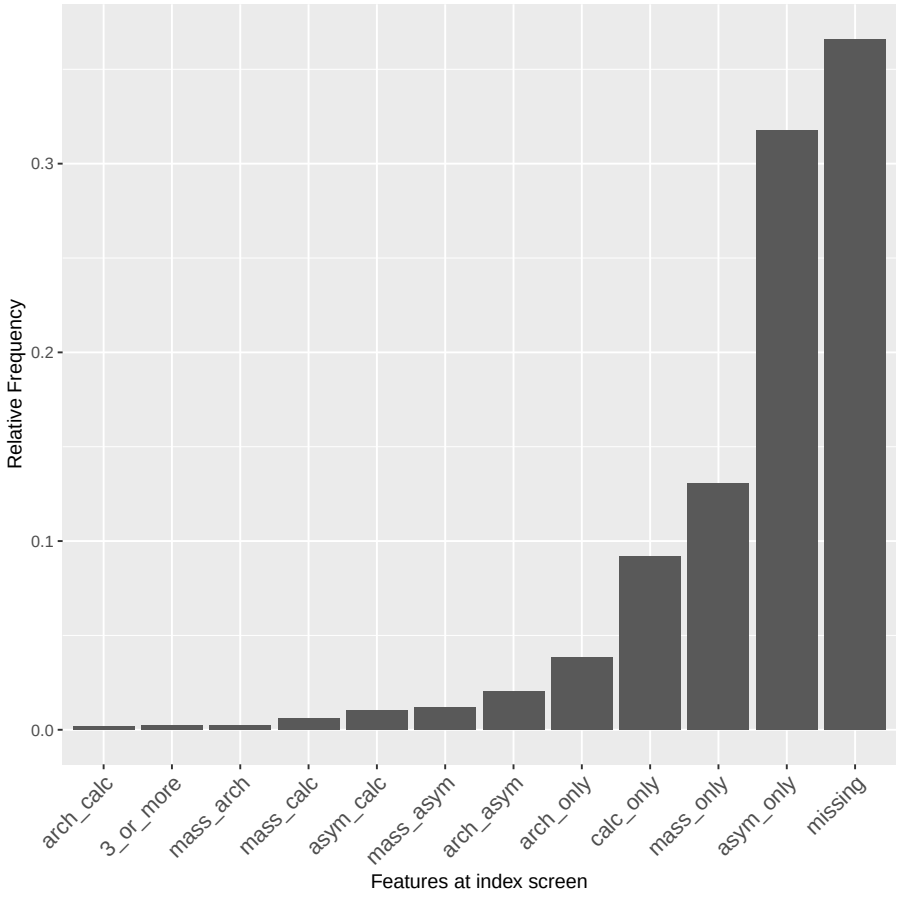

Figure 3.3: Frequency of mammographic features for the index screens where there was an abnormality indicated.

('arch_calc') occurs least often (0.2%). Examining the distribution of the mammographic features using all the features from all screens during follow-up revealed a nearly identical frequency distribution to the distribution of features on the index screen.

When examining education, it was found that 50% of the women had some college or completed a post-secondary degree, 36% had graduated high school or completed some high school, and 6% had no high school education. On further inspection, the average age of those women who reported they had no high school was 58.5 at the index screen, compared to the average age of 53.3 for those who had a high school education and 49.7 for those who had post-secondary education. Education was also found to be mildly related to the number of children. Figure 3.4 shows that as the education level increases the number of children decreases on average. We can also observe the increase in children for older women. When compared with birth year, it was found that women born in the 1920s and 1930s had the most children on average. It appears that both education and number of children are associated with age at index screen and birth cohort.

With respect to ethnicity, 47% reported one of the ethnicities that were used to create the Caucasian group, 10% reported East Asian, 3% South Asian and only 1.2% and 0.2% First Nations and African, respectively. The remaining 38% of the women did not have an ethnicity recorded, with 4.1% responding 'other', 0.3% 'refused to answer', and 33.2% stated their ethnicity was 'unknown'. These were grouped into the category 'Other', making this 'Other' group comprised of 11.0% 'other', 0.7% 'refused to answer', and 88.3% 'unknown'. This category may be thought of as all individuals who did not place themselves into one of the predefined ethnicities.

We confirmed that the breast density decreases with age and is stratified by ethnicity (Figure 3.5). Breast density is a measure of how much fibroglandular tissue is seen on the mammogram, represented as a percentage of total breast area. On a mammogram, breast tissue appears white and fat appears dark [MRMTB14]. Of particular note is that the East Asian women have the highest breast density, while the First Nations women have the lowest.

When examining associations between age, mammographic breast density and the type of feature, we found that calcifications were associated with dense breast. Mass and calcification ('mass_calc') falls in between calcifications ('calc_only') alone and mass ('mass_only') alone. Figure 3.6 shows select features.

Table 3.4 provides a summary of the risk factor values for the cohort used in this analysis. For mammographic breast density and features, the values used represent the values recorded at the index screens. A summary

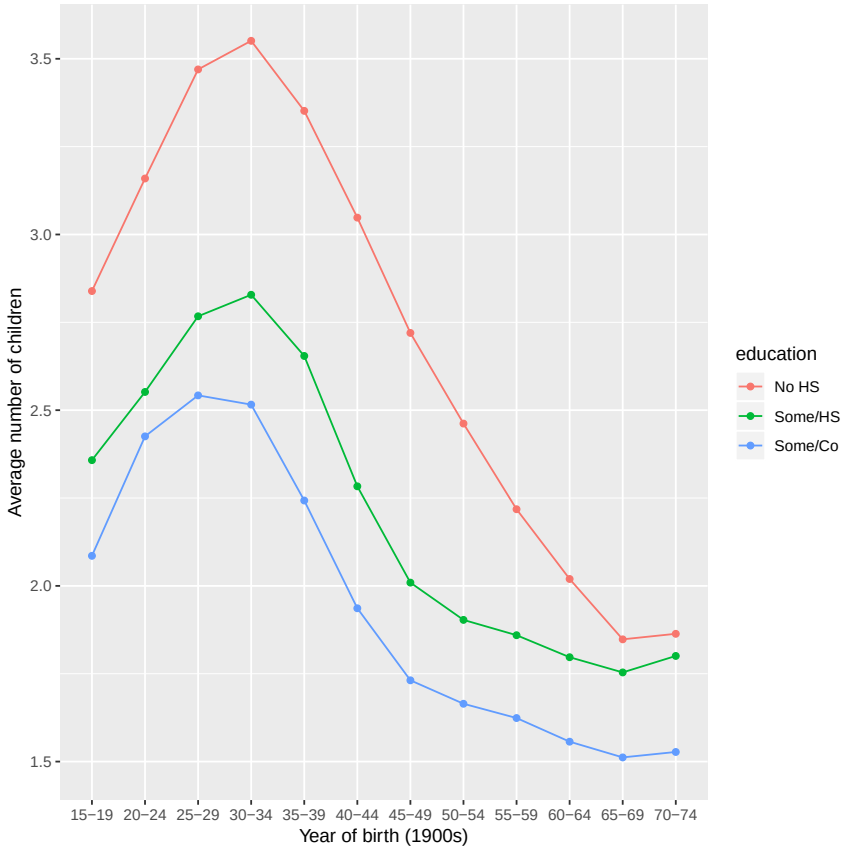

Figure 3.4: Average number of children per half decade of birth stratified by self-reported education level.

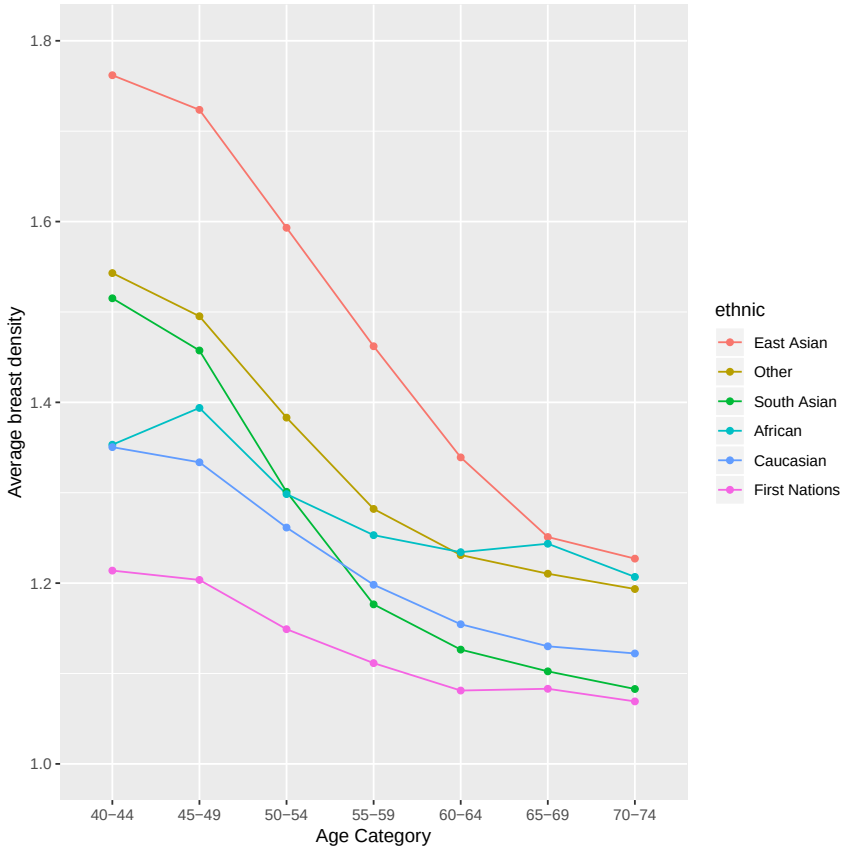

Figure 3.5: Average density for each age group at index screen stratified by ethnicity.

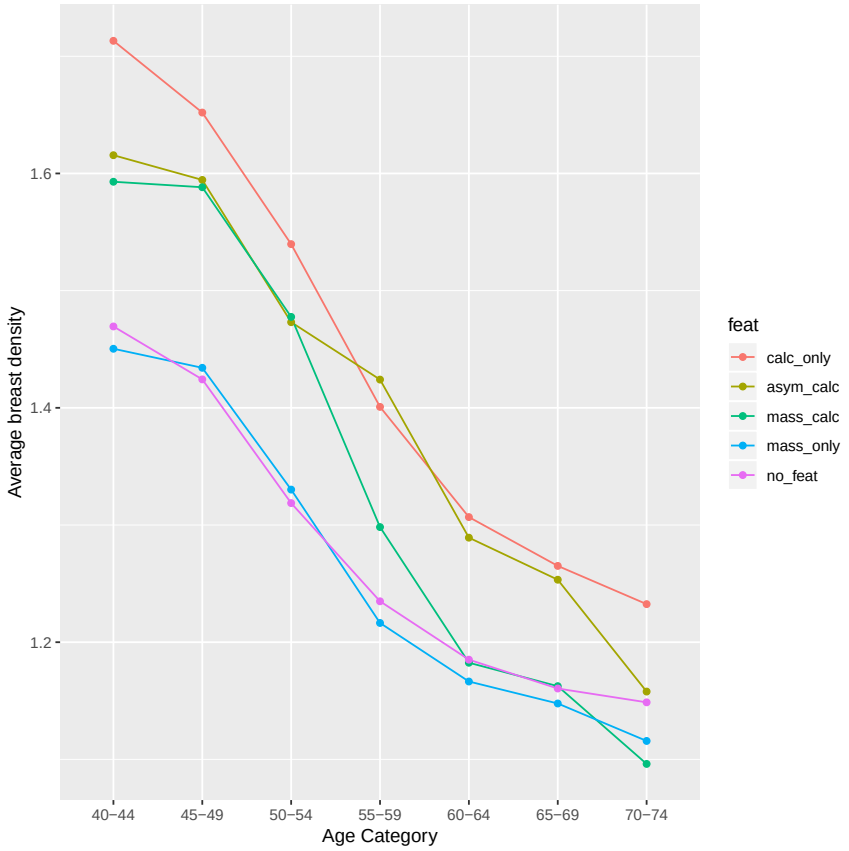

Figure 3.6: Average density for each age group at index screen stratified by select mammographic features.

Table 3.4: Table showing the values of the risk factors in the data

Variable	Missing (%)	Average or count (%)
age	0	51.8
ethnic	305,596 (37.5)	
First Nations		9,911 (1.2)
African		1,794 (0.2)
East Asian		85,512 (10.5)
Caucasian		382,942 (47.0)
South Asian		27,525 (3.4)
education	60,897 (7.5)	
No HS		49,542 (6.1)
Some/Co		407,344 (50.1)
Some/HS		295,497 (36.3)
num_full_term_deliveries	2,617 (0.3)	1.9
ever_mom	0	
0		753,566 (92.7)
1		59,714 (7.3)
ever_sis	0	
0		779,616 (95.9)
1		33,664 (4.1)
Ever_Estrogen	6588 (0.8)	
0		702,591 (86.4)
1		104,101 (12.8)
Ever_Biopsy	5010 (0.6)	
0		730,691 (89.8)
1		77,579 (9.5)
result	0	
0		553,486 (68.1)
1		259,794 (31.9)
density	143,807 (17.7)	
1		433,171 (53.3)
2		236,302 (29.1)
feat	95,275 (11.7)	
3_or_more		578 (0.1)
arch_asym		5,350 (0.7)
arch_calc		452 (0.1)
arch_only		9,941 (1.2)
asym_calc		2,685 (0.3)

Variable	Missing (%)	Average or count (%)
feat		
asym_only		82,510 (10.1)
calc_only		23,823 (2.9)
mass_arch		582 (0.1)
mass_asym		3,135 (0.4)
mass_calc		1,547 (0.2)
mass_only		33,916 (4.2)
no_feat		553,486 (68.1)

of the subsequent screens will be provided when these screens are used in analysis.

3.4 Cox proportional hazards model

To model the risk of a future breast cancer we first use a Cox proportional hazards model (Equation 3.1) [Cox72]. The first point of this model is to determine which risk factors are important to breast cancer risk and by what magnitudes, an insight of clinical relevance. The second point is to obtain a risk estimate for each observation, which can be used to predict breast cancer risk at an individual level [Col15].

The overall framework of the model is as follows. Predictors, in our case, risk factors, are measured at an initial point. Here, that initial point is the index screening mammogram. Individuals are followed for a period of time, in our case, 5 years, and the outcome, diagnosis or no diagnosis, at the end of the 5 years and when it happened, is determined.

In this way, this semi-parametric model allows for incorporating the time to diagnosis while taking into account the censored observations. It has a baseline hazard function, $h_0(t)$, which makes no assumptions on the form of its distribution. The baseline hazard is the value of the hazard function, $h_i(t)$, when all the predictor values are set to zero. The predictors are an exponentiated linear combination, where x_j is the j^{th} covariate in the model and β_j the j^{th} coefficient. The hazard function, $h_i(t)$, is the hazard at time t. This is the outcome of interest, or response [All10]. The assumptions of the model are proportional hazards, independent observations and that the predictors are multiplicative on the baseline hazard.

