Prognostic Modeling in the Presence of Competing Risks: An Application to Cardiovascular and Cancer Mortality in Breast Cancer Survivors

Nicole M. Leoce

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Public Health
in the Mailman School of Public Health

COLUMBIA UNIVERSITY
2016

ABSTRACT

Prognostic Modeling in the Presence of Competing Risks: an application to Cardiovascular and Cancer Mortality in Breast Cancer Survivors

Nicole M. Leoce

Currently, there are an estimated 2.8 million breast cancer survivors in the United States. Due to modern screening practices and raised awareness, the majority of these cases will be diagnosed in the early stages of disease where highly effective treatment options are available, leading a large proportion of these patients to fail from causes other than breast cancer. The primary cause of death in the United States today is cardiovascular disease, which can be delayed or prevented with interventions such as lifestyle modifications or medications. In order to identify individuals who may be at high risk for a cardiovascular event or cardiovascular mortality, a number of prognostic models have been developed. The majority of these models were developed on populations free of comorbid conditions, utilizing statistical methods that did not account for the competing risks of death from other causes, therefore it is unclear whether they will be generalizable to a cancer population remaining at an increased risk of death from cancer and other causes.

Consequently, the purpose of this work is multi-fold. We will first summarize the major statistical methods available for analyzing competing risk data and include a simulation study comparing them. This will be used to inform the interpretation of the real data analysis, which will be conducted on a large, contemporary cohort of breast cancer survivors. For these women, we will categorize the major causes of death, hypothesizing that it will include cardiovascular failure. Next, we will evaluate the existing cardiovascular disease risk models in our population of cancer survivors, and then propose a new model to simultaneously predict a survivor's risk of death due to her breast cancer or due to cardiovascular disease, while accounting for additional competing causes of death. Lastly, model predicted outcomes will be calculated for the cohort, and evaluation methods will be applied to determine the clinical utility of such a model.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my deepest gratitude to my advisors, Dr. Zhezhen Jin and Dr. Mary Beth Terry, for all their guidance, caring, patience, and encouragement. Your ideas, mentorship, and example has taught me more than any class or textbook and will remain invaluable as I continue my career.

I would like to thank Dr. Lawrence Kushi and the research staff at Kaiser Permanente for their collaboration, suggestions, and data sharing. Without them, this work would have remained hypothetical, and lacking the real-world implications so crucial to the field of biostatistics.

I would also like to thank Dr. Wei-Yann Tsai for both his expert advice on this work, and the years spent as a mentor to me through the Cancer Training Fellowship. I am also exceptionally grateful to Dr. Al Neugut for accepting me to the Cancer Training Fellowship and the four years we spent together researching initiation and adherence.

I would also like to thank Dr. Ying Wei for her valuable input and recommendations to improving this work, and the rest of the faculty of the Department of Biostatistics, from whom I have had the pleasure of learning during my coursework years. I would also like to thank Justine Herrera for all of her patience and guidance.

Lastly, I would like to thank my colleagues from the Department of Biostatistics at Memorial Sloan Kettering, especially Katherine Panageas, Chaya Moskowitz, Elena Elkin, Sujata Patil, Mithat Gönen, and Colin Begg, who mentored, supported, and encouraged me in my early career. My respect and admiration for you all led me to where I am today.

Dedication

I dedicate this dissertation to my family. First, to my husband, Craig, who has supported me in every way imaginable through this long endeavor, constantly giving me the pep talks I needed to reach the finish line. I would never have been able to complete this without your unconditional love and support. I also dedicate it my parents and my brother, who always encouraged my curious spirit, taught me to never give up on my dreams, and have shown me that with hard work and perseverence, anything can be achieved.

Lastly, I dedicate this to my two children, who were not around when this journey began, and may be too young to ever remember the time "when mommy was in school." Without you, I probably would have been done 2 years sooner, but I could not have imagined it any other way. I hope that when you are older, you learn to always set goals for yourselves and never give up your dreams, no matter how diffult the challenges may seem.

Chapter 1

Introduction

Currently, there are an estimated 2.8 million breast cancer survivors in the United States [ACS, 2014]. Under modern screening practices, the majority of breast cancer cases are caught in the early stages, where effective, nearly curative, treatment therapies may exist. Subsequently, a large proportion of breast cancer survivors will ultimately fail from something other than breast cancer. At present, cardiovascular disease (CVD) is the leading cause of mortality in the United States, including for early stage breast cancer survivors [Singla et al., 2012]. While focusing on their risk of cancer recurrence, it is possible that many breast cancer survivors are underestimating their risk of cardiovascular disease and cardiovascular mortality[Ganz, 2009]. Evidence has also shown that several of the standard breast cancer treatment regimens are associated with additional cardio-toxic side effects [Gagliardi et al., 1998, Nichols et al., 2013, Azim Jr et al., 2011], and whether they may contribute to an increased risk of cardiovascular events, has not been well studied.

Given the expected long term survival, and possible increase in risk of cardiovascular events due to cancer treatments, primary care physicians and oncologists should be aware of the need to counsel breast cancer survivors for increased CVD risk. Currently, there are several CVD risk prediction models that have been developed for the general population that are used in clinical practice [Wilson et al., 1998, Adult Treatment Panel (ATP), 2001,

Ridker et al., 2007, Conroy et al., 2003], however, a breast cancer survivor-specific model may more accurately predict a survivor's risk by combining known CVD risk factors with breast cancer disease and treatment information, while also accounting for the competing risk of cancer death.

While the literature describing cardiovascular disease and mortality following breast cancer treatment is limited, there have been a handful of studies that have examined these outcomes in this patient population. In 2004, Schairer et al. [Schairer et al., 2004] were among the first to examine cause of death for breast cancer patients using SEER data from 1973-2000. In a sample of over 430,000 women, they found that probability of death from breast cancer versus other causes varied substantially according to stage, tumor size, ER status, and age at diagnosis in both white and black patients. Greater probability of breast cancer death was associated with higher stage disease, ER negative tumors, and African American race. Other major causes of death included cardiovascular disease (33%), other cancers (20%), circulatory disease (11%), respiratory diseases (9%), and other diseases (26%, excludes aforementioned disease categories). While the primary focus of this study was not to examine cardiovascular disease mortality, it was one of the earliest ones to demonstrate that the leading cause of death for breast cancer patients, especially those diagnosed with localized disease, may in fact be cardiovascular disease.

In 2011, two additional studies were published using SEER data [Patnaik et al., 2011] . Patnaik et al. used the SEER-Medicare linked database data for all breast cancer diagnoses between 1992 and 2000, examining causes of death and comorbid conditions related to cause of death. In a study of 63,566 women with a median follow-up of 105 months, 47% were still alive at the time of analysis, 15% were deceased due to breast cancer and 36% due to other causes. While the primary focus of the study was to determine how comorbid conditions influence cause of death, they also found that cardiovascular disease was the primary cause of death in the study population, with 15.9% (95% CI 15.6% to 16.2%), failing from the cause. The study was not, however, limited to early stage disease, for which the rates of

death from breast cancer are expected to be even lower. The authors also showed that within each cancer stage, as age increased, women were more likely to die of cardiovascular disease.

Schonberg et al. [Schonberg et al., 2011] used SEER-Medicare linked data, matching 66,039 breast cancer cases (including DCIS) with non-cancer controls in an effort to see how a breast cancer diagnosis impacts overall survival. The data showed that women diagnosed with DCIS or stage I cancer had slightly lower overall mortality than the controls, however, cardiovascular disease was the most common cause of death, as well as for women ages ≥ 80 years with stage II disease. They also found that women with DCIS or stage I disease had no differences in 10-year survival from non-cancer controls, but those with stage II disease or higher had worse overall survival than their matched controls.

Bardia et al. [Bardia et al., 2012] used data from the ELPH trial, whose primary goal was to look at the effect of 2 years of aromatase inhibitors on surrogate markers of response, to compare the 10 year predicted risk of breast cancer recurrence and 10 year predicted risk of cardiovascular disease in 415 postmenopausal, HR+, non-metastatic breast cancer survivors. They prospectively collected data on age, family history of heart disease, smoking, Body Mass Index (BMI), diabetes status, systolic blood pressure and whether on antihypertensive medications, HDL cholesterol, and total cholesterol, to sufficiently calculate a modified Framingham risk score for cardiovascular disease. A breast cancer recurrence risk score was calculated using Adjuvant! Online. The endpoint of interest was an indicator variable of whether a patient had a greater predicted risk of a CVD event than breast cancer recurrence at 10 years.

The results showed that 38% of the study had a low (<10%) 10 year risk of CVD, 50% had a moderate (10-25%) risk, and only 12% were considered high risk (>25%). In comparing the model based risks of CVD and breast cancer recurrence, 43% of women had a 10-year predicted risk of a CVD event equivalent to their risk of a breast cancer recurrence, and 37% had a predicted risk of a CVD event higher than the risk of a breast cancer recurrence. Specifically, women with stage I disease (67.5% of cohort), tumors less than 2 cm, grade 1

or 2 disease, negative lymph nodes, or heart age over 65 years were more likely to have a higher 10 year predicted risk of CVD than cancer recurrence.

A limitation of the study was that the endpoints were based on model-predicted outcomes, rather than prospectively collected follow-up information. Additionally, CVD risk estimates were based on a tool that was developed for use in the general population, and does not take into account any increased risk associated from the use of breast cancer treatments, or modifications of traditional risk factors from their use. Thus, it may have overestimated the 10-year cardiovascular disease risk. No model evaluation was performed, possibly due to the fact that cardiovascular event data was not collected, or due to low numbers of events.

In 2012, Singla et al.[Singla et al., 2012] re-capitulated the above information in an editorial piece. This piece also highlighted the potential cardio-toxic effects related to several breast cancer therapies, in order to raise awareness of the need to counsel breast cancer survivors, particularly those receiving several therapies, for their risk cardiovascular disease, in addition to their potential cancer recurrence.

In summary, while the mortality due to cardiovascular disease among breast cancer survivors has been documented, the literature is somewhat limited and data sources are largely from a claims-based population. There has been only one study [Bardia et al., 2012], that has calculated a cardiovascular disease (Framingham) risk score in a cancer population, but it was not able to validate the risk score with actual follow-up data, therefore its clinical utility remains unknown. In addition, there have been no studies to examine whether cancer characteristics or treatments may modify the predicted risk of cardiovascular disease and cardiovascular related mortality in this population. Development of a breast cancer survivor-specific cardiovascular disease risk model may aid oncologists and primary care physicians in counseling survivors who would especially benefit from increased awareness of their CVD risk and possible risk-lowering interventions.

In order to develop such a model, one must first acknowledge that the risk of death from competing causes remains high in this patient population. As the current aim is to

create such a model on a cohort of breast cancer survivors, I first review the statistical methods available for competing risk modeling, in Chapter 2. In Chapter 3, I summarize the published risk models for CVD, as well as the current models for predicting breast cancer recurrence and mortality. Focus will be placed on the statistical methodology used, handling of competing events, inclusion of covariates, and prediction for a new patient. Chapter 4 is reserved for prognostic model validation and evaluation techniques, specifically in the area of predicting disease risk.

Chapter 5 focuses on the comparison of the two most popular model choices, the cause-specific hazards (CSH) model and the subdistribution hazards (SH) model, and review published simulations that have evaluated the two approaches side by side. Before creating my own simulation study to compare the two methods, as well as include newly proposed multi-state (MS) framework [Putter, 2014], I review the available approaches for generating data for such a simulation. Due to the inconsistency of the data generation methods used in previous studies, the simulation is repeated for several methods of data generation, over a range of censoring percentages.

Finally, in Chapter 6, I perform an analysis on a real data set consisting of over 20,000 women diagnosed with stage I-III breast cancer within the Kaiser Permanente Northern California (KPNC) health care delivery system between January 1, 2000 and December 31, 2010. I first summarize the cohort's breast cancer and cardiovascular disease risk factors and outcomes, before evaluating the performance of the current cardiovascular risk models in this population. Next, I develop a new model that can be used to predict a patient's risk of failure from cardiovascular disease versus her breast cancer, and evaluate the model performance. Conclusions and limitations are summarized in Chapter 7.

Chapter 2

Statistical Methods for Modeling Disease Risk

A disease prediction model is defined in a general sense as a mathematical equation used to quantify the risk that an individual will experience a particular disease outcome in a specified time-period. Such models are useful tools for clinicians and patients and have a variety of purposes, including counseling patients on general disease risk, screening for high risk patients who could benefit from further testing, surveillance, or interventions, or even to select patients for randomized controlled trials. After disease onset, they can also be useful for patient prognosis or survival estimation.

The choice of statistical model depends on how risk is to be described and the handling of competing events. *Pure* risk assumes that there are no other competing events, and models the probability of an event occurring in a specified time frame, conditioning on the person being alive at the beginning of the interval, in the presence of known risk factors. *Absolute* risk, models the probability that an event will occur in a certain time interval, given that the person is alive and event free at the start of the interval, with certain risk factors, in the

presence of competing risks. This is also called the "crude risk" or "cumulative incidence". When the probability of competing events is rare, the pure and absolute risks are nearly equal. Models can be further complicated in the presence of right censoring, which may be due to death, drop-out, or loss to follow-up.

Popular methods for disease-specific risk models include, but are not limited to: Logistic regression models, Relative Risk Models, Parametric Accelerated Failure Time (AFT) Model (Weibull Regression or modification), Cox Proportional Hazards Models, and Semi-Parametric AFT Models. These models can include censored data, but do not specifically model competing events. Some models that can account for competing outcomes include the Cause Specific Hazards (CSH) Model, the Fine and Gray Subdistribution Hazards (SH) model, and a multi-state (MS) model.

2.1 Concepts and Notation

First I introduce some concepts and notation that are essential for the discussion of time inherent risk prediction modeling. Note that the terms "failure" and "event" may be used interchangeably in the proceeding sections.

Let T be a continuous variable representing survival (or failure) time, which may be observed or censored.

Let X be a vector of covariates.

To handle censoring, T_i is a bivariate vector (Y_i, δ_i), where

$$Y_i = \min(T_i, C_i)$$

and,

$$\delta_i = \begin{cases} 1 & \text{if } Y_i = T_i, \text{ (event observed)} \\ 0 & \text{if } Y_i = C_i \text{ , (event censored)} \end{cases}$$

2.1.1 Cumulative Distribution Function

The cumulative distribution function, or cdf, is a function that tells us the probability that a variable will be less than or equal to a value t. In survival analysis, it can be thought of as the probability of an event by time t, and is sometimes referred to as the cumulative incidence. In its simplest form, ignoring covariates, it can be denoted:

$$F(t) = Pr\{T \leq t\}$$

2.1.2 Survivor Function

In survival analysis, it is more useful to talk about the survivor function, which is the probability of surviving beyond a time, t, is the complement of the cumulative distribution function, and is usually denoted:

$$S(t) = Pr\{T > t\} = 1 - F(t)$$

Since T cannot be negative, $S(0) = 1$. S is also non-increasing, meaning that as time increases, the probability of survival cannot increase.

2.1.3 Probability Density Function

The probability density function, or pdf, is the derivative of the cdf, and is a way of describing the distribution of a variable. In survival analysis, the pdf, cdf, and survivor function are related as follows:

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$$

2.1.4 Hazard Function

In survival data, it is more common to describe the distribution using the hazard function. The hazard rate is defined as the instantaneous risk of failure, or event rate, at time t, conditioning on having survived until t, which makes it a conditional density, and without covariates, is usually denoted:

$$h(t) = \lim_{dt \to 0} \frac{Pr(t \leq T \leq t + dt | T \geq t)}{dt}$$

It must be non-negative. It is related to the survivor function and cumulative density function by:

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} = -\frac{d}{dt} \log S(t)$$

such that additionally,

$$S(t) = \exp\big[- \int_0^t h(u)du \big] = \exp[-H(t)]$$

where

$$H(t) = \int_0^t h(u)du$$

is the cumulative hazard function (which is a measure of risk, not a probability)

2.2 Relative Risk Model

Ignoring the time aspect, if we wish to model the probability of disease $Pr(Y = 1)$, we can treat the outcome as a binary variable. While the logistic model has been the most commonly implemented model for binary data, it has been argued that relative risks can provide a more interpretable description of associations between disease prevalence and risk factors, than an odds ratio [Lumley et al., 2006]. While in the case of rare events, the odds ratio approximates the relative risk, that does not justify its use for studies of more common

disease outcomes. One of the historical reasons for preference of logistic regression may have been the result of lack of computing capabilities, but recently the relative risk model has been implemented in standard software packages such as SAS, SPSS, Stata, and R.

The relative risk model is a generalized linear model of the form:

$$\log(Pr[Y = 1|X]) = \log(\mu) = X\beta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

with a log link and $V(\mu) = \mu(1 - \mu)$.

Its estimation can be done via maximum likelihood with some additional constraints such that the predicted value remains ≤ 1. There has been some debate over which optimization method to use, and it has been argued that when the MLE sits on the boundary of the parameter space, nonlinear least squares is a better option. Other methods of tricking the software by using a Poisson working model or Cox regression have also been applied.

Some important limitations of the Relative Risk model for prediction are that it does not include information about time to an event, but rather a binary outcome of whether the disease has occurred or not, presumably in a fixed interval with no complications of censoring. For this reason, it is important that it only be used for categorizing associations with disease and not for individual risk predictions.

2.3 Time to Event Models

2.3.1 Parametric Models

Parametric Accelerated Failure Time (AFT) Model

Often referred to as the AFT model [Wei, 1992], this class of models are fully parametric models that can be used to model an individual's survival time, $T_i > 0$.

If $X_i = (X_{i1}, \ldots, X_{ik})$ is a vector of covariates, then the model is:

$$\log(T_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \sigma\epsilon_i,$$

where ϵ_i is a random error term, and $\beta_0, \ldots, \beta_p$ and σ are parameters to be estimated. This models a linear relationship between the log of time and the covariates.

Exponentiating both sides gives an alternate way of expressing the model:

$$T_i = \exp[\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \sigma\epsilon_i],$$

though it is preferable to use the log transformation to ensure that the predicted values of T remain positive.

The model can also be written in terms of the hazard:

$$h(t|X) = h_0(t\exp(X\beta))\exp(X\beta)$$

Table 2.1: Choice of error distribution and corresponding distribution of time, AFT Model

Distribution of ϵ	**Distribution of T**
Extreme value (2 parameter)	Weibull
Extreme value (1 parameter)	Exponential (constant hazard)
Log-gamma	Gamma
Logistic	Log-logistic
Normal	Log-normal

While it is generally estimated by maximum likelihood, when there is no censoring, the AFT model can also be estimated by least squares. Similarly, if the ϵ_i are assumed to be normally distributed, the OLS estimates will also be the ML estimates (called the log-normal model). In the presence of non-normally distributed ϵ_i, or censoring, maximum likelihood estimation is used. Generally, the choice of error distribution (ϵ_i) has an associated T

distribution, which gives the model its name. Some commonly used distribution choices and corresponding models are listed in table 2.1.

In analysis of *human* failure time data, the Weibull model has been the most popular AFT model. So much so, in fact, that if the distribution is not specified in the literature, a Weibull model is assumed for an AFT model. The appeal of using a Weibull distribution (of which the exponential distribution is a special case with shape parameter equal to 1) is that it is the only family of distributions that is closed under both the AFT and proportional hazards models. This means that it can be estimated either in terms of the survival times or the hazard, and get equivalent results. In this way, parameters can be interpreted in terms of either their effect on survival time, or on the hazard.

The effect of explanatory variables in the original time scale (survival time) is to change the time scale by a factor $\gamma = \exp[X\beta]$ which will either accelerate (or degrade) the failure time, depending on the sign of $X\beta$. A γ greater than one will increase the survival time, while a factor less than one is harmful on survival.

A major disadvantage is the parametric assumption on the survival time (regression diagnostics or tests for deviations from Weibull can suggest whether assumption is violated). The advantages are that it relaxes the assumption of proportional hazards, the formula for predicted probabilities is simple, and when the fit is good, they can provide more precise estimates. Because it directly models the effect of covariates on survival time, interpretation of effects is considered more straightforward [Orbe et al., 2002]. This method was used in the 1991 Framingham Model for cardiovascular disease. It is available in most standard software packages (e.g. Survreg in R, Stata, Proc lifereg in SAS).

Parametric Proportional Hazards Model

Similar to a Cox model, except the baseline hazard is specified, rather than estimated from the data. The rationale is that if an adequately specified baseline hazard could be found, it would improve estimation efficiency and enable full likelihood estimation to be done. The

Weibull distribution has been a popular choice. This method also allows for time-varying hazard ratios. [Royston and Mahesh, 2002]. Interpretation of coefficients is the same as in the Cox model. This is used in many of the Breast Cancer prediction models, where the baseline hazard is calculated from population incidence data, but coefficient estimation is done on the observed cohort.

2.3.2 Semi-Parametric Models

Semi-Parametric AFT Model

The semi-parametric AFT model also linearly relates the covariates to the log of the failure time:

$$\log(T_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

However, it differs from the fully parametric AFT model in that it does not specify a distribution for the error term, ϵ_i. It is also assumed that censoring is independent of failure time. Note that an intercept term is not calculated in the presence of censoring. Estimation can be done using either a rank-based approach [Jin et al., 2003] or an extension of the least squares approach [Jin et al., 2006].

The semiparametric AFT model provides an attractive alternative because it directly relates the effect of explanatory variables on the survival time allowing an easier interpretation of the results. However, its use has been limited due to its computational difficulties, especially in presence of censoring, and further analytical and graphical methods are needed for model checking and outlier detection. Presently, an R package has been written [Huang and Jin, 2007] that corresponds to the method of Jin, 2006 [Jin et al., 2006]. This method can also accommodate multiple failure times and clustered data.

Cox Proportional Hazards Model

In 1972, (Sir) David Cox published his paper "Regression Models and Life Tables", in which he details what has come to be known as the Cox Proportional Hazards Model or Cox Regression [Cox, 1972]. Two major contributions from this paper include the concept of proportional hazards (though the model can also be extended to accommodate non-proportional hazards), and a new estimation method known as partial likelihood.

The model takes the following form:

$$h_i(t) = h_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}),$$

where $h_0(t)$ is the baseline hazard (unspecified, non-negative, and constant over time), and $(x_{i1}, x_{i2}, \ldots, x_{ik})$ is a vector of covariates. It can be alternatively written in integrated form as:

$$H(t) = \left(\int_0^t h_0(u) du \right) \exp(X\beta) = H_0(t) \exp(X\beta)$$

Estimation is done by partial likelihoods, rather than the full likelihood, because when the likelihood function is factored into 2 parts, one depending on $h_0(t)$ and the β vector, and the other only depending on the β vector, partial likelihoods only maximizes the portion depending on β alone. Partial likelihood is valid when there are no ties in event time. In the presence of ties, Breslow or Efron approximations to the partial log-likelihood can be used. This is a product, over the event times, of a quotient that compares the hazard of the individual with the event at t_j to the hazard of all the individuals at risk at t_j

$$L(\beta) = \prod_{Y_i uncensored} \frac{\exp X_i\beta}{\sum_{Y_j \geq Y_i} \exp(X_j\beta)}$$

Such that the log of the partial likelihood is:

$$l(\beta) = \log L(\beta) = \sum_{Y_i uncensored} \{X_i\beta - \log \sum_{Y_j \geq Y_i} \exp(X_j\beta)\}$$

Coefficients can be interpreted as their effect on the hazard function, $h_i(t)$. The effect of the covariate is multiplicative on the hazard, meaning, values greater than 1 indicate an increase in the hazard rate, while those between 0 and 1 indicate a decrease in the hazard rate.

Some of its advantages over the parametric models are that it does not make any assumptions on the distribution of survival times, can accommodate time dependent covariates, can allow for stratified analyses, can allow for discrete or continuous measures of time, and can accommodate tied data. The interpretation of covariate effects is a bit less intuitive since it is on the hazard of an event, rather than directly on survival time. Since it has become standard in most software packages, predictions can also be made easily. This method has been used extensively in disease prediction modeling, one such example being the 1998 Framingham model for cardiovascular disease.

A disadvantage of this method is that it makes the assumption of proportional hazards, meaning, that the effect of each covariate is the same at any time-point. If the effect of a variable is known to vary with time, then it would violate the proportional hazards assumption. This can be checked using residuals (Martingale or Schoenfeld), or adding a time-dependent covariate to the model [Kalbfleisch and Prentice, 2002]. If the proportional hazards assumption is violated, then a stratified analysis, or inclusion of a time-dependent variable should satisfy the violation. Interpretation of the time-dependent variable would also be altered. A second disadvantage is that in the presence of competing terminal events, the competing event would be treated as a censored observation, which has been shown to lead to over-estimation of a predicted risk [Pepe and Mori, 1993, Dignam and Kocherginsky, 2008].

Conclusion

The above models all focus on the analysis of a single endpoint, and do not take into account that there may be other outcomes whose occurrence would preclude the occurrence of the event of interest, which is known as a "competing event". It has been shown

[Dignam and Kocherginsky, 2008] that when there is a nontrivial percentage of competing risks, failure to account for them could lead to overestimation of the results. Therefore, when competing risks are present, every effort should be made to include them in the analysis.

2.4 Models for Competing Risks

Very often in studies of human disease processes, there are multiple possible endpoints or events. Under certain scenarios it is likely that a patient may experience more than one event, for example, a disease recurrence, hospitalization for treatment complication, or additional unrelated illness. The fact that the subject could experience any number of these events, in any order, means that they are not necessarily competing events. In an analysis where a subject may die from a different event prior to experiencing the outcome of interest, such as a cardiovascular death before a cancer recurrence, the event that would preclude the event of interest is considered the "competing event". This specific type of competing event is also called a *terminal* event, one that would make the patient no longer at risk for experiencing any further outcomes. Analyses which involve a terminal event and non-terminal event are also referred to as *semi-competing* risks.

2.4.1 Definitions and notation

Competing risk framework requires some additional definitions and notations beyond traditional survival, or time-to-event, analyses.