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}) h_0(t) \tag{3.1}$$

We can write Equation 3.1 more concisely as in Equation 3.2.

$$h_i(t) = \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}\right\} h_0(t) \tag{3.2}$$

The hazard associated with a covariate is provided by $\exp(\beta_j)$. If it takes a value >1, then there is an increase in hazard, if it is <1 then there is a reduction in hazard. For continuous variables, $\exp(\beta_j)$ is the increase in hazard with one unit increase in the predictor, holding the other predictors constant (Equation 3.3).

$$\frac{\exp\{\beta(x+1)\}}{\exp(\beta x)} = e^{\beta} \tag{3.3}$$

For categorical variables, a reference category must be chosen to take the baseline value of $\gamma_1 = 0$ and the hazard estimates are made with respect to that reference category. So for $j = 1, \ldots, m$ groups, where $j = 1$ is the reference category, the hazard function for an individual in the j^{th} group is

$$h_j(t) = \exp(\gamma_j) h_0(t) \tag{3.4}$$

Rearranging, we find that the relative hazard for someone in group j relative to group 1, is e^{γ_j}.

$$\frac{h_j(t)}{h_0(t)} = e^{\gamma_j} \tag{3.5}$$

Prior to model fitting, missing values for all variables were imputed with SAS's surveyimpute procedure [SAS19, Inc15]. This function imputes missing values by sampling from other values that are present for that variable. Next, the data was split into a randomly selected training and testing set where 70% was used for training and 30% for testing. To reduce the number of variables in the model, variable selection was performed using backwards selection through the selectCox() function in R, which uses the reduction in AIC as a stopping rule [R C19, MIG12, SIK86]. The resulting model is shown in Table 3.5.

From the model in Table 3.5, we find that the features that have the largest increase in risk compared to the reference category of no features ('no_feat') are when 3 or more ('3_or_more'), architectural distortion and calcifications ('arch_calc'), mass and architectural distortions ('mass_arch') and calcifications alone ('calc_only') are seen. We also note that the confidence intervals for the features are quite large, making it difficult to deduce how much hazard is actually attributed by the presence of these features.

Table 3.5: Cox proportional hazards model

	coef	exp(coef)	p-value	2.5 %	97.5 %
age	0.04	1.03	0.00	1.03	1.04
featno_feat (ref)	0.00	1.00			
feat3_or_more	0.84	2.31	0.01	1.28	4.17
featarch_asym	-0.04	0.96	0.81	0.71	1.31
featarch_calc	1.17	3.23	0.00	1.88	5.57
featarch_only	0.03	1.03	0.74	0.85	1.26
featasym_calc	0.42	1.53	0.01	1.09	2.15
featasym_only	-0.12	0.88	0.00	0.81	0.96
featcalc_only	0.70	2.01	0.00	1.82	2.22
featmass_arch	0.96	2.62	0.00	1.52	4.52
featmass_asym	0.43	1.53	0.01	1.11	2.11
featmass_calc	0.46	1.58	0.03	1.04	2.41
featmass_only	0.03	1.03	0.56	0.92	1.16
featmissing	0.16	1.17	0.00	1.10	1.25
density1 (ref)	0.00	1.00			
density2	0.39	1.48	0.00	1.41	1.55
ethnicCaucasian (ref)	0.00	1.00			
ethnicFirst Nations	0.21	1.23	0.04	1.01	1.50
ethnicAfrican	0.21	1.24	0.33	0.80	1.90
ethnicEast Asian	-0.05	0.95	0.20	0.88	1.03
ethnicOther	0.01	1.01	0.58	0.96	1.06
ethnicSouth Asian	-0.32	0.73	0.00	0.62	0.85
Ever_Estrogen0 (ref)	0.00	1.00			
Ever_Estrogen1	0.16	1.17	0.00	1.10	1.24
Ever_Biopsy0 (ref)	0.00	1.00			
Ever_Biopsy1	0.35	1.42	0.00	1.34	1.51
ever_mom0 (ref)	0.00	1.00			
ever_mom1	0.51	1.66	0.00	1.55	1.78
ever_sis0 (ref)	0.00	1.00			
ever_sis1	0.41	1.50	0.00	1.38	1.64
educationNo HS (ref)	0.00	1.00			
educationSome/HS	0.27	1.31	0.00	1.19	1.45
educationSome/Co	0.32	1.38	0.00	1.25	1.53

Surprisingly, asymmetry ('asym_only') has a negative coefficient. Since the upper bound of the 95% confidence interval is very close to 1.0, it is likely that this is not really significant, as found in [CTRP+16], but the large number of screens causes it to appear significant. Another interesting observation is that the First Nations group when compared to Caucasian have an increase in risk. Recall from Figure 3.5 that the First Nations had the lowest breast density. As such, this may suggest that a high breast density should be taken into account with ethnicity and other factors.

It is also intriguing to find that the education status shows up as significant, while none of the breast cancer risk prediction models we are aware of include it. It is possible that there is still some overfitting occurring or some behaviour that has been unaccounted for to make it appear as a significant risk factor. Alternatively, this education variable may be more of a measure of lifestyle than just strictly education. For example, we found that women who had more children also had the first child at a younger age. We also saw in Figure 3.4 that education is associated with number of children. It could be that higher educated women are having less children and at a later age, thus increasing their risk of breast cancer, resulting in education level as a significant risk factor [DQ20].

To assess whether this model adequately describes the data, we check for proportional hazards, influential observations, and non-linearity. Proportional hazards were checked using cox.zph() in the survival package in R which checks the Schoenfeld residuals [Sch82, R C19, The15]. A global p-value of $0.025 < 0.05$ shows us that this assumption may not be handled correctly. A graphical check in Figure 3.7 of the variables which failed the test reveal that high school education, missing features, family history and breast density don't seem to have any strong patterns with time [R C19, KK19]. A possible concern might be breast density, which shows a slight decrease until about 2 years after the index mammogram.

Influential observations were assessed using dfbetas [W+86, R C19, The15]. The cut-off for influential outliers was 1.0, but no values above 1.0 were observed. As well, non-linearity in the only numeric variable, age, was assessed using Martingale residuals, showing no concerning patterns [TGF90, R C19, The15].

The concordance for training and test set were 0.637 (95% CI 0.631, 0.643) and 0.627 (95% CI 0.621, 0.633), respectively [R C19, The15]. From Figure 3.8, we can see that there is a serious amount of overlap between the risk distributions of those with a diagnosis and those without a diagnosis, a model producing a higher concordance value would reduce the overlap, allowing for better risk stratification. The calibration was assessed at 0.96

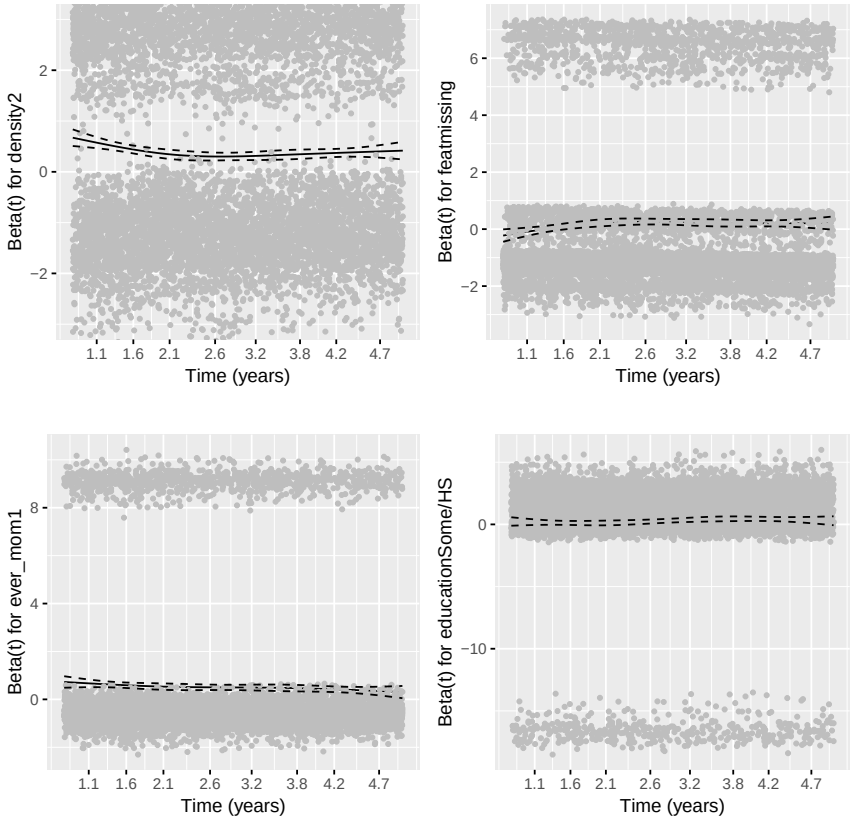

Figure 3.7: Checking the proportional hazards assumptions for the variables and variable categories that showed a significant p-value in the Schoenfeld residuals.

and 1.04 for 5 and 2-year risk, respectively, a value of 1.0 would indicate perfect agreement between observed outcomes and predictions [SVC+10, R C19, Har19]. The overall 5-year mean absolute risk was 1.61%, while the mean absolute risk for those with no diagnosis and those with a diagnosis was 1.6% and 2.0%, respectively ($p < 0.0001$) [R C19, KW52].

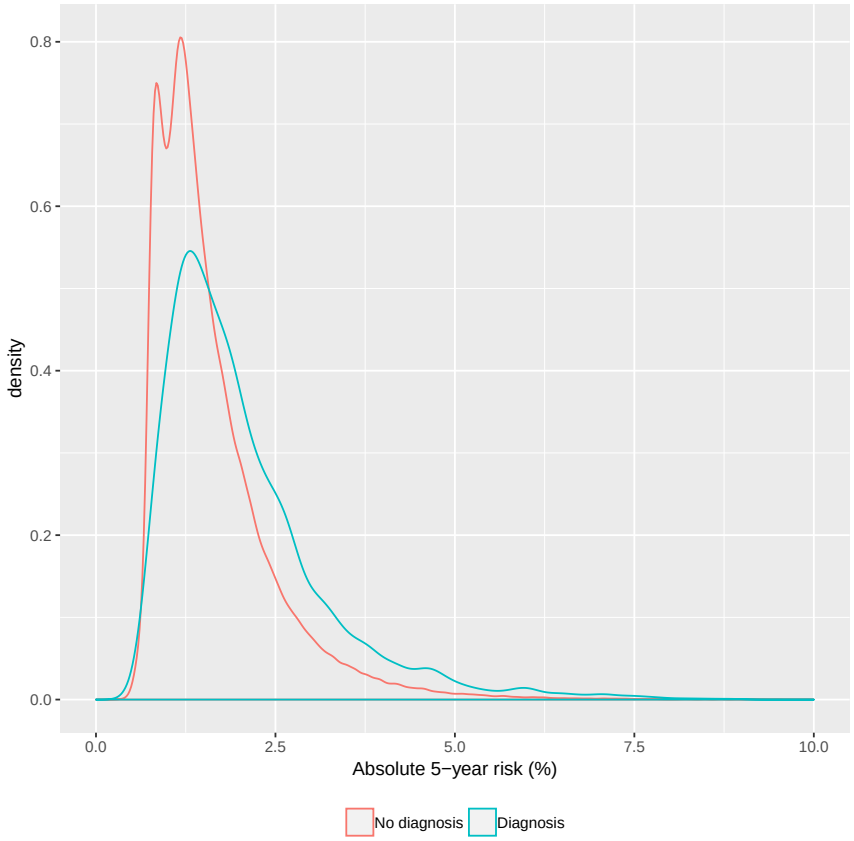

Figure 3.8: The discrimination from the Cox model is quite low, as seen in the overlap of the curves for those who had a diagnosis and those who did not.

Table 3.6 shows the absolute risk predictions on the test set for 2 and 5 year absolute and stratified risk. Stratified risk was calculated by dividing the mean risk in each group by the mean risk in the low risk group and is used to observe how much more risk an individual above low risk has relative to an individual in the low risk group. We have broken the risk predictions

into risk groups corresponding to the NICE guidelines [fCEN$^+$]. The NICE guidelines specify 10-year risk divided into general, moderate and high at <3%, 3-8% and >8%. To get 5 and 2-year risk cut-offs we divided these by 2 and 5, respectively. An additional low risk group was created since most of the risk fell in the general group. We observe that compared to the lowest risk group, the highest risk group has about 8 times more risk for 2 and 5 years. We also notice that a small proportion are in the high risk group, with most individuals classified as general risk. We can interpret the mean risk as saying that on average an individual in the moderate risk group has a 2.2% and 0.82% probability of developing breast cancer in the next 5 and 2 years, respectively.

Table 3.6: 2 and 5-year absolute risk predicted from Cox proportional hazards model

Risk group	Risk range	% at risk	Mean risk	Stratified risk
		5-Year		
low	0-0.75	3.0	0.67	1.0 (reference)
general	0.75-1.5	55	1.1	1.6
moderate	1.5-4	40	2.2	3.2
high	>4	2.5	5.3	7.8
		2-Year		
low	0-0.3	34	0.24	1.0 (reference)
general	0.3-0.6	50	0.41	1.7
moderate	0.6-1.6	14	0.82	3.4
high	>1.6	0.43	2.0	8.6

We compare this risk distribution to [ECP$^+$17], who used a logistic regression model with age, family history, BMI, estrogen use, and number of masses and microcalcifications interpreted by fully automated computer-aided detection and found a concordance of 0.71 (0.69-0.73). From Table 3.7, we note that the increase in concordance allowed them to find a low risk group, whereas our concordance lumps most of our observations into the general risk group. They also determined a larger proportion to be at a high risk. While the proportions for the groups are different, the mean risk for these groups in our model and in [ECP$^+$17] are astoundingly similar.

Since the concordance value is quite low, we move on to some alternative models to see if the performance may be improved.

Table 3.7: Comparison of 2-year risk predicted from our Cox model to [ECP+17]

Risk group	Risk range	% at risk	Mean risk	Stratified risk
[ECP+17] (c=0.71)				
low	0-0.15	10	0.12	1.0 (reference)
general	0.15-0.6	65	0.33	2.8
moderate	0.6-1.6	23	0.82	6.8
high	≥1.6	2	1.95	16.2
Our Cox model (c=0.64)				
low	0-0.15	0.37	0.14	1.0 (reference)
general	0.15-0.6	85	0.34	2.4
moderate	0.6-1.6	14	0.82	5.8
high	≥1.6	0.43	2.00	14.3

3.5 Time-dependent Cox regression model

From clinical knowledge, we know that the mammographic features can change over time. From the literature, we know that these changes are associated with an increased risk of developing breast cancer [CTRP+16]. Therefore, we fit a time-dependent Cox regression model to incorporate this changing behaviour of the mammographic features [FL99, R C19, The15]. Since the previous model measured hazard from the index screen, incorporating the more recent features from screens closer to a diagnosis and the changing behaviour may improve the model performance.

This time-varying model allows for the value of x to change over time, denoted by $x(t)$ (Equation 3.6). This means that if the features change from one screen to the next, the value of x changes. Note that this is the only change in the model equation from Equation 3.1 in the previous section. Baseline hazard, $h_0(t)$, is defined as the hazard function for someone who has a value of zero for every variable for all time [Col15].

$$h_i(t) = \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}(t)\right\} h_0(t) \tag{3.6}$$

The relative hazard e^{β_j} can be interpreted as follows. For two individuals s and r at a given time t whose values differ by one unit, with the same values of the other predictors, the relative hazard is given by Equation 3.7 below.

$$\frac{h_r(t)}{h_s(t)} = \exp[\beta_1\{x_{r1}(t) - x_{s1}(t)\} + \ldots + \beta_p\{x_{rp}(t) - x_{sp}(t)\}] \tag{3.7}$$

In order to capture the time-varying behaviour, the false positive abnormal screens and normal screens were selected. True positive and false negative screens were removed as the features on these screens are indicative of a cancer that is already present.

In order to set the data up for input into the tmerge() and coxph() functions in R, we had to take a subset of screens from the data according to a specific algorithm [R C19, The15]. All index screens were included. A subsequent screen was included if the values of at least one of the density or feature variables was different from the screen immediately preceding it. The missing features were treated as generic abnormal results, so if someone had an abnormal mammogram but the feature was missing, a subsequent abnormal with the feature present was included as there was no way to tell if the features were the same or not. Figure 3.9 shows the remaining screens used in this portion of the analysis. An individual with one screen means that they did not have any changes in their screens. A woman with 2 remaining screens had 1 change at some point during follow-up, and so on.

Figure 3.10 shows the distribution of features on the screens used in this analysis which had features present. 78% of the remaining 914,111 screens did not have features, leaving 101,441 screens with features. The relative frequency is nearly identical to when looking at the entire dataset, despite all the modifications made to get the data into the proper format. This was checked because it was unknown what the effect would be when moving from the original dataset with all screens to the dataset with only screens that were different from the previous screen.

Using the same methods as in the previous section, we fit a time-varying Cox model [FL99, R C19, The15]. We assessed goodness of fit using dfbetas and Martingale residuals, but there is no longer the assumption of proportional hazards, so we did not need to check that this time [Col15, TCA17, TGF90, BP88]. The dfbetas and Martingale residual plots are similar to the Cox proportional hazards model, so are not shown here [R C19, The15]. The resulting model is in Table 3.8. The concordance was 0.643 (95% CI 0.637, 0.649) for the training set.

We again see that 3 or more features, architectural distortion with calcification and mass with architectural distortion have large coefficients, indicative of an increase in risk. The 95% confidence intervals for the exponentiated coefficients are again quite large, making the estimates subject to

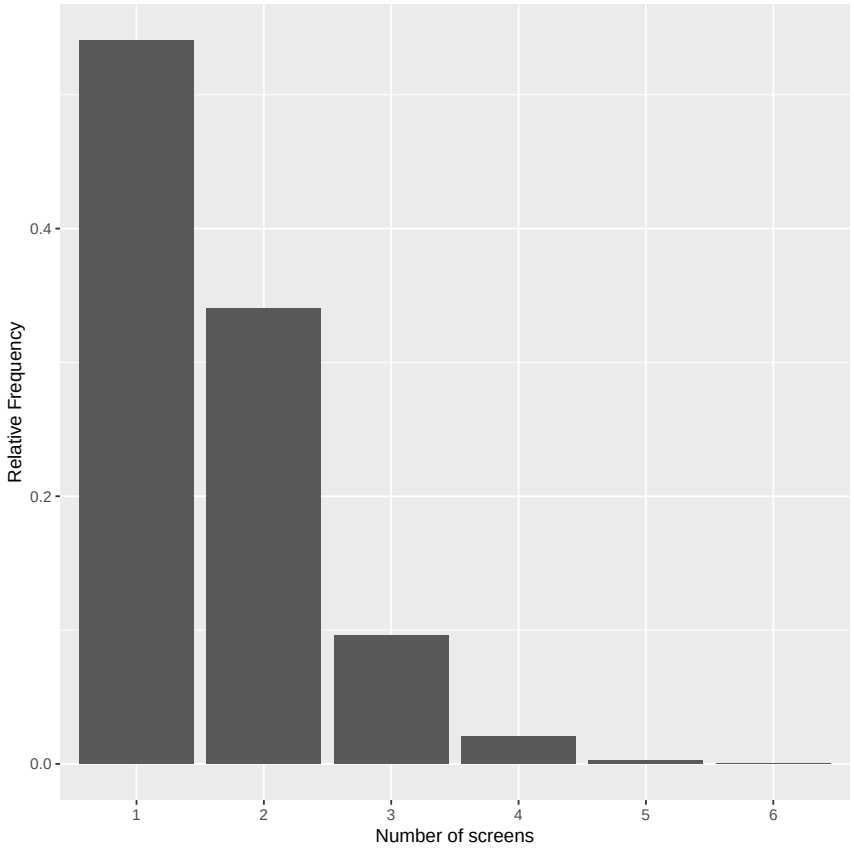

Figure 3.9: Number of screens remaining to use in the model that changed from the previous screen.

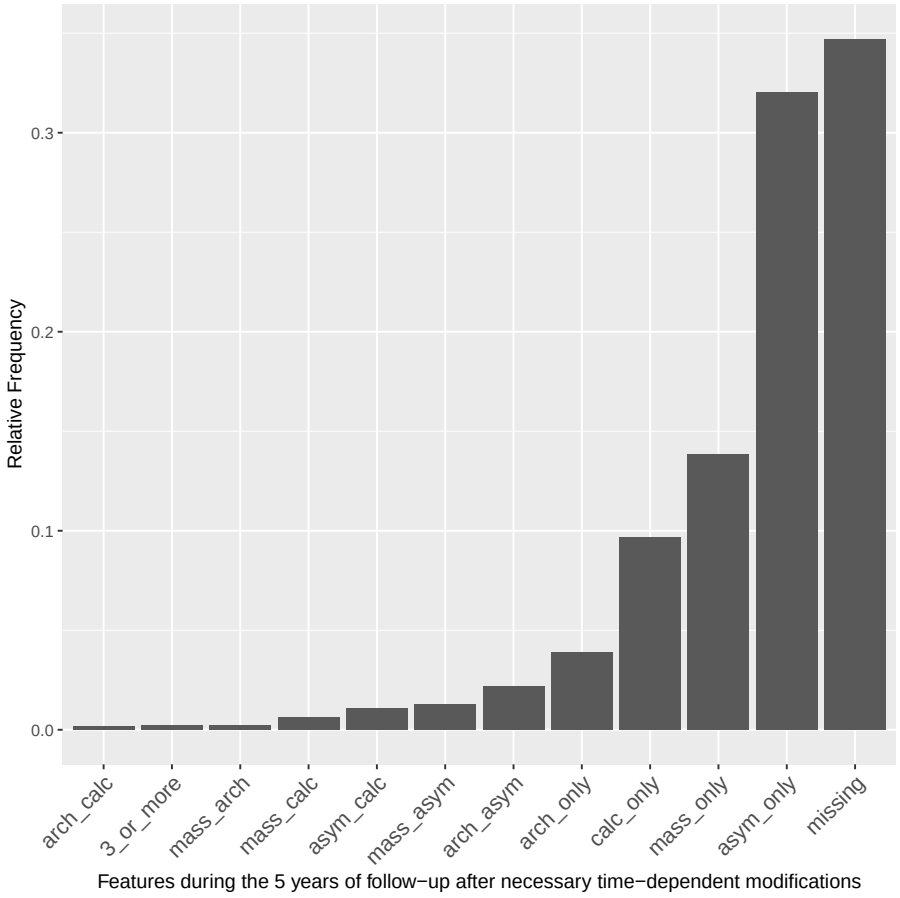

Figure 3.10: Remaining features on screens with a feature identified after formatting the data prior to fitting the Cox regression model.

Table 3.8: Cox time varying regression model

	coef	exp(coef)	p-value	2.5 %	97.5 %
age	0.03	1.04	0.00	1.03	1.04
featno_feat (ref)	0.00	1.00			
feat3_or_more	1.45	4.25	0.00	2.21	8.21
featarch_asym	0.56	1.75	0.00	1.21	2.54
featarch_calc	1.70	5.46	0.00	2.94	10.11
featarch_only	0.53	1.70	0.00	1.29	2.24
featasym_calc	0.73	2.08	0.00	1.35	3.20
featasym_only	0.39	1.48	0.00	1.33	1.65
featcalc_only	1.05	2.84	0.00	2.51	3.22
featmass_arch	1.52	4.56	0.00	2.45	8.48
featmass_asym	0.69	1.99	0.00	1.31	3.02
featmass_calc	1.05	2.86	0.00	1.83	4.47
featmass_only	0.51	1.66	0.00	1.43	1.92
featmissing	0.56	1.74	0.00	1.60	1.89
density1 (ref)	0.00	1.00			
density2	0.42	1.52	0.00	1.45	1.59
ethnicCaucasian (ref)	0.00	1.00			
ethnicFirst Nations	0.21	1.23	0.04	1.01	1.50
ethnicAfrican	0.22	1.25	0.31	0.81	1.92
ethnicEast Asian	-0.04	0.96	0.34	0.89	1.04
ethnicOther	0.03	1.03	0.29	0.98	1.08
ethnicSouth Asian	-0.32	0.73	0.00	0.62	0.84
Ever_Estrogen0 (ref)	0.00	1.00			
Ever_Estrogen1	0.16	1.17	0.00	1.10	1.24
Ever_Biopsy0 (ref)	0.00	1.00			
Ever_Biopsy1	0.35	1.42	0.00	1.33	1.51
ever_mom0 (ref)	0.00	1.00			
ever_mom1	0.51	1.66	0.00	1.55	1.78
ever_sis0 (ref)	0.00	1.00			
ever_sis1	0.41	1.50	0.00	1.38	1.64
educationNo HS (ref)	0.00	1.00			
educationSome/HS	0.27	1.30	0.00	1.18	1.44
educationSome/Co	0.31	1.36	0.00	1.23	1.50

much variability. While slightly better than the previous model, since the concordance is again quite low, namely too low to be used in individualized risk prediction, we abandon the Cox models and move on to other models to see if there is a better option available.

3.6 Accelerated Failure Time model

Accelerated failure time (AFT) models are a parametric alternative to the Cox regression model [Wei92, KP11, Col15, Moo16]. While they require stronger assumptions to be met, they can be used to predict the length of time to an event. The model can be described as in Equation 3.8, where $\eta_i = \alpha_1 x_{1i} + \ldots + \alpha_p x_{pi}$ is the linear combination of predictors, $h_0(t)$ is the baseline hazard function, and t is time. The baseline hazard can be thought of as the hazard of diagnosis at time t for someone who has the baseline value of zero for all predictors.

$$h_i(t) = \exp(-\eta_i) h_0(\frac{t}{\exp(\eta_i)}) \tag{3.8}$$

A somewhat simpler way to view this model is the log-linear form as in Equation 3.9, where σ is the scale parameter, μ is the intercept, and ϵ is the error which has a certain distribution depending on the type of model chosen. Interpreting the coefficients, we get that a positive α increases survival time, while negative values decrease survival time. Note that this is an exactly opposite interpretation to the previous Cox models, where a positive coefficient was indicative of an increase in risk and negative values decrease the risk.

$$\log T_i = \mu + \alpha_1 x_{1i} + \ldots + \alpha_p x_{pi} + \sigma \epsilon_i \tag{3.9}$$

Using the same training and testing set and variable selection method as before, we fit an accelerated failure time model for several different types of AFT models, to determine which will satisfactorily fit the data [R C19, The15]. According to Table 3.9, the Weibull, lognormal or loglogistic are best suited as they have the lowest AIC (Akaike Information Criterion) [SIK86]. AIC was calculated with R's AIC() function and it is merely a way to compare models where the smaller the AIC, the better the fit [R C19].

We first examine how well the data follow a Weibull distribution. We can do this graphically, by plotting the log of survival time and the log-log of the estimated survival (Figure 3.11) [Moo16]. The Kaplan-Meier estimate of survival is computed using the survfit() function in the survival

Table 3.9: AIC of AFT models

Model	AIC
Lognormal	102,521
Loglogistic	102,604
Weibull	102,612
Exponential	103,952
Gaussian	105,298

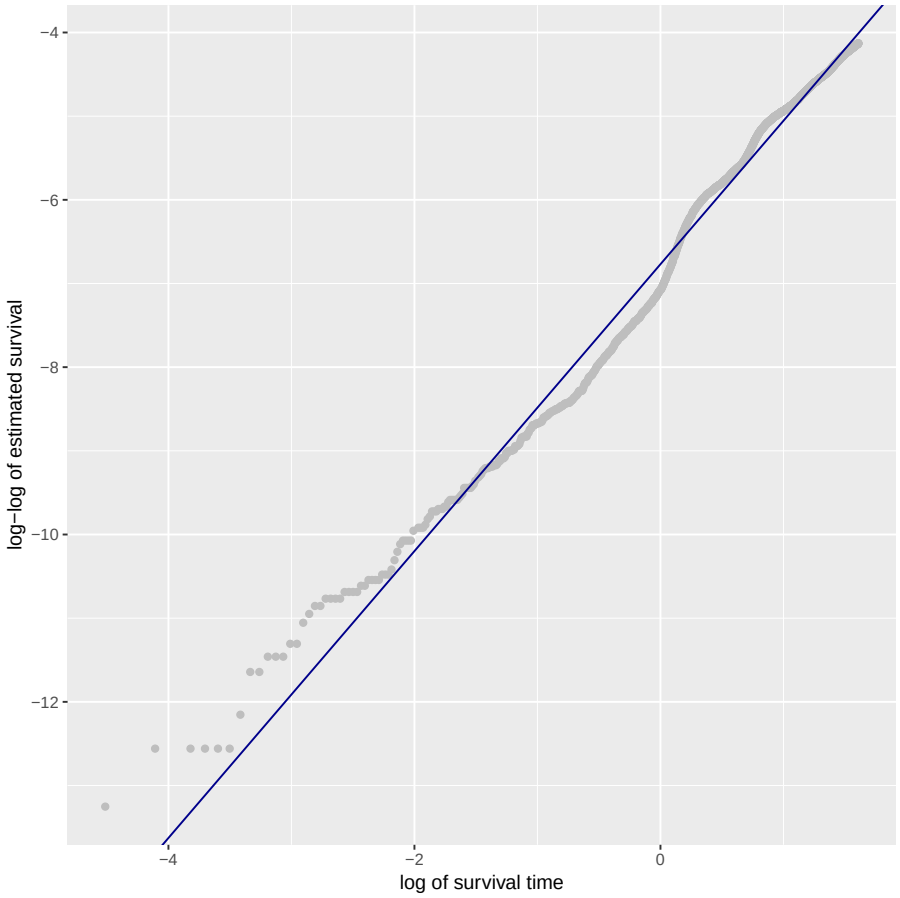

Figure 3.11: Plot of the log-log transformation of the Kaplan-Meier survival vs the log transformation of time. Used to assess whether the Weibull distribution is appropriate for the data.

package in R [R C19, The15]. The Kaplan-Meier survival estimates survival empirically without taking into account any predictors [KM58]. The log-log transformation of the survival estimate is plotted vs the log transformation of the time variable and a regression line through the points is fitted (Figure 3.11). The closer the points are to the straight line, the better. There is some concern in the bottom left of the graph where the points deviate from the line, but as this only concerns a few points, we proceed with this model.