Cause-specific density

A cause-specific density represents the unconditional probability that a subject fails at time t of cause j, in the presence of covariates x, and can be written

$$f_j(t,x) = \lim_{dt \to 0} \frac{Pr(t \leq T \leq t + dt, J = j|x)}{dt}$$

The **overall density** of deaths at time t is the summation of the j individual causes of death at time t:

$$f(t,x) = \sum_{j=1}^{m} f_j(t,x)$$

Cause-specific hazards

A cause-specific hazard rate represents the instantaneous risk of dying at time t, of cause j, conditioning on having survived to time t, and in the presence of covariates, x, is denoted:

$$h_j(t,x) = \lim_{dt \to 0} \frac{Pr(t \leq T \leq t+dt, J = j | T \geq t, x)}{dt}$$

Similarly, the **overall hazard** , $h(t)$, is the summation of the m cause-specific hazards, up to time t:

$$h(t,x) = \sum_{j=1}^{m} h_j(t,x)$$

The cumulative cause-specific hazard, $H_j(t)$ equals the cause-specific hazard, h_j summed from the start of observation time, until time t. Then, the cumulative (overall) hazard, is the sum of all mutually exclusive hazards, such that:

$$H(t) = \sum_{j=1}^{m} H_j(t)$$

and has corresponding survivor function $S(t)$, which is the probability of remaining event free past time t, and $S(t) = \exp[-H(t)]$.

Cumulative Incidence Function

The cumulative incidence is defined as the probability of failing from cause, j, by a specified time, t, and is denoted:

$$F_j(t) = Pr(T \leq t, J = j) = \int_0^t f_j(u)du = \int_0^t S(u)h_j(u)du$$

for $t > 0$.

It is the integration up to t of the individual cause-specific densities. It involves the probability of having not failed from some other event up to time t, denoted $S(u)$, and the cause-specific hazard for the event of interest, $h_j(u)$. It is also known as the sub-distribution, marginal probability function, crude incidence, and absolute cause-specific risk.

2.4.2 Cause Specific Hazards Model

In the presence of covariates, the standard method for analyzing competing events data would be to model the cause-specific hazard functions of the different failures under a proportional hazards assumption. This is also sometimes referred to as a stratified Cox model.

The model takes the form:

$$h_j(t; X) = h_{0j}(t) \exp(X\beta_j)$$

Here, the cause-specific hazard is modeled as a function of a baseline hazard, h_{0j}, and a set of covariates, X. When failure type j is modeled, the other failure types are regarded as censored at the time of failure and standard Cox methods are used. This method assumes that causes of failure are not interrelated, an assumption that is difficult to verify statistically, and requires background knowledge on the subject matter from experts in the field [Kalbfleisch and Prentice, 2002].

A stratified accelerated failure time model can also be used.

$$h_j(t;X) = h_{0j}\left\{ \int_0^t \exp\left[X\beta_j\right]du \right\} \exp\left[X\beta_j\right]$$

If it is assumed that the hazards of the various causes are proportional, then a proportional risk model can also be used:

$$h_j(t;X) = h_{0j}(t) \exp\left[\gamma_j + X\beta_j\right]$$

This assumption can be checked graphically by plotting $\log h_{0j}(t)$, versus t.

Under this model, the covariate effects on the cause-specific hazard of the event of interest cannot be directly interpreted in terms of the cumulative incidence function, but are rather only on the hazard of the failure type being modeled. This may be particularly useful when studying the effect of a treatment that is specific to the outcome of interest, but not the other failures, such as a chemotherapy agent that may improve survival from cancer but should not have an effect on deaths from other causes.

Advantages of this model are that it can be fit on any standard software that can fit a Cox model, by stratifying on cause of failure or event type. A disadvantage is that it does not have a direct interpretation in terms of survival probabilities for each event type.

2.4.3 Fine and Gray Model of the Subdistribution Hazard

In the presence of covariates, the cumulative incidence function (probability that an event of type j has occurred by time t) can be written:

$$F_j(t;X) = \int_0^t f_j(u;X)du$$

Using the relationship between the survival function, hazard function , and the cumulative incidence function

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)},$$

the subdistribution hazard (hazard of the cumulative incidence) for cause j, thought of as the hazard for an individual who either fails from cause j or does not, can be written:

$$h_j^*(t;X) = \frac{f_j(t)}{1 - F_j(t)}$$

By mimicking a proportional hazards model, the Fine and Gray model is specified as:

$$h_j^*(t;X) = h_{0j}^*(t)\exp[X\beta_j],$$

where $h_{j0}^*(t)$ is the baseline subdistribution hazard for failures of type j, and $\exp[X\beta_j]$ is the relative risk associated with the covariates.

Estimation follows the partial likelihood approach used in a standard Cox model since a proportional hazards assumption is imposed on the subdistribution hazards.

$$L(\beta) = \prod_{i=1}^{n} \Big(\frac{\exp(X_i\beta)}{\sum_{j \in \mathscr{R}_i} w_{ij} exp(X_j\beta)} \Big)^{\delta_i = 1}$$

The weights, w_{ij}, are used when censoring occurs. Patients who do not experience the event of interest before t are given a weight of 1, whereas patients who experience competing events before t are given a weight:

$$w_{ij} = \frac{\tilde{G}(T_i)}{\tilde{G}(\min(T_j, T_i))}$$

where $\tilde{G}$ is the Kaplan-Meier estimate of the survival function of the censoring distribution, which is the cumulative probability that a patient is still being followed at time t.

This looks very similar to the Cox proportional hazards model, except that the modified risk set, $\mathscr{R}$, includes both those who are alive and at risk for the event of interest at time t, and also those who have failed prior to time t from some other cause, and thus have an infinite failure time for cause j:

$$h_j^*(t;X) = \lim_{dt \to 0} \frac{Pr(t \leq T \leq t + dt \text{ and } J = j | T > t \text{ or } T \leq t \text{ and } J \neq j)}{dt}$$

For patient prediction, the baseline subdistribution hazard can be obtained using a variation of Breslow's estimator that incorporates the modified risk sets and the gradual reduction of weights for those who failed from competing causes and were retained in the risk set [Kohl et al., 2015], estimated by

$$\hat{H}_{1,0}(t) = \sum_{i=1}^{n} \int_0^t \frac{1}{\sum\limits_{a \in \mathscr{R}(s)} w_a(s) \exp(x_a \hat{\beta})} dN_i(s),$$

where $\mathscr{R}(s)$ denotes the risk set defined above, $w_a(s)$ are the weights, $\hat{\beta}$ are the coefficient estimates from the SH model, and $dN_i(s)$ is the increment in the counting process $N_i(t) = I(t_i \leq t | j = 1)$, describing the status of subject i with respect to event type 1 in the interval $[t, t + dt]$. (This counting process changes from 0 to 1 at the event time t_i if the event type 1 occurred at that time.)

For the cause-specific hazard, the risk set decreases at each time point at which there is a failure of another cause, while the Fine and Gray risk set continues to include those who have failed from other causes prior to time t.

Though this may not seem to be a very intuitive way of classifying the at-risk set, for mutually exclusive event types, those who fail from one cause are not vulnerable to fail from another. It is also a "convenient" way to model the cumulative incidence function due to the log-log transformation, corresponding to a proportional hazards model

$$\log(-\log(1 - F_j(t, x))) = \log(-\log(1 - F_{j0}(t))) + X\beta_j$$

which means that the covariate effect is to shift the transformed cumulative incidence function up or down by an amount, such that the interpretation becomes more straightforward: positive coefficients indicate increases in the cumulative incidence function, and negative coefficients indicate decreases in the CIF.

The model is semi-parametric, and when censoring is absent, or always observed (such

as in a finite time-fixed study with complete follow-up through the end of study), can be estimated by a partial likelihood approach, analogous to the proportional hazards model, with the only difference being in the definition of the risk set. In the presence of right censoring, an inverse probability of censoring weighting (IPCW) approach is used to construct the score function for the partial likelihood.

It is important to note that under this method, any covariate effects on the cumulative incidence may either be direct effects, such as strictly lowering the cumulative incidence of that specific failure type, or indirect effects, such as the cumulative incidence of the failure of interest is lowered due to increases in competing failure types. Currently, there are packages in R (cmprsk) and Stata (stcrprep) to estimate the model, and SAS macros (cuminc, cumincv, and PSHREG [Kohl et al., 2015]) to estimate the predicted cumulative incidence function.

2.4.4 Semi-Competing Risks Model

When studies involve a terminal event that censors a non-terminal event, for example death from any cause (terminal) and disease relapse (non-terminal), this is called *semi-competing* risks data. With such data, it may not be possible to make non-parametric inferences on the non-terminal event without imposing some additional assumptions, due to the dependent censoring. If only the time and type of the first event are recorded, then the data can be analyzed using traditional approaches. However, it will often be the case that the dependence between the two events is of scientific interest, and inferences on the **non-terminal event** are desired. Additionally, the joint distribution is only important (can only be identified) in the"lower wedge" (the space where the event of interest is not censored by the terminal event).

Several methods for this type of analysis have been examined [Fine et al., 2001, Peng and Fine, 2007, Hsieh and Huang, 2012]. If the two events are positively correlated, then a gamma frailty model can be used [Fine et al., 2001].

In the presence of a discrete covariate, Hsieh et. al. [Hsieh et al., 2008] proposed a two-stage approach in which the first stage is to specify an Archimedean copula model for the dependence structure between the two event times. Here, separate copula models are assumed for separate classes of a discrete covariate to allow for different dependence structures between groups. It is further assumed that censoring is not related to the covariates. In the second stage, the regression parameter is modeled in terms of the marginal distribution for the non-terminal event using the following model:

$$h(t) = -X'\theta + \epsilon,$$

where X is the $p \times 1$ discrete covariate vector, θ is the $p \times 1$ parameter vector, $h(t)$ is a monotonic increasing function, and ϵ is the error term. If the error term follows an extreme value distribution, then the model becomes the Cox proportional hazards model. If it follows the standard logistic distribution, then it becomes the proportional odds model. If instead, $h(t) = \log(t)$, then it is an accelerated failure time model. Regardless of model choice, estimating equations are proposed to test for covariate effects.

Another analytic choice is to use a time-varying effect in a Cox proportional hazards model for the non-terminal event [Hsieh and Huang, 2012], while assuming a copula model for the dependence structure between the two events. Under the copula model, the joint distribution of the two events can be expressed as a function of the marginal distributions and the association parameter through the copula function C. This makes it possible to discuss the marginal distributions of the terminal and non-terminal events separately, as well as their dependence.

$$Pr(T_1 > s, T_2 > t|X) = \mathbf{C}Pr(T_1 > s|X), Pr(T_2 > t|X), \alpha_0(s,t), 0 \leq s \leq t$$

Where T_1 and T_2 are the non-terminal and terminal event times, respectively, $\mathbf{C}(T_1, T_2, \alpha)$ is the copula function, and α is the association parameter. The parameters are estimated

using conditional likelihood

The method of [Hsieh and Huang, 2012] can be performed using nlminb in R. One controversy over this method is whether the marginal distribution of the non-terminal event is meaningful, such as a disease progression in the absence of death [Kalbfleisch and Prentice, 2002]. However, there remain specific instances where this information may be of value, such as evaluating a new drug or treatment, especially on the subset of patients that would remain alive long enough to benefit from it. Additionally, the method presented in [Hsieh and Huang, 2012] does not include covariates, but instead focuses on estimating the marginal distributions of the two events. To make comparisons between the placebo and treatment groups, they use plots of the estimated survival of each group, which may be valid in a randomized, controlled, double-blind trial, but not necessarily in an observational study. Prior work by the same authors presented an analytic method for discrete covariates only, but variance estimation was predominantly based on resampling techniques [Hsieh et al., 2008].

2.5 Multi-State Models

Multi-state models can be considered as a generalization of competing risk models, and can also be used to analyze the transition to an intermediate event in an illness-death process, such as the movement to disease recurrence prior to death. These intermediate events may provide additional information on the disease/recovery process that can aid in treatment, and also allow for greater prognostic accuracy for patients. In this type of model, movement to event states are called *transitions*, and can be modeled using standard statistical software, with some extra data preparation [Putter et al., 2007, Putter, 2014]. Data need to be entered in the counting process style, in which a period of observation (episode) is described by a starting time, an ending time, and a reason for ending (transition or censoring) [Willekens and Putter, 2014]. Specifically, for a multi-state model this means that

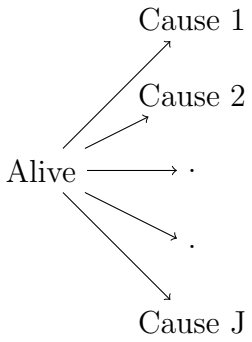


Figure 2.1: Competing Risks as a multi-state formulation

each subject has as many rows as the number of transitions for which they are at risk [de Wreede et al., 2010]. The data preparation and analysis methods discussed here can be found in the *mstate* package in R.

In the multi-state setting, a competing risk model can be represented as a series of transitions from an initial state, to a number of different endpoints, as illustrated in Figure 2.1.

A fundamental concept in multi-state models is the transition intensity or hazard rate $\alpha_{gh}(t)$, which denotes the instantaneous risk of a transition from state g into state h at time t. It is defined as:

$$\alpha_{gh}(t) = \lim_{dt \to 0} \frac{Pr(X(t+dt) = h | X(t) = g)}{dt}$$

These intensities can be gathered into an $S \times S$ matrix, $A(t)$, with diagonal elements $A_{gg}(t) = \sum_{h \neq g} A_{gh}(t)$. Individuals who have no transition (do not experience any of the endpoints) remain in state g.

Then transition probability matrix, $P(s,t)$ is the quantity of interest. It has elements $P_{gh}(s,t) = P(X(t) = h | X(s) = g)$, which denote the transition probability from state g to state h in the time interval $(s,t]$. Specifically for the case of competing risks models, the cumulative incidence function expresses the probability of failing of cause j before time t and is given by $P_{1j}(0,t)$. Classical survival models that have only a single transition, from

being alive to death, can be written as a two-state model $S(t) = 1 - P_{12}(0,t) = P_{11}(0,t)$

The effect of covariates can be modeled using the semi-parametric Cox proportional hazards model, treating each transition as its own stratum:

$$\alpha_{gh}(t|\mathbf{X}) = \alpha_{gh,0}(t)\exp(\boldsymbol{\beta}_{gh}^T\mathbf{X}),$$

where $\alpha_{gh,0}(t)$ is the baseline transition hazard from g to h, and the vector $\boldsymbol{\beta}_{gh}$ represents the covariate effects on transition gh. This is equivalent to performing separate Cox regressions for each of the transitions $g \to h$. In the above model, the covariates are fixed at baseline and allowed to have a different effect on each transition. However, it is also possible to specify time-varying covariates, as well as forcing them to have an identical effect on each transition, if biologically meaningful. It is also possible to specify that the transition hazards remain proportional. Using subject-specific transition rates and the Cox model, the package also gives estimates of subject-specific transition probabilities and the associated variance-covariance matrices, based on the Aalen-Johansen estimator.

In recent years the number of packages for the estimation of multistate models has increased rapidly, with R being the language of choice. The use of multi-state models for modeling competing risk data has also been detailed by Beyersmann and Allignol [Beyersmann et al., 2012] (package *MVNA*), Jackson [Jackson et al., 2003] (package *msm*), and Ferguson [Ferguson et al., 2012](package *MsSurv*), though methods of data entry, model specification, and estimation differ between authors. None of these packages has been extended to cover the Fine and Gray model in a multi-state setting since standard errors of the estimated hazards and cumulative incidence functions are not available [de Wreede et al., 2010].

Discussion

In addition to the aforementioned methods for handling competing events, there have been other methods proposed by Klein and Andersen [Klein and Andersen, 2005], Zhang and Scheike [Zhang et al., 2008], Gerds and Scheike [Gerds et al., 2012], and Ishwaran and Gerds[Ishwaran et al., 2014] that examine additive hazards, pseudovalues, binomial regression, or random forests. Quantile regression [Portnoy, 2003] and transformation models [Cheng et al., 1995, Chen et al., 2002] including those with a cure fraction [Zeng et al., 2006] are other examples of methods that have also been used to analyze right-censored survival data, but have not been adapted to the competing risk setting. Due to the lack of readily available software to implement some of these more complex methods, combined with the user-friendly interpretation of the subdistribution hazards model and cause specific hazards, in a way that is analogous to the Cox PH model, they have remained the most widely applied models in the field of competing event analysis to date.

Chapter 3

Established Disease Risk Models

3.1 Models for Predicting Risk of Cardiovascular Disease and Mortality

Focusing on externally validated, widely used risk models applicable to females, there are several major cardiovascular disease (CVD) risk models used in practice in the U.S. today: the Framingham Risk Score (FRS) [D'Agostino Sr et al., 2001] , FRS Adult Treatment Panel III (ATP-III) for hard CHD [Wilson et al., 1998], the Systematic Coronary Risk Evaluation (SCORE) model for CVD mortality [Conroy et al., 2003], and the Reynolds risk score [Ridker et al., 2007]. Due to the age distribution of our population of breast cancer survivors, it may also be useful to investigate models that were developed on an older population, or one with a prior history of CVD, such as the SCORE OP (older persons) [Cooney et al., 2015], the Coronary Risk in Elderly (CORE) model [Koller et al., 2012], and the Framingham risk score for those with a history of a CVD event [D'Agostino et al., 2000]. This is not a comprehensive list of available cardiovascular disease risk models, but rather a summary of models that may be most useful for our patient age and geographical distribution, as well as containing covariates that can easily be obtained from patient electronical medical records. The use of genetic assays or complex biomarkers as predictors for these

mortality endpoints, while currently being pursued in the literature, is beyond the scope of this work.

3.1.1 Framingham Models

Perhaps the most widely used cardiovascular disease risk model in the United States remains the Framingham Risk Score. It was first established in the 1950s and 1960s, when investigators from the Framingham Heart Study developed a series of equations that could be used by clinicians to predict a person's risk of coronary heart disease (CHD), which included the following clinical endpoints: coronary heart disease, brain infarction (stroke), intermittent claudication (atherosclerosis), and congestive heart failure (in the absence of coronary or rheumatic heart disease) [Gordon et al., 1973]. Risk factors examined were: age, gender, systolic blood pressure (BP), serum cholesterol, cigarette smoking, glucose intolerance, and left ventricular hypertrophy (LVH). In 1973, a handbook containing tables from these equations was produced, followed shortly by a pocket reference card for physicians.

1976 Model

The 1976 function [Kannel et al., 1976] was based on a logistic regression model, in which the endpoint included any occurrence of coronary heart disease, congestive heart failure, cerebrovascular disease, or intermittent claudication. Individual models were also presented for each type of heart disease, but use of the conglomerate endpoint was recommended for screening for high risk individuals. The model was developed on an all-white cohort aged 37 to 75, and separate models were calculated for men and women. A person's risk of disease in the next 8 years is calculated based on a set of variables measured at baseline, conditional on being free of disease at baseline. The female specific equation for the overall risk of CHD in the next 8 years was based on 320 female cases of CHD during follow-up, and contained

the following coefficients and covariates:

$$\begin{aligned} P(\text{CHD}) &= [1+\exp\{-(-16.46+0.27*\text{Age}+-0.001*\text{Age}^2 \\ &\quad +0.016*\text{Serum cholesterol}\ +0.014*\text{SBP}\ +0.04*\text{smoking} \\ &\quad +0.875*\text{LVH}+0.682*\text{glucose intolerance} \\ &\quad -0.0002*\text{cholesterol}*\text{age})]^{-1}\} \end{aligned}$$

Authors recommended that persons in the highest decile be considered as high-risk. No model validation or checking techniques were presented in this paper.

1991 Model

In 1991, the Framingham investigators updated the equations with a larger, more contemporary data set, with baseline evaluations between 1968-1975 [Anderson et al., 1991b, Anderson et al., 1991a]. They combined the original cohort with new offspring data, and created models that could predict outcomes from ages 30 to 74. The study included 5,573 persons (2,983 women and 2,590 men). In addition, information was now collected on high density lipoprotein (HDL), instead of just total serum cholesterol. Risk factors examined were age (years), sex (female=1), SBP (average of 2 measurements, mm Hg), DBP (average of 2 measurements, mm Hg), cholesterol (total serum cholesterol), HDL cholesterol (mg/dl), smoking (1= current smoking or quit within past year), diabetes (treatment with insulin or oral agents or fasting glucose of 140 mg/dl or above10), and ECGLVH.

In the updated analysis, an accelerated failure time (AFT) model (Weibull distribution) was used to create the model, with two variations. Let T denote the time from baseline until the first occurrence of any CHD, μ a location parameter, and σ a scale parameter for $\log(T)$. Then

$$\frac{\log(T)-\mu}{\sigma} \sim \text{ extreme value distribution}$$

implying that T follows the Weibull distribution. Then assume that μ depends on risk factors

as:

$$\mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

and that

$$\log(\sigma) = \theta_0 + \theta_1 \mu$$

The models were estimated using maximum likelihood in Proc NLIN in SAS, and all events occurring in the first four years were coded as 4 years to avoid influence of early onset outliers.

Two forms were presented for the results: the actual mathematical equations, updated to include HDL, and also a point scoring system for 5 or 10-year predictions, based on the equations 7. The authors also note that either SBP or DBP is to be used, not both, and recommend SBP for users without a strong preference. In order to use the model to make predictions, several interim calculations are necessary, and will be presented for females, using SBP. The first step is to calculate an interim value, a, that is based on the risk factors:

$$\begin{aligned} a &= 11.1122 - 0.9119 * \log(\text{SBP}) - 0.2767 * \text{smoking} \\ &\quad -0.7181 * \log(\text{cholesterol/HDL}) - 0.5865 * \text{ECGLVH} \end{aligned}$$

Next, a second interim value, m, is calculated based on a, log of age, and diabetes, and again has different values for men and women. For women the calculation is:

$$m = a - 5.8549 + 1.8515 * [\log(\text{age}/74)]^2 - 0.3758 * \text{diabetes}$$

Then, μ and σ are computed (for women) as:

$$\mu = 4.4181 + m$$

$$\sigma = \exp(-0.3155 - 0.2784 * m)$$

Lastly, select a prediction time frame between 4 and 12 year, denoted t and compute:

$$u = \frac{\log(t) - \mu}{\sigma}$$

and then the predicted probability for t is:

$$p = 1 - \exp(-e^u)$$

To simplify predictions, a worksheet was also provided, and can be found in Appendix: 7.

While no specific model validation is presented in the original paper, the authors explain that the variable selection was based on previous studies and that log transformation of continuous covariates, as well as a quadratic term for age in the female model, were chosen for "improved model fit", though the evaluation metric was not explicitly stated. They warn that the equation is to be used only when there is information on all the risk factors, and while some important risk factors such as family history and obesity were not included, they should be taken into account when counseling patients. Additionally, the risk for patients with very high values for SBP or fasting glucose (the upper percentiles) may not be as accurately predicted by the model, as well as for populations with very low CHD incidence.

A companion paper [Anderson et al., 1991a] was also published in the same year that presents detailed equations for each of the individual endpoints, separately for men and women, and uses the same cohort and statistical methods. In this paper, the details for the confidence interval calculation, hazard ratios, and excess risk are also outlined. Comparisons with the overall model for CHD conclude that, with the exception of stroke, the individual models perform similarly to the overall model in identifying the highest risk decile.

1998 Model - ATP III Model

In 1998, Framingham investigators again modified the equation to include the Joint National Committee (JNC-V) blood pressure categories and the National Cholesterol Education Program (NCEP) ATP III cholesterol categories [Wilson et al., 1998]. This model is sometimes referred to as the Framingham-ATP III model or modified Framingham risk score. The analysis was based on the same cohort used in the 1991 paper, with 12 years of follow-up. The endpoint of CHD included angina pectoris, recognized and unrecognized myocardial infarction, coronary insufficiency, and coronary heart disease death. Hard CHD events included total CHD without angina pectoris.

In this analysis, Cox models were used, though AFT models were also considered (results not shown), with separate models for each gender. Again, corresponding score sheets were presented for simplification in risk calculation, and are in appendix 7. The overall C index [Pencina and DAgostino Sr, 2004] was used to evaluate discrimination and whether individual covariates could be considered independent predictors of CHD. Results were nearly identical when blood pressure and cholesterol were entered as continuous or categorical covariates, thus for ease of application, categories were presented.

Similarly to the 1991 equations, predictions are calculated in a step-wise fashion where first the β coefficients from the model are used to calculate a value, L, for an individual as follows:

$$
\begin{aligned}
L = {} & 0.33766 * \text{age} - 0.00268 * \text{age}^2 - 0.26138(\text{chol} < 160) \\
& + 0.20771(\text{chol 200 - 239}) + 0.24385(\text{chol 240 - 279}) + 0.53513(\text{chol} \geq 280) \\
& + 0.84312(\text{HDL} < 35) + 0.37796(\text{ HDL 35 - 44}) + 0.19785(\text{HDL 45 - 49}) \\
& - 0.42951(\text{HDL} \geq 60) - 0.53363(\text{BP optimal}) - 0.06773(\text{BP high normal}) \\
& + 0.26288(\text{BP stage I hypertension}) + 0.46573(\text{if BP stage II-IV hypertension}) \\
& + 0.59626(\text{diabetes}) + 0.29246(\text{smoker})
\end{aligned}
$$

Next, the function is evaluated at the mean level of each variable, call it G. For women, using cholesterol, the value is 9.92545, using LDL, it was 9.914136. This value is subtracted from L to get the value A. Next, A is exponentiated such that:

$$B = \exp^A$$

Lastly, the value is inserted into a survival function to calculate the 10 year probability of CHD:

$$P = 1 - [s(t)]^B$$

where $s(t) = 0.96246$ if using cholesterol and 0.9628 if using LDL. This calculation can be done much more concisely using the conversion to a point system in the provided score sheet.