We compared the plotted deviance residuals as a model diagnostic and found the distributions were reasonably comparable (Figure 3.12) [R C19, The15]. We also examined the dfbeta residuals for influential points, and

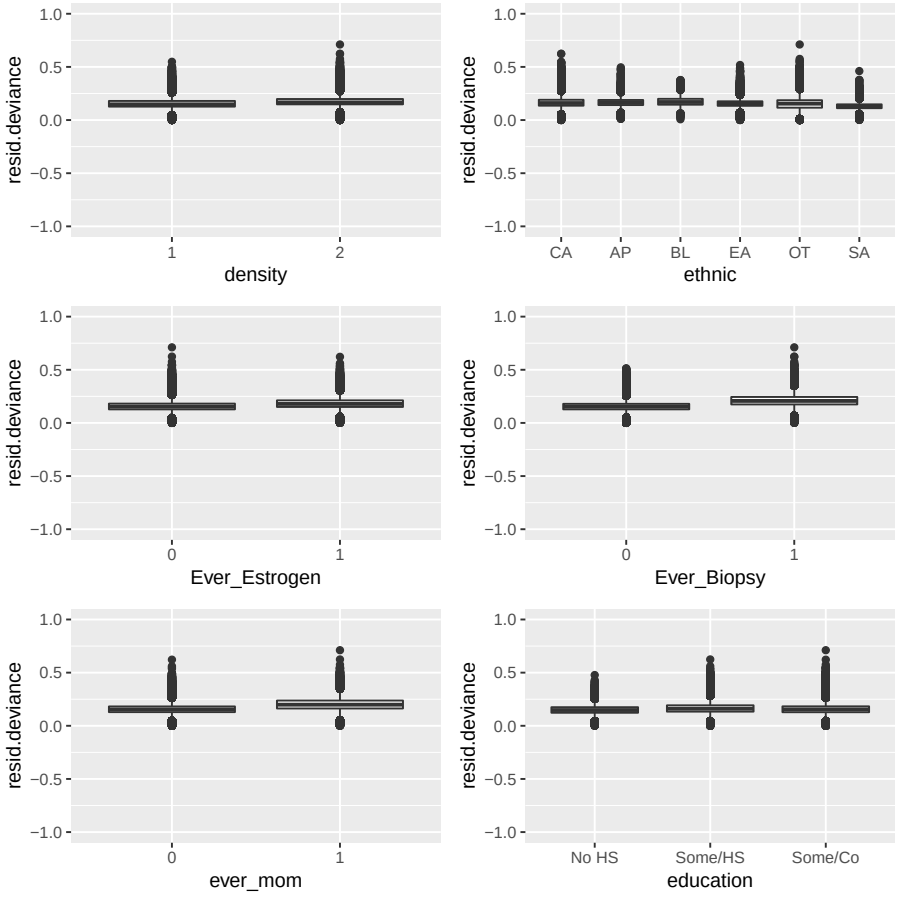

Figure 3.12: The deviance residuals of select variables in the AFT model. The distributions are similar.

found several influential points [W+86, R C19, The15].

Table 3.10 shows the coefficients. We see that the survival time for individuals with architectural distortion and calcification ('arch_calc') have a decrease in survival time by a factor of $\exp(-0.64)$, the most devastating impact of all the features. This is followed by the mass-architectural distortion combination and calcifications alone.

Table 3.11 shows some of the risk percentiles predicted by the model on test data. We decided to hold several variables constant and vary the mammographic features to examine the time to event prediction for the various features. Holding the other variables constant at age 40, Caucasian, mother who had breast cancer, high school education, has never used estrogen or had a biopsy, and a low breast density, we see how different features alter the time prediction. The median time to a breast cancer diagnosis for all features is quite high, and indicates that these women will not likely be diagnosed with breast cancer in their lifetime. However, women in the 1st percentile, or 1% of women, with architectural distortion and calcification on their mammogram, are predicted to have a breast cancer diagnosis in 2 years (95% CI: 1.3, 2.7). This agrees with the findings from the proportional hazards model, as there is a large majority of women who are not likely to be diagnosed with breast cancer, but a small proportion are at risk of diagnosis within a few years. The training concordance of this model was 0.637 (0.631, 0.643), which is again too low to be used clinically.

3.7 Poisson regression model

We now turn to Poisson regression. Poisson regression is used when the response variable is a count of a relatively rare event [MMVR12]. Since out of 813,280 individuals only 11,166 (1.4%) had a diagnosis, the event can be considered as relatively rare. The point of using this model is to see if our risk factors influence the diagnosis count. The model is shown in Equation 3.10 where the βs are the coefficients and $g(\mu)$ is the link function. A positive coefficient indicates that the variable increases the number of diagnoses while a negative coefficient decreases the number of diagnoses.

$$g(\mu_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{3.10}$$

Prior to fitting the model, a Lexis split was run on the data [PC+11, R C19]. The Lexis split allows us to follow individuals on different time scales. Several different splits and combinations were tried, including age groups, calendar time and time since false positive. It was determined that

Table 3.10: Weibull AFT model

	coef	exp(coef)	p-value	2.5 %	97.5 %
(Intercept)	5.93	375.53	0.00	324.31	434.85
age	-0.02	0.98	0.00	0.97	0.98
featno_feat (ref)	0.00	1.00			
feat3_or_more	-0.42	0.65	0.05	0.43	1.00
featarch_asym	-0.04	0.96	0.65	0.79	1.16
featarch_calc	-0.64	0.53	0.00	0.36	0.78
featarch_only	-0.04	0.96	0.57	0.85	1.10
featasym_calc	-0.35	0.70	0.00	0.57	0.87
featasym_only	0.06	1.06	0.03	1.00	1.12
featcalc_only	-0.45	0.64	0.00	0.60	0.68
featmass_arch	-0.49	0.61	0.02	0.41	0.92
featmass_asym	-0.18	0.84	0.13	0.67	1.05
featmass_calc	-0.30	0.74	0.03	0.56	0.97
featmass_only	-0.07	0.93	0.06	0.87	1.00
featmissing	-0.12	0.89	0.00	0.85	0.92
density1 (ref)	0.00	1.00			
density2	-0.26	0.77	0.00	0.75	0.79
ethnicCaucasian (ref)	0.00	1.00			
ethnicFirst Nations	-0.14	0.87	0.03	0.76	0.99
ethnicAfrican	-0.03	0.97	0.84	0.72	1.31
ethnicEast Asian	0.05	1.05	0.04	1.00	1.11
ethnicOther	0.01	1.01	0.48	0.98	1.04
ethnicSouth Asian	0.17	1.19	0.00	1.08	1.31
Ever_Estrogen0 (ref)	0.00	1.00			
Ever_Estrogen1	-0.10	0.90	0.00	0.86	0.94
Ever_Biopsy0 (ref)	0.00	1.00			
Ever_Biopsy1	-0.22	0.80	0.00	0.77	0.83
ever_mom0 (ref)	0.00	1.00			
ever_mom1	-0.32	0.72	0.00	0.69	0.76
ever_sis0 (ref)	0.00	1.00			
ever_sis1	-0.22	0.80	0.00	0.75	0.85
educationNo HS (ref)	0.00	1.00			
educationSome/HS	-0.13	0.88	0.00	0.83	0.94
educationSome/Co	-0.17	0.84	0.00	0.79	0.90

Table 3.11: Time predictions for different percentiles

	feature/percentile	1st	2.5th	10th	50th
1	no_feat	4.4	8.0	20.4	69.9
2	3_or_more	2.5	4.6	11.8	40.3
3	arch_asym	4.5	8.2	20.8	71.2
4	arch_calc	2.0	3.7	9.5	32.4
5	arch_only	4.3	7.9	19.9	68.2
6	asym_calc	3.3	6.1	15.4	52.7
7	asym_only	4.8	8.7	22.1	75.5
8	calc_only	2.8	5.1	12.9	44.1
9	mass_arch	2.3	4.3	10.8	37.0
10	mass_asym	3.3	6.1	15.4	52.7
11	mass_calc	3.2	5.9	15.1	51.6
12	mass_only	4.3	7.8	19.9	68.1
13	missing	4.0	7.3	18.4	63.1

splitting on time since false positive and 10-year age groups were most relevant to the problem and the model was adjusted for these variables. We note that architectural distortion with calcification and mass with architectural distortion have the largest impact on the diagnosis count (Table 3.12). The log link was used as this ensures that the predicted values will not be negative, which is an impossible count option. An offset of log-time was used to account for differing amounts of time contributed to each time band. The model was assessed for goodness of fit using deviance and influential diagnostics (Figure 3.13) [R C19].

The plots in Figure 3.13 show that the Poisson model is not a good fit as there is a pattern in the residuals [PS86, TC90]. Since the pattern involves a lot of zero values, a transformation of the data would result in a large amount of transformed zeros, which may not address the problem. Over and underdispersion were assessed using the dispersiontest() in R's AER package and a non-significant p-value meant the null hypothesis of the true dispersion equalling 1 could not be rejected [DL89, R C19, KZ08]. However, the deviance divided by the degrees of freedom showed a value of 0.29, which is not close to 1 and thus this model is underdispersed. Various combinations of Lexis split were tried and the resulting models fitted, but none passed the diagnostic tests. Predictions made from this model would not be reliable. Future work could examine whether a generalized or zero-inflated Poisson model can accommodate for the problems in this model [CF92, Lam92].

Table 3.12: Poisson model

	coef	exp(coef)	p-value	2.5 %	97.5 %
(Intercept)	-11.96	0.00	0.00	0.00	0.00
Age40-50	0.00	1.0			
Age50-60	0.25	1.29	0.00	1.23	1.35
Age60-70	0.61	1.84	0.00	1.74	1.94
Age70+	0.90	2.47	0.00	2.31	2.64
featno_feat (ref)	0.00	1.00			
feat3_or_more	0.64	1.89	0.02	1.04	3.12
featarch_asym	-0.02	0.98	0.89	0.76	1.25
featarch_calc	0.87	2.38	0.00	1.34	3.86
featarch_only	0.06	1.06	0.51	0.89	1.25
featasym_calc	0.51	1.66	0.00	1.25	2.16
featasym_only	-0.17	0.84	0.00	0.78	0.91
featcalc_only	0.62	1.85	0.00	1.70	2.02
featmass_arch	0.85	2.33	0.00	1.37	3.67
featmass_asym	0.38	1.46	0.01	1.10	1.91
featmass_calc	0.31	1.36	0.12	0.90	1.95
featmass_only	0.06	1.06	0.25	0.96	1.17
featmissing	0.24	1.28	0.00	1.21	1.35
density1 (ref)	0.00	1.00			
density2	0.35	1.42	0.00	1.37	1.48
ethnicCaucasian (ref)	0.00	1.00			
ethnicFirst Nations	0.20	1.22	0.02	1.02	1.44
ethnicAfrican	0.01	1.01	0.95	0.65	1.49
ethnicEast Asian	-0.07	0.94	0.06	0.87	1.00
ethnicOther	-0.06	0.94	0.00	0.90	0.98
ethnicSouth Asian	-0.31	0.74	0.00	0.64	0.83
Ever_Estrogen0 (ref)	0.00	1.00			
Ever_Estrogen1	0.18	1.20	0.00	1.14	1.26
Ever_Biopsy0 (ref)	0.00	1.00			
Ever_Biopsy1	0.38	1.46	0.00	1.38	1.54
ever_mom0 (ref)	0.00	1.00			
ever_mom1	0.45	1.57	0.00	1.48	1.67
ever_sis0 (ref)	0.00	1.00			
ever_sis1	0.39	1.47	0.00	1.36	1.58
educationNo HS (ref)	0.00	1.00			
educationSome/HS	0.22	1.25	0.00	1.15	1.36
educationSome/Co	0.23	1.26	0.00	1.16	1.37

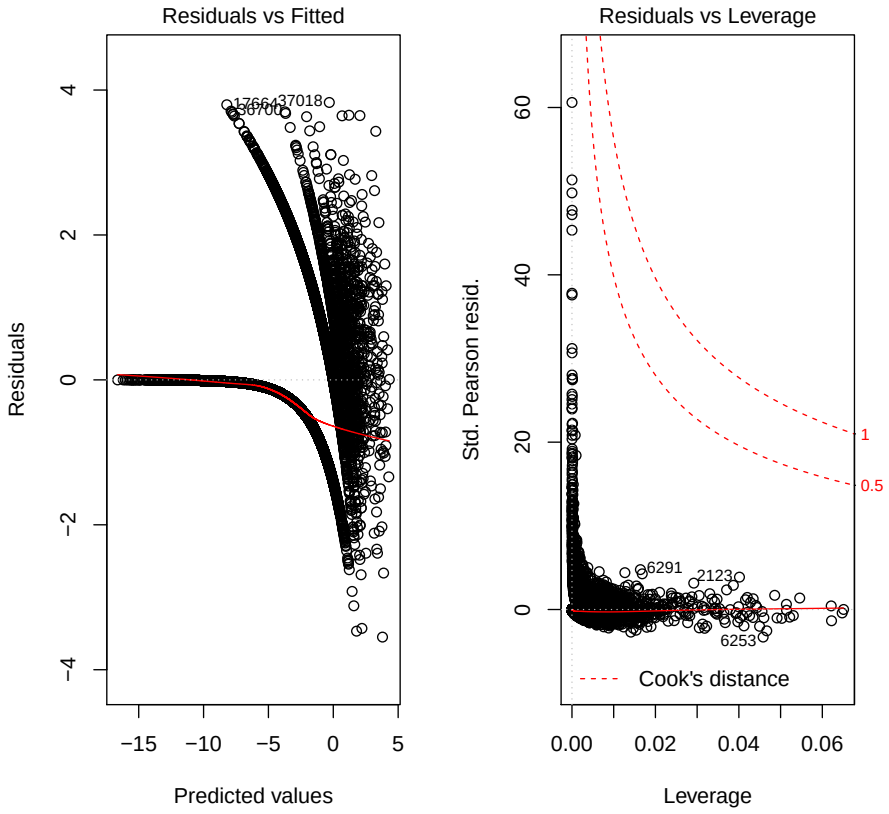

Figure 3.13: Diagnostics of Poisson model.