Framingham 2000 and 2001

The original Framingham models were developed on a population that was considered healthy at baseline, but growing interest in predicting recurrent events, or events in an already high-risk individual, began to emerge, such that in 2000 a model was developed that could be applied to persons with a previous history of cardiovascular disease [D'Agostino et al., 2000]. The subsequent CHD endpoints are mainly hospitalizations for any of: myocardial infarction, coronary insufficiency, angina pectoris, and coronary death. Authors again chose an accelerated failure time (AFT) model with a Weibull distribution, and the following risk factors were included in the final model (for women): age, log ratio of total cholesterol to HDL cholesterol, diabetes, log-transformed SBP, and smoking. This study also examined alcohol consumption and triglyceride levels, but neither was found to significantly improve prediction. This paper presented a worksheet for predicting a CHD even in the next 2 years, as well as a description for calculating a 4 year risk based on the model estimates.

Other concerns over the earlier versions of the Framingham risk scores were its inclusion

of non-fatal endpoints, inclusion of endpoints not routinely examined in clinical trials, and generalizability to other populations. To address this, in 2001, D'Agostino and colleagues presented an analysis in JAMA in which they evaluated the Framingham model on 6 new cohorts from a variety of populations: Atherosclerosis Risk in Communities Study (ARIC ,black and white), Physicians Health Study (PHS, white, men only), Honolulu Heart Program (HHP, Japanese American, men only), the Puerto Rico Heart Health Program (PR, Hispanic, men only), the Strong Heart Study (SHS, native American), and the Cardiovascular Health Study (CHS, white) [D'Agostino Sr et al., 2001]. They also altered the endpoint definition to include only "hard CHD" events, namely, coronary death or myocardial infarction. For each cohort they evaluated the prediction using the betas from the 1998 Framingham Model and also calculated new betas for the same risk factors but on each of the current cohorts, calling it a "Framingham recalibrated" model. They again evaluated the best Cox model using the C-index. For female specific models, a number of coefficients were not statistically significant, possibly due to low numbers of observed events. Several differences in RRs were seen in a few of the non-Framingham cohorts, specifically Japanese American and Hispanic men and Native American women. As a result of this paper, it was advised that when the Framingham model did not seem to fit a particular cohort well, one should recalibrate the existing risk factor categories to the cohort under investigation, in order to produce more accurate predictions.

Framingham 2008

In 2008, the investigators presented an analysis in Circulation, in which they created a single endpoint that is the *first* of any CVD event, including CHD coronary death, myocardial infarction, coronary insufficiency, angina, cerebrovascular events (including ischemic stroke, hemorrhagic stroke, and transient ischemic attack), peripheral artery disease (intermittent claudication), and heart failure [D'Agostino Sr et al., 2008]. The authors explain that while individual risk factor effects may vary from one specific CVD endpoint to another, there is

enough commonality of risk factors to allow for a single, general, risk prediction tool that can predict overall risk as well as the individual endpoints. They also created a tool that can calculate a person's "heart age", which would be the age of a person with the same predicted CVD risk, but with all risk factors in normal ranges. Authors believed that this simplified tool would be of greater use in a general practice setting, rather than examining several individual assessment methods. For this tool they elected to use a Cox proportional hazards model, and evaluated predictions at 10 years.

To create the prediction model, information about CVD events on follow-up was obtained with from medical histories, physical examinations in clinic, hospitalization records, and communication with personal physicians. Sex-specific functions were created, and all continuous variables were log-transformed. An additional modification was to model the effect of systolic blood pressure differently for those already on anti-hypertensive medication. Two risk scores were presented, one based on all traditional risk factors (age, HDL cholesterol, total cholesterol, SBP, BP treatment, smoking and diabetes), and another on nonlaboratory-based predictors (age, body mass index, systolic blood pressure, antihypertensive medication use, current smoking, and diabetes status). The rationale was that this tool could be used in settings where laboratory based testing was not always feasible, such as in populations without regular access to healthcare, or a quick primary care consultation. Of the 4,522 women, 456 developed a cardiovascular event during follow-up, and the equations to predict 10-year CVD risk were calculated using the following:$\hat{p} = 1 - S_0(t)^{\exp\left[\sum_{i=1}^{p} \beta_i X_i - \sum_{i=1}^{p} \beta_i \bar{X}_i\right]}$

Where, in the lab-based model for women; $S_0(t) = 0.95012$, $\sum_{i=1}^{p} \beta_i \bar{X}_i = 26.1931$, and

$$\begin{aligned}\sum_{i=1}^{p} \beta_i X_i \quad &= 2.32888 * \log(\text{age}) + 1.20904 * \log(\text{total chol}) - 0.70833 * \log(\text{HDL}) \\ &+0.52873 * \text{smoker} + 2.76157 * \log(\text{SBP}) * (\text{not on BP meds}) \\ &+2.82263 * \log(\text{SBP}) * (\text{on BL meds}) + 0.69154 * \text{diabetes}\end{aligned}$$

and in the non-laboratory based model for women: $S_0(t) = 0.94833$, $\sum_{i=1}^{p} \beta_i \bar{X}_i$ was not

provided, and

$$\begin{aligned}\sum_{i=1}^{p}\beta_i X_i &= 2.72107 * \log(\text{age}) + 0.51125 * \log(\text{BMI}) \\ &\quad +2.81291 * \log(\text{SBP}) * (\text{not on BP meds}) + 0.61868 * \text{smoker} \\ &\quad +2.88267 * \log(\text{SBP}) * (\text{on BL meds}) + 0.69154 * \text{diabetes}\end{aligned}$$

Models were evaluated for discrimination and calibration by use of the C-statistic and Hosmer-Lemeshow statistic, respectively. The C-statistic for women was 0.793, indicating good discrimination, while the Hosmer-Lemeshow statistic of 7.79, on 9 degrees of freedom, indicated excellent calibration. Calibration plots also showed good calibration for both sexes. Current models were also compared to previous ones using net reclassification improvement based on patient's predicted risk categories, and the current model showed a nearly 8% improvement over the previous Framingham model.

Other Framingham based publications of note

While there have been numerous validation studies of the Framingham models, there are two worth pointing out due to their use of publicly available NHANES data [Liao et al., 1999][Gaziano et al., 2008].

In 1999, Liao et al. used pooled National Health and Nutrition Examination Survey (NHANES) I and II data, and the variables used in Framingham Model, to fit separate models for blacks and whites. This was the first study to examine the performance of the newest Framingham model in an African American population. They found similar results overall, with only age risk significantly different between black and white women. When the model for white women was used to predict risk for black women , the predicted cumulative CHD mortality was very similar to the observed values, suggesting there may be a smaller difference in CHD deaths by race for women than was observed in men (where the model developed on a predominantly white population overestimated black CHD mortality).

In 2008, Gaziano and colleagues [Gaziano et al., 2008] also used NHANES data to assess whether a risk algorithm that did not include laboratory based tests could predict as accurately as one that did, in order to have a screening tool for low-income settings. The results showed that using BMI instead of cholesterol can as accurately predict risk of CHD events.

Summary

Though it has been widely accepted, modified, and validated, the main criticism of the Framingham models is its generalizability, due to their development on a predominantly white cohort. Other limitations are the exclusion of other known risk factors including, but not limited to: family history, obesity, physical activity, use of hormonal therapy and antihypertensive therapy. However, the risk factors in the models remain the most important risk factors associated with CVD, and recalibration of the model parameters has proven useful for populations that may not be directly comparable to the one on which it was developed. It remains the most widely used CVD risk model in the United States today, both clinically and for research purposes.

3.1.2 ATP III guidelines

The third Adult Task Force (ATP) updated the guidelines for management of high cholesterol in 2001 [Adult Treatment Panel (ATP), 2001]. Their goal was to prevent coronary heart disease in patients who are considered high risk for CHD, and recommend assessment of LDL, HDL and total cholesterol every five years. They define cutoffs for normal range for each type of cholesterol and list other high-risk factors:

- Cigarette smoking
- Hypertension ($BP \geq 140/90 mmHg$ or on antihypertensive medication)
- Low HDL cholesterol (<40 mg/dL)

- Family history of premature CHD (<55 in male first-degree relative, <65 in female first-degree relative)
- Age (men $\geq$ 45 years; women $\geq$ 55 years)

If a person has a high HDL cholesterol ($\geq 60 mg/dL$), then one risk factor is subtracted from the count. If a person has type 2 diabetes, they are classified as having a CHD risk equivalent. Presence of any 2 or more of these risk factors would then lead to calculation of a person's 10-year risk via the 1998 Framingham model for hard coronary heart disease endpoints (myocardial infarction + CHD death)[Wilson et al., 1998]. The rationale for the cutoff of two risk factors is that persons with zero or one risk factor, would most likely have a 10-year model-based risk of CHD less than 10%, and therefore it is not necessary to obtain the full information for the Framingham calculation. The remainder of the publication lists guidelines for clinical management of cholesterol in the various combinations of risk groups, as well as recommendations to improve adherence, and does not focus on validation of the risk prediction algorithm.

3.1.3 SCORE model

The goal of the Systematic Coronary Risk Evaluation (SCORE) project was to develop a risk assessment method for cardiovascular risk in Europe. Prior to its development, risk assessment in Europe had been carried out using Framingham scoring, however, there was some concern that the model developed on U.S. data would overestimate the risk of CHD in Europe, as was shown to be the case in both German and Dutch studies [Thomsen et al., 2002, Hense et al., 2003]. Additionally, the Framingham definition of non-fatal endpoints is different than what is used in many other cohort studies and clinical trials, and thus can be difficult to replicate. The SCORE project elected to examine only fatal cardiovascular disease events for ease in model development, as well as reproducibility across studies, citing issues with recreating the Framingham endpoint as a primary rea-

son. The justification was that most countries reliably capture mortality data, but do not uniformly collect data on non-fatal endpoints. In this study, cardiovascular mortality was defined as ICD-9 codes 401-414, 426-443, 798.1, and 798.2, excluding codes for "definitely non-atherosclerotic causes of death" (ICD-9: 426.7, 429.0, 430.0, 432.1, 437.3, 437.4, and 437.5

Data from 12 cohort studies in Europe were combined, yielding sample sizes of 88,080 women and 117,098 men. Risk functions were calculated using a Weibull proportional hazards model that had two parts: the first modeled the shape of the baseline survivor function (which was calculated separately for each of cohort and gender), and the second calculated the relative risk for each of the risk factors (which was assumed to not vary from cohort to cohort or by gender). The Weibull model was chosen over the Cox model so that the risk estimation equation could be written as a formula, but the authors verified the results with a Cox model to ensure that the additional assumption of the Weibull distribution did not significantly alter model performance. Lastly, the investigators decided not to model age as a risk factor, but rather construct the hazard function based on the person's age, so that survival could be estimated for the full range of observed ages in the study, rather than just the length of follow-up on study. (This is similar to what is done in many breast cancer risk models) The other risk factors included were sex, smoking, systolic blood pressure, and cholesterol (entered 2 ways: either total cholesterol or total/HDL ratio).

The female cohort experienced 7,934 cardiovascular deaths, of which 5,652 were deaths from coronary heart disease. Separate estimation equations were calculated for coronary heart disease and for non-coronary cardiovascular disease so that a person's total risk could be partitioned into its coronary and non-coronary components. It also allows one to calculate the reduction in end-points of each type resulting from specific risk factor interventions. Separate models were built for high and low-risk populations, but both contained the same set of covariates; gender, smoking status, total cholesterol or ratio of total to HDL, and systolic blood pressure. Discrimination was assessed using the area under the ROC curve,

and diagnostic performance was assessed by examining the positive clinical (diagnostic) likelihood ratios for various thresholds of risk. Charts were also presented for ease of risk estimation and correspond to a persons 10-year risk. Equations for calculations of other lengths of time were provided in the appendices and appear below:

Step 1: Calculate the underlying risks for coronary heart disease and for non-coronary cardiovascular disease for the person's current age and their age in ten years, using the values for α and p provided in table 3.1. The underlying survival probability, S_0,is given by:

$$S_0(\text{age}) = \exp\left[-(\exp(\alpha))(\text{age} - 20)^p\right]$$

$$S_0(\text{age} + 10) = \exp\left[-(\exp(\alpha))(\text{age} - 10)^p\right]$$

Table 3.1: α and p Coefficients for Step 1 - Women

	CHD		Non-CHD CVD	
	α	p	α	p
low-risk	-29.8	6.36	-31.0	6.62
high-risk	-28.7	6.23	-30.0	6.42

Step 2: Using the coefficients in table 3.2, calculate the weighted sum,w, of the risk factors cholesterol, smoking and systolic blood pressure. Two weighted sums will have to be calculated, one for CHD and one for non-CHD CVD.

$$w = \beta_{chol}(\text{cholesterol} - 6) + \beta_{SBP}(\text{SBP} - 120) + \beta_{smoker} * \text{smoker}$$

Step 3: Combine the underlying risks for CHD and for non-coronary CVD, at the person's age and at their age ten years from now (four calculations) which were calculated in step 1 with the weighted sum of risk factors from step 2 for the two end-points, to get the

Table 3.2: β Coefficients for Step 2 - Women

	CHD	Non-CHD CVD
Current smoker	0.71	0.63
Cholesterol (mmol/L)	0.24	0.02
Systolic BP (mmHg)	0.018	0.022

survival probability at each age for each cause:

$$S(\text{age}) = [S(\text{age})]^{\exp(w)}$$

$$S(\text{age} + 10) = [S(\text{age} + 10)]^{\exp(w)}$$

Step 4: For each cause, calculate the 10-year survival probability based on the survival probability for the current age and their age plus 10:

$$S_{10}(\text{age}) = S(\text{age} + 10)/S(\text{age})$$

Step 5: Calculate the 10 year risk for each endpoint:

$$Risk_{10} = 1 - S_{10}(\text{age})$$

Step 6:Combine the 10-year risks for CHD and non-coronary CVD by adding:

$$\text{CVDRisk}_{10}(\text{age}) = \text{CHDRisk}(\text{age}) + \text{Non-CVDRisk}(\text{age})$$

Some limitations of the SCORE models are that they were only calculated on the age range from 45 to 64, and do not include information on diabetes, due to the lack of uniformity with which the information was collected across cohorts. They also only use a single blood pressure and cholesterol value for each person which is thought to reflect their usual levels, but may not always be the case. Strengths of the model are the size of the cohort upon which it was developed, as well as the ability to separately model CHD and non-CHD CVD

risk. The utility of this model in a US female population is yet to be evaluated.

SCORE OP Model

Due to the limitation of its development on a population less than 65 years of age, in 2015, Cooney and colleagues re-evaluated the SCORE model on a pooled European cohort of over 40,000 individuals over 65 years of age [Cooney et al., 2015]. They excluded those with a history of myocardial infarction at baseline, and kept the existing covariates and endpoint definitions from the original SCORE model. In terms of statistical method, the model for older persons was fit using Cox proportional hazards regression, rather than a Weibull model. Five and ten year survival estimates, as well as beta coefficients, were provided in the write-up, so as to allow for calculation of an individual's risk. Two dimensional worksheets were also made available with a point scoring system. Internal validation of the model yielded an AUC of 0.74 overall and of 0.78 on women only. However, the model did not evaluate the use of blood pressure lowering medication, and did not account for deaths from other causes.

3.1.4 CORE Model

Due to the growing evidence that traditional CVD risk models extrapolated poorly to an elderly (over 70) population, Koller and colleagues [Koller et al., 2012] sought to create a model that could be used in an aging, possibly frailer population, that must also account for the increased risk of death from competing causes. Data were drawn from 2 large and similarly designed cohort studies on cardiovascular disease in elderly persons, the Cardiovascular Health Study (CHS) in the United States and Rotterdam Study (RS) in the Netherlands. The endpoint definition for this study was time to first CHD event, which included nonfatal MI and fatal CHD. Sex-specific models were created using age, systolic blood pressure (BP), use of blood pressure lowering medication, diabetes, total and HDL cholesterol, and smoking. Several new biomarkers and risk factors were assessed for incremental value in predictive accuracy, under a competing-risk model based on the methods of

Fine and Gray. Prognostic accuracy was assessed at 10 years with a modified Harrell's C-statistic [Wolbers et al., 2009, Wolbers et al., 2014], as well as a calibration plot of observed versus predicted risk.

The women's model was based on 5,878 individuals and the following is the linear predictor calculation used in predicting a woman's risk: linear predictor (LP) = $\sum_{k=1}^{p} \beta_k X_{ik}$

$$\begin{aligned} LP &= [0.463 * (\text{age} - 75)/10)] - 0.262 * [((\text{age} - 75)/10)^2] + [0.288 * \text{bp treatment}] \\ &\quad + [\text{bp treatment} * 0.08 * (\text{SBP} - 130)/10] + 0.330 * \text{diabetes} \\ &\quad + [\text{bp NO treatment} * 0.127 * (\text{SBP} - 130)/10] + 0.125 * \text{eversmoke} \\ &\quad + 0.041 * (\text{total chol} - 5) - 0.432 * (\text{HDL chol} - 1) \end{aligned}$$

In order to calculate a risk at a specific time t, one must have an estimate of the linear predictor, as well as the baseline cumulative subdistribution hazard, to use the following formula:

$$I_i(t|x_i) = 1 - \exp\Big(- \exp\Big(\sum_{k=1}^{p} \beta_k X_{ik} \Big) \cdot \int_0^t \bar{\lambda}_{1,0}(s) ds \Big)$$

The paper provides this value for several timepoints in the appendix. Then, for U.S. women, whose 10 year cumulative subdistribution hazard was 0.125, and the calculation for the 10 year predicted risk becomes:

$$1 - \exp[-0.125 * \exp(LP)]$$

Although this model demonstrated generalizability in aging U.S. and European populations, it did not demonstrate any statistically significant improvements over the Framingham model (only 0.02 to 0.03 changes in the C-statistic). This exemplifies the need for further work in cardiovascular risk prediction in the elderly.

3.1.5 Reynolds Risk Score

Despite its widespread adoption, it has been shown that up to 20% of female instances of heart disease occur in the absence of traditional Framingham risk factors [Khot et al., 2003]. In a 2007 article published in JAMA, Ridker, Cook, and colleagues examined a panel of new and traditional cardiovascular risk factors in a cohort of 24,558 healthy US female participants 45 years or older from the Womens Health Study (WHS) [Ridker et al., 2007, Cook et al., 2006]. Endpoints examined included incident myocardial infarction, stroke, coronary revascularization, or cardiovascular death, of which there were 766 such events over a median follow-up of 10.2 years.

Cox proportional hazards models were built on two-thirds of the data set ($n = 16,400$; 504 events) and evaluated on the remainder ($n = 8,158$; 262 events). Variables were selected via stepwise methods and with the use of regression trees, while final inclusion criteria was based on Bayes Information Criteria (BIC). Additionally, the ATP-III and Framingham models were both examined and re-calculated (original variables used, but new regression coefficients calculated), and the performance of these models was assessed in the current cohort, using both new and previously published coefficients.

The investigators presented two new models, and compared to the ATP-III and Framingham equations, the new models showed improvement in each of the summary statistics examined (Entropy, Yates Slope, Brier Score, and C-statistic), over both the originally published and re-calculated models. Reclassification was also examined compared to the ATP-III model. Additionally, while both of the novel models had very similar performance, the clinically simpler one was selected, and contained the following risk factors: age, systolic blood pressure, hemoglobin A1c (if diabetic), current smoking, total and HDL-C, high-sensitivity C-reactive protein (hsCRP), and parental history of myocardial infarction before age 60 years. This model was named the Reynolds Risk Score. Authors propose using the new model for women due to its large, prospectively collected sample, and inclusion of markers shown elsewhere to be more accurate predictors of risk. The model can be used to pre-

dict a woman's 10-year risk of disease and is simple and free to implement on the website `http://www.reynoldsriskscore.org`, and appears below:

$$Risk = 1 - 0.98634^{\exp[B - 22.325]}$$

where

$$\begin{aligned} B &= 0.0799 * \text{age} + 3.137 * \log(\text{SBP}) + 0.180 * \log(\text{hsCRP}) \\ &\quad +1.382 * \log(\text{total cholesterol}) 1.172 * \log(\text{HDL}) 0.438 * (\text{famhist}) \\ &\quad +0.134 * \text{hemoglobin A1c (if diabetic)} + 0.818 * (\text{smoker}) \end{aligned}$$

Limitations of the Reynolds Risk Score may include its generalizability, as it was developed on a predominantly white, middle-class population, and reliance on self-report for several key variables.

3.1.6 Summary of CVD Models

While there has been widespread public health interest in evaluating a person's risk of cardiovascular disease for at least the last fifty years, the current prediction algorithms are subject to several areas of limitations.

In terms of risk factors and data collection, many of the cohorts upon which models were derived contained only a single measurement for key risk factors such as blood pressure, cholesterol, or C-reactive protein, and would be more accurately represented by an average of two or more measurements, or collection of measurements over time. Secondly, other risk factors may be based on patient self-report, such as family history or smoking behaviors, which are subject to patient bias or human error.

Additional challenges include the generalizability of the models to minority racial and ethnic groups, as many were developed on a predominantly white, middle-class population. Prospectively collected data sets of both sufficient size and containing all key risk factors

and endpoint information are limited, and validation data sets may be underpowered. While the risk factors may be the same, risk functions derived from one population may not be applicable to a second population in terms of calculating predicted absolute risk. It has also been shown that several models may under-perform for observations in the highest and lowest percentiles of some of the risk factors.

A further limitation arises with trying to directly compare the performance across models due to the lack of uniformity of endpoint definition. The endpoint of the original Framingham models includes soft CVD endpoints, which are not often captured in other cohorts or clinical trials. A remedy to this would be to use a straightforward and uniformly defined endpoint such as cardiovascular mortality. Another inconsistency across studies is in the choice of statistical model used, and time frame for which predictions are valid.

Issues in model validation include inconsistent reporting and use of model validation metrics, ranging from C-statistics and indices, to AUCs, reclassification percentages, and several others. Additionally, some validation studies will use the exact coefficients from the original models, while others will use the same variables but recalculate, or "recalibrate" them to fit their data set. The latter should not be considered a true validation, but rather an extension of a previously published model to a new population.

Some concerns for a model's use in clinical practice should also be mentioned. Many of the models suggest using a cut-off of twenty percent or higher to indicate a person who is at "high-risk", however, as it has been shown that several of these models can give a variable range of results, perhaps a more rigorous cut-point analysis is warranted in order to correctly classify patients into high and low risk categories.

Lastly, the majority of these models have been developed on middle-aged, healthy cohorts, but their applicability in older, or frailer populations, with comorbid conditions, has not been routinely assessed. There have been few validation studies on older populations and while they were restricted to those who were considered healthy at baseline, the models were still found to have poor performance in those greater than 65 years of age.

Given that there is an increase in cancer survivorship, it may be useful to evaluate the current models in a cohort of survivors, as well as propose a new model for CVD risk, that may also include information on cancer disease and treatment characteristics that could alter a person's risk and improve prediction over the established risk factors alone. This would require the use of competing risk methodology, to account for deaths from non-CVD causes, which is not routinely practiced in the current literature on CVD risk modeling.

Figure 3.1: Summary of CVD Risk Models

Model	Cohort information	Endpoint(s), follow-up	Single model?	Statistical method	Covariates	Covariate inclusion criteria	Model Performance	Clinical use
Framingham 1991 (Anderson 1991)	5,573 oringinal and offspring Framingham cohorts Ages 30-74	Time to event of: MI, CHD, CHD death, stroke, CVD, CVD death 10 years	One model for each endpoint	Weibull Accelerated Failure time model	age, gender, SBP, DBP, total cholesterol, HDL cholesterol, smoking, diabetes, left ventricular hypertrophy	comparison of log likelihood between nested models	none described	Point scoring algorithm worksheet available
Framingham 2000 (D'Agostino 2000)	5,333 women 458 with prior event Ages 30-74	MI, angina, coronary insufficiency, CHD death Up to 4 years Subsequent event in 2 years	Sex specific models, Model for those with prior event	Weibull AFT Model	Age, menopause, log (tot/HDL cholesterol), log(SBP), on BP lowering meds, diabetes, smoker	Prior models, p<0.05	none	Point score chart included
Framingham 2001 (Wilson1998, D'Agostino 2001)	2,812 women Ages 30-74	Myocardial infarction or coronary death 5 or 10 years	Sex specific models	Age-adjusted Cox proportional hazards model and accelerated failure models (chose Cox)	Age, SBP, DBP, total cholesterol (or LDL-C), HDL cholesterol, smoking, diabetes Uses categories for cholesterol and BP	In previous FRS *also recommend recalculating coefficients for new cohorts	Validated in 4 more studies, C-statistics Hosmer-Lemeshow	Model based calculation or Point scoring system
Framingham 2008 (D'Agostino 2008)	4,522 women Ages 30-74	First of any event (MI, coronary death, coronary insufficiency, PAD, stroke, angina) 10 years	Sex specific models	Cox PH model	age, total cholesterol, HDL cholesterol, systolic blood pressure, antihypertensive medication, smoking, diabetes	In prior versions of FRS	C-statistics Hosmer-Lemeshow	Risk score sheets with points per risk factor provided, also calculation for "heart age"
SCORE (Conroy 2003)	12 European cohorts, 88,080 women) Analysis 45-64	Fatal cardiovascular events only 10 years	Separate models for high and low risk countries and genders	Weibull proportional hazards model, Age as time scale	Total cholesterol or total/HDL ratio, SBP, current smoker	No details provided	AUC, positive clinical diagnostic likelihood ratios	Easy to use worksheet provided for calculating 10 year risk
SCORE OP (Cooney 2015)	20,121 women Ages 65+	Fatal cardiovascular events only 10 years	Sex specific models	Cox proportional hazards model, age as a covariate	Age, Total cholesterol, HDL cholesterol, SBP, current smoker, diabetes	Based on prior models	AUC, Harrell's C, Hosmer Lemeshow goodness of fit	Easy to use worksheet provided for calculating 10 year risk
CORE (Koller2012)	Ages 65-80"+" 3,029 U.S. women	MI or fatal CHD	Sex and country specific models	Fine and Gray subdistribution hazards model	Age, Total cholesterol, HDL cholesterol, SBP, BP lowering medication, current smoker, diabetes	Based on traditional risk factors	Harrell's C statistic, calibration plots, net reclassification index	Sample calculation in Appendix. First to use competing risk method.
Reynolds Risk Score (Ridker 2007)	Women's Health Study 24,558 women 10 years	MI, Ischemic stroke, coronary revascularization procedures, death from cardiovascular	yes	Cox proportional hazards	age, SBP, hemoglobin A1c (if diabetic), current smoking, total and HDL-C, high-sensitivity C-reactive protein (hsCRP), parental history of MI< age 60 years	stepwise selection, regression trees, Bayes Information Criteria (BIC)	Entropy, Yates Slope, Brier Score, and C-statistic, reclassification	Website with calculator www.reynoldsriskscore.org

3.2 Models for Predicting Breast Cancer Recurrence or Mortality

An accurate estimate of disease prognosis is an essential piece in deciding which breast cancer treatment(s) an individual should undergo. In the recent era of multi-modal adjuvant therapies, a patient must weigh the risks of treatment with the benefit, usually given in terms of survival. Recently, several prognostic models have been created to aid clinicians and patients in making informed decisions regarding treatment options. While their generalizability has been variable, especially when transporting across countries and age groups, they are similar in the inclusion of pathological covariates, as well as their utility in research and patient decision making.