Chapter 4

Application

In this chapter, we refit some of the published models on our data and make comparisons to our models of Chapter 3, focusing on the concordance, risk distributions and calibration [R C19, The15, MIG12, Har19]. To that end, in this chapter we did not perform variable selection as the goal was to see how the combinations of predictors from published models perform on our data, and not to find them.

4.1 Concordance comparison to other models

4.1.1 Cox proportional hazards model concordance comparison

In this section we compare the concordance from our models to the published models. We begin with the model developed by [TCSB$^+$08], which included only age, breast density and ethnicity and achieved a concordance of 0.670. When fitting this model to our data, we find that the concordance is only 0.609 (Table 4.1).

Table 4.1: Training and testing concordance values for published models, published model on our data, and the same model with mammographic features added

Model	Published	Our data	Our data with features
Training concordance			
[TCSB$^+$08]	0.660 (0.650, 0.670)	0.609 (0.603, 0.615)	0.618 (0.612, 0.623)
[TCZK05] no BD	0.670 (0.650, 0.680)	0.620 (0.614, 0.626)	0.628 (0.622, 0.634)
[TCZK05] with BD	0.680 (0.660, 0.700)	0.632 (0.626, 0.638)	0.638 (0.632, 0.645)
Testing concordance			
[TCSB$^+$08]	NA	0.608 (0.599, 0.618)	0.614 (0.604, 0.624)
[TCZK05] no BD	NA	0.607 (0.597, 0.617)	0.614 (0.604, 0.624)
[TCZK05] with BD	NA	0.620 (0.610, 0.630)	0.625 (0.615, 0.635)

Aside from fitting the models on different training populations, two differences may account for some of the concordance difference. First, they had 4-category breast density, while we have 2-category. Second, the ethnicity categories were slightly different as they could not be recoded in the

same way. They had 'Caucasian', 'African American', 'Latina', 'Asian' and 'Other'. We combined 'East Asian' and 'South Asian' into 'Asian' and 'First Nations' with 'Other', but we could not create a 'Latina' group as there was no 'Latina' category in our ethnicity variable. We fit the model using our previously coded ethnicity and compared it to the results with the adjusted coding and found that there was little difference in our model concordance with the recoding of the ethnicity.

Next, we move on to [TCZK05], which fit a Cox Gail model to their data with and without the addition of breast density. They had a training set of 81,777 women with 955 diagnoses, compared to our training set of 569,296 women and 7,859 diagnoses, which indicate a cancer rate of 1.2% and 1.4%, respectively. This model uses age at menarche, number of biopsies, age, age at first child, and number of relatives with breast cancer, with interactions between family history and age at first birth and age and biopsy. In order to make a more fair comparison, some alterations to the coding had to be made to create similar levels and coding style as used in the paper. Age at menarche was recoded to a categorical variable with '$\leq$12', '12-13' and '$\geq$14'. Age at first child was converted to categorical with '$\leq$20', '20-24', '25-29', and '30+'. For women with no children, we set their age at first child to '25-29', as done in the original Gail model [GBB$^+$89] and the models we are comparing to [TCZK05]. We converted family history into a categorical variable with three levels, '0', '1', '2+', representing the number of first degree female relatives who have had a diagnosis of breast cancer. We again used our 2-category density and our coding of ethnicity.

We found that the published models did not perform as well when refitted to our data as seen in our training concordance that is lower than the concordance values published (Table 4.1). The published training models had 95% confidence intervals of (0.65, 0.67), (0.65, 0.68) and (0.66, 0.70) for the [TCSB$^+$08] model, [TCZK05] model and [TCZK05] with mammographic density, respectively, while our confidence intervals didn't even reach 0.64. Adding the mammographic features resulted in improvement in the concordance for all models on the training set, but only by about 0.008, while adding breast density increased the concordance by about 0.01. For comparison, in our Cox proportional hazards model, age increases concordance by 0.1, family history by 0.01, breast density by 0.01, mammographic features by 0.006, and the remaining predictors by less than 0.002, when all the other predictors are in the model. When mammographic features are the only predictor in the Cox model, they increase concordance by 0.04, which is more than the 0.02 increase from a mother or sister with a history of breast cancer. We conclude that the addition of mammographic features was

statistically significant, as seen in comparing the models with and without features using anova() in R [R C19], and shows a small increase in concordance. While the overall concordance is low, this increase in concordance may be of clinical interest.

Examining the testing concordance for the 6 models (3 models with mammographic features and the same 3 without mammographic features) could be more enlightening as this gives us a sense of how the model would predict for future observations, since the concordance values are estimated on data not used in the fitting of the model [R C19, The15]. In Table 4.1 we see that the test concordance with the mammographic features is higher than without the features, by about 0.006. Therefore, the addition of mammographic features does appear to improve the concordance. We also note that the simple model's [TCSB$^+$08] test concordance is in the confidence interval of the training concordance, but the other two models' test concordance falls out of the training concordance confidence interval [TCZK05], suggesting overfitting of the latter two models.

4.1.2 Cox proportional hazards model comparison of absolute risk predictions

We now examine the 5-year absolute risk predictions. Predicted 5-year absolute risk is the probability of getting a breast cancer diagnosis within 5 years. We use the model fitted on the training set to predict the absolute risk for the test set. We compare the 5-year absolute risk predicted for the test set for these 6 models to our Cox model shown in Figure 3.8 (Figure 4.1). In this figure, we are plotting on the x-axis the predicted 5-year absolute risk from our Cox model. The 5-year predicted absolute risk from the published models is on the y-axis, where the left column y-axes in the figure contain the published models, while the right column y-axes contain the published models with the addition of mammographic features. The 5-year absolute risk was calculated using predictSurvProb() function in the pec package in R [R C19, MIG12]. This function calculates the survival probability at specific time points. The resulting predictions were subtracted from 1 to get the risk, and multiplied by 100% to get the risk value as a percentage. The dotted red line represents the line of best fit, calculated using a simple linear regression with our model risk predictions as the predictors and the published models risk predictions as the response. The solid blue line is the 45° line. If the predictions from our model and a given published model were exactly the same, all the points in the plot would fall on this blue line. In other words, it is a visual measure of how different the predicted values

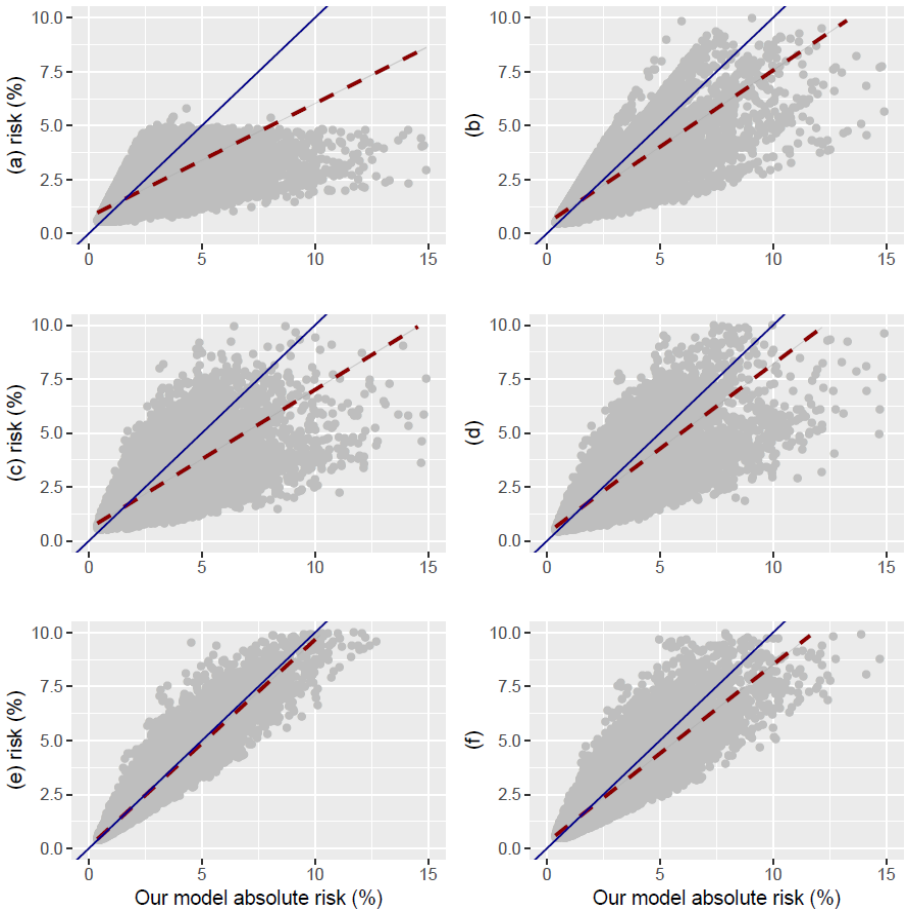

Figure 4.1: (a) [TCSB$^+$08], (b) [TCSB$^+$08] with features, (c)[TCZK05] no BD, (d) [TCZK05] no BD, with features, (e) [TCZK05] with BD, (f) [TCZK05] with BD and features. Comparison of the absolute risk predicted by our model and the published models. The blue line is the 45° line. The red dashed line is the line of best fit.

are for each individual in the test set.

Examining Figure 4.1 we note that the [TCZK05] model with mammographic density included, graph (e), performs most similarly in terms of predicted absolute risk compared with the proportional hazards model of Section 3.4, as seen by the tight grouping of predictions in (e) to the 45° line and how closely the line of best fit matches the 45° line. The least similar appears to be (a), the [TCSB$^+$08] model with only age, ethnicity and density. We note that this model predicts most risk to be below 5%. Why the predicted risk only falls from 0-5% when using these predictors leads one to question whether the risks in the 5-15% range are exaggerated or if these 3 variables are not enough for individual risk stratification.

Upon comparing the risk distribution of these models (Table 4.2) by breaking risks into 4 risk groups, we note that most of the individuals fall in the general and moderate risk groups. When comparing these proportions between the models with and without features, there does not seem to be much, if any, increase in risk stratification when features are included. Despite using different variables, the results are similar to our model in Table 3.6.

Repeating the process for 2-year predicted risk shows similar behaviour, with (a) again not predicting hardly anyone to be at a high risk and model (e) predicting most similarly to our model. The risk distributions held the majority of observations in the general group, however, the low risk group held roughly 30% of the observations and the moderate groups held about 14%. These are similar to the findings of Table 3.6.

Subsetting the test data to only those with complete follow-up, i.e. those with a diagnosis or those with a follow-up time of 5 years, showed no differences in the absolute risk plots and risk distribution. This subset held 184,007 observations with 3,307 diagnoses, compared with 243,984 observations and 3,307 diagnoses in the full test set.

4.1.3 Time-dependent Cox model concordance comparison

We follow a similar process and fit these models using time varying breast density and features (Table 4.3). For this, we take the published Cox models re-fitted to our data and convert it into a time varying model following the same techniques used in Section 3.5, including allowing for time varying breast density and mammographic features. We note that there is not much difference in concordance, as they are all in the 0.61-0.64 range. However, we do see an improvement in concordance with the addition of mammographic features of about 0.01. As well, the simple model [TCSB$^+$08]

Table 4.2: Risk distribution for published models with and without mammographic features added.

Risk group	Risk range	% at risk	Mean risk	Stratified risk
Without features				
(a) [TCSB$^+$08]				
low	0-0.75	1.0	0.6	1.0 (reference)
general	0.75-1.5	51.6	1.2	1.8
moderate	1.5-4	46.7	2.1	3.3
high	>4	0.8	4.4	6.7
(c) [TCZK05]				
low	0-0.75	0.8	0.7	1.0 (reference)
general	0.75-1.5	56.6	1.2	1.7
moderate	1.5-4	41.0	2.2	3.2
high	>4	1.6	4.9	7.2
(e) [TCZK05] with BD				
low	0-0.75	4.3	0.7	1.0 (reference)
general	0.75-1.5	53.4	1.1	1.6
moderate	1.5-4	39.8	2.2	3.2
high	>4	2.6	5.3	7.8
With features				
(b) [TCSB$^+$08]				
low	0-0.75	1.4	0.7	1.0 (reference)
general	0.75-1.5	53.4	1.1	1.7
moderate	1.5-4	43.7	2.1	3.3
high	>4	1.5	5.0	7.6
(d) [TCZK05]				
low	0-0.75	3.1	0.7	1.0 (reference)
general	0.75-1.5	53.0	1.1	1.7
moderate	1.5-4	41.9	2.2	3.2
high	>4	2.0	5.0	7.3
(f) [TCZK05] with BD				
low	0-0.75	1.3	0.7	1.0 (reference)
general	0.75-1.5	57.0	1.1	1.7
moderate	1.5-4	39.5	2.2	3.2
high	>4	2.2	5.2	7.7

again has a lower concordance than the other two models, but the addition of mammographic features brings the concordance into the same range as the more complicated models, and uses only four predictors to do it.

Table 4.3: AIC of our time varying Cox model and the time varying Cox models using the published predictors.

	Model	AIC	training c	2.5%	97.5%
	Our time varying Cox model	203,702	0.643	0.637	0.649
(a)	[TCSB$^+$08]	204,657	0.612	0.606	0.618
(b)	[TCSB$^+$08] with features	204,166	0.625	0.619	0.631
(c)	[TCZK05] no BD	204,474	0.620	0.614	0.626
(d)	[TCZK05] no BD, features	203,980	0.632	0.626	0.638
(e)	[TCZK05] with BD	204,152	0.634	0.628	0.640
(f)	[TCZK05] with BD, features	203,696	0.644	0.638	0.650

4.1.4 AFT model concordance comparison

We now take the 3 published models and use them to fit AFT models with and without mammographic features. For this, we are still using the published model set-up (predictors and interactions), but this time we are using them in an AFT model with the Weibull baseline hazard distribution, as in Section 3.6. With respect to the AIC in Table 4.4, we note that the models have a similar AIC, although our model is among the lowest two. With respect to the concordance, which was calculated using the concordance() function in the survival package in R [R C19, The15], we note that the c values are again in the 0.61-0.64 range and that mammographic features increase the concordance by about 0.008, on average.

Table 4.4: AIC of our AFT model and the AFT models using the published predictors.

	Model	AIC	training c	2.5%	97.5%
	Our AFT model	103,081	0.637	0.631	0.643
(a)	[TCSB$^+$08]	103,800	0.609	0.603	0.615
(b)	[TCSB$^+$08] with features	103,552	0.618	0.612	0.624
(c)	[TCZK05] no BD	103,574	0.620	0.614	0.626
(d)	[TCZK05] no BD, features	103,334	0.628	0.622	0.634
(e)	[TCZK05] with BD	103,293	0.632	0.626	0.638
(f)	[TCZK05] with BD, features	103,076	0.638	0.632	0.644

Overall, the concordance values among the Cox proportional hazards, time varying Cox and AFT models do not go past 0.64 and are in fact quite similar between models. Since we have compared different models, including different predictors with some models having interactions and some without, as well as different types of models, and found similar results among all of them, it may not be possible to improve the concordance using these data. To be used clinically, we would prefer to have concordances above 0.7, which we have not seen with these data and these models.

4.2 Calibration comparison to other models

We now examine the calibration. Calibration looks at the agreement between predictions and observations. In other words, if our model predicts a 20% probability of a breast cancer diagnosis for some individuals, then we should see 20/100 of those individuals diagnosed with breast cancer. Further, a perfectly calibrated model will have a slope of 1 and an intercept of 0 when plotting the observed outcomes and predictions [S$^+$19].

4.2.1 Cox proportional hazards model calibration comparison

First we check the calibration on the Cox model developed in Section 3.4. Figure 4.2 shows the calibration on the training and test set, on the left and right respectively for 2, 4 and 5 year survival, on the top, middle and bottom rows, respectively. Using calPlot() in R, survival probabilities were grouped into quantiles and the averages for the predicted and observed proportions were calculated [R C19, MIG12]. The blue line is the linear regression line of best fit. The red line is a flexible calibration line fitted to the data. The black line is a 45° line which is the ideal calibration. The closer the other two lines are to the black line, the better the calibration [S$^+$19].

The calibration appears satisfactory for the training and test set, even while a larger standard error is present for the test set. A larger standard error on the test set is expected, as the test set observations were not used in the fitting of the model. Calibration on the training set is not particularly informative even with cross-validation measures, thus we examine it only as a baseline to compare with. Since the red and blue lines overlap with the black line, the training calibration is 1, as it should be [S$^+$19]. The bottom left corner of the test plots suggest that we are overestimating risk for the high risk individuals.

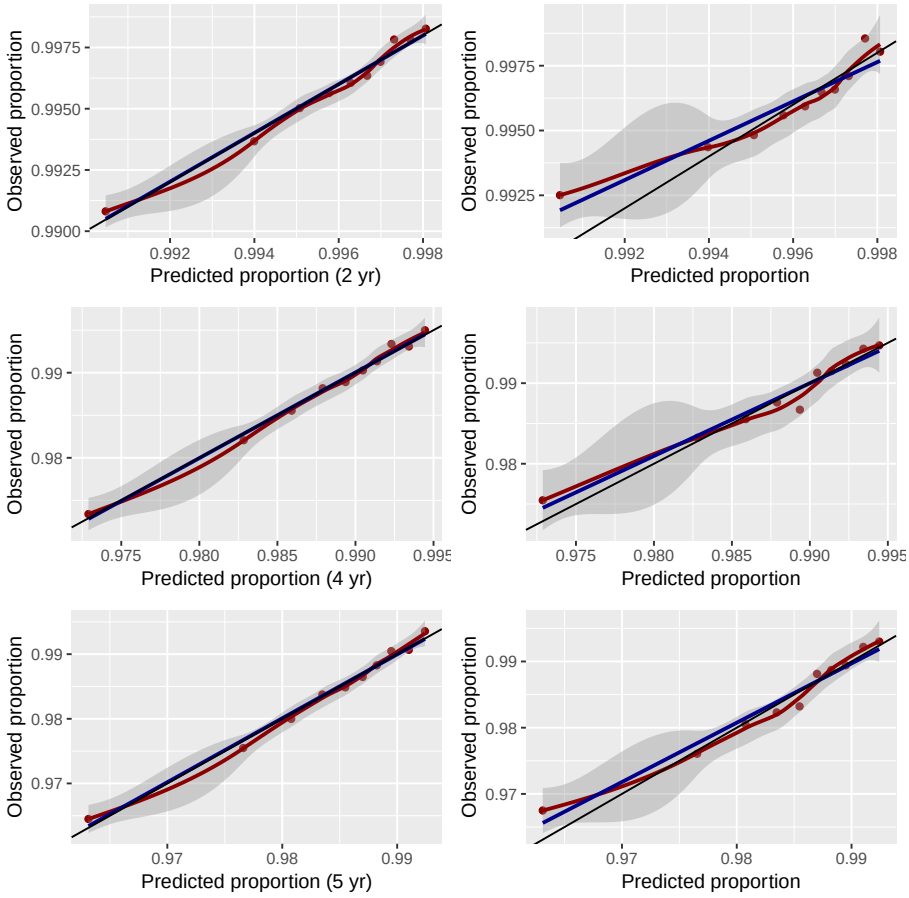

Figure 4.2: Calibration of our Cox model for training set (left) and testing set (right) for 2, 4 and 5 year survival. The blue line is the linear model line. The red line is a flexible calibration line fitted to the data. The black line is a 45° line, the ideal calibration.

Figure 4.3 shows the calibration on the test set to the comparison models for 5 year risk. We note that the calibration appears satisfactory for the higher survival quantiles, or individuals with lower risk, but the lower survival quantiles, or individuals at higher risk, show that the calibration is not ideal. We see this in graphs (c)-(f) where the black line is not inside the shadowed area. Models (a) and (b) appear to have the best calibration graphically and are most similar to our training calibration, suggesting a lack of overfitting.

To quantify the calibration, we examine the coefficients of the blue line in Table 4.5. The ideal line would have a slope of 1.0 and an intercept of 0. We note that the models appear to be satisfactorily calibrated, though the

Table 4.5: Comparisons of Cox model calibration on test set

	model		coef	2.5 %	97.5 %
0	ideal	intercept	0.00		
		slope	1.00		
(a)	[TCSB$^+$08]	intercept	-0.01	-0.10	0.07
		slope	1.02	0.93	1.10
(b)	[TCSB$^+$08] with features	intercept	0.04	-0.05	0.14
		slope	0.96	0.86	1.05
(c)	[TCZK05] no BD	intercept	0.11	-0.03	0.24
		slope	0.89	0.75	1.03
(d)	[TCZK05] no BD, features	intercept	0.11	-0.03	0.25
		slope	0.89	0.75	1.03
(e)	[TCZK05] with BD	intercept	0.14	0.01	0.28
		slope	0.85	0.72	0.99
(f)	[TCZK05] with BD, features	intercept	0.12	-0.01	0.25
		slope	0.87	0.74	1.01
	our model	intercept	0.10	-0.04	0.24
		slope	0.90	0.75	1.04

slopes are a bit on the lower side. A calibration coefficient much less than 1 is an indicator of overfitting [RACOO17]. Except for model (a), our point estimates are lower than 1, but the confidence intervals all contain 1, except for model (e), which has an upper bound of 0.99. Thus, the calibration measures are reasonable. The amazing calibration in model (a) suggests a lack of overfitting.

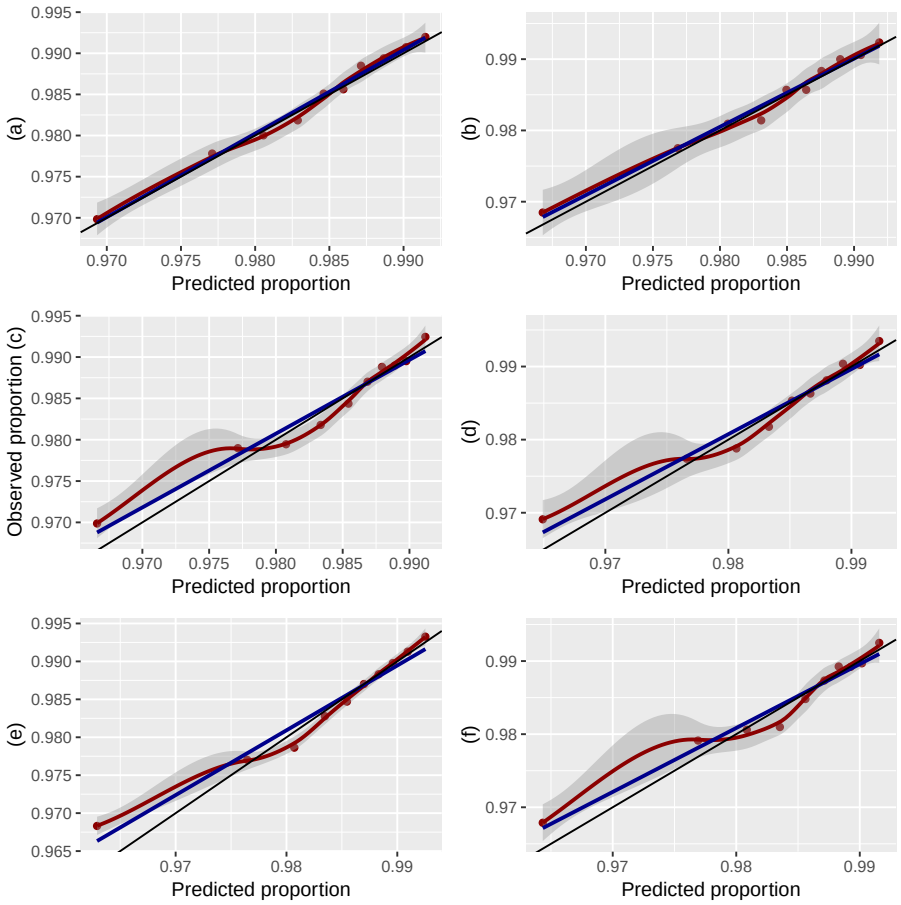

Figure 4.3: Calibration of the comparison Cox models. (a) [TCSB$^+$08], (b) [TCSB$^+$08] with features, (c)[TCZK05] no BD, (d) [TCZK05] no BD, with features, (e) [TCZK05] with BD, (f) [TCZK05] with BD and features. The blue line is a linear model line. The red line is a flexible calibration line fitted to the data. The black line is a 45° line, the ideal calibration.

4.2.2 AFT model calibration comparison

Next, we examine the calibration slope and plots for the AFT model for 2, 4 and 5 year survival on the training and testing set (Figure 4.4) [R C19, MIG12]. We note that the calibration plots are quite similar to our Cox model. As before, there is a larger standard error on the testing calibration than the training and the predictions are better for the lower risk (higher survival) quantiles. Next we examine the calibration plot for

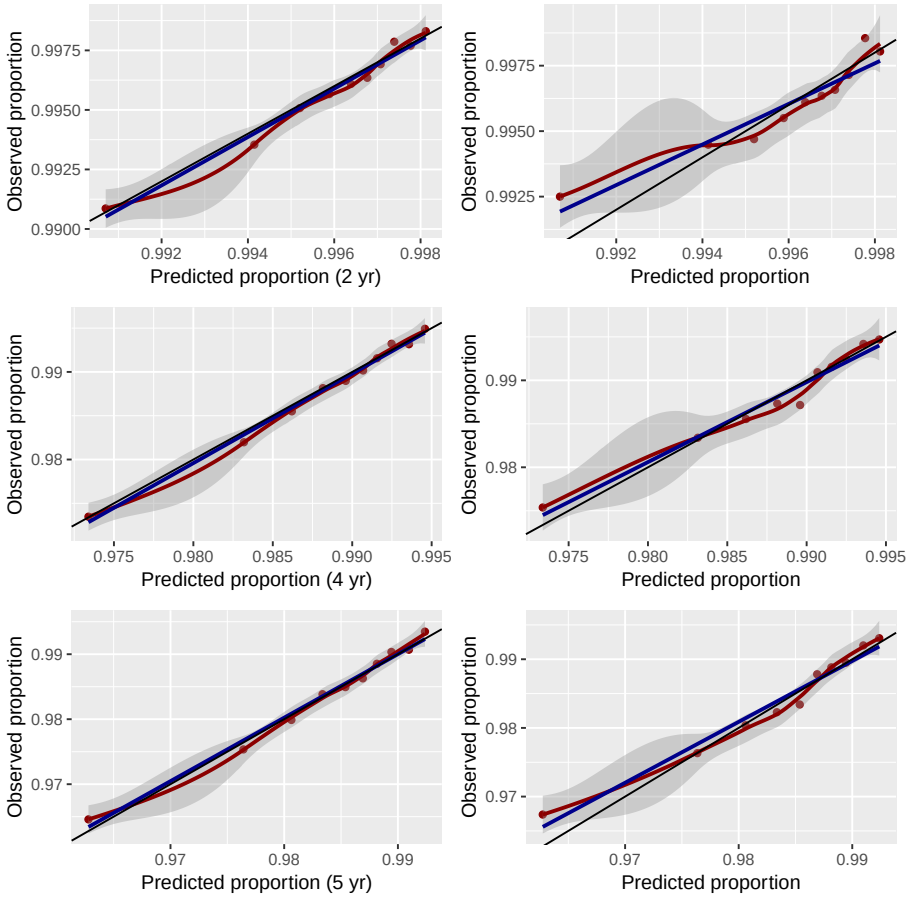

Figure 4.4: Calibration of our AFT model for training set (left) and testing set (right) for 2, 4 and 5 year survival. The blue line is the linear model line. The red line is a flexible calibration line fitted to the data. The black line is a 45° line, the ideal calibration.

the comparison models on the test set (Figure 4.5). To quantitatively examine the calibration, we see in Table 4.6 that the simple model of only age, ethnicity and breast density has an amazing calibration, with a slope of 1 and intercept of 0. The other models' calibration is satisfactory, according to the same reasoning as used previously.

Table 4.6: Comparisons of AFT model calibration on test set

	model		coef	2.5 %	97.5 %
0	ideal	intercept	0.00		
		slope	1.00		
(a)	[TCSB$^+$08]	intercept	0.00	-0.08	0.08
		slope	1.00	0.92	1.09
(b)	[TCSB$^+$08] with features	intercept	0.05	-0.04	0.14
		slope	0.95	0.85	1.04
(c)	[TCZK05] no BD	intercept	0.11	-0.03	0.25
		slope	0.89	0.74	1.03
(d)	[TCZK05] no BD, features	intercept	0.13	0.00	0.26
		slope	0.87	0.74	1.00
(e)	[TCZK05] with BD	intercept	0.12	-0.03	0.27
		slope	0.88	0.73	1.03
(f)	[TCZK05] with BD, features	intercept	0.15	0.02	0.29
		slope	0.84	0.70	0.98
	our model	intercept	0.11	-0.02	0.24
		slope	0.89	0.76	1.02

4.3 Birth cohort and risk

Recent work has shown that predictions of breast cancer risks may be influenced by calendar time [SRC$^+$15]. Using all 813,280 participants, we adjusted for birth cohort by breaking birth year into whether they were born prior to 1946, between 1946 and 1965, and after 1965, corresponding to the trend seen in Figure 3.4. We also adjusted for number of births by grouping this variable into 0 children, 1 child, and 2 or more children. We found that First Nations have a larger relative risk compared to Caucasian despite adjusting for birth cohort and number of children (Table 4.7). We also note that for this group of women it takes two or more children to see a protective effect compared to having no children.