Below, some of the more commonly referenced models for use in breast cancer prognosis are summarized. These models were chosen due to their inclusion of traditional disease risk factors, that can be obtained from administrative and registry based data, and do not rely on genetic assay or microarray data. A more comprehensive overview of breast cancer recurrence and mortality models, including those that also examine tumor genetic information, can be found in Engelhardt et. al. [Engelhardt et al., 2014]

While these models have been validated and are widely used, some have been shown to have poorer performance in specific subgroups of women, such as those less than 35 or over 75 years old[Wishart et al., 2010, Olivotto IA, 2005].

3.2.1 Nottingham Prognostic Index

Using data from 500 consecutive patients treated with mastectomy for primary breast cancer in Nottingham City Hospital, a Cox model of overall survival was employed to determine which breast cancer factors could best determine prognosis. The model was developed

on 387 patients with complete data, and the following index was proposed:

$$I = (0.17 * \text{size}) + (0.76 * \text{nodes}) + (0.82 * \text{grade}),$$

where larger values of I indicate poorer prognosis [Haybittle et al., 1982]. The performance of the index was evaluated at 5 years post-diagnosis, for both recurrence and survival, and categories of risk were proposed based on the following index values: high risk (>4.4), medium risk (2.8 - 4.4), low risk(<2.8). This model assumes that adjuvant chemotherapy was not used.

However, this initial index was published in 1982 and based on a very small data set. Since then, survival for breast cancer has greatly improved, and an update to the NPI was published in 2007, based on cases diagnosed from 1990-1999 [Blamey et al., 2007a].

The updated NPI is calculated quite simply as:

$$\text{lymph node (LN) stage (1 to 3)} + \text{Grade (1 to 3)} + \text{maximum diameter (cm)} * 0.2$$

which yields a range of NPI from 2.08 (node negative, grade 1, 0.4 cm) to 6.8 (node stage 3, grade 3, size 4.9 cm). Three prognostic groups were still constructed: poor prognosis group (5.5 - 7.0), medium prognosis group (3.5 - 5.4), and excellent prognosis group (2.0 - 3.4). One criticism of the index was that it did not provide estimates for a survival probability or absolute risk [Lundin, 2007], rather just a general "prognosis". It is able to stratify patients into 3 groups with distinct differences in survival, but it may not be as useful for predicting an individual patient survival or absolute risk. However, a companion paper [Blamey et al., 2007b] was published that described how to calculate individual 10 year survival estimates, based on a curve fitting technique that resulted in the following formula:

$$3.0079 * \text{NPI}^2 + 12.295 * \text{NPI} + 83.84$$

This could be calculated for any 0.1 change in NPI, to provide better counseling at the individual level.

3.2.2 Kattan Nomogram

In 2004, Kattan and colleagues sought to create a continuous prognostic model (rather than a risk-group based model) that could better predict breast cancer mortality risk [Kattan et al., 2004]. Five-hundred and nineteen women who had been treated with mastectomy and axillary lymph node dissection, without chemotherapy, at Memorial Sloan-Kettering Cancer Center between 1976 and 1979 were included. The competing-risk method (Fine and Gray) was used to predict disease-specific survival, as a large number of the cohort had died from other causes. The accuracy of the new prognostic model was evaluated using the concordance index.

Age, multifocality, tumor size (cm), grade, lymph node involvement, and staining (IHC, H&E), were included as risk factors in the model. For ease of use, the model was presented in nomogram form, whereby points could be assigned to the value of each risk factor and the total score corresponded to a predicted 15-year disease specific survival. Estimates from the new model were compared to the NPI and found to be more accurate (higher C-index). Since the cohort on which the tool was developed did not receive any chemotherapy, the authors stated that it could provide a decision aid to women considering adjuvant chemotherapy, by providing an estimate of disease specific survival in the absence of treatment.

3.2.3 Adjuvant! Online

Another prognostic model/decision tool to aid women in the decision of whether to undergo adjuvant chemotherapy, Adjuvant! Online, was created in 2001 as an internet-based computer program providing 10-year prognosis predictions for early breast cancer patients [Ravdin et al., 2001], (http://www.adjuvantonline.com). Its primary goal was to give an estimate of the net benefit in survival of an individual patient's decision to undergo adjuvant

chemotherapy, based on their personal and disease related risk factors and the known efficacy of different adjuvant chemotherapy regimens. The tool was derived from women ages 35 to 59 who were diagnosed between 1988 and 1992.

Simply described, the risk calculation is done in two steps. The first is to estimate a patient's risk of failure (defined as either death or relapse) and the second is to multiply that risk by the proportion of failures that a given adjuvant therapy is known to prevent, while accounting for competing risk of death from non-breast cancer related causes. Actuarial methods were used iteratively, at yearly increments post-diagnosis, to estimate patient overall survival (based on patient age from U.S. mortality statistics) and breast cancer survival (based on SEER registry data). The estimates of breast cancer survival are based on an individual's tumor size, the number of involved lymph nodes, and estrogen receptor (ER) status. A limitation of the SEER data is that it does not reliably capture relapse or cause of death information, therefore, estimates of breast cancer related mortality were derived indirectly from total age adjusted survival using SEER Stat2 software, while estimates of recurrence were assumed to be slightly higher than breast cancer mortality at any given time since it was assumed that there would always be a proportion who had relapsed but not yet died. Estimates can also be adjusted for the presence of comorbid conditions. Information on additional prognostic factors that arise in the literature but were not included in the development of the initial tool, can be used to refine the 10 year estimates under the section labeled "Prognostic Factor Impact Calculator", which uses a Bayesian method to make adjustments based on relative risks and prevalence of positive test results.

Once the individual survival calculations are made, the program will also produce a series of predictions and graphs based on the use of endocrine therapy (5 years of adjuvant tamoxifen) with or without chemotherapy (either CMF-like, anthracycline-based, or anthracycline- and taxane-based). Estimates of improvement in survival attributed to the various therapies are based on randomized controlled clinical trials from the Early Breast Cancer Trialists Collaborative Group.

Individual risk calculations must be done using the online tool, by a trained professional, making it a bit of a "black box" for users, unlike some of the simpler tools such as the NPI, or paper based worksheets such as the Kattan nomogram. While its primary purpose was to aid in patient treatment decision making, it has also been used in patient selection for randomized controlled trials, and has begun to be validated in breast cancer populations in other countries [Olivotto IA, 2005, Campbell et al., 2009, Mook et al., 2009], though prognosis may be overestimated in certain high risk groups (young age, high-grade and HER2-positive patients) [Hajage et al., 2011].

3.2.4 CancerMath

Michaelson and Chen [Michaelson et al., 2011] developed the SNAP (size, nodes, and prognostic factors) method from 1,352 breast cancer patients treated at the University of Southern California Van Nuys Breast Center between 1966 and 2006. It uses the "binary biological model of cancer metastasis", that is a complex set of equations, combined with national mortality data, to predict the risk of death for the first 15 years after diagnosis. It incorporates data from the breast cancer trials to assess the survival impact of various adjuvant therapies. This model includes the risk factors from previous prognostic models (age, grade, tumor size, nodal involvement, ER, PR, HER2), and also includes histologic subtype, as the authors postulate that the type of cell has a distinct biological pathway that plays a role.

Though the web based calculator is extremely user friendly, and the output easy to interpret, the mathematics underlying the screens are complex and likely not reproducible by a third party (http://www.lifemath.net/cancer/index.html). To date there has been one published validation of SNAP [Michaelson et al., 2011], found in the paper detailing its development. They found that SNAP provided accurate estimates of risk of death when examined in large patient data sets.

3.2.5 Oxford or Options Model

In 2010, after discovering that Adjuvant! Online was not directly transferable to the UK population, Campbell and colleagues [Campbell et al., 2010] sought to estimate and externally validate a new UK-specific prognostic model for predicting recurrence free survival in women with early stage breast cancer. Data consisted of 1,844 consecutive women diagnosed at a single hospital between January 1986 and January 2001. A parametric, accelerated failure time model was used to model the time until the first recurrent event with the following risk factors: number of positive axillary lymph nodes, tumor grade (1, 2, or 3), tumour size (cm), ER status (positive or negative), and patient age (years). In addition, indicator variables for use of adjuvant radiation, chemotherapy, and hormonal therapy were included to control for the effects of the different treatment options patients could have received, as well as be able to estimate the survival times when used as a prognostic model, in the absence of such treatments. The rationale for the use of the AFT model over the more commonly used Cox proportional hazards model, was that it is more straightforward to calculate a predicted time until event on an external observation when a distributional assumption is imposed, as well as make predictions beyond the range of the observed data, rather than the use of a "baseline hazard" as is done in the Cox model. In addition, 573 women (31% of the sample) had experienced a recurrent event in the follow-up time period.

Assuming a gamma distribution for survival times, with ancillary parameters 1.698 for shape and 0.567 for scale, the fitted AFT model was as follows:

$$\begin{aligned}\log T_i = {} & 0.402 * \ln(\text{positive nodes}) + 0.898 * \text{tumor size}^2 + 1.045 * \text{tm size}^2 * \ln(\text{tm size}) \\ & + 0.647 * \text{grade} + 1.015 * \text{age} + 1.209 * \text{ER} + 1.543 * \text{radiation} + 1.23 * \text{hormonal therapy} \\ & + 0.357 * \text{chemotherapy} + 1.226 * \text{ER} * \text{hormonal therapy} + 1.023 * \text{chemotherapy} * \text{age} \\ & + 1.418 * \ln(\text{positive nodes}) * \text{chemotherapy}\end{aligned}$$

Table 3.3: Hazard Ratios from the PREDICT Model for Breast Cancer Survival

Prognostic variable (category)	HR (ER+)	HR (ER-)
Nodes positive (0, 1, 2 to 4, 5 to 9,10+)	1.75	1.55
Tumor size mm (<10, 10 to 19, 20 to 29, 30 to 49, 50+)	1.43	1.44
Grade (Low, intermediate, high)	2.33	1.50
Screen detected	0.70	0.86
Chemotherapy	0.73	0.82
Hormonal therapy	0.95	1.43

When evaluated on an external data set from the same country of origin, the authors found that the model performed well, having an overall C-index of 0.76.

3.2.6 PREDICT Model

Another web-based prediction tool developed in the UK, PREDICT, was launched in 2010 (www.predict.nhs.uk) and can estimate 5 and 10 year breast cancer specific survival following primary surgery [Wishart et al., 2010]. The model was derived from the cancer registry information on 5,694 breast cancer patients treated in East Anglia from 1999 to 2003 and used the following factors: age at diagnosis, tumour size (categorical), tumor grade (Low, Intermediate, High), number of positive nodes (categorical), ER status, mode of detection (screening versus not), and use of adjuvant chemotherapy or hormonal therapy. A Cox proportional hazards model was used to model the time until breast cancer death, and competing causes of death were modeled separately, adjusted for age at diagnosis. Separate models were created for ER positive and negative subgroups, due to the differing effectiveness of treatments.

An external data set of 5,468 patients was used for validation, and the model was well calibrated and showed good discrimination ($AUC = 0.79$, analogous to C-index). It has since been further externally validated and updated to include information on HER2 [Wishart et al., 2012] and Ki67 [Wishart et al., 2014].

3.2.7 Summary of BC Models

There has been increasing development and validation of breast cancer prognostic models in the last decade, with the main goal of helping to aid a patient's decisions on whether or not to undergo adjuvant chemotherapy, in early stage disease. The models have been developed using a variety of statistical techniques, some quite complex, while still presenting results in a manner that is understandable to a clinician, and possibly, to a patient. Model validation continues to be assessed, especially in populations outside of where the models were developed, with mixed results. In the era of microarrays, genomic information is also being used to predict survival, and combining genetic with clinical factors should be the next steps in prognostic model building. While newer biomarkers may show only small incremental improvements in prediction, the risk factors of grade, tumor size, and nodal involvement remain the strongest clinical risk factors. Age at diagnosis also plays an important role, especially through its importance in predicting competing risk of death from other causes.

Figure 3.2: Summary of Breast Cancer Prognosis and Decision Making Models

Model, year published	**Cohort information**	**Endpoint(s)**	**Statistical method**	**Covariates**	**Covariate inclusion criteria**	**Model Performance (in original paper)**	**Clinical use**
Nottingham Prognostic Index (NPI), 1982	UK N=387 (1973-1979)	Recurrence Free Survival	Cox PH	Nodes, grade, tumor size	P<0.05	Classification table, Stratified survival curves	Simple calculation, Cut-points for risk groups provided
Kattan Nomogram 2004	US N=519 (1976-1979)	Disease Specific Survival	Fine and Gray model	Age, focality, nodes, grade, tumor size, staining (H&E, IHC)	P<0.05	C-index at 15 years	Nomogram worksheet provided
Adjuvant! Online 2001	US (N=37,968) 1988-1995 +US mortality statistics	Recurrence Free Survival, Overall survival	"Actuarial methods" combining SEER data with national mortality data	Age, comorbidities, grade, tumor size, nodes, ER, chemotherapy, hormonal therapy	SEER collected variables	none	Online calculator https://www.adjuvantonline.com/
Oxford 2010	UK (N=1,844) 1986-2001	Recurrence Free Survival	Accelerated Failure Time Model (gamma distn)	Nodes, grade, tumor size, age, ER, radiation, chemotherapy, hormonal therapy	Used in prior models	Brier score, overall C	Score sheet in supplemental materials
PREDICT (2010) **PREDICT+** (2012)	UK (N=5,694) 1999-2003	Overall Survival	Cox cause specific hazards models, stratified by ER	Age, nodes, tumor size, grade, ER, chemotherapy, hormonal therapy, screen detected	Used in prior models	AUC	Online tool http://www.predict.nhs.uk/predict.html
CancerMath (2011)	US (N=1,352) 1966-2006 + US mortality statistics	Overall survival, Disease specific survival	"Binary biological model of cancer metastases"	Tumor size, nodes, age, grade, ER, PR, HER2, histologic type	Prior risk factors plus cellular level data	C-index at 15 years	Online tool http://www.lifemath.net/cancer/

Chapter 4

Disease Risk Model Evaluation

Before a model can be adopted into clinical practice, it is important to evaluate its utility on independent samples of patients, called validation. Validation can only be performed when the true outcome is already known. There are numerous methods and performance metrics available for model evaluation, and it is important that one first assess the goals of the model in order to choose the best evaluation procedure. Two important concepts in model evaluation are calibration and discrimination. Calibration refers to how closely the predicted outcomes are to the observed outcomes. Discrimination refers to the ability of a model (or medical test) to differentiate between individuals that experience and do not experience the outcome of interest, and has many summary measures to choose from.

Generally, the predicted output from a disease risk model will be on a continuous scale, either predicting a person's remaining lifetime risk, or risk in a specified time period (e.g. 5-year risk, 10-year risk, etc.). Very often, clinicians and investigators have dichotomized or established cut-points that take the continuous outcome measure and classify patients into risk groups, such as high, medium, and low-risk, or high versus all else. Risk categories are easier to interpret and convey meaningful results to patients, despite the potential loss of

information. Additionally, many of the established performance evaluation metrics are for binary or categorical outcomes, rather than a continuous scale.

In the following sections, common methods for quantifying accuracy of a diagnostic test, predictive model, or biomarker are summarized, and these terms are used interchangeably, though the main application will be for quantifying output from a predictive model, rather than a biomarker or medical test.

4.1 Traditional Measures of Overall Performance

Traditional measures of overall model performance quantify how close the predictions are to the actual observed outcomes and have included measures explained variation (R^2) and the Brier score [Steyerberg et al., 2010].

4.1.1 Explained Variation or R^2

Variations of R^2 have been proposed for different model types, but generally take the form of $(Y - \hat{Y})^2$, or the squared distance between observed and predicted, where lower values indicate a better fitting model. It is well established for linear regression, and has been adapted by Nagelkerke [Nagelkerke, 1991] to more general models that are fit using maximum likelihood methods. It is bounded by 1 and is interpreted as the "power of explanation" of the model.

$$R^2 = 1 - \exp\Big[-\frac{2}{n}\{l(\hat{\beta}) - l(0)\}\Big]$$

where $l(\hat{\beta}) = \log L(\hat{\beta})$ and $l(0) = \log L(0)$ denote the log likelihoods of the fitted and the null model, respectively. For certain models, such as logistic regression and the Cox model, where the maximum R^2 cannot achieve a value of 1 using the above definition, it is redefined to:

$$\bar{R}^2 = R^2 / \max(R^2)$$

where $\max(R^2) = 1 - \exp\{2n^{-1}l(0)\} = 1 - L(0)^{2/n}$.

4.1.2 Brier Score

The Brier score [Brier, 1950], or mean probability score, is the average difference between the observed and predicted (or expected) events. It was originally proposed for weather forecasting to include multiple categories

$$\frac{1}{n}\sum_{j=1}^{r}\sum_{i=1}^{n}(f_{ij} - E_{ij})^2,$$

where the f_{ij} are the forecasted probabilities, E_{ij} is an indicator for whether the event occurred in class j or not, and the r classes are chosen to be mutually exclusive such that

$$\sum_{j=1}^{n} f_{ij} = 1, i = 1, 2, 3, \ldots n$$

In disease prediction modeling, it has been adapted to measure the average discrepancies between true disease status and estimated predictive values, providing an overall "goodness of fit" for the model.

$$B = \sum_{i=1}^{n}\frac{(f_i - O_i)^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(f_i - O_i)^2$$

where f_i and O_i are the forecasted and observed values, respectively.

The Brier score ranges from 0 to 0.25 when there is 50% disease incidence, where 0 indicates a perfect prediction and 0.25 indicates useless. It can also be scaled to range between 0 and 1, as Nagelkerke's approach to R^2, and this scaled measure becomes is very similar to Pearson's correlation coefficient[Steyerberg et al., 2010]. It has also been adopted for survival outcomes [Gerds and Schumacher, 2006].

4.2 Calibration

As stated previously, a well-calibrated model is one in which the predicted values do not systematically differ from the observed values. Assessing calibration can be done visually with the use of a calibration plot, where the observed values are plotted against the predicted values. While somewhat arbitrary, this is usually done by grouping observations by decile of risk [Steyerberg et al., 2010]. A perfectly calibrated model would have values that all fall on the 45-degree line.

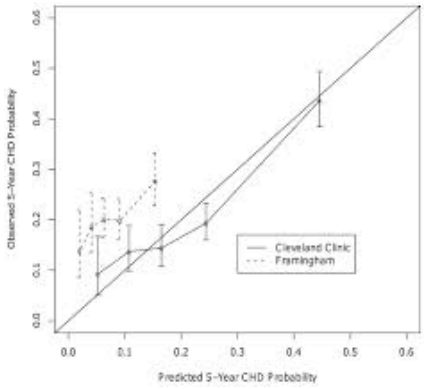


Figure 4.1: Sample calibration plot for cardiovascular risk models

4.2.1 Hosmer-Lemeshow Statistic

Originally developed for logistic regression, the Hosmer-Lemeshow statistic measures goodness of fit by grouping the data into a specified number of equally sized subgroups (usually deciles) and compares the observed and predicted outcomes across the groups [Agresti, 2002]. Well-calibrated models should have similar observed and expected event rates in each subgroup. The statistic follows a Chi-squared distribution with $(G-2)$ degrees of freedom, where G denotes the number of subgroups (usually deciles) created for model assessment:

$$H = \sum_{g=1}^{G} \frac{O_g - E_g}{N_g \pi_g (1 - \pi_g)}$$

and O_g and E_g denote the observed and expected event rates, respectively, in each subgroup g. In a formal test, the null hypothesis is no difference in observed and predicted outcomes, and large values of the test statistic indicate a poor fitting model. However, with large samples, it is common to reject the null hypothesis, even though the model may appear to fit the data well.

4.3 Discrimination

4.3.1 Binary Data

In the simplest case of a binary test and outcome, where patients are either diseased ($D = 1$) or healthy ($D = 0$), and censoring and time-dependence are not issues, there are several standard metrics for quantifying predictive ability.

Let the binary variable, D, denote a person's (known) disease status:

$$D = \begin{cases} 0 & \text{for healthy} \\ 1 & \text{for diseased} \end{cases}$$

Let Y denote the result of the test, and assume that higher values of the test are indicative of disease:

$$Y = \begin{cases} 0 & \text{negative for disease} \\ 1 & \text{positive for disease} \end{cases}$$

For a binary test, there are four possible outcomes, as detailed in the table below:

	$D = 0$	$D = 1$
$\mathbf{Y = 0}$	True negative	False negative
$\mathbf{Y = 1}$	False positive	True positive

In two of these scenarios, the test performs correctly; the test can be positive for a patient with disease (true positive), and can be negative for a patient without disease (true negative).

However, the test can have two types of errors: calling a patient positive in the absence of disease (false positive), and calling a diseased patient negative (false negative).

Sensitivity, Specificity, PPV, and NPV

Sensitivity and specificity are conditional probabilities that quantify how well a test can differentiate between diseased and non-diseased individuals. If a test (or predictive model) has a dichotomous output, or an established cutpoint that dichotomizes patients into "diseased" or "healthy", sensitivity refers to the probability that a test is positive, given that a patient has the disease. It is also called the true positive fraction, or TPF.

$$\text{Sensitivity} = Pr(\text{Test} = 1|\text{D} = 1)$$

Specificity is defined as the probability that the test is negative, given that the patient does not have disease.

$$\text{Specificity} = Pr(\text{Test} = 0|\text{D} = 0)$$

Another commonly used measure is the false positive rate, or FPR, which is 1-Specificity. This refers to the probability that the patient did not have disease in the presence of a positive test result $Pr(D = 0|Test = 1)$. Generally, the goal of any test is to maximize the sensitivity, or ability to detect disease in diseased individuals, while minimizing the FPR.

While the sensitivity and specificity describe how well the *test* performs, in the real-world setting, for counseling patients on their test results when disease status is not yet known, it may be more helpful to understand the positive and negative predictive values (PPV and NPV) of a test. PPV refers to the probability that a patient has disease, given that the test was positive:

$$\text{PPV} = Pr(\text{D=1}|\text{Test=1})$$

NPV refers to the probability that a patients does not have disease, given that they have a

negative test result:

$$\text{NPV} = Pr(\text{D=0}|\text{Test=0})$$

These also depend on the population prevalence of the disease, and cannot be calculated in case-control studies.

Diagnostic Likelihood Ratios

Diagnostic likelihood ratios, or DLRs, are an alternative to PPV and NPV that summarize predicting disease status from a test result, without depending on disease prevalence. This is extremely useful for screening, when the true disease status is not yet known. They are defined as:

$$\text{positive DLR} = \text{DLR}^{+} = \frac{Pr\left[Y = 1|D = 1\right]}{Pr\left[Y = 1|D = 0\right]}$$

$$\text{negative DLR} = \text{DLR}^{-} = \frac{Pr\left[Y = 0|D = 1\right]}{Pr\left[Y = 0|D = 0\right]}$$

They give the ratio of the likelihood observed test result in diseased versus healthy patients. DLRs can take any values between zero and infinity. A perfect test would have a DLR^{+} value of infinity and a DLR^{-} value of zero.

Discrimination or Yates slope

Discrimination slope in the binary context is defined as difference of mean predicted probabilities of events and non-events. The mean probability score is written:

$$\bar{PS}(f, d) = (\frac{1}{N}) \sum_{i=1}^{N} (f_i - d_i)^2$$

where f_i is the forecast, or predicted probability of an event, and d_i denotes whether it is an event (1) or nonevent (0). This is the same as the Brier score, but for the specific case of binary data.

4.3.2 Continuous Data

The output from regression models is very often a continuous value. The same is true for many medical tests, including markers measured from human serum, blood pressure, temperature, and others. Some tests or evaluations may also be ordinal in nature and take on categories such as "severe", "moderate", or "mild". If we again assume that larger values are more indicative of disease, it is possible to create a binary test result from a continuous or ordinal test using a cutpoint of c as follows:

$$Y = \begin{cases} \text{positive} & \text{if } Y \geq c \\ \text{negative} & \text{if } Y < c \end{cases}$$

and then use any of the measures described previously for binary data.

Receiver Operating Characteristic (ROC) Curve

The ROC curve is the set of all TPFs and FPFs calculated at every possible value of Y. It is a monotone, increasing function in the positive quadrant. A useless test would be one for which $TPF(c) = FPF(c)$ for all values of c, and would be the 45-degree line, while a perfect test would have some value of c such that $TPF(c) = 1$ and $FPF(c) = 0$ and would lie along the upper and left borders of the quadrant. Better tests lie are ones that lie closer to the upper left corner.

The ROC curve can be summarized using numerical indices. The most common is called the area under the curve, or AUC, and is defined as:

$$AUC = \int_0^1 ROC(t)dt$$

The AUC can take values between 0.5 (useless test) and 1.0 (perfect test). The value itself, can be interpreted as the probability that in two randomly chosen individuals (one diseased, one non-diseased), the value for the diseased individual will be higher than the

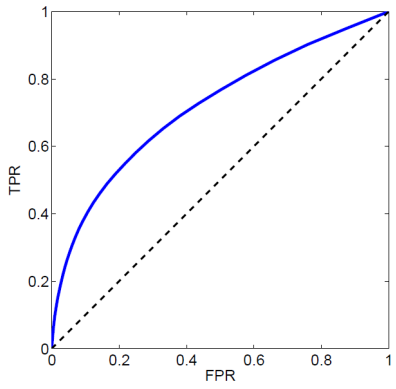


Figure 4.2: Sample ROC curve

non-diseased individual.

Another summary index for the ROC curve is the partial AUC, or $pAUC$. If it is only desirable to compare tests within a specific acceptable false positive rate, say t_0, then the AUC calculation can be restricted to:

$$pAUC(t_0) = \int_0^{t_0} ROC(t)dt$$

A third commonly used way of comparing ROC curves is to choose a specific false positive fraction and look at the corresponding true positive fraction value. However, this ignores a large amount of information from the curves.

Additionally, when comparing two models on the same set of patients, it is possible to compare the 2 AUCs [DeLong et al., 1988], or the change in AUC, ΔAUC.

Some advantages of ROC curves are that one does not need to define cutoffs for the continuous value, but rather can examine accuracy across the entire range of values. It is also possible to compare results across different markers or predictive models on the same cohort, does not depend on disease prevalence, and can be estimated for case-control and cohort studies.

Some disadvantages are that the AUC may be an oversimplification of a test's performance and may not be relevant for clinical practice. Additionally, it has been shown

[Vickers et al., 2011] that a statistically significant improvement in predictive accuracy will not always yield a statistically different AUC value when comparing two models.

C-statistic, Probability, or Index

Equivalent to the Area Under the ROC Curve, the C-statistic can be interpreted as the probability that a subject from the event group has a higher predicted probability of having an event than a subject from the non-event group [Pencina and DAgostino Sr, 2004]. Let Y be a random variable describing the predicted probabilities of having an event for subjects who had events, and V be another random variable describing the predicted probabilities for those who did not have an event. Then $C = P(Y \geq V)$ for continuous V and Y, and $C = P(Y > V) + 0.5 * P(Y = V)$ for discrete V and Y.

Hanley and McNeil [Hanley and McNeil, 1982] first noticed the relationship between this measure and the Mann Whitney statistic, which can be re-defined as follows: let $Y_1, Y_2, \ldots, Y_k$ be the predicted probabilities of having an event in the event group, and $V_1, V_2, \ldots, V_n$ be the predicted probabilities in the non-event group, such that for each pair of subjects (i, j), where the first one comes from the event-free group and the second one comes from the event group, assign a 1 if $Y_j > V_i$, a 0.5 if $Y_j = V_i$ and 0 otherwise. Then, summing over all possible pairs of subjects will yield the Mann-Whitney statistic, denoted W_{VY}. The area under the curve is equal to $\frac{1}{kn} W_{VY}$.

4.4 Reclassification Methods

Another important area of research is to determine whether existing risk prediction models can be improved with the addition of new biomarkers or other risk factors. The improvement in risk prediction is often referred to as the incremental value or prediction increment, and recently has been evaluated by use of the net reclassification index [Pencina et al., 2008].

Risk Reclassification

Consider an established model, such as the Framingham risk score, or Gail model (denoted *Model 1*, and then consider a scenario in which a single, new predictor is added to the established model (denoted *Model 2*). If cutpoints have been established that classify patients into high risk, intermediate risk, low risk, etc., a two-way table can be constructed that shows the distribution of patients by each model. Furthermore, the cells that are off-diagonal are the ones in which reclassification has occurred.

	Risk Model 2		
Risk Model 1	Low	Intermediate	High
Low	n_1	n_2	n_3
Intermediate	n_4	n_5	n_6
High	n_7	n_8	n_9

This makes it possible to calculate the percent reclassified, as the sum of all the off-diagonal cells, divided by the total number of subjects. This summary measure may be useful in that it gives some indication of how many persons would change risk categories, and possibly treatment decisions, under a new model.

Cook and Ridker [Cook and Ridker, 2009], proposed calculating a reclassification calibration statistic, analogous to the Hosmer-Lemeshow goodness-of-fit statistic but using reclassified categories instead of deciles. It is calculated as follows:

$$X^2_{RC} = \sum_{k=1}^{K} \frac{(O_k - n_k\bar{p_k})^2}{n_k\bar{p_k}(1-\bar{p_k})},$$

where n_k is the number of observations in cell k, O_k is the observed number of events in cell k, and $\bar{p_k}$ is the averaged predicted risk in cell k for the model. The statistic follows an approximate chi-squared distribution with $K-2$ degrees of freedom.

However, it has been shown [Pepe, 2011] that in most instances, as long as the original model is an adequate fit of the data, the enhanced model will accept the null hypothesis of a good fit, and this test is useless. Moreover, the reclassification calibration test of the

baseline model should reject the null hypothesis, as it has already been established that the new biomarker or predictor, has an association with the outcome. In fact, the reclassification calibration statistic for the original model is the same as the Pearson Chi-squared statistic for association between the new risk factor, controlling for the risk factors in the original model.

In summary, if the new biomarker has been shown in preliminary analyses to be a risk factor, the reclassification calibration test of the baseline model will be rejected when there is a sufficiently large sample size, and therefore it has been argued that there is "no point in performing this test" [Pepe, 2011].

Net Reclassification Index

Another way of comparing two models, on the basis of reclassification, is to compute the Net Reclassification Index or Improvement, or NRI.

$$\begin{aligned} NRI = \quad & P(\text{up}|\text{disease}) - P(\text{down}|\text{disease}) \\ & + P(\text{down}|\text{nondiseased}) - P(\text{up}|\text{nondiseased}), \end{aligned}$$

where "up" refers to a person being moved into a higher risk category under the new model, and "down" refers to a person being placed in a lower risk category under the new model. This general formula can be used for any number of categories. It is similar to the percent reclassified, but it distinguishes between movement in the correct and incorrect directions. It can also be rewritten into the sum of improvements for diseased (moving into higher categories) and nondiseased (moving into lower categories):

$$\begin{aligned} NRI &= [P(\text{up}|\text{disease}) - P(\text{down}|\text{disease})] \\ &\quad + [P(\text{down}|\text{nondiseased}) - P(\text{up}|\text{nondiseased})] \\ &\quad = \text{relative improvement(diseased)} + \text{relative improvement(nondiseased)} \end{aligned}$$

Interpretation of the NRI is difficult, however, because by summing the 2 components, it is not possible to see the relative contributions of each, and it has been argued that it may be better to report the two components separately, or the changes in proportions of subjects in each of the risk categories, with "+" or "-" to denote movement into a higher or lower category [Pepe, 2011, Kerr et al., 2014].

Some limitations of the use of risk categories in general are that it is assumed that the categories have been defined based on rigorous and uniformly agreed upon criteria, though this is often not the case. Additionally, it has been shown that substantial reclassification can occur even without substantial improvement in model performance.

Category-Free Net Reclassification Index

A category-free net reclassification index was proposed [Pencina et al., 2008] to deal with the criticism of arbitrarily defined categories. It is calculated as the net proportion of diseased patients for whom the risk under the new model is higher than the risk under the baseline model plus the net proportion of nondiseased individuals for whom the risk under the new model is lower than the risk from the baseline model:

$$\begin{aligned} NRI \quad &= [\text{Proportion}(\text{risk}^{new} > \text{risk}^{base} | \text{disease}) \\ &\quad -\text{Proportion}(\text{risk}^{new} < \text{risk}^{base} | \text{disease})] \\ &+[\text{Proportion}(\text{risk}^{new} < \text{risk}^{base} | \text{non-diseased}) \\ &-\text{Proportion}(\text{risk}^{new} > \text{risk}^{base} | \text{non-diseased})] \end{aligned}$$

While use of the NRI gained momentum in the last few years, a recent paper by Pepe et al. [Pepe et al., 2014] showed that it yielded false positive conclusions at an alarmingly high rate, compared to the change in AUC and a likelihood ratio statistic, which showed a positive effect in 9.8% and 5% of simulations, respectively. It warned against its use to draw conclusions about predictive capability of new biomarkers and instead recommends the use

of a standard test, such as a likelihood ratio test, to test the statistical significance of the new risk factor in the predictive model. Then they recommend using more clinically meaningful ways of describing the improvement in prediction such as net benefit and relative utility.

Integrated Discrimination Index

The integrated discrimination improvement (IDI) integrates the NRI over all possible cut-offs [Pencina et al., 2008]. It is the difference in Yates, or discrimination slopes between 2 models, which is the mean difference in predicted probabilities between events and nonevents.

$$\begin{aligned} \text{IDI} &= (\bar{\hat{p}}_{\text{events}} - \bar{\hat{p}}_{\text{nonevents}})_{\text{new model}} \\ &\quad -(\bar{\hat{p}}_{\text{events}} - \bar{\hat{p}}_{\text{nonevents}})_{\text{old model}}, \end{aligned}$$

where $\bar{\hat{p}}_{\text{events}}$ is the mean of the model-based predicted probabilities of an event for those who develop events, and $\bar{\hat{p}}_{\text{nonevents}}$ is the mean of the model-based predicted probabilities of an event for those who do not develop an event. An asymptotic test for the null hypothesis of $IDI = 0$ can be constructed as follows, since the actual events do not depend on the model:

$$z = \frac{\hat{\text{IDI}}}{\sqrt{(\hat{se}_{\text{events}})^2 + (\hat{se}_{\text{nonevents}})^2}}$$

Values of the IDI can be interpreted as the improvement in discrimination of the new model, and be formally tested for the null hypothesis of no improvement. However, because these measures all condition on event status, they cannot be used for time to event data.

4.5 Methods for Censored and Survival Data

In survival analysis, discrimination is defined as the model's ability to separate those with longer event-free survival from those with shorter event-free survival within some time frame of interest [Pencina et al., 2012a]. Discrimination of statistical models with dichotomous

outcomes is most commonly quantified by the AUC, sometimes called the C or concordance index, but this does not take into account any time component or the fact that observations may be censored. Several different versions of the C-index for survival data have been proposed.

In general, for any pair of bivariate observations (X_1, Y_1) and (X_2, Y_2), the concordance probability is defined as the probability that $(Y_2 > Y_1)$ given that $(X_2 > X_1)$. This idea can be extended to survival analysis, where T is the length of time until the event of interest and X is a predictor variable, such that the observed data are (Y, δ, X), where $Y = min(T, Q)$, Q is the censoring time and $\delta = I\{T \leq Q\}$. Then the concordance probability is defined as:

$$C_{T,X} = P(X_2 > X_1 | T_2 > T_1)$$

Several variations of this measure have been proposed.

Harrell's C-index

Harrell's C-index was first proposed in 1996 [Harrell Jr et al., 1996] and extends the idea of the AUC to the case of right-censored survival time outcomes and assesses the amount of agreement or concordance between predictions and outcomes comparing not only events and nonevents but also two events that happened at different points in time. Briefly, 2 subjects are described as comparable if we can determine which one survived longer (i.e. they have not both been censored, or if one has been censored, it takes place after the other has experienced the event). The pair is considered concordant if their predicted probabilities of survival and observed survival times go in the same direction, meaning, that for every comparable pair, the person with the longer survival time, also has the higher predicted probability of survival. Then, the overall C-index is defined as the probability of concordance given comparability:

$$C_{Y,X} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} [\delta_i I\{Y_i < Y_j\} I\{X_i < X_j\} + \delta_j I\{Y_j < Y_i\} I\{X_j < X_i\}]}{\sum_{i=1}^{n} \sum_{j=1}^{n} [\delta_i I\{Y_i < Y_j\} + \delta_j I\{Y_j < Y_i\}]}$$

In order for two subjects to be considered comparable, the subject with the shorter observed time must be the one who experienced an event, while the subject with longer observed time could either have experienced the event at a later time or have been censored. However, the exclusion of pairs where the person with shorter follow-up time was censored, means that this version of the C-Index may depend on censoring and caution should be used when there is a fair amount of censoring present in the data set [Liu and Jin, 2009].

Chambless and Diao's C statistic

Chambless and Diao [Chambless and Diao, 2006] defined a time-dependent AUC as:

$$AUC(t) = P(Z_i > Z_j | D_i(t) = 1, D_j(t) = 0)$$

Where Z_i and Z_j are two predicted probabilities for an event and non-event. It is the probability that in two randomly chosen subjects, one who experienced the event and one who did not, the model predicted risk will be higher for the person who experienced the event.

They derive the sample estimator using Bayes' formula to obtain:

$$C_{CD} = \frac{E((1 - S(_t|P_j))S(_t|P_i)I(P_i < P_J))}{E((1 - S(_t|P_j))E(S(_t|P_i)))}$$

Where S and I are the survival and indicator functions and E denotes expectation.

It can be thought of as a standardized average of event probabilities multiplied by survival probabilities, from all (discordant) pairs where$Z_i > Z_j$. Persons that are censored before time t do not contribute to the calculation and it focuses only on event versus nonevent comparisons. Survival times enter the definition only implicitly, by calculation of the index only at a specified time, t.

Gönen and Heller's Concordance Probability

Gonen and Heller proposed a method to estimate a variant of the c-statistic that is not sensitive to the degree of censoring, is a simple function of the Cox model, and does not require imputation of survival times [Gönen and Heller, 2005]. Unlike the previously mentioned c-indices, this one calculates the probability that of any two subjects, the one with the worse model-based risk profile will have the shorter survival time. It is the probability of time, based on risk profile, which is the reversal of the other C-indices:

$$K = P(T_j > T_i | P_i \geq P_j)$$

This definition is applied to Cox regression models and uses the proportional hazards notion to derive an estimator for K. Assuming that subjects are ordered according to increasing linear predictors $\beta^T X$, (or equivalently, decreasing predicted probabilities of survival) it can be formulated as

$$K = \frac{2}{N(N-1)} \sum_{i<j} \left\{ \frac{I(\beta^T X_i > \beta^T X_j)}{1 + \exp(\beta^T X_j - \beta^T X_i)} + \frac{I(\beta^T X_i < \beta^T X_j)}{1 + \exp(\beta^T X_i - \beta^T X_j)} \right\}$$

Features of this estimator are that it depends only on the linear predictors values, so it can be calculated from only the Cox regression coefficients and risk factor levels, and that is handles censoring automatically through the Cox regression coefficients, estimated using the full survival data. For any two subjects, the expression under summation takes values that range from 0.5 to 1.0. When linear predictors are very close to each other it approaches 0.5 and when they are further apart it approaches 1. This gives this index an interpretation of a measure that assigns a distance to each pair of subjects and then averages these distances over all pairs. Higher values indicate better discrimination.

Discussion

In survival data, the C-index summarizes the models ability to discriminate those with longer event-free survival from those with shorter event-free survival in a given time, T. Chambless and Diao's definition is the closest to the definition of the AUC in the binary case, but does not include censored observations in its calculation. Gönen and Hellers K, deviates from the classical definition in that it focuses on the distance between the model based predictions rather than true outcomes, and also must fulfill the proportional hazards assumption. Of all the C-indices proposed, only Harrell's is invariant to transformations and is fully nonparametric, based only on ranks.

In a 2012 paper by Pencina et al. [Pencina et al., 2012b], the various C-indices were calculated for the Framingham risk score on both real and simulated data, and were found to have very different results. Thus, the authors concluded that different versions of the C-index should not be compared across studies and that care should be taken in interpreting results. Similarly, the choice of index should be made *a priori*, and the authors recommend that in situations where prediction is long-term and in which not all subjects experience the event of interest, Harrell's C would be the best choice. However, in clinical trials or other studies of shorter duration where the focus is completely on event status (which is close to binary), Chambless and Diao's may be a good choice.

Recently, Wolbers and colleagues [Wolbers et al., 2014] have described a method for adapting Harrell's C to a subdistribution hazards model, however, software is not currently available.

4.6 Decision Analysis

Sometimes the goal of a test or prediction model is to inform clinical decision making, such as whether to initiate a medical intervention, referral for more invasive testing, or beginning a preventive strategy or treatment. In these circumstances, where there is

a clinical decision to be made, perhaps more important than the statistical measures of model performance, are evaluating the real-world impacts the model would have. Additional methods for evaluating models in the context of decision making and clinical utility have been established and include net benefit and decision (or relative utility) curves [Kerr et al., 2014, Pepe et al., 2013, Vickers et al., 2008].

Net Benefit

Net benefit, or NB, refers to the change associated with the use of the new marker or risk factor. It is the average benefit of the intervention among those who otherwise would have an event, B, minus the cost of the intervention to nonevents, C. For example, if the risk model has a threshold used to classify patients into high and low risk, where those in the high risk group would receive an intervention, then the net benefit is:

$$NB = B * P(\text{event}) * P(\text{high risk}|\text{event}) - C * P(\text{nonevent}) * P(\text{high}|\text{nonevent})$$

A rational choice for a threshold is Cost/Benefit, when one does not already exist.

Predictiveness Curve

Another graphical approach to evaluating the performance of a model, called the predictiveness curve, was developed by Pepe et al. in 2008 [Pepe et al., 2008]. This work was motivated by the apparent contradictions seen in evaluating biomarkers in that a marker that is strongly related to risk, even after controlling for other risk factors, often seems to add little improvement in classification performance.

The predictiveness curve combines concepts from both risk modeling and performance evaluation and can simultaneously assess model fit and clinical utility in the population. To create it, the individual predicted values from the model are ordered from lowest to

highest and plotted on the curve by percentile of distribution. This is analogous to a plot of goodness-of-fit, similar to concepts for the Hosmer-Lemeshow statistic, which formally tests it.

For example, in the plot below from Pepe [Pepe et al., 2008], at 90 percent on the x-axis, the risk value is roughly 0.10, indicating that 90 percent of the cohort have predicted values lower than 0.10. The graph can also be used to fix a risk percentage and determine what percentage of the population have at least that as their risk score. Using 0.20 risk as an example, about 98 percent of the population have a score lower than 0.20.

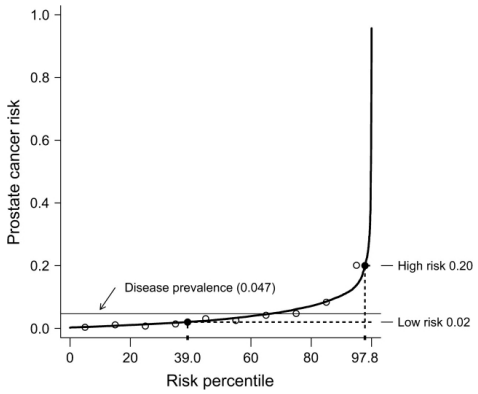


Figure 4.3: Sample Predictiveness curve

Another useful property of the predictiveness curve is that the explained variation, or R^2 can be calculated from it as:

$$R^2 = \int_0^1 (\text{pred}(\nu) - \rho)^2 d\nu / \rho(1-\rho)$$

where ρ = disease prevalence in the study population, and pred(ν) is the risk value at the ν^{th} percentile. The denominator is so that the value of R^2 remains in the range from 0 (a useless prediction) to 1 (predicts perfectly). It is also helpful to add a horizontal line to the plot at the disease prevalence, which would be a model in which all patients were assigned the same risk, and thus would be a useless model. The total gain can also be visually assessed as the area between the predictiveness curve and the useless model reference line.

Strengths of the predictiveness curve are that once a model has been developed, it is easy to plot, and intuitive to understand for clinicians. It has been extended for case control studies, and can also be used in the setting of a time dependent outcome by defining the event in a fixed interval of time, or plotting several possible time intervals.

Decision Curve Analysis

Another method for evaluating and comparing prediction models that incorporates clinical consequences is decision curve analysis [Vickers and Elkin, 2006]. Examples of clinical consequences are a false positive, where a patient would be referred for additional, possibly invasive, medical testing, who does not have disease, or a false negative, in which a patient who has the disease, misses the opportunity to be put on medication or other intervention. This method only requires the data set on which the models are tested, and can be applied to models that have either continuous or dichotomous results.

It starts by assuming that the threshold probability of a disease at which a patient would consider an intervention is informative of how the patient weighs the relative consequences of a false-positive and a false-negative prediction. At a predicted risk of 1, the patient would likely opt for treatment, whereas at a probability near 0, the patient will forgo it. Define a threshold between 0 and 1, p_t at which a patient will be uncertain of whether or not to treat. Then, define the Net Benefit as:

$$\text{Net Benefit} = \frac{\text{true positive count}}{n} - \frac{\text{false positive count}}{n}\left(\frac{p_t}{1-p_t}\right)$$

In this formula, true positive count and false positive count are the number of patients with true and false positive results at the threshold of p_t, and n is the total number of patients.

In order to create the decision curve, the net benefit must be calculated for a range of corresponding thresholds, p_t, and the authors provide a simple to follow, step by step algorithm for doing so:

1. Chose a value for p_t.

2. Calculate the number of true and false positive results using p_t as the cut-point for determining a positive or negative result.

3. Calculate the net benefit of the prediction model.

4. Vary p_t over an appropriate range and repeat steps 2 and 3.

5. Plot net benefit on the y-axis against p_t on the x-axis.

6. Repeat steps 1 through 5 for each model under consideration.

7. Repeat steps 1 through 5 for the strategy of assuming all patients are positive.

8. Draw a straight line parallel to the x-axis at $y = 0$ representing the net benefit associated with the strategy of assuming that all patients are negative, ("treat none").

Plotting the net benefit against threshold probability yields the decision curve. Below is a sample curve from the original paper, and note how the lines corresponding to the "treat all" (45-degree line) and "treat none" intersect at disease prevalence.

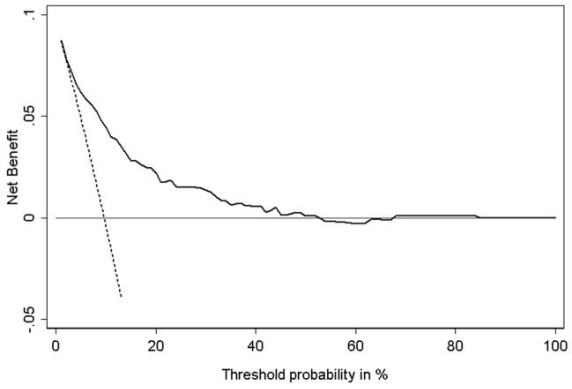


Figure 4.4: Sample Decision curve

The authors also provide a modification to the net benefit calculation that subtracts off a penalty for "test harm" in the case of an invasive medical test, or way of quantifying the

consequence of a false negative relative to a false positive (for example it being considered 50 times worse to miss a diseased patient than to send a healthy patient for additional testing or intervention). It can also be considered as the number of patients a clinician would feel comfortable testing, in order to find one who has disease.

The net benefit at a particular threshold has a simple interpretation as: compared to assuming all patients are free of disease, it is the number of additional diseased patients treated per 100 patients, without an increase in the number of false positive results. Other benefits of the decision curve are that it can also be used to compare several different models, and methods have been developed to correct for overfitting, estimate confidence intervals, apply to censored and competing risk data (including competing risk), and to calculate decision curves directly from predicted probabilities [Vickers et al., 2008]. Packages are available in Stata and R and are located at `http://www.decisioncurveanalysis.org`.

Discussion

Though there is much debate over how best to assess the performance and utility of a predictive model, one general consensus is that any new marker or risk factor must first show statistical significance in order to be considered. However, as statistical significance does not necessarily imply clinical significance or improvement in model prediction, the goal of the model and how it will be used clinically, should be taken into account when assessing its performance. Both the AUC and Net Reclassification Index have been criticized as they are not clinically meaningful and are broad summaries of changes in risk models [Kerr et al., 2014]. It has been suggested that it may be best to present several, complementary evaluation methods, such as first assessing the statistical significance of the risk factors, then presenting measures such as the AUC or C-index, and following up with an example of clinical implications for decision making [Vickers et al., 2011]. Regardless of the evaluation method chosen, model validation is always best performed on an external data source, meaning one

separate from the data set upon which it was created.

In our analysis of real data we will present the AUCs when validating the established CVD risk models in our cohort of breast cancer survivors, and present calibration plots. When developing our new models, we will first present tables of risk factor estimates, in order to demonstrate their statistical significance, and thus inclusion in the multivariate model. Once a final model has been selected, we will calculate the AUCs, and corresponding 95% confidence intervals, for a range of time points of prediction, noting that they will likely be an overstatement of model performance until they can be validated on an external data source. Areas of future work can exam the model performance in a decision analysis framework, once we have obtained adequate clinical input in regards to potential "treatment" options and thresholds.

Chapter 5

Comparison of Competing Risk Models

In order to use the competing risks methods for the analysis of the breast cancer data, I first evaluate and compare the models under consideration through simulations, for their robustness and estimation precision. With competing risks data, patients are at risk for a number of possible causes of failure or events, differing from traditional survival analysis under which only a single failure is analyzed, and competing failures are treated as censored observations. In the last two decades, there have been several noteworthy analysis methods proposed for analyzing multiple covariates in the presence of competing risks. The simplest way is to analyze each endpoint in its own Cox model, called a cause-specific hazards model. To analyze the effect of covariates on the cumulative incidence function, Fine and Gray proposed a method that is based on the proportional subdistribution hazards of the cumulative incidence function, whereas Klein and Andersen's method [Klein and Andersen, 2005] is based on calculating pseudovalues from a jackknife statistic from the cumulative incidence curve. Klein [Klein, 2006] has also proposed an additive hazards model. More recently, several groups have been favoring a multi-state modeling approach, including Beyersmann, Allignol and Schumacher [Beyersmann et al., 2012], as well as Putter, de Wreede and Fiocco

[Putter et al., 2007, de Wreede et al., 2010], where failures are modeled as transitions from one state to another.