When fitting a separate model for each birth cohort, we found that the

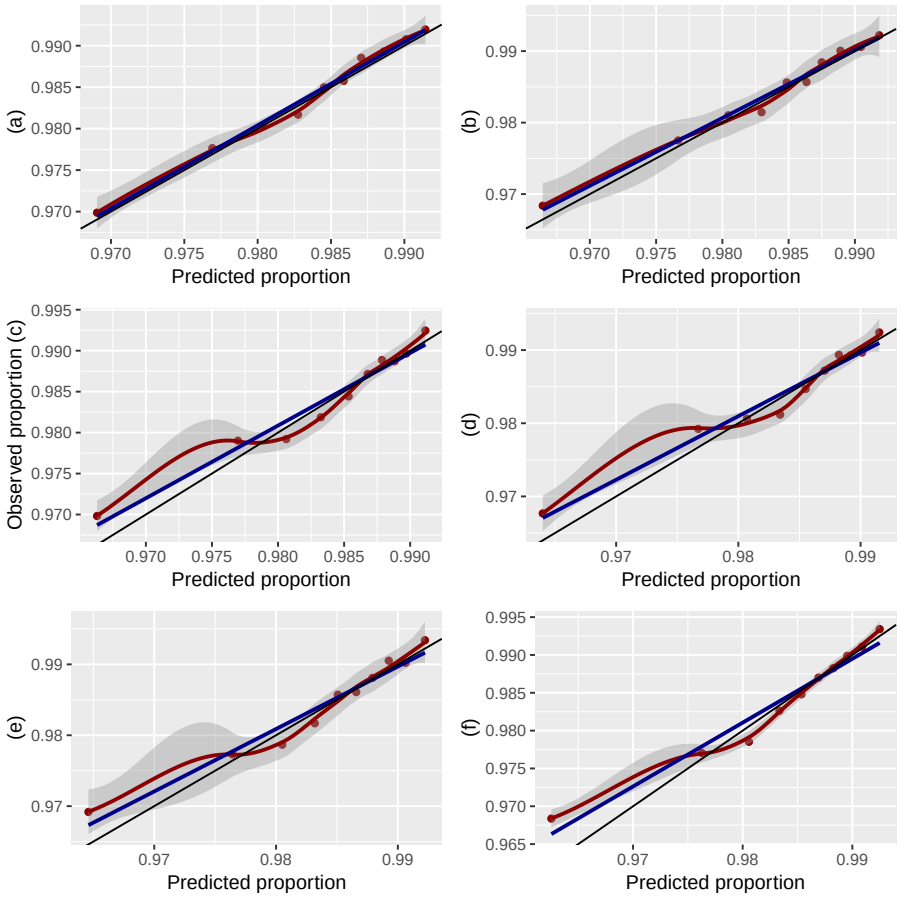

Figure 4.5: Calibration of the comparison AFT models. (a) [TCSB$^+$08], (b) [TCSB$^+$08] with features, (c)[TCZK05] no BD, (d) [TCZK05] no BD, with features, (e) [TCZK05] with BD, (f) [TCZK05] with BD and features. The blue line is a linear model line. The red line is a flexible calibration line fitted to the data. The black line is a 45° line, the ideal calibration.

Table 4.7: Proportional hazards model adjusted for birth cohort and number of births

	coef	exp(coef)	p-value	2.5 %	97.5 %
age	0.03	1.03	0.00	1.03	1.03
featno_feat (ref)	0.00	1.00			
feat3_or_more	0.67	1.96	0.02	1.14	3.37
featarch_asym	0.06	1.06	0.64	0.83	1.36
featarch_calc	1.01	2.75	0.00	1.68	4.50
featarch_only	0.07	1.07	0.40	0.91	1.27
featasym_calc	0.59	1.81	0.00	1.38	2.37
featasym_only	-0.10	0.91	0.01	0.85	0.98
featcalc_only	0.67	1.96	0.00	1.80	2.14
featmass_arch	0.92	2.51	0.00	1.56	4.04
featmass_asym	0.41	1.51	0.00	1.15	1.98
featmass_calc	0.39	1.47	0.04	1.02	2.14
featmass_only	0.10	1.11	0.04	1.00	1.22
featmissing	0.17	1.18	0.00	1.12	1.24
density1 (ref)	0.00	1.00			
density2	0.39	1.47	0.00	1.42	1.53
ethnicCaucasian (ref)	0.00	1.00			
ethnicFirst Nations	0.22	1.25	0.01	1.06	1.47
ethnicAfrican	-0.00	1.00	0.99	0.67	1.49
ethnicEast Asian	-0.05	0.95	0.13	0.89	1.01
ethnicOther	-0.00	1.00	0.88	0.95	1.04
ethnicSouth Asian	-0.28	0.75	0.00	0.66	0.85
Ever_Estrogen0 (ref)	0.00	1.00			
Ever_Estrogen1	0.17	1.18	0.00	1.12	1.24
Ever_Biopsy0 (ref)	0.00	1.00			
Ever_Biopsy1	0.34	1.40	0.00	1.33	1.47
ever_mom0 (ref)	0.00	1.00			
ever_mom1	0.48	1.61	0.00	1.52	1.71
ever_sis0 (ref)	0.00	1.00			
ever_sis1	0.39	1.48	0.00	1.37	1.59
educationNo HS (ref)	0.00	1.00			
educationSome/HS	0.24	1.27	0.00	1.17	1.38
educationSome/Co	0.28	1.33	0.00	1.23	1.44
num_deliv0 (ref)	0.00	1.00			
num_deliv1	-0.01	0.99	0.69	0.93	1.05
num_deliv2+	-0.09	0.92	0.00	0.88	0.96
birth_yearpre1946 (ref)	0.00	1.00			
birth_year1946-65	-0.11	0.90	0.00	0.84	0.95
birth_year1965+	-0.12	0.88	0.02	0.80	0.98

relative risk of 0.74 for East Asian compared to Caucasian found by [Hoe13] was corroborated with a relative risk of 0.73 for East Asian women in only the cohort born prior to 1946 (Table 4.8) [R C19, The15].

Table 4.8: Cox proportional hazards model stratified by birth cohort. Only showing the coefficients for ethnicity as the rest remained stable.

ethnicity	coef	exp(coef)	p-value	2.5 %	97.5 %
born prior to 1946					
First Nations	0.13	1.14	0.36	0.86	1.51
African	0.20	1.22	0.51	0.68	2.21
East Asian	-0.32	0.73	0.00	0.64	0.83
Other	-0.02	0.98	0.57	0.93	1.04
South Asian	-0.41	0.66	0.00	0.53	0.83
born between 1946 and 1965					
First Nations	0.19	1.21	0.08	0.97	1.51
African	-0.64	0.53	0.09	0.25	1.11
East Asian	0.06	1.06	0.17	0.97	1.15
Other	0.01	1.01	0.85	0.94	1.08
South Asian	-0.20	0.82	0.01	0.70	0.96
born after 1965					
First Nations	0.69	1.99	0.01	1.16	3.42
African	1.22	3.38	0.00	1.50	7.61
East Asian	0.15	1.16	0.22	0.91	1.49
Other	-0.02	0.98	0.78	0.83	1.15
South Asian	-0.55	0.58	0.02	0.36	0.94

This is not surprising as [SRC+15] found that the incidence rates of East Asian and Western women are converging. They suggest that the previously seen protective effect of Asian ethnicity was likely due to lifestyle. Asian women living in BC are subject to the same change as they adapt a more Western lifestyle. These changes include earlier age at menarche, later age at first birth, low number of children, rising body mass index, and shifts in diet. Further, women born prior to 1946 had access to screening at an older age as BC Cancer Breast Screening began operations in 1988. These women would not have had the advantages that come with screening for the same length of time as women born in later years.

Further, as these women born prior to 1946 have become too old to participate in screening mammography, the influence from this cohort on predictions will continue to diminish. In the same way, over 80% of women attending screening born after 1975 have at least some post-secondary education. Thus, including education with 'no high school' as the reference may not be relevant in a few years. A more useful predictor would be a more

comprehensive measure of lifestyle, that is left for future work.

Chapter 5

Conclusion

This research attempted to predict risk of breast cancer for BC Cancer Breast Screening participants using several different models. Compared to the literature, the performance metrics of calibration and concordance were similar. We did find that the mammographic features seen on a false positive increased the risk of breast cancer, but since the numbers of women who have these features was not very large and the increase in risk was not substantial enough, they did not make a meaningful contribution to the model discrimination.

The goal of providing a risk prediction model using the mammographic features was not achieved. While the calibration was fair, the concordance was too low to be used clinically.

The strengths of the research included the large number of individuals and many predictors to examine. Limitations included how the risk factors were collected; personal risk factors were from the participants' memory and the mammographic information was dependent on the radiologists' interpretation.

The models should not be applied in clinical settings. Since the ability of the models to distinguish who will be diagnosed with breast cancer is unsatisfactory, the mammographic features cannot be recommended to BC Cancer Breast Screening for use in their screening eligibility guidelines. However, as there was a small increase in the concordance with the addition of mammographic features and most of the mammographic features had significant hazard ratios with increased risks, further investigation into the mammographic features may be recommended instead.

Radiologists only point out the features that are concerning on an abnormal mammogram, thus the data available in the Breast Screening registry does not provide a complete picture of the evolution of mammographic features. Possible future research may include machine learning techniques on serial mammographic images themselves [YLS+19], which would provide a more detailed picture of the evolution of the mammographic features.

Bibliography

[AALY+18] Kawthar Al-Ajmi, Artitaya Lophatananon, Martin Yuille, William Ollier, and Kenneth R Muir. Review of non-clinical risk models to aid prevention of breast cancer. *Cancer Causes & Control*, 29(10):967–986, 2018. → pages 4, 7

[All10] Paul D Allison. *Survival analysis using SAS: a practical guide*. Sas Institute, 2010. → pages 24

[ATW+12] Thunyarat Anothaisintawee, Yot Teerawattananon, Chollathip Wiratkapun, Vijj Kasamesup, and Ammarin Thakkinstian. Risk prediction models of breast cancer: a systematic review of model performances. *Breast cancer research and treatment*, 133(1):1–10, 2012. → pages 7

[ATW+14] Thunyarat Anothaisintawee, Yot Teerawattananon, Cholatip Wiratkapun, Jiraporn Srinakarin, Piyanoot Woodtichartpreecha, Siriporn Hirunpat, Sansanee Wongwaisayawan, Panuwat Lertsithichai, Vijj Kasamesup, and Ammarin Thakkinstian. Development and validation of a breast cancer risk prediction model for Thai women: a cross-sectional study. *Asian Pac J Cancer Prev*, 15(16):6811–7, 2014. → pages 8

[BJS+17] Hannah R Brewer, Michael E Jones, Minouk J Schoemaker, Alan Ashworth, and Anthony J Swerdlow. Family history and risk of breast cancer: an analysis accounting for family structure. *Breast cancer research and treatment*, 165(1):193–200, 2017. → pages 2

[BP88] William E Barlow and Ross L Prentice. Residuals for relative risk regression. *Biometrika*, 75(1):65–74, 1988. → pages 32

[BWBB+06] William E Barlow, Emily White, Rachel Ballard-Barbash, Pamela M Vacek, Linda Titus-Ernstoff, Patricia A Carney, Jeffrey A Tice, Diana SM Buist, Berta M Geller, Robert Rosenberg, et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. *Journal of the National Cancer Institute*, 98(17):1204–1214, 2006. → pages 9

[BWD+20] Darren R Brenner, Hannah K Weir, Alain A Demers, Larry F Ellison, Cheryl Louzado, Amanda Shaw, Donna Turner, Ryan R Woods, and Leah M Smith. Projected estimates of cancer in Canada in 2020. *CMAJ*, 192(9):E199–E205, 2020. → pages 1

[Car07] Bendix Carstensen. Age–period–cohort models for the Lexis diagram. *Statistics in medicine*, 26(15):3018–3045, 2007. → pages 7

[CF92] PoC Consul and Felix Famoye. Generalized Poisson regression model. *Communications in Statistics-Theory and Methods*, 21(1):89–109, 1992. → pages 41

[CG01] Mark Clemons and Paul Goss. Estrogen and the risk of breast cancer. *New England Journal of Medicine*, 344(4):276–285, 2001. → pages 10

[CGBB+17] Jessica A Cintolo-Gonzalez, Danielle Braun, Amanda L Blackford, Emanuele Mazzola, Ahmet Acar, Jennifer K Plichta, Molly Griffin, and Kevin S Hughes. Breast cancer risk models: a comprehensive overview of existing models, validation, and clinical applications. *Breast cancer research and treatment*, 164(2):263–284, 2017. → pages 4, 7, 8

[CGK+19] Tess V Clendenen, Wenzhen Ge, Karen L Koenig, Yelena Afanasyeva, Claudia Agnoli, Louise A Brinton, Farbod Darvishian, Joanne F Dorgan, A Heather Eliassen, Roni T Falk, et al. Breast cancer risk prediction in women aged 35–50 years: impact of including sex hormone concentrations in the Gail model. *Breast Cancer Research*, 21(1):42, 2019. → pages 8

[Col15] David Collett. *Modelling survival data in medical research*. Chapman and Hall/CRC, 2015. → pages 24, 31, 32, 36

[Cox72] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. → pages 24

[CP12] Andrew J Coldman and Norman Phillips. False-positive screening mammograms and biopsies among women participating in a Canadian provincial breast screening program. *Canadian Journal of Public Health*, 103(6):e420–e424, 2012. → pages 10

[CRR+13] X Castells, M Roman, A Romero, J Blanch, R Zubizarreta, N Ascunce, D Salas, A Burón, M Sala, Cumulative False Positive Risk Group, et al. Breast cancer detection risk in screening mammography after a false-positive result. *Cancer epidemiology*, 37(1):85–90, 2013. → pages 2

[CRT93] Elizabeth B Claus, Neil Risch, and W Douglas Thompson. The calculation of breast cancer risk for women with a first degree family history of ovarian cancer. *Breast cancer research and treatment*, 28(2):115–120, 1993. → pages 8

[CTRP+16] Xavier Castells, Isabel Torá-Rocamora, Margarita Posso, Marta Román, Maria Vernet-Tomas, Ana Rodríguez-Arana, Laia Domingo, Carmen Vidal, Marisa Baré, Joana Ferrer, et al. Risk of breast cancer in women with false-positive results according to mammographic features. *Radiology*, 280(2):379–386, 2016. → pages 2, 11, 27, 31