While the availability of software for these methods is variable, the most popular methods presented in the literature for analyzing data with competing risk endpoints remain the cause-specific hazards (CSH) and subdistribution hazards (SH) (also known as the Fine-Gray) models. Both of these models extend a proportional hazards model, however, their interpretation is slightly different, due to the way the competing events are handled. The cause specific hazards model studies the effect of a covariate only on the hazard of the event of interest, and treats all other failures as censored observations. The subdistribution hazards model studies the effect of the covariate on the cumulative hazard function, or cumulative event probability, defined as the probability of failing from cause k at a specific time, given that that person has not yet failed from cause k. While the results of both models are reported as a "hazard ratio", the estimates of the hazard ratio from each model on a single data set can be very different, and even opposite in direction. In addition, a data set cannot simultaneously satisfy the proportionality assumptions of both models [Latouche et al., 2007, Beyersmann et al., 2007], and could be further complicated by time-varying effects. While the choice of model depends on the scientific question of interest, it has been argued recently that there is merit in presenting the results of both models side-by-side [Latouche et al., 2013, Dignam et al., 2012] to get a complete understanding of the effect of covariates, especially if the research question is for prognosis. Even when the subdistribution hazards model has been incorrectly specified, the hazard ratio can be interpreted as a time averaged effect of the covariate on the cumulative event probability [Beyersmann et al., 2009, Grambauer et al., 2010].

The following chapter will first review the existing literature that compares the CSH and SH models through simulations, highlighting important findings, as well as gaps. Next, in order to create my own simulation to compare the performance of the two models, I must also include data generation techniques, as there has been no consensus on how to most

accurately generate data that simulate a competing risk process. Lastly, I evaluate the two models, using a number of data generation methods, and include a simulation of a multi-state model for the generation process that most closely resembles the real data of Chapter 6.

5.1 Previous Simulation Studies of Competing Risks

Though the use of competing risks analysis methods has become widespread in the literature, simulation studies evaluating their properties, as well as directly comparing the results of CSH and SH models is scarce [Allignol et al., 2011, Beyersmann and Schumacher, 2007]. Several papers have been presented examining various facets of a competing risk analysis, however, most have only examined a single binary covariate, while altering other components such as dependence between outcomes, degree of censoring, and magnitude of covariate effects. The relationship between competing causes of failure and covariates that may act on one or more causes, can become rather complex, and the results from different analytic methods may yield differing, sometimes contradictory results.

The first paper to investigate the subdistribution hazards model through simulations was the original paper by Fine and Gray in 1999. The goals of the simulations they presented were to compare the results of their censoring complete estimation (when potential censoring time is always observable, i.e. no loss to follow-up, such as in clinical trials) with an adapted IPCW (inverse probability of censoring weighting) estimating function for right-censored data estimators. However, it was also the first paper to describe a method for simulating data that follow this type of model.

The authors ran two sets of simulations. Each set contained two covariates which were used to generate two independent causes of failure. In the first set of simulations, the covariates were standard normal, and in the second they were Bernoulli. Censoring percentages ranged from no censoring to 68% censored, and were examined for data sets of N=200, for

1,000 replicates. The simulation results are only presented for the first cause of failure, as the second cause was not generated under a subdistribution satisfying the model's proportionality assumption. The results showed that the two estimating techniques presented in the paper provided nearly identical results, thus even in the presence of censoring, the weighted estimating equation proved to be as efficient as the score function based on partial likelihood.

The paper then compared the results of a SH model with a CSH model in a data set of breast cancer patients treated with tamoxifen, in which the goal was predicting breast cancer recurrence while accounting for non-cancer death without recurrence. The results of the two models were quite similar in the real data set, largely because the covariates associated with an increased risk of breast cancer recurrence do not have an effect on the cause-specific hazard of death without recurrence, but the authors state that the advantage of the subdistribution hazards approach is that one can see the effect of each covariate directly on the cumulative incidence.

The next paper evaluating the subdistribution hazards model through simulations was in 2007 by Latouche and colleagues [Latouche et al., 2007]. The goal of the paper was to evaluate the "prognostic influence of an exposure on a specific cause of failure." To do so, the authors focus on a single failure type of interest and a single binary covariate, where the covariate relationship was generated from a cause-specific hazards, but a subdistribution hazards model is fit. Data for a second cause of failure were generated under a latent failure time model using an absolutely continuous bivariate exponential distribution, denoted (T_1, T_2) $ACBVE(a_1, a_2, a_{12})$, where $a_{12} = 0$ denotes independent failure times. Each data set contained 400 observations divided into two equally balanced groups, representing a single binary covariate under investigation.

When the covariate did not have an effect on the competing cause of failure, the CSH estimates were close to the true parameter values for the cause of failure under investigation, especially for smaller values of the effect. When the covariate had a protective effect on the competing cause of failure (negative coefficient), this led to an increase in the hazard ratio

for the first cause of failure. By decreasing the risk of the competing cause, it increases the risk of the primary cause of failure, thus inflating the hazard ratio. Conversely, when the covariate effect on the competing cause is positive, indicating an increase in the risk of the competing event, this will lead to a decrease in the coefficient estimate on the primary cause of failure. This is because observing an increase in events in the competing group will lead to fewer failures from the primary cause. Thus the authors conclude that when the covariate has an effect on the competing cause of failure, the parameter estimate for the effect on the primary event under investigation may be incorrect under the subdistribution hazards model.

In a letter to the editor regarding the Latouche et al. paper, Beyersmann and colleagues [Beyersmann and Schumacher, 2007] build upon the previous work, and through simulations, show that even when the SH models are incorrectly specified, they may still be useful. They begin by detailing the relationship between the subdistribution hazards α_1 and α_2 and the cause specific hazards, λ_1 and λ_2, in terms of the cumulative incidence function, as follows:

$$F_1(t) = 1 - \exp\left(- \int_0^t \alpha_1(u)du \right) = \int_0^t P(T > u) \cdot \lambda_1(u)du, \tag{5.1}$$

Where $P(T > u)$ depends on both cause specific hazards:

$$P(T > u) = \exp\left(- \int_0^u \lambda_1(\nu) + \lambda_2(\nu)d\nu \right).$$

The right hand side of equation (5.1) is the usual definition of the cumulative incidence function, while the middle part is the representation of failure times in terms of the subdistribution hazard. Differentiating equation (5.1) with respect to t gives the following formula, a relationship between the cause specific hazard and subdistribution hazard

$$\lambda_1(t) = \left(1 + \frac{P(T \leq t, \epsilon = 2)}{P(T > t)}\right) \cdot \alpha_1(t) = \left(1 + \frac{F_2(t)}{P(T > t)}\right) \cdot \alpha_1(t) \tag{5.2}$$

The authors state that this relationship has two important implications. The first is that the right hand side of equation (5.2) is time dependent, meaning that a proportional model for the subdistribution hazard, will not imply a proportional model for the cause-specific hazard, and vice versa. This was shown in the simulation study by Latouche *et al.*[Latouche et al., 2007].

The second implication is that the impact of a covariate may be different on the cause-specific hazard and the subdistribution hazard. For example, in the case where a binary covariate has an increase on λ_1, and an even stronger positive effect on λ_2, this will usually lead to a decreasing effect on the subdistribution hazard α_1, because more people will fail from the second cause, leaving fewer to fail from the first. This phenomenon was also exemplified in the simulation results of Latouche *et al.*[Latouche et al., 2007].

In conclusion, the authors state that the proportional subdistribution hazards model is appealing because it allows for interpretation of covariate effects in terms of the cumulative incidence function. It also seems that even when the model is mis-specified, that is, the proportional cause specific hazards model holds, it can still be used to offer an average effect of a covariate on the cumulative incidence function.

Prompted by the results of their 2007 papers, Beryersmann, Latouche, et al. wrote a follow-up paper on simulating competing risks data in survival analysis [Beyersmann et al., 2009]. They explain that the majority of simulation studies in the competing risks setting have utilized the latent failure time approach, rather than simulating data based on (possibly time dependent) cause specific hazards, the quantities to be modeled. A thorough literature search from 2000 to 2008 revealed that even the few articles that simulated data based on cause-specific hazards, assumed constant hazards, which may not always be realistic in the medical setting. Additionally, they notice a trend in the literature to present the results of both analyses side by side and wanted to examine the results in a controlled setting.

They illustrate their method by examining the least false parameter of a subdistribution

hazards model. The definition of a least false parameter is an estimator from a mis-specified model, or more simply put: when a particular model is specified and known to be true, if we fit a different class of models (e.g. fitting a proportional subdistribution hazards model when a proportional cause-specific hazards model is known to be true) but still estimate the parameter from the incorrectly specified model. While it has been shown that under a proportional cause-specific hazards model, the subdistribution hazards model will be mis-specified, the results from this model are still useful, and the least false parameter may be interpreted as a *time-averaged* effect on the cumulative event probability.

Using a single, binary covariate and hazard rates chosen to mimic data from their ONKO-KISS study, the authors found that after 500 simulations of 10,000 individuals each, the hazard estimate for the least false parameter from the SH model was within 0.1 percent of the effect simulated under the CSH model. In a second simulation of 1,000 runs, where each run contained a data set comprised of 1,616 individuals, as in the actual study, the estimate of the least false parameter was asymptotically the same as in the first set of simulations, and the empirical coverage of the 95% confidence interval was 95.3%, indicating accurate results, despite a misspecified model. Thus they conclude that even under a misspecified SH model, the estimate $\hat{\beta}$, is asymptotically consistent for the least false parameter $\tilde{\beta}$, and can be interpreted as a time-averaged subdistribution hazard ratio. However, these simulations only evaluated a single, binary covariate.

Building upon the work of the previous paper, Grambauer, Schumacher and Beyersmann [Grambauer et al., 2010] present a simulation study in 2010 to further examine the properties of the least false parameter (LFP) from a subdistribution hazards model, when the cause specific hazards model is the correct model. They once again study a single binary predictor and two competing events, with data generated to mimic the ONKO-KISS study of bloodstream cancer patients undergoing peripheral blood stem-cell transplantation. It advances the previous work by examining a range of censoring percentages, and four different scenarios of covariate effect, aiming to connect the interpretation from proportional CSH

and proportional SH models. To the best of the authors' knowledge, theirs and the 2009 Beyersmann paper are the only ones to study the effects from CSH and SH models side by side through simulations, at time of publication.

The four different scenarios of the effect of a binary covariate and two competing events are as follows:

1. effect only on the CSH of interest
2. effect only on the competing CSH
3. effect on both CSHs in the same direction
4. effect on both CSHs in opposite directions

For each scenario, N=1000 data sets were generated under a proportional cause specific hazards model, with either n=200 or n=1000 observations, and results were presented for 60 per cent censoring. The authors examined a range of censoring, but stated that the degree of censoring did not greatly alter the results, and thus were not shown. For scenario 1, where the covariate only has an effect on the event of interest, the LFP is close to the true β values for all values examined, with very little bias.

For scenario 2, where the covariate has an effect on only the competing event type, instead of finding LFP estimates of zero when modeling the effect of a covariate on the primary failure of interest, they take a sign opposite of the effect on the competing cause of failure. For instance, if the covariate is known to increase the hazard of the second cause of failure, the β estimate will have a negative sign, indicating a decrease in risk, for the hazard of the first cause of failure. Conversely, if the covariate is known to have a protective effect for the competing cause of failure, the LFP will demonstrate an increase in risk for the cause of failure of interest. This is due to the fact that as individuals are at an increased risk, and subsequently fail from one cause, they will not be able to experience the competing event, and thus their risk for the competing event will decrease, and vice versa.

For scenario 3, where the covariate has the same directional effect on both competing events, the results from the SH hazards model were generally somewhat attenuated, but

in the expected direction and with very little bias. Results were not presented depicting whether this effect was also seen when modeling for the competing event. Conversely, for scenario 4, where the directional effect was opposite on the two competing causes, the hazard ratio estimates for the first cause of failure are somewhat overestimated, which is expected under the same explanation given for scenario 2. The results for the second cause of failure are not presented, but one would expect to see slightly overestimated results in the correct direction.

In conclusion, the authors state that even when a subdistribution hazards models is misspecified, it can be interpreted as a time averaged effect on the cumulative incidence function and offers a complementary analysis to the traditional cause specific hazards approach.

In 2012, Dignam et al. [Dignam et al., 2012] present a simulation study to further elucidate when the results of the CSH and SH may differ. They expand the literature by also examining the results when there is dependence between events. For the simulation, two event times were generated from a bivariate survival distribution, and a censoring time was generated independently. They also examine a single binary predictor, under four different covariate-outcome relationships, with a range of censoring percentages, and dependence (correlation of 0.60) between the competing risks, versus independent events. The four scenarios examined were:

1. equal hazards in both groups (null covariate effect)
2. only the first cause of failure is affected by the covariate, and the hazard of group B being twice as high as group A
3. both events influenced by group, hazard of group B twice that of group A for both events
4. both events influenced by group, but the effect is greater for the first cause of failure than the second

The results presented were based off of 3,000 simulated data sets, each of 250 observations with approximately 33% censoring. Findings for independent competing risks were as

follows: under scenario 1, null covariate effect, both models performed as expecting, finding no covariate effect. Under scenario 2, the CSH model gives a hazard ratio estimate for group B that is twice that of group A for the first cause of failure, and equivalent for the second cause of failure, as expected. However, the SH hazard model gives a slightly smaller group effect for the first cause of failure (1.79) and also indicates that group B is less likely than group A to fail from the second cause of failure, despite the fact that the data were generated under equal hazards for cause 2. The explanation for this is that because fewer people in group A fail from the first cause, they are more likely to fail from the second cause. For scenario 3, the CSH found both hazards twice as high for group B as expected. However, the SH model had attenuated hazards for both causes (HR=1.29), due to the fact that for each cause of failure, a number of individuals fail from the competing cause. For scenario 4, the CSH model shows the correct effects for both causes of failure, but the SH model finds a reduced HR for the first cause of failure and a null effect for the second cause of failure. The explanation for the null effect is that because fewer group A individuals fail from the first cause, there are more available to fail from the second cause, thus the cumulative incidence of the second cause is similar between the groups.

When the simulations were repeated with moderate correlation between the event times for the causes of failure, the results for scenarios 1 and 3 were similar to those of independent competing risks. For scenario 2, the CSH and SH have very similar hazard ratios, showing a favorable effect for group B. The authors postulate that the dependence between the events causes the group effect for failure 1 to also be seen in failure 2. Similarly, for scenario 4, the two models give very similar estimates of group effect.

The last set of simulations presented investigated the impact of censoring, specifically for scenario 2, where group B has twice the hazard of group A but only for the first cause of failure. When there is no censoring, the SH find a significantly lower hazard for group B for event 2 because more of these patients have failed from the first cause. However, as censoring increases, the estimates become closer to the CSH estimates, and eventually show

a null effect.

Based on their simulation study and real data example, the authors conclude that the effect of a covariate on either the cause-specific hazard of an event or the cumulative incidence of that event can differ for the same set of data. Covariate effects in the CSH model are only on the failure type of interest, regardless of their effect on the competing events, since those who fail from the other events are removed from the at-risk set. However, in the SH model, the effect of the covariate is on the overall cumulative incidence, and may be a result of its direct effect on the event of interest, or an indirect effect of making the competing causes of failure more (or less) likely to occur. Both results have merit, and the choice of analytic method depends on the scientific question of interest. Dignam et al. suggest that for prognostic modeling, it is valid to model the cumulative incidence. He and other authors propose presenting both analyses together, to gain a true understanding of the effects of covariates on the various types of failure [Dignam et al., 2012, Latouche et al., 2013, Beyersmann et al., 2007, Wolbers et al., 2009].

Building upon the previous work, the current simulation will seek to compare the 2 model types when there is more than one covariate present, and when there is interest in predicting for more than one specific cause of failure. I will generate two covariates (risk factors) for each of 2 causes of failure, and assume that the risk factors for one cause of failure are not associated with the competing cause. Instead of examining a single binary covariate, I will generate one binary and one standard normal covariate for each endpoint. As the previous papers did not follow a uniform method to generate their competing risk data, I will include results for a variety of data generation techniques, which I will outline in the next section. Lastly, I will suggest an optimal method for data generation when there is more than one failure type of interest to be modeled, and using that method, extend the simulation to a multistate modeling approach described in [Putter, 2014].

5.2 Generating Competing Risks Data

Competing risk analyses are based on cause or event-specific hazards, which completely determine the competing risk process [Putter et al., 2007]. However, in a review of simulation studies of competing risks in the literature, Beyersmann et al. [Beyersmann et al., 2009] found that the majority of studies generated data that was not based on the cause specific hazards, but rather used the flawed latent failure time approach, or the unit exponential mixture distribution from Fine and Gray's original paper. A review of these methods follows.

5.2.1 Simulating using a unit exponential mixture distribution

In their original article presenting the proportional subdistribution hazards regression model, Fine and Gray [Fine and Gray, 1999] used simulated data to investigate their estimation. They limited the simulation to only two causes of failure and define the cumulative incidence function for the event of interest ($\epsilon = 1$) as

$$F_1(t; \mathbf{Z}) = Pr(T \leq t, \epsilon = 1 | \mathbf{Z}) = 1 - \Big(1 - p(1 - \exp(-t))\Big)^{\exp(\beta \mathbf{Z})}$$

where p is the probability of a type 1 failure for an individual with all covariate values equal to zero, and β represents the vector of proportional subdistribution log-hazard ratios for the failure time $\epsilon = 1$.

For failures of type 1, the subdistribution was generated using a unit exponential mixture with mass $(1 - p)$ at ∞ when $\mathbf{Z}_i = (0, 0)$, and the following proportional subdistribution hazards model when the covariates had nonzero values.

$$Pr(T_i \leq t, \epsilon_i = 1 | \mathbf{Z}_i) = 1 - \left[1 - p\left\{1 - \exp(-t)\right\}\right]^{\exp(Z_{i1}\beta_{11} + Z_{i2}\beta_{12})}$$

The subdistribution for failures of type 2 were then generated by setting $Pr(\epsilon_i = 2 | \mathbf{Z}_i) = 1 - Pr(\epsilon_i = 1 | \mathbf{Z}_i)$ and using an exponential distribution with rate parameter $\exp(Z_{i1}\beta_{21} +$

$Z_{i2}\beta_{22})$ for $Pr(T_i \leq t, \epsilon_i = 2, \mathbf{Z}_i)$. When censoring was included, censoring times were generated from the uniform $[a, b]$ distribution.

Data generation is performed by first generating the covariate values, and then using the marginal event type probability to determine the cause of failure. Next, the event time is derived based on event type and covariate values, when $\mathbf{Z} \neq 0$. In the original paper, each simulated data set contained 2 causes of failure and two covariates $\mathbf{Z}_i = (Z_{i1}, Z_{i2})$.

5.2.2 Simulating from a proportional hazards model

The simulation of data from a Cox proportional hazards model was first proposed by Leemis in 1987, [Leemis, 1987] and again by Bender et al. in 2005 [Bender et al., 2005]. Generating data from this model is not straightforward for several reasons, the first being that the Cox model is formulated through the hazard function, and not on individual survival times. Furthermore, the effect of the covariates is also in terms of the hazard function and must be translated into survival time. When the baseline hazard is constant, as in a Cox model, it can be calculated fairly easily using the relationship between survival time and the hazard function, as detailed below. While a bit more complex, the authors also provide details for simulating from the more Cox-Weibull and Cox-Gompertz models.

Recall the survival function of the Cox proportional hazards model

$$S(t|x) = \exp[-H_0(t)\exp(\beta' x)]$$

where

$$H_0(t) = \int_0^t h_0(u)du$$

is the cumulative baseline hazard. Using the relationship between the survival function and the distribution function yields

$$F(t|x) = 1 - \exp[-H_0(t)\exp(\beta' x)]$$

If Y is a random variable with distribution function F, then $U = F(Y)$ follows a uniform distribution on the interval $[0, 1]$, as does $(1 - U)$. Then

$$U = \exp[-H_0(T)\exp(\beta' x)]$$

If $h_0(t) > 0$ for all t then H_0 can be inverted, and the survival time T of the Cox model can be written as

$$T = H_0^{-1}[-\log(U)\exp(-\beta' x)]$$

where U is a random variable from the Uniform [0,1] distribution, and H_0^{-1} is the inverse of a cumulative baseline hazard function. For the exponential distribution, the survival times can be simulated through the following formula:

$$T = -\frac{\log(U)}{\lambda \exp(\beta' x)}$$

where the corresponding hazard function is given by

$$h(t|x) = \lambda \exp(\beta' x)$$

and covariates can easily be incorporated through $\exp(-\beta' x)$. This has become known as simulating via the "inversion method" or the method of Bender.

The inversion method was adapted for simulation of competing risks by simply repeating the data generating process for the number of failures under investigation, generally two. In this way there would now be a number of generated survival times, $(Y_{i1}, Y_{i2}, \cdots, Y_{ij})$, however, in the case of competing risks, only one can be observed. In order to decide on the observed failure, a latent failure time approach was used. Briefly, the latent failure time model was first conceptualized by Prentice et al. in 1978 [Prentice et al., 1978] and states that each failure type has its own associated failure time but only the first can be observed. Therefore $T = \min(Y_{i1}, Y_{i2}, \cdots, Y_{ij})$ and Y_{ij}, the time of failure from cause j, which would

be observed if the possibility of failure from causes other than j were removed.

5.2.3 Simulating from a Multi-State framework

After increasing criticism of the latent failure time model in conjunction with the Bender approach, Beyersmann et al [Beyersmann et al., 2009], described how to simulate competing risk data from a proportional cause-specific hazards model. For simplicity, they only present the method for two competing causes, but mention that it can be extended to more than two. They first present the competing risk setting as a simple multistate model with a beginning state and two absorbing states:

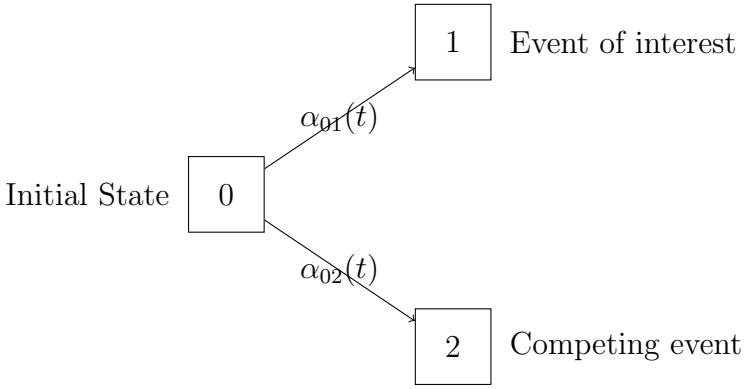


Figure 5.1: Competing Risks as a Multi-State Formulation

In this setting, the competing risk process is completely determined through the cause specific hazards (CSHs)

$$\alpha_{0i}(t)dt = P(T \in dt, X_T = i | T \geq t), i = 1, 2$$

which, in turn, completely determine the survival function

$$P(T > t) = \exp(-\int_0^t \alpha_{01}(u) + \alpha_{02}(u)du)$$

and the sum of the CSHs equal the all cause hazard:

$$\alpha_{01}(t) + \alpha_{02}(t) \cdot dt = P(T \in dt | T \geq t)$$

Therefore, when running a simulation, to determine which event occurs at survival time T, one can use the following probability: given that an individual fails at time T, the probability that the individual fails from cause 1 is

$$P(X_T = 1 | T \in dt, T \geq t) = \frac{P(T \in dt, X_T = 1 | T \geq t)}{P(T \in dt | T \geq t)} = \frac{\alpha_{01}(t)}{\alpha_{01}(t) + \alpha_{02}(t)}$$

Then, CSH driven competing events data can be simulated in four simple steps:

1. Define the CSHs ($\alpha_{01}(t)$ and $\alpha_{02}(t)$) as functions of time, and possibly depending on covariates [Bender et al., 2005].

2. Simulate the survival times T with all cause hazard $\alpha_{01}(t) + \alpha_{02}(t)$.

3. After the times have been generated, run a binomial experiment that decides on failure cause 1 with probability $\alpha_{01}(T)/((\alpha_{01}(T) + \alpha_{02}(T))$.

4. If desired, generate independent censoring times, C.

It is important to mention that under a proportional CSH model, a subdistribution hazards model will be misspecified. Thus data simulated from one may not yield accurate results when analyzed via the other. The authors provide details in the appendix for simulating under a SH model, which they purport is less biologically plausible and more challenging. In addition, the subdistribution hazards model only analyzes a single event of interest, while all other failures are subsumed into a single competing event. If the goal is to assess the effect of covariates on a specific failure type, a CSH model would be the more appropriate choice.

5.3 New Simulation Results

Here I examine the performance of the subdistribution hazards (SH) model, and cause-specific hazards model (CSH) side by side, using several methods of data generation. For simplicity, I include only two causes of failure, however, results for both causes are presented. To build upon previous work, I also include 2 distinct covariates (one standard normal, and one Bernoulli), for each cause of failure, as the primary interest is in the ability to use the coefficients from the models to make individual predictions. This is meant to mimic the real data analysis that will be presented in the following chapter. The simulations are repeated for a range of censoring distributions, and repeated to include the newly proposed multi-state model [Putter, 2014].

5.3.1 Fine and Gray Approach

In the first set of simulations, data were generated following the process described in the original Fine and Gray paper, and recreate the simulations from their Tables 1 and 2. Two causes of failure were generated, each sharing the same set of covariates $\mathbf{Z}_i = (Z_{i1}, Z_{i2})$. In the first table, the covariates were assumed to be i.i.d. standard normal and the true parameter values were $(p, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22},) = (0.3, 0.5, 0.5, -0.5, 0.5)$. Four different censoring scenarios were examined, and for the uncensored observations, roughly one-third failed from type 1 and two-thirds from failure type 2. Each simulated data set contained 200 observations, and results are presented for 1,000 runs.

Building upon the original simulations, results are presented for both the failure type "of interest" (failure 1), and the competing failure type (failure 2). Additionally, results from two cause-specific proportional hazards models (one for each failure cause) are presented. For the cause-specific models, the competing cause is treated as a censored observation when modeling the cause of interest, and vice versa.

For completeness, the second set of simulation parameters presented in Fine and Gray's

Failure Modeled	Censored	Fail from cause	Covariate	SH E(β)(SE)	Empirical Coverage	CSH E(β)(SE)	Empirical Coverage
Failure1	0%	33%	$\beta_{11} = 0.5$	0.51 (0.13)	94.2%	0.20 (0.14)	38.6%
			$\beta_{12} = 0.5$	0.50 (0.14)	93.1%	0.65 (0.13)	81.7%
Failure 2		67%	$\beta_{21} = -0.5$	-0.46 (0.10)	98.2%	-0.53 (0.10)	91.6%
			$\beta_{22} = 0.5$	-0.16 (0.09)	0%	0.18 (0.09)	10.7%
Failure1	25%	25%	$\beta_{11} = 0.5$	0.50 (0.14)	94.3%	0.29 (0.15)	68.0%
			$\beta_{12} = 0.5$	0.51 (0.14)	94.6%	0.60 (0.14)	92.0%
Failure 2		50%	$\beta_{21} = -0.5$	-0.59 (0.11)	92.6%	-0.57 (0.11)	89.9%
			$\beta_{22} = 0.5$	-0.06 (0.10)	1.9%	0.20 (0.10)	19.0%
Failure1	50%	16.7%	$\beta_{11} = 0.5$	0.51 (0.17)	94.7%	0.36 (0.18)	86.6%
			$\beta_{12} = 0.5$	0.51 (0.16)	95.5%	0.56 (0.17)	94.5%
Failure 2		33.3%	$\beta_{21} = -0.5$	-0.64 (0.13)	90.5%	-0.62 (0.13)	86.2%
			$\beta_{22} = 0.5$	0.15 (0.12)	18.1%	0.23 (0.12)	39.0%
Failure1	68%	11%	$\beta_{11} = 0.5$	0.50 (0.24)	92.1%	0.41 (0.23)	90.8%
			$\beta_{12} = 0.5$	0.51 (0.22)	94.0%	0.55 (0.22)	95.1%
Failure 2		22%	$\beta_{21} = -0.5$	-0.67 (0.17)	89.2%	-0.64 (0.16)	84.8%
			$\beta_{22} = 0.5$	0.21 (0.16)	47.7%	0.26 (0.16)	61.6%

Table 5.1: Results using Fine and Gray Simulation Plan: Standard Normal Covariates

paper are also presented below in Table 5.2, along with the results from corresponding cause-specific hazards models. Here the covariates were assumed to be i.i.d. Bernoulli (0.5) , and the true parameter values were $(p, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22},) = (0.6, 1, -1, 1, 1)$. Four different censoring scenarios were again examined, with 200 observations per data set, over 1,000 runs.

In Fine and Gray's original paper, simulation results were only presented for cause 1, where the data were generated from an exponential mixture model, and are very similar to the $E(\beta)$ and (SE) shown in Table 5.1. Additionally, the empirical coverage, defined as the number of 95% confidence intervals that contain the true parameter value, is close to 95% for failure 1, under all censoring scenarios. Omitted from the authors' original paper, is the performance of the model when modeling a second failure cause, Failure 2. As the model is only designed to show the relationship of covariates on the overall cumulative incidence curve, it was not developed to estimate the relationship of individual covariates on more

than one failure type's cumulative incidence.

As can be seen from the results, the β estimates for failure 2 are not as close to the true value for the beta coefficients. For β_{21}, where the covariate effect is protective for failure 2 but increases the risk for a failure 1 event, the estimates are slightly biased, though the empirical coverage is still good. For β_{22}, where there is an increased risk on both causes of failure, the estimates are not accurate for the effect on failure 2, and the empirical coverage ranges from zero to extremely low at best. There are two explanations for this: 1) the data for cause 2 were not generated from an exponential distribution and do not follow a proportional subdistribution hazards model, and 2) when a covariate has an effect on both causes of failure, under the subdistribution hazards model it is an estimate of the effect via a direct effect on the primary cause of failure and an indirect effect, by its effect on the competing cause [Dignam et al., 2012].

In terms of the performance of the CSH models, as the data for failure 1 were generated under the subdistribution hazards model, they will violate the assumption of the proportional hazards model [Grambauer et al., 2010], and we see biased estimates for the CSH models and poor empirical coverages, as expected. For failure cause 2, where the data were generated under an exponential distribution, the estimates are slightly inflated for β_{21}, and slightly attenuated for β_{22}, thus further demonstrating that treating competing failures as censored observations can give biased results.

When the simulations were repeated using the Fine and Gray method of data generation, but with Bernoulli covariates (Table 5.2), similar results were observed: performance of the subdistribution hazards model was excellent for failure cause 1, but poor for failure cause 2, due to the assumptions under which each cause of failure was generated, and the fact that each covariate was related to both causes of failure. The performance of the cause specific hazards model was similar to its performance under the first set of simulations; the parameter estimates were inconsistent, inaccurate, and highly variable across the different levels of censoring.

Failure Modeled	Censored	Fail from cause	Covariate	SH E(β)(SE)	Empirical Coverage	CSH E(β)(SE)	Empirical Coverage
Failure1	0%	60%	$\beta_{11} = 1.0$	1.01 (0.19)	95.3%	0.83 (0.20)	83.6%
			$\beta_{12} = -1.0$	-1.00 (0.19)	96.1%	-0.39 (0.20)	12.8%
Failure 2		40%	$\beta_{21} = 1.0$	-0.99 (0.27)	0%	-0.76 (0.25)	0%
			$\beta_{22} = 1.0$	1.20 (0.23)	86.6%	1.14 (0.25)	94.1%
Failure1	25%	45%	$\beta_{11} = 1.0$	1.02 (0.23)	94.9%	0.93 (0.24)	93.9%
			$\beta_{12} = -1.0$	-1.02 (0.22)	96.3%	-0.54 (0.24)	48.9%
Failure 2		30%	$\beta_{21} = 1.0$	-0.75 (0.27)	0%	-0.58 (0.26)	0%
			$\beta_{22} = 1.0$	1.49 (0.27)	60.4%	1.39 (0.29)	77.8%
Failure1	50%	30%	$\beta_{11} = 1.0$	1.03 (0.30)	96.2%	0.98 (0.30)	95.3%
			$\beta_{12} = -1.0$	-1.02 (0.29)	95.1%	-0.64 (0.31)	77.7%
Failure 2		20%	$\beta_{21} = 1.0$	-0.51 (0.30)	0%	-0.38 (0.29)	0%
			$\beta_{22} = 1.0$	1.66 (0.34)	54.5%	1.57 (0.35)	67.1%
Failure1	68%	19%	$\beta_{11} = 1.0$	1.04 (0.49)	96.6%	1.04 (0.46)	96.7%
			$\beta_{12} = -1.0$	-1.05 (0.47)	96.2%	-0.83 (0.46)	93.7%
Failure 2		13%	$\beta_{21} = 1.0$	-0.12 (0.37)	28.1%	-0.06 (0.36)	16.4%
			$\beta_{22} = 1.0$	1.98 (0.94)	59.6%	1.98 (1.62)	68.6%

Table 5.2: Results using Fine and Gray Simulation Plan: Bernoulli Covariates

5.3.2 Proportional Hazards with Latent Failure Approach

In order to more accurately assess the side-by-side performance of the CSH and SH models, especially with multiple covariates per failure type, it was necessary to create new simulations. The first set of new simulations generates two failure times using a unit exponential mixture distribution for each, examining two distinct covariates for each type of failure, one standard normal and one Bernoulli (0.5). This is equivalent to a real world setting in which there are two (or more) causes of failure, but each has a different set of known risk factors. For generating cause of failure, it assumes a latent failure time model where the observed failure is the minimum of the two independent failure times: $T_i = \min(T_{i1}, T_{i2})$. Results for $N = 500$ observations per data set, $k = 1,000$ runs, appear in Tables 5.3 and 5.4.

In Table 5.3 only the relevant covariates are included for each failure type, meaning, that when failure 1 is being modeled, only Z_1 and Z_2 are included as covariates in the model. Similarly, when modeling failure type 2, I only include Z_3 andZ_4 in the model. For the CSH models this is appropriate, as the coefficients are interpreted in terms of the change on the hazard of a specific event type. For the SH model, however, the effect of covariates is on the cumulative hazard, which can be directly, through their effect on the cause of failure of interest, or indirectly, by their effect on competing causes of failure. This illustrated by Table 5.4, which shows results for all covariates in all models.

Failure Modeled	Censored	Fail from cause	Covariate	SH E(β)(SE)	Empirical Coverage	CSH E(β)(SE)	Empirical Coverage
Failure1	0%	40%	$\beta_{11} = 0.5$	0.43 (0.07)	84.2%	0.50 (0.07)	95.4%
			$\beta_{12} = 1.0$	0.87 (0.15)	84.4%	1.00 (0.15)	95.5%
Failure2		60%	$\beta_{21} = 0.75$	0.42 (0.07)	0.5%	0.75 (0.07)	96.0%
			$\beta_{22} = -0.5$	-0.28 (0.12)	56.6%	-0.50 (0.12)	95.7%
Failure1	3%	41%	$\beta_{11} = 0.5$	0.43 (0.07)	84.9%	0.50 (0.07)	95.9%
			$\beta_{12} = 1.0$	0.87 (0.15)	85.3%	1.00 (0.15)	95.6%
Failure2		56%	$\beta_{21} = 0.75$	0.45 (0.07)	1.5%	0.75 (0.07)	95.7%
			$\beta_{22} = -0.5$	-0.30 (0.12)	63.8%	-0.51 (0.12)	95.2%
Failure1	16%	37%	$\beta_{11} = 0.5$	0.44 (0.08)	88.4%	0.50 (0.08)	96.3%
			$\beta_{12} = 1.0$	0.89 (0.16)	87.2%	1.00 (0.16)	95.5%
Failure2		47%	$\beta_{21} = 0.75$	0.57 (0.07)	32.9%	0.75 (0.07)	96.4%
			$\beta_{22} = -0.5$	-0.38 (0.13)	85.2%	-0.51 (0.13)	94.8%
Failure1	32%	32%	$\beta_{11} = 0.5$	0.46 (0.08)	93.1%	0.50 (0.08)	95.6%
			$\beta_{12} = 1.0$	0.93 (0.17)	92.3%	1.01 (0.17)	95.7%
Failure2		36%	$\beta_{21} = 0.75$	0.65 (0.08)	78.2%	0.75 (0.08)	96.4%
			$\beta_{22} = -0.5$	-0.44 (0.15)	93.0%	-0.50 (0.15)	95.9%
Failure1	48%	25%	$\beta_{11} = 0.5$	0.48 (0.09)	94.5%	0.50 (0.09)	95.4%
			$\beta_{12} = 1.0$	0.96 (0.19)	94.2%	1.00 (0.20)	95.4%
Failure2		27%	$\beta_{21} = 0.75$	0.70 (0.09)	92.5%	0.75 (0.09)	96.2%
			$\beta_{22} = -0.5$	-0.48 (0.18)	95.7%	-0.51 (0.18)	95.8%

Table 5.3: Latent Failure time approach, exponential distributions, separate covariates

Looking in greater detail at Table 5.3, for the first cause of failure, we see that the coefficients from the SH model and the empirical coverages are slightly lower than expected. As censoring increases, we see improvement in both the estimates and the coverage, as subjects are less likely to fail from the competing cause of failure as well. Results are similar when modeling the second cause of failure, with biased estimates and poor coverage, especially at low levels of censoring. For the CSH model, the coefficient estimates and the empirical coverages all perform well for both causes of failure, regardless of the censoring. This further illustrates the point made by Beyersmann [Beyersmann and Schumacher, 2007] and Grambauer [Grambauer et al., 2010], that when the cause specific hazards model is correctly specified, the subdistribution hazards model will be incorrect.

In Table 5.4, all covariates were included in all models to illustrate the indirect effect of a covariate on the subdistribution hazard. Looking at the first rows in the table, where we model the hazard of failure type 1, with known associated covariates Z_1 and Z_2 we see similar results to Table 5.3, with attenuated estimates and reduced coverage for the SH model, but excellent performance of the CSH model. However, when we look at the covariates that are not associated with the first cause of failure, Z_3 and Z_4, we see nontrivial coefficient estimates, and in the opposite direction of their known effect on the hazard of the second cause of failure. Specifically, the covariate Z_3 was generated such that it increases the hazard of the second cause of failure, but we observe a negative coefficient when modeling the first cause of failure. As explained in Dignam 2012 [Dignam et al., 2012], this is considered an indirect effect on the hazard for failure type 1 because as it increases subjects' risk for failing from the second cause of failure, it makes them less likely to fail from the first cause, thus appearing as a *protective* effect for that failure type. Conversely, Z_4, was generated with a protective effect on the hazard of failing from cause 2, but when modeling the first cause of failure, we see that it increases the hazard for failure type 1. This is also due to the indirect effect, whereby subjects who are at a decreased risk for failing from cause 2, have an increased risk of failing from the other cause of failure.

Failure Modeled	Censored	Fail from cause	Covariate	SH E(β)(SE)	Empirical Coverage	CSH E(β)(SE)	Empirical Coverage
Failure1	0%	40%	$\beta_{11} = 0.5$	0.44 (0.07)	87.4%	0.50 (0.07)	95.8%
			$\beta_{12} = 1.0$	0.89 (0.15)	86.9%	1.01 (0.15)	95.5%
			$\beta_{13} = 0$	-0.30 (0.07)	-	0.002 (0.08)	-
			$\beta_{14} = 0$	0.20 (0.14)	-	-0.003 (0.14)	-
Failure2	0%	60%	$\beta_{21} = 0$	-0.29 (0.06)	-	-0.003 (0.06)	-
			$\beta_{22} = 0$	-0.60 (0.13)	-	0.005 (0.12)	-
			$\beta_{23} = 0.75$	0.45 (0.07)	0.5%	0.76 (0.07)	95.3%
			$\beta_{24} = -0.5$	-0.30 (0.12)	60.8%	-0.51 (0.12)	95.6%
Failure1	3%	41%	$\beta_{11} = 0.5$	0.44 (0.07)	88.3%	0.50 (0.07)	95.9%
			$\beta_{12} = 1.0$	0.89 (0.15)	87.9%	1.01 (0.15)	95.5%
			$\beta_{13} = 0$	-0.30 (0.07)	-	0.002 (0.07)	-
			$\beta_{14} = 0$	0.20 (0.14)	-	-0.003 (0.14)	-
Failure2	3%	56%	$\beta_{21} = 0$	-0.28 (0.06)	-	-0.003 (0.06)	-
			$\beta_{22} = 0$	-0.57 (0.13)	-	0.004 (0.13)	-
			$\beta_{23} = 0.75$	0.48 (0.07)	2.0%	0.76 (0.07)	95.4%
			$\beta_{24} = -0.5$	-0.32 (0.12)	67.7%	-0.51 (0.12)	94.9%
Failure1	16%	37%	$\beta_{11} = 0.5$	0.45 (0.08)	90.6%	0.50 (0.08)	95.8%
			$\beta_{12} = 1.0$	0.91 (0.15)	89.2%	1.01 (0.16)	95.5%
			$\beta_{13} = 0$	-0.27 (0.07)	-	0.003 (0.08)	-
			$\beta_{14} = 0$	0.18 (0.15)	-	-0.003 (0.15)	-
Failure2	16%	47%	$\beta_{21} = 0$	-0.23 (0.07)	-	-0.003 (0.07)	-
			$\beta_{22} = 0$	-0.46 (0.14)	-	0.0008 (0.14)	-
			$\beta_{23} = 0.75$	0.59 (0.07)	37.4%	0.76 (0.07)	95.7%
			$\beta_{24} = -0.5$	-0.40 (0.13)	87.9%	-0.51 (0.13)	95.0%
Failure1	32%	32%	$\beta_{11} = 0.5$	0.47 (0.08)	94.2%	0.51 (0.08)	96.0%
			$\beta_{12} = 1.0$	0.94 (0.17)	92.2%	1.01 (0.17)	95.7%
			$\beta_{13} = 0$	-0.22 (0.08)	-	0.0007 (0.09)	-
			$\beta_{14} = 0$	0.15 (0.16)	-	- 0.0005 (0.16)	-
Failure2	32%	36%	$\beta_{21} = 0$	-0.17 (0.07)	-	-0.001 (0.08)	-
			$\beta_{22} = 0$	-0.34 (0.15)	-	0.0009 (0.15)	-
			$\beta_{23} = 0.75$	0.66 (0.08)	79.3%	0.75 (0.08)	96.2%
			$\beta_{24} = -0.5$	-0.45 (0.15)	94.2%	-0.51 (0.15)	95.7%
Failure1	48%	25%	$\beta_{11} = 0.5$	0.48 (0.09)	94.7%	0.50 (0.09)	95.1%
			$\beta_{12} = 1.0$	0.97 (0.19)	94.5%	1.01 (0.20)	95.5%
			$\beta_{13} = 0$	-0.16 (0.09)	-	0.001 (0.10)	-
			$\beta_{14} = 0$	0.11 (0.18)	-	0.002 (0.19)	-
Failure2	48%	27%	$\beta_{21} = 0$	-0.12 (0.09)	-	-0.0005 (0.09)	-
			$\beta_{22} = 0$	-0.23 (0.17)	-	0.0003 (0.18)	-
			$\beta_{23} = 0.75$	0.71 (0.09)	92.4%	0.76 (0.09)	96.3%
			$\beta_{24} = -0.5$	-0.48 (0.18)	95.7%	-0.51 (0.18)	96.0%

Table 5.4: Latent Failure time approach, exponential distributions, separate covariates, all variables in both models

In the CSH models, for the variables that are not associated with the specific cause being modeled, we see a non-significant coefficient, and they would ultimately be removed from the model. The results remain fairly consistent over the range of censoring distributions, with the exception being when we have nearly 50% censored observations, we see very similar results between the CSH and SH models, with accurate estimates and excellent empirical coverages. While statistically, a 50% censored sample seems undesirable, this is very common in clinical studies, especially in the setting where patients may fail quickly from their disease, or go on to be cured and remain alive until the end of study follow-up, when they would be censored.

5.3.3 Multi-State Approach

However, despite the excellent performance of the CSH models under the previous data generation methods, use of the latent failure time model for the generation of competing risk data has come under much criticism recently, for both the difficult to verify assumption of independence between events, and the lack of biological plausibility [Beyersmann et al., 2009]. Therefore, we also investigated the performance of the SH and CSH models under the same set of model and simulation parameters, but instead using the Beyersmann approach of data generation, as detailed in section 5.2.2.

Using these methods, we once again define two covariates (one Bernoulli and one standard normal) for each of 2 causes of failure, with the true $\beta = (0.5, 1.0, 0.75, -0.5)$. Results for $N = 500$ observations, over 1,000 simulations, appear in Tables 5.5 and 5.6.

Once again, we see excellent estimation and empirical coverage for the CSH models, across the entire range of censoring distributions. This is to be expected, as data were generated under a proportional hazard model. Since the CSH model holds, the SH model will be incorrectly specified, and we see this in the attenuated estimates and poor coverage of the SH models. Similarly to the latent failure approach, we see improvement in the SH model as the amount of censoring increases, and by approximately 50% censored observations, the

SH model performs nearly identically to the CSH model.

The results in Table 5.6, where all covariates are included in all models, again illustrate the indirect effect that a risk factor may have on the subdistribution hazard. Once again, β_3 and β_4, were generated such that they were only associated with failure 2, however, when we model failure 1 using the SH model, we see an effect for each that is opposite in sign to their effect on the second cause of failure. That is, β_3 was generated to have an increased risk on the hazard of failure type 2, but under a SH model for failure 1, we see a negative sign (protective effect), indicating a reduced risk of failure from cause 1. This is due to the fact that as subjects are at an increased risk to fail from cause 2, they are less likely to fail from cause 1. This also resulted in an attenuated estimate for β_3 when modeling failure 2 as the event type of interest under a SH model. Similarly, β_4 was generated to have a protective effect on the hazard of failure 2, and when modeling failure 1, we see that it increases the hazard of that failure, due to the fact that subjects who are less likely to fail from cause 2, indirectly become at an increased risk for failure 1. This is not seen in the CSH models, where only the relevant covariates have a statistically significant association with the failure cause being modeled.

Results from the multi-state model are identical to the CSH models, as they were analyzed using a Cox proportional hazards model with 2 absorbing states, and no intermediate events. Thus, the results are shown in Table 5.5 but are not repeated in Table 5.6, as the columns would be identical to those for the CSH results.

Failure Modeled	Censor	Fail from cause	True β	SH E(β)(SE)	Empirical Coverage	CSH E(β)(SE)	Empirical Coverage	Multistate E(β)(SE)	Empirical Coverage
Failure1	0%	57%	$\beta_{11} = 0.5$	0.36 (0.07)	40.8%	0.50 (0.06)	94.5%	0.50(0.06)	94.5%
			$\beta_{12} = 1.0$	0.71 (0.12)	34.2%	1.01 (0.13)	95.7%	1.01 (0.13)	95.7%
Failure2	0%	43%	$\beta_{21} = 0.75$	0.57 (0.08)	22.8%	0.76 (0.08)	94.9%	0.76 (0.08)	94.9%
			$\beta_{22} = -0.5$	-0.38 (0.14)	83.5%	-0.50 (0.14)	95.3%	-0.50 (0.14)	95.3%
Failure1	12%	50%	$\beta_{11} = 0.5$	0.39 (0.07)	62.4%	0.50 (0.07)	95.5%	0.50 (0.07)	95.5%
			$\beta_{12} = 1.0$	0.78 (0.13)	61.1%	1.01 (0.14)	95.9%	1.01 (0.14)	95.9%
Failure2	12%	38%	$\beta_{21} = 0.75$	0.61 (0.08)	46.9%	0.76 (0.08)	95.5%	0.76 (0.08)	95.5%
			$\beta_{22} = -0.5$	-0.40 (0.15)	89.3%	-0.50 (0.15)	95.8%	-0.50(0.15)	95.8%
Failure1	32%	39%	$\beta_{11} = 0.5$	0.44 (0.08)	86.4%	0.50 (0.08)	95.3%	0.50 (0.08)	95.3%
			$\beta_{12} = 1.0$	0.89 (0.15)	89.2%	1.01 (0.15)	95.3%	1.01 (0.15)	95.3%
Failure2	32%	29%	$\beta_{21} = 0.75$	0.68 (0.09)	78.1%	0.76 (0.09)	94.8%	0.76 (0.09)	94.8%
			$\beta_{22} = -0.5$	-0.45 (0.17)	94.0%	-0.50 (0.17)	95.0%	-0.50 (0.17)	95.0%
Failure1	50%	29%	$\beta_{11} = 0.5$	0.47 (0.09)	94.7%	0.50 (0.09)	94.1%	0.50 (0.09)	94.1%
			$\beta_{12} = 1.0$	0.95 (0.18)	94.2%	1.01 (0.18)	95.0%	1.00 (0.18)	95.0%
Failure2	50%	22%	$\beta_{21} = 0.75$	0.72 (0.10)	87.4%	0.76 (0.10)	94.7%	0.76 (0.10)	94.7%
			$\beta_{22} = -0.5$	-0.48 (0.20)	95.2%	-0.50 (0.20)	95.4%	-0.50 (0.20)	95.4%

Table 5.5: Beyersmann Approach, separate covariates

Failure Modeled	Censored	Fail from cause	Covariate	SH E(β)(SE)	Empirical Coverage	CSH E(β)(SE)	Empirical Coverage
Failure1	0%	57%	$\beta_{11} = 0.5$	0.37 (0.06)	49.4%	0.51 (0.06)	94.5%
			$\beta_{12} = 1.0$	0.74 (0.12)	42.6%	1.01 (0.13)	95.8%
			$\beta_{13} = 0$	-0.36 (0.06)	-	-0.003 (0.07)	-
			$\beta_{14} = 0$	0.24 (0.12)	-	0.002 (0.12)	-
Failure2	0%	43%	$\beta_{21} = 0$	-0.29 (0.07)	-	-0.0007 (0.07)	-
			$\beta_{22} = 0$	-0.61 (0.15)	-	0.00003 (0.15)	-
			$\beta_{23} = 0.75$	0.59 (0.08)	28.1%	0.76 (0.08)	94.7%
			$\beta_{24} = -0.5$	-0.39 (0.14)	86.9%	-0.50 (0.14)	95.3%
Failure1	12%	50%	$\beta_{11} = 0.5$	0.40 (0.07)	69.2%	0.50 (0.07)	95.5%
			$\beta_{12} = 1.0$	0.81 (0.13)	68.8%	1.02 (0.14)	95.8%
			$\beta_{13} = 0$	-0.31 (0.06)	-	-0.003 (0.07)	-
			$\beta_{14} = 0$	0.20 (0.13)	-	0.0002 (0.13)	-
Failure2	12%	38%	$\beta_{21} = 0$	-0.25 (0.07)	-	-0.0004 (0.08)	-
			$\beta_{22} = 0$	-0.52 (0.15)	-	0.0008 (0.15)	-
			$\beta_{23} = 0.75$	0.63 (0.08)	50.3%	0.76 (0.08)	95.0%
			$\beta_{24} = -0.5$	-0.42 (0.15)	90.2%	-0.50 (0.15)	94.8%
Failure1	32%	39%	$\beta_{11} = 0.5$	0.45 (0.08)	87.6%	0.50 (0.08)	95.5%
			$\beta_{12} = 1.0$	0.91 (0.15)	90.7%	1.01 (0.15)	95.4%
			$\beta_{13} = 0$	-0.22 (0.07)	-	-0.002 (0.08)	-
			$\beta_{14} = 0$	0.15 (0.14)	-	-0.001 (0.15)	-
Failure2	32%	29%	$\beta_{21} = 0$	-0.18 (0.08)	-	-0.0005 (0.09)	-
			$\beta_{22} = 0$	-0.37 (0.17)	-	-0.001 (0.17)	-
			$\beta_{23} = 0.75$	0.69 (0.09)	77.9%	0.76 (0.09)	94.7%
			$\beta_{24} = -0.5$	-0.45 (0.17)	94.4%	-0.50 (0.17)	95.2%
Failure1	50%	29%	$\beta_{11} = 0.5$	0.48 (0.09)	95.0%	0.50 (0.09)	94.6%
			$\beta_{12} = 1.0$	0.96 (0.18)	94.8%	1.01 (0.18)	95.3%
			$\beta_{13} = 0$	-0.15 (0.08)	-	0.0009 (0.09)	-
			$\beta_{14} = 0$	0.10 (0.17)	-	- 0.002 (0.17)	-
Failure2	50%	22%	$\beta_{21} = 0$	-0.12 (0.10)	-	0.0002 (0.10)	-
			$\beta_{22} = 0$	-0.24 (0.20)	-	0.001 (0.20)	-
			$\beta_{23} = 0.75$	0.73 (0.10)	87.1%	0.76 (0.10)	94.6%
			$\beta_{24} = -0.5$	-0.48 (0.20)	95.0%	-0.51 (0.20)	95.0%

Table 5.6: Beyersmann approach, all variables in both models

5.4 Discussion

In this new set of simulations, I compared the results from a set of CSH and SH models, when there were two causes of failure of interest. The results were examined under three different methods of data generation, and assumed each failure type to have its own subset of known risk factors. The risk factors were either standard normal or Bernoulli distributed, and results were repeated a range of censoring, from 0% up to 50%.