[DL89] Charmaine Dean and Jerald Franklin Lawless. Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association*, 84(406):467–472, 1989. → pages 41

[DQ20] Jia-Yi Dong and Li-Qiang Qin. Education level and breast cancer incidence: a meta-analysis of cohort studies. *Menopause*, 27(1):113–118, 2020. → pages 10, 27

[DYFC15] Sara W Dyrstad, Yan Yan, Amy M Fowler, and Graham A Colditz. Breast cancer risk associated with benign breast

disease: systematic review and meta-analysis. *Breast cancer research and treatment*, 149(3):569–575, 2015. → pages 10

[EBM+98] Joann G Elmore, Mary B Barton, Victoria M Moceri, Sarah Polk, Philip J Arena, and Suzanne W Fletcher. Ten-year risk of false positive screening mammograms and clinical breast examinations. *New England Journal of Medicine*, 338(16):1089–1096, 1998. → pages 1, 17

[ECP+17] Mikael Eriksson, Kamila Czene, Yudi Pawitan, Karin Leifland, Hatef Darabi, and Per Hall. A clinical model for identifying the short-term risk of breast cancer. *Breast cancer research*, 19(1):29, 2017. → pages viii, 30, 31

[EF15] Christoph Engel and Christine Fischer. Breast cancer risks and risk prediction models. *Breast care*, 10(1):7–12, 2015. → pages 7

[fCEN+] National Institute for Clinical Excellence (NICE) et al. Familial breast cancer: Classification, care and managing breast cancer and related risks in people with a family history of breast cancer (cg164). → pages 30

[FL99] Lloyd D Fisher and Danyu Y Lin. Time-dependent covariates in the Cox proportional-hazards regression model. *Annual review of public health*, 20(1):145–157, 1999. → pages 31, 32

[GBB+89] Mitchell H Gail, Louise A Brinton, David P Byar, Donald K Corle, Sylvan B Green, Catherine Schairer, and John J Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute*, 81(24):1879–1886, 1989. → pages 7, 45

[Gor13] Leon Gordis. *Epidemiology 5th Edition with Student Consult Online Access*. Elsevier Saunders, Philadelphia, PA, 2013. → pages 1

[gro19] BMJ Publishing group. How to calculate risk [online]. 2019 [cited December 23, 2019]. → pages 4

[GSW+17] Paula Grabler, Dominique Sighoko, Lilian Wang, Kristi Allgood, and David Ansell. Recall and cancer detection rates for screening mammography: finding the sweet spot. *American Journal of Roentgenology*, 208(1):208–213, 2017. → pages 1

[Har19] Frank E Harrell Jr. *rms: Regression Modeling Strategies*, 2019. R package version 5.1-3.1. → pages 29, 44

[HHS+15] Louise M Henderson, Rebecca A Hubbard, Brian L Sprague, Weiwei Zhu, and Karla Kerlikowske. Increased risk of developing breast cancer after a false-positive screening mammogram. *Cancer Epidemiology and Prevention Biomarkers*, 24(12):1882–1889, 2015. → pages 2

[HJ15] Frank E Harrell Jr. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015. → pages 5

[HJLC+84] Frank E Harrell Jr, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2):143–152, 1984. → pages 4

[HJLM96] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996. → pages 4

[Hoe13] Tanja Hoegg. Statistical modelling of breast cancer risk for British Columbian women. Master's thesis, University of British Columbia, 2013. → pages 7, 8, 15, 59

[Inc15] SAS Institute Inc. *SAS/STAT 14.1 User's Guide*. SAS Institute Inc., Gary, NC, USA, 2015. → pages 5, 25

[Kei90] Niels Keiding. Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 332(1627):487–509, 1990. → pages 7

[KGR+95] Karla Kerlikowske, Deborah Grady, Susan M Rubin, Christian Sandrock, and Virginia L Ernster. Efficacy of screening mammography: a meta-analysis. *Jama*, 273(2):149–154, 1995. → pages 1

[KK19] Alboukadel Kassambara and Marcin Kosinski. *survminer: Drawing Survival Curves using 'ggplot2'*, 2019. R package version 0.4.4. → pages 27

[KKLL18] Sohyun Kim, Yeonsook Ko, Hwa Jeong Lee, and Jung-eun Lim. Menopausal hormone therapy and the risk of breast cancer by histological type and race: a meta-analysis of randomized controlled trials and cohort studies. *Breast cancer research and treatment*, 170(3):667–675, 2018. → pages 10

[KM58] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958. → pages 38

[KP11] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011. → pages 36

[KW52] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952. → pages 29

[KZ08] Christian Kleiber and Achim Zeileis. *Applied Econometrics with R*. Springer-Verlag, New York, 2008. ISBN 978-0-387-77316-2. → pages 41

[Lam92] Diane Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992. → pages 41

[LPB+19] Javier Louro, Margarita Posso, Michele Hilton Boon, Marta Román, Laia Domingo, Xavier Castells, and María Sala. A systematic review and quality assessment of individualised breast cancer risk prediction models. *British journal of cancer*, 2019. → pages 7

[LZW+12] Lihua Liu, Juanjuan Zhang, Anna H Wu, Malcolm C Pike, and Dennis Deapen. Invasive breast cancer incidence trends by detailed race/ethnicity and age. *International Journal of Cancer*, 130(2):395–404, 2012. → pages 10

[MCL+70] Brian MacMahon, Ph Cole, TM Lin, CR Lowe, AP Mirra, B Ravnihar, EJ Salber, VG Valaoras, and S Yuasa. Age at first birth and breast cancer risk. *Bulletin of the world health organization*, 43(2):209, 1970. → pages 10

[MIG12] Ulla B. Mogensen, Hemant Ishwaran, and Thomas A. Gerds. Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software*, 50(11):1–23, 2012. → pages 25, 44, 46, 51, 55

[MMVR12] Raymond H Myers, Douglas C Montgomery, G Geoffrey Vining, and Timothy J Robinson. *Generalized Linear Models.: with Applications in Engineering and the Sciences*, volume 791. John Wiley & Sons, 2012. → pages 39

[Moo16] Dirk F Moore. *Applied survival analysis using R*. Springer, 2016. → pages 36

[MRMTB14] DS AL Mousa, EA Ryan, C Mello-Thoms, and PC Brennan. What effect does mammographic breast density have on lesion detection in digital mammography? *Clinical radiology*, 69(4):333–341, 2014. → pages 19

[MSG02] Jenny McCann, Diane Stockton, and Sara Godward. Impact of false-positive mammography on subsequent screening attendance and risk of cancer. *Breast Cancer Research*, 4(5):R11, 2002. → pages 1

[oC11] Public Health Agency of Canada. Organized breast cancer screening programs in Canada: Report on program performance in 2005 and 2006, 2011. → pages 1

[oCS12] Canadian Cancer Society's Steering Committee on Cancer Statistics. Canadian cancer statistics 2012, 2012. → pages 1

[oHFiBC+12] Collaborative Group on Hormonal Factors in Breast Cancer et al. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *The lancet oncology*, 13(11):1141–1151, 2012. → pages 10

[PC+11] Martyn Plummer, Bendix Carstensen, et al. Lexis: An r class for epidemiological studies with long-term follow-up. *Journal of Statistical Software*, 38(5):1–12, 2011. → pages 7, 39

[PDD+97] Paul DP Pharoah, Nicholas E Day, Stephen Duffy, Douglas F Easton, and Bruce AJ Ponder. Family history and the risk of breast cancer: a systematic review and meta-analysis. *International journal of cancer*, 71(5):800–809, 1997. → pages 10

[PS86] Donald A Pierce and Daniel W Schafer. Residuals in generalized linear models. *Journal of the American Statistical Association*, 81(396):977–986, 1986. → pages 41

[R C19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. → pages 25, 27, 29, 31, 32, 36, 38, 39, 41, 44, 46, 50, 51, 55, 59

[RACOO17] M Shafiqur Rahman, Gareth Ambler, Babak Choodari-Oskooei, and Rumana Z Omar. Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC medical research methodology*, 17(1):60, 2017. → pages 53

[RCW94] Bernard Rosner, Graham A Colditz, and Walter C Willett. Reproductive risk factors in a prospective study of breast cancer: the nurses' health study. *American journal of epidemiology*, 139(8):819–835, 1994. → pages 10

[S+19] Ewout W Steyerberg et al. *Clinical prediction models*. Springer, 2019. → pages 5, 51

[SAS19] SAS Institute Inc. *SAS University Edition*. SAS Institute Inc., Cary, NC, USA, 2019. → pages 5, 25

[SBSTL15] Sepideh Saadatmand, Reini Bretveld, Sabine Siesling, and Madeleine MA Tilanus-Linthorst. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients. *Bmj*, 351:h4901, 2015. → pages 1

[Sch80] David Schoenfeld. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, 67(1):145–153, 1980. → pages 5

[Sch82] David Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241, 1982. → pages 6, 27

[Scr18] BC Cancer Breast Screening. BC Cancer Breast Screening 2017 program results, 2018. → pages 1, 2

[SIK86] Yosiyuki Sakamoto, Makio Ishiguro, and Genshiro Kitagawa. Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, 81, 1986. → pages 7, 25, 36

[SMvdW+13] Ewout W Steyerberg, Karel GM Moons, Danielle A van der Windt, Jill A Hayden, Pablo Perel, Sara Schroter, Richard D Riley, Harry Hemingway, Douglas G Altman, PROGRESS Group, et al. Prognosis research strategy (progress) 3: prognostic model research. *PLoS Med*, 10(2):e1001381, 2013. → pages 4

[SPF+97] Catherine Schairer, Ingemar Persson, Margareta Falkeborn, Tord Naessen, Rebecca Troisi, and Louise A Brinton. Breast cancer risk associated with gynecologic surgery and indications for such surgery. *International journal of cancer*, 70(2):150–154, 1997. → pages 10

[SRC+15] Hyuna Sung, Philip S Rosenberg, Wan-Qing Chen, Mikael Hartman, Wei-yen Lim, Kee Seng Chia, Oscar Wai-Kong Mang, Chun-Ju Chiang, Daehee Kang, Roger Kai-Cheong Ngan, et al. Female breast cancer incidence among asian and western populations: more similar than expected. *Journal of the National Cancer Institute*, 107(7):djv107, 2015. → pages 56, 59

[SVC+10] Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010. → pages 29

[TC90] William Thomas and R Dennis Cook. Assessing influence on predictions from generalized linear models. *Technometrics*, 32(1):59–65, 1990. → pages 41

[TCA17] Terry Therneau, Cindy Crowson, and Elizabeth Atkinson. Using time dependent covariates and time dependent coefficients in the Cox model. *Survival Vignettes*, 2017. → pages 32

[TCSB+08] Jeffrey A Tice, Steven R Cummings, Rebecca Smith-Bindman, Laura Ichikawa, William E Barlow, and Karla Kerlikowske. Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Annals of internal medicine*, 148(5):337, 2008. → pages xi, 44, 45, 46, 47, 48, 49, 50, 53, 54, 56, 57

[TCZK05] Jeffrey A Tice, Steven R Cummings, Elad Ziv, and Karla Kerlikowske. Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population. *Breast cancer research and treatment*, 94(2):115–122, 2005. → pages xi, 7, 44, 45, 46, 47, 48, 49, 50, 53, 54, 56, 57

[TG00] Terry M Therneau and Patricia M Grambsch. Statistics for biology and health. In *Modeling survival data: extending the Cox model*, pages 87–152. Springer-Verlag, 2000. → pages 5, 6

[TGF90] Terry M Therneau, Patricia M Grambsch, and Thomas R Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990. → pages 6, 27, 32

[The15] Terry M Therneau. *A Package for Survival Analysis in S*, 2015. version 2.38. → pages 27, 31, 32, 36, 38, 39, 44, 46, 50, 59

[UTT+03] Kimiko Ueda, Hideaki Tsukuma, Hideo Tanaka, Wakiko Ajiki, and Akira Oshima. Estimation of individualized probabilities of developing breast cancer for Japanese women. *Breast Cancer*, 10(1):54–62, 2003. → pages 9

[VdSKDDVR11] AFW Van der Steeg, CMG Keyzer-Dekker, J De Vries, and JA Roukema. Effect of abnormal screening mammogram on quality of life. *British Journal of Surgery*, 98(4):537–542, 2011. → pages 1

[vECKV14] My von Euler-Chelpin, Megumi Kuchiki, and Ilse Vejborg. Increased risk of breast cancer in women with false-positive test: the role of misclassification. *Cancer epidemiology*, 38(5):619–622, 2014. → pages 2

[vECRTV12] My von Euler-Chelpin, Louise Madeleine Risør, Brian Larsen Thorsted, and Ilse Vejborg. Risk of breast cancer after false-positive test results in screening mammography. *Journal of the National Cancer Institute*, 104(9):682–689, 2012. → pages 2

[vH00] Hans C van Houwelingen. Validation, calibration, revision and combination of prognostic survival models. *Statistics in medicine*, 19(24):3401–3415, 2000. → pages 5

[W+86] Chien-Fu Jeff Wu et al. Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, 14(4):1261–1295, 1986. → pages 6, 27, 39

[Wei92] Lee-Jen Wei. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879, 1992. → pages 36

[WGB+14] Yuan Wang, Ying Gao, Munkhzul Battsend, Kexin Chen, Wenli Lu, and Yaogang Wang. Development of a risk assessment tool for projecting individualized probabilities of developing breast cancer for Chinese women. *Tumor Biology*, 35(11):10861–10869, 2014. → pages 9

[WRB+13] Christina R Weisstock, Rasika Rajapakshe, Christabelle Bitgood, Steven McAvoy, Paula B Gordon, Andrew J

Coldman, Brent A Parker, and Christine Wilson. Assessing the breast cancer risk distribution for women undergoing screening in British Columbia. *Cancer Prevention Research*, 6(10):1084–1092, 2013. → pages 2

[WvECL+17] Rikke Rass Winkel, My von Euler-Chelpin, Elsebeth Lynge, Pengfei Diao, Martin Lillholm, Michiel Kallenberg, Julie Lyng Forman, Michael Bachmann Nielsen, Wei Yao Uldall, Mads Nielsen, et al. Risk stratification of women with false-positive test results in mammography screening based on mammographic morphology and density: a case control study. *Cancer epidemiology*, 49:53–60, 2017. → pages 2

[YLS+19] Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 292(1):60–66, 2019. → pages 61

[ZRT+18] Xuehong Zhang, Megan Rice, Shelley S Tworoger, Bernard A Rosner, A Heather Eliassen, Rulla M Tamimi, Amit D Joshi, Sara Lindstrom, Jing Qian, Graham A Colditz, et al. Addition of a polygenic risk score, mammographic density, and endogenous hormones to existing breast cancer risk prediction models: A nested case–control study. *PLoS medicine*, 15(9):e1002644, 2018. → pages 8