In terms of the subdistribution hazards model, most of the methods of data generation available, do not follow the proportionality assumptions of the subdistribution hazards model and lead to incorrect β estimates and poor empirical coverage. One exception to this is the original Fine and Gray method of data generation, however, the model only performs well for the first cause of failure that is analyzed. When the second cause of failure was modeled, coefficient estimates were not close to the true parameter values. This is due to the fact that the data for failure 2 were not generated under a proportional hazards assumption, but when the simulation was repeated with failure 2 generated from a unit exponential distribution, the results did not improve. Under these simulations, when the performance of the SH model was improved, the results from the CSH model were poor. However, as censoring increased, the results from the two models became very similar. Both models seemed to perform slightly better when the covariates were standard normal versus Bernoulli.

Using the latent failure time approach led to improved results for the CSH models, where the estimated values were equal to the true parameter values for over 95% of the simulations, regardless of censoring amount, or distribution of the covariate. The corresponding SH results showed more accurate results for the failure type that had fewer events (and thus greater censoring), but coverage was only moderate, until censoring neared about 50%. At this percentage of censoring, the CSH and SH models gave nearly identical results.

Lastly, when the exercise was repeated using the Beyersmann approach, the CSH models again performed as expected, while the SH model estimates were attenuated under the case of no censoring, but asymptoted to the CSH estimates as the amount of censoring increased,

especially once there was 50% censoring present.

The major findings are that the simulation results are highly sensitive to the method used to generate the data, and echo what was found by Latouche, Beyersmann and others, that if the CSH model is correctly specified, the SH model will not be, and the results may be very different between the two. However, we found that even when the model may be incorrectly specified, when there is a high degree of censoring present, the two methods give nearly equivalent results. This may be useful in the analysis of real data. We also demonstrated that the results of the multi-state model with absorbing states, were identical to the cause-specific hazards model.

It is important to note that aside from the original Fine and Gray method, none of the simulation results presented had data generated from a subdistribution hazards model. To the best of our knowledge, Beyersmann (2009) was the first and only author to describe how it could be done, but due to the complex programming and lack of biological plausibility, did not pursue it. A Pubmed search of citations revealed only 12 citing the original article, however, none of them generated data from the subdistribution hazards model.

While the goal of our real data analysis is risk prediction, rather than estimation of risk factor effect on a specific cause of failure, based on the similar results between the two models studied through simulation, given the high degree of censoring, we expect to see very similar results between the two methods.

Chapter 6

Application to Real Data

For the data analysis, a cohort was obtained from Kaiser Permanente Northern California (KPNC). This study included all women who were diagnosed with stage I to III breast cancer in the KPNC health care delivery system from January 1, 2000 to December 31, 2010, with follow-up through April 30, 2015. The KPNC coverage area includes 23 counties in the San Francisco Bay Area and the Central Valley of California, from Sacramento to Fresno [Kurian et al., 2013].

The first aim is to assess the performance of the existing cardiovascular disease risk models in a cohort of breast cancer survivors. To do this, the model predicted risk was calculated for each woman across a number of previously published, female-specific, risk models. The performance of the model prediction was assessed using the area under the curve (AUC) as a measure of discrimination, and model fit was assessed by calibration plots.

For the second aim, a model will be created, using established risk factors, to predict a woman's risk of breast cancer mortality and cardiovascular disease mortality, following a stage I-III breast cancer diagnosis. This will use available competing risk methodology discussed in the previous chapters, to examine two competing causes of failure, while also accounting for all other competing causes of death, collectively included as a "death from other" category.

Electronic medical records (EMR) were used to obtain patient demographic and disease specific information. Covariates were selected based on the existing literature of cardiovascular and breast cancer prognostic models, as well as their availability in the KPNC EMR database. Patient age was age at first diagnosis of breast cancer, race was coded as White, Black, Asian/Pacific Islander, or Other. Breast cancer covariates included type of surgery (lumpectomy, mastectomy, bilateral mastectomy), stage, grade, lymph nodes (number examined and status), estrogen receptor (ER) and progesterone receptor (PR) positivity, HER2 status, tumor size (cm), and laterality. Treatment variables included whether or not the patient received any radiation, chemotherapy, or hormonal therapy. A modified Charlson Comorbidity Index was used to summarize the patients' other comorbidities.

Cardiovascular disease risk factors corresponding to the Framingham models were extracted for this patient population from the time frame of 6 months post cancer diagnosis, up to 18 months post diagnosis. For those who still had missing risk factor information, data prior to breast cancer diagnosis was included, in order to obtain more complete risk factor data. Covariates included total cholesterol, HDL cholesterol, LDL cholesterol, systolic blood pressure (average of 2 measurements), smoking status, diabetes, and whether prescribed blood pressure lowering medication. History of cardiovascular disease was also collected and was defined as the occurrence of any of the ICD-9 codes in Table 6.1 at any time prior to, up through 6 months post, breast cancer diagnosis [D'Agostino et al., 2000].

Table 6.1: Cardiovascular Event ICD-9 Codes

Event Type	ICD9 Code
Myocardial infarction	410, 412
Coronary insufficiency	411, 414
Angina pectoris	413
Ischemic stroke	434
Hemorrhagic stroke	431, 432
Transient ischemic attack	435
Congestive heart failure	428
Intermittent claudication	440.21

Patient outcome data included disenrollment from KPNC (defined as >90 day lapse in

enrollment), and date and cause of death. Patients who were still alive were censored at their disenrollment date or on April 30, 2015, the last date of data collection. For all analyses, cause of death was coded as breast cancer death (ICD 10 code: C50), cardiovascular death (ICD 10 codes: I00-I99) or other. All analyses were approved by the Institutional Review Board of the KPNC Division of Research.

6.1 Statistical Methods Summarized

Patient and disease characteristics were summarized with descriptive statistics including means, standard deviations, and percentages. Cumulative incidences of death by cause of death were plotted over time to graphically display the outcomes, as well as estimate the incidence of death from each cause.

In the first set of analyses, the performance of previously developed cardiovascular disease models was assessed. Patient-level risk scores were calculated based on the originally published models either by use of score sheets, or direct calculation using model based parameter estimates. Discrimination ability of each model was assessed at the recommended time frame for prediction (e.g. 5 or 10 years) by the AUC (area under the ROC curve), with 95% confidence intervals calculated using the methods of DeLong [DeLong et al., 1988]. Calibration was assessed by plotting the observed and predicted number of events for each decile of risk. In order to create the Framingham recalibrated model, the published risk factor categories from the 2008 Framingham Model [D'Agostino Sr et al., 2008] were used in a Cox proportional hazards model where the parameter estimates were calculated for the current cohort. For this model, discrimination was also assessed on a continuous time scale using Harrell's C-index [Harrell Jr et al., 1996].

In the next set of analyses, prognostic models were created on the same set of patients, to predict the risk of death from cardiovascular disease and breast cancer, examining three types of models: a cause specific hazards model (CSH), subdistribution hazards model (SH)

(also known as Fine and Gray models), and a multi-state model (MS). In this analysis all states are absorbing states, meaning once a patient transitions into the state, they cannot move from it. This is logical since we are only modeling the transition from disease diagnosis to death. The transition intensities (hazards) are calculated using a proportional intensity regression model, which is similar to the Cox regression model, provided the data have been entered in the counting process style of format, and are presented as hazard ratios.

All three model types were run on both the full cohort, and a reduced cohort who did not contain missing covariate information, with the exception of smoking. This is called the complete case (CC) analysis. All possible two-way interactions were examined, and final multivariate models only contained covariates whose p-value remained ≤ 0.05. Due to the non-linear effect of age, it was modeled categorically. For ease of prediction age was not used as the time scale, but rather included as a covariate. For missing covariates, a "missing" category was included, in order to analyze the full data set, as well as examine whether there was any relationship between missing data and outcome. A separate analysis utilizing the reweighting approach described in Xu et. al. [Xu et al., 2009, Xu et al., 2011], was performed to assess the effect of the missing smoking variable on the parameter estimates in the Cox model. Models stratified by age and breast cancer stage were also examined.

All analyses were performed in SAS version 9.4 for windows (SAS Institute, Cary, NC), or R version 3.1.2 (http://cran.us.r-project.org/), using the *cmprsk*, *mstate*, *cmisc* and *survival* libraries.

6.2 Results

6.2.1 Description of Cohort and Outcomes

There were 20,462 women diagnosed with stage I-III breast cancer between 2000 and 2010 in KPNC. Demographic and breast cancer characteristic appear in Table 6.3. The mean age at diagnosis was 60 (range 21 to 103). The majority of the sample was white (79%) and

non-Hispanic (90%). Over half the sample (52%) was diagnosed with stage 1 disease, and the majority were treated with surgery (57% lumpectomy and 41% mastectomy).

Table 6.2: Summary of Outcomes and Event data

Event Type	**N(%)**	**Median time to event in months, (IQR)**
Disenroll from KP	209 (1.0%)	28 (15 - 50)
Death	2,729 (18.2%)	44 (23 - 69)
Breast Cancer death	842 (5.3%)	40 (38 - 43)
CVD death	696 (4.7%)	43 (40 - 46)
Other cancer	321 (2.3%)	51 (47 - 55)
All other causes	870 (5.9%)	45 (42 - 48)
Survivors	17,733 (81.8%)	7.2 years (5.2 - 10.5)
>5 years follow-up	13,685 (66.9%)	
>10 years follow-up	5,106 (25.0%)	
Summary of Non-fatal events		
Any Non-fatal CVD event	8,023 (74.2%)	52 (22 - 90)
CHD event	2,395 (22.5%)	35 (15 - 68)
Stroke	2,637 (30.9%)	56 (28 - 91)
Peripheral artery disease	5,738 (71.3%)	79 (49 - 117)
Heart failure	2,277 (23.2%)	40 (16 - 76)

There were 17,773 (81.8%) women who were alive at last follow-up, and median follow-up for survivors was 7.5 years. The cumulative incidence curves appear in Figure 6.1. There were 2,729 (18.2%) deaths overall, 842 (5.3%) deaths due to breast cancer, 696 (4.7%) due to cardiovascular events, and 321 (2.3%) due to other cancers. Of the 870 (5.9%) other deaths, the largest subgroups of causes included respiratory unspecified (198, 7.3%) septicemia (92, 3.4%), respiratory arrest (91, 3.3%), pneumonia (70, 2.6%) and Alzheimer's (37, 1.4%). Data

on intermediate cardiovascular events (non-fatal) following a breast cancer diagnosis were also collected and appear in Table 6.2.

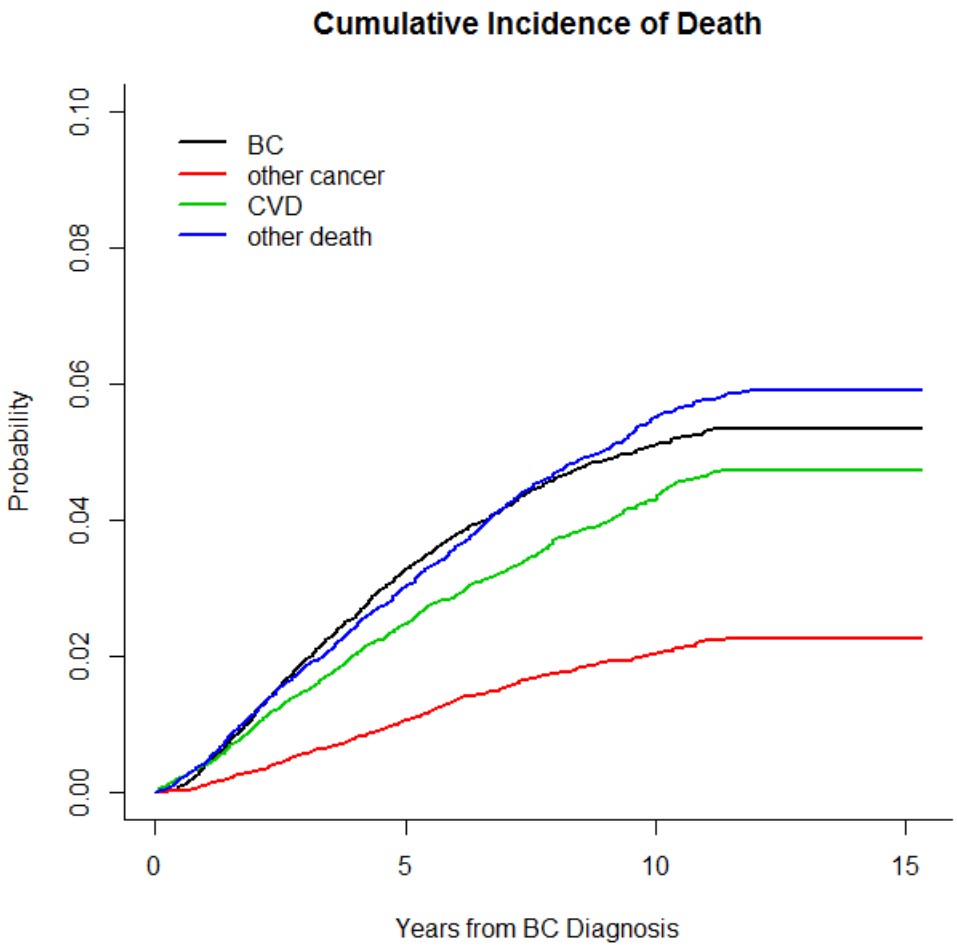
Table 6.3: Patient and Breast Cancer Characteristics (N=20,462)

	N	**%**
Age (at cancer diagnosis)		
21-34	302	1.48
35-39	571	2.80
40-44	1340	6.55
45-49	2167	10.59
50-54	2604	12.72
55-59	2962	14.47
60-64	2866	14.01
65-69	2482	12.13
70-74	2004	9.79
75+	3164	15.46
Race		
White	16166	79.0
Black	1507	7.36
Asian/Pacific Islander	2719	13.29
Other/Unknown	70	0.34
Hispanic		
Yes	1999	9.77
No	18455	90.19
Unknown	8	0.04
Smoking status		
Current smoker	2430	11.88
Former smoker	3095	15.13
Non-smoker	7389	36.11
Unknown	7548	36.89
Charlson comorbidity		
0	13130	64.17
1-2	6137	29.99
3+	1163	5.68
unknown	32	0.16
Surgery		
None	560	2.74
Lumpectomy	11583	56.60
Mastectomy	8309	40.60
Unknown	10	0.05

	N	**%**
Bilateral Cancer	138	0.67
Grade		
1 well differentiated	4734	23.13
2 - moderately differentiated	8336	40.74
3 poorly differentiated	5481	26.79
4 diffuse	259	1.27
Unknown	1652	8.07
Stage		
1	10843	52.99
2	7806	38.15
3	1813	8.86
Tumor size		
$\leq$ 2 cm	13727	67.09
(2,5] cm	5723	27.97
>5 cm	838	4.10
Diffuse or Unknown	174	0.85
ER/PR		
positive	11917	58.24
Negative	2689	13.14
Unknown/not done	5856	28.62
HER2		
Positive	1,862	9.10
Negative	11,938	58.34
Unknown/not done	6,662	32.56
Positive Lymph Nodes		
0	14150	69.15
1-3	4429	21.65
>3	1851	9.05
None examined/unknown	32	0.16
Treatments received		
Any chemotherapy	8683	42.43
Any hormonal therapy	8199	40.07
Any radiation	6621	32.36

Table 6.4 gives the distribution of cardiovascular disease risk factors in our population, as

Figure 6.1: Cumulative Incidence of Mortality by Cause of Failure



well as the female population from which the original Framingham equations were derived. As the original cohort was young and healthy at baseline, while the current cohort is selected based on their diagnosis of cancer, it is not surprising that the distribution of age is quite different between the two populations (mean age of 60 versus 49.6 in Framingham). The proportion of diabetes was also higher in the KPNC population (13% versus 4%), as well as those with a prior history of a CVD event (14% versus 0% in Framingham). The proportion of smoking was higher in the Framingham cohort, however, this could be due to the nearly one third that are missing smoking information. In addition, roughly 10% of the sample are missing cholesterol information, while 27% have been prescribed blood pressure lowering medication.

Looking in greater detail at the timing of the blood pressure and cholesterol information, relative to date of breast cancer diagnosis, the majority of blood pressure measurements were taken within a year of cancer diagnosis, with only 10% occuring a year after diagnosis. The cholesterol information was slightly more difficult to obtain, with 1,622 (8%) having information prior to breast cancer diagnosis and another 11% missing it altogether.

6.2.2 Evaluation of Published Cardiovascular Disease Risk Models

Currently, the majority of cardiovascular disease prediction models have been developed and validated on cohorts that are considered healthy at baseline. Some have been evaluated on older (greater than 65 years) populations, and found to have poor performance [Koller et al., 2012, Cooney et al., 2015], with AUCs in the range (0.67-0.68) for women. Their performance in a population with comorbid conditions, or at high risk for death from competing events, has not been studied. The first set of analyses focuses on evaluating the major cardiovascular disease prediction models in our cohort of cancer survivors.

In evaluating these prediction models on our population, it is important to note that not all models predict for the same outcomes or on the same time period. For example, the Framingham models predict for "hard CHD events", which they define as coronary death or

Table 6.4: Cardiovascular disease risk factors compared to Framingham women

	KPNC		Framingham
	N	% or SD	%
Age	60.0 (mean)	(SD=12.8)	49.6 (SD=11.1)
HDL Cholesterol			
<35	7338	35.9	4
35-44	4957	24.2	15
45-49	2342	11.5	12
50-59	2928	14.3	28
60+	515	2.5	41
Missing	2382	11.6	0
Total Cholesterol			
<160	2350	11.5	8
160-199	6431	31.4	30
200-239	6191	30.3	33
240-279	2571	12.6	20
280+	750	3.7	9
missing	2169	10.6	0
Systolic blood pressure			
<120 (optimal)	5145	25.1	35
120-129 (normal)	4967	24.3	21
130-139 (high normal)	4600	22.5	15
140-149 (stage I hypertension)	2179	10.7	19
150-159 (stage I hypertension)	1499	7.3	
160+ (stage II-IV hypertension)	914	4.5	10
missing	1158	5.7	0
On blood pressure lowering medication	5499	26.9	n/a
Smoker			
No	7389	36.1	62
Yes (current)	2430	11.9	38
Quit	3095	15.1	n/a
Missing	7548	36.9	0
Diabetic	2669	13.0	4
History of any CVD event	2854	14.0	0

CHD events (ICD codes 402, 410-414, 429.2, and 429.9), while the 2008 model can be used to predict the "earliest of any event", including stroke, peripheral artery disease, and congestive heart failure. They can predict either a 5 or 10-year risk. The European models (CORE and SCORE) use a cardiovascular mortality endpoint, as mortality is the least subjective endpoint to capture. In order to validate the model performance for a given time point, we can only include patients who have been followed for the corresponding length of time. Any competing events occurring prior to the follow-up period are also removed from the data set used for evaluation. Although this is known to cause bias, these models were created using methodology that did not account for competing events, and must therefore be evaluated in the same manner.

Our data set includes cases diagnosed from 2000-2010, thus at least half of the study will not have been followed for 10 years, a common prediction horizon for a number of the cardiovascular models. Therefore, it is first important to understand whether there are major differences between the groups with and without 10 years of follow-up. In our study there were only 25% survivors who were alive and followed for at least 10 years. Compared to the survivors with less than 10 years of follow-up, the women with longer follow-up had significant differences ($p<0.01$) in race (more white, fewer Asian), smoking (more likely to be current and former smokers than never smokers (counter-intuitive), Charlson score (lower scores), systolic blood pressure (slightly higher BP), whether on blood pressure lowering medication (less likely to be on it), breast cancer stage (fewer stage 3), ERPR positivity (more positives), and whether they received radiation or chemotherapy (more received both). This comparison may be further biased because the group of survivors with less than 10 years of follow-up includes women who may still experience one of the fatal endpoints prior to 10 years post diagnosis, whereas the group of women surviving for 10 years or more are known to be event free at 10 years. Therefore, the results of the model performance may not be generalizable to the cohort as a whole.

Table 6.5 lists the cardiovascular disease risk models evaluated on the current data set. It

Table 6.5: Summary of CVD Model performance in KP BC survivors

Risk Model	**Endpoint definition**	**Time period predicts for**	**Sample size**	**Events**	**AUC (95% CI)**
Framingham 2000	Hard CHD	2 years	10,211	304	0.70 (0.67, 0.73)
	Hard CHD	4 years	9,529	596	0.74 (0.72, 0.76)
Framingham 2001	Hard CHD	5 years	8,236	699	0.71 (0.69, 0.73)
	Hard CHD	10 years	1,976	952	0.64 (0.62, 0.67)
CORE	Hard CHD	5 years	8,180	692	0.75 (0.73, 0.77)
	Hard CHD	10 years	1,963	943	0.75 (0.73, 0.78)
Framingham recalibrated	Hard CHD	Continuous	11,019	966	0.62 (0.56, 0.68)
	Hard CHD	5 years	8,236	699	0.78 (0.76, 0.80)
	Hard CHD	10 years	1,976	952	0.76 (0.74, 0.79)
Framingham 2008	Earliest event	10 years	4,478	3,734	0.66 (0.64, 0.68)
SCORE	CVD death	10 years	1,308	188	0.73 (0.69, 0.78)
SCORE OP	CVD death	10 years	1,300	184	0.76 (0.72, 0.80)

gives the time period for which the model predicts, the definition of an event for that model, and the sample size from the current cohort on which the evaluation is based. The results are grouped by outcome definition. The model evaluation metric that is reported is the area under the curve (AUC). This is the most commonly reported measure of discrimination in the cardiovascular disease risk modeling literature, with AUCs usually in the (0.66-0.86) range [D'Agostino Sr et al., 2001], when evaluated on independent cohorts who are healthy at baseline and similar in age distribution.

Generally, the models demonstrated moderate discrimination on our cohort, with AUCs ranging from 0.64 (Framingham model for hard events evaluated at 10 years) to 0.78 (Framingham model recalibrated to our data set, Table 6.6, evaluated at 5 years). However, it is important to note that these validations exclude women who have died from breast cancer or other causes, prior to any cardiovascular event, and may therefore over-estimate risk, by artificially reducing the at-risk set. This can be seen when we evaluate the model continuously, rather than at a fixed time-point, such as in the Framingham recalibrated model using Harrell's C-index [Harrell Jr et al., 1996] (which is analogous but not equal to an AUC), yielding a C-index of only 0.62. The SCORE OP model which was developed specifically for

Table 6.6: Parameter Estimates and Relative Risks for Framingham model Recalibrated to KP data set, complete cases (N=11,019)

	Framingham Women		**KP BC Women**	
	Beta	**RR (95% CI)**	**Beta**	**RR (95% CI)**
Age	0.17	1.19 (0.97-1.45)	0.04	1.04 (0.99, 1.10)
Age squared	0.001		0.0002	
Blood Pressure				
<120 (optimal)	0.74	0.48 (0.22-1.05)	-0.14	0.87 (0.72, 1.06)
120-129 (normal)	reference		reference	
130-139 (high normal)	0.37	0.69 (0.34-1.42)	0.06	1.06 (0.90, 1.26)
140-159 (stage I hypertension)	0.22	1.24 (0.69-2.24)	-0.027	0.97 (0.81, 1.17)
160+ (stage II-IV hypertension)	0.61	1.84 (1.00-3.39)	0.299	1.35 (1.01, 1.79)
Total Cholesterol				
<160	0.21	1.23 (0.27-5.64)	-0.009	0.99 (0.82, 1.19)
160-199	reference		reference	
200-239	0.44	1.55 (0.81-2.96)	-0.177	0.84 (0.71, 0.99)
240-279	0.56	1.74 (0.90-3.40)	0.058	1.06 (0.86, 1.30)
280+	0.89	2.44 (1.21-4.93)	-0.045	0.96 (0.68, 1.36)
HDL Cholesterol				
<35	0.73	2.08 (1.00-4.31)	0.513	1.67 (1.22,2.29)
35-44	0.60	1.82 (1.05-3.16)	0.252	1.29 (1.07,1.55)
45-49	0.60	1.82 (1.05-3.14)	0.121	1.13 (0.92,1.39)
50-59	reference		reference	
60+	0.54	0.58 (0.33-1.02)	-0.138	0.87 (0.74,1.03)
Diabetes	0.87	2.38 (1.40-4.06)	0.808	2.24 (1.94,2.60)
Smoker	0.98	2.65 (1.77-3.97)	0.529	1.70 (1.46,1.98)

an older population, had an AUC of 0.76, but did not significantly outperform any of the other models.

Additionally, plots of observed versus expected events were examined by decile of risk for each established model. Overall, the plots indicated poor calibration in our data set. The majority of the models underestimated the actual number of events, with the exception of the recalibrated Framingham model evaluated at 5 years, which overestimated risk except for the highest decile. The Framingham models that predicted in the 2 and 4 year horizon, were the best suited to our data set, likely due to the large percentage of the cohort having not been followed for the full 5 or 10 years, as well as the exclusion of competing breast cancer and other deaths.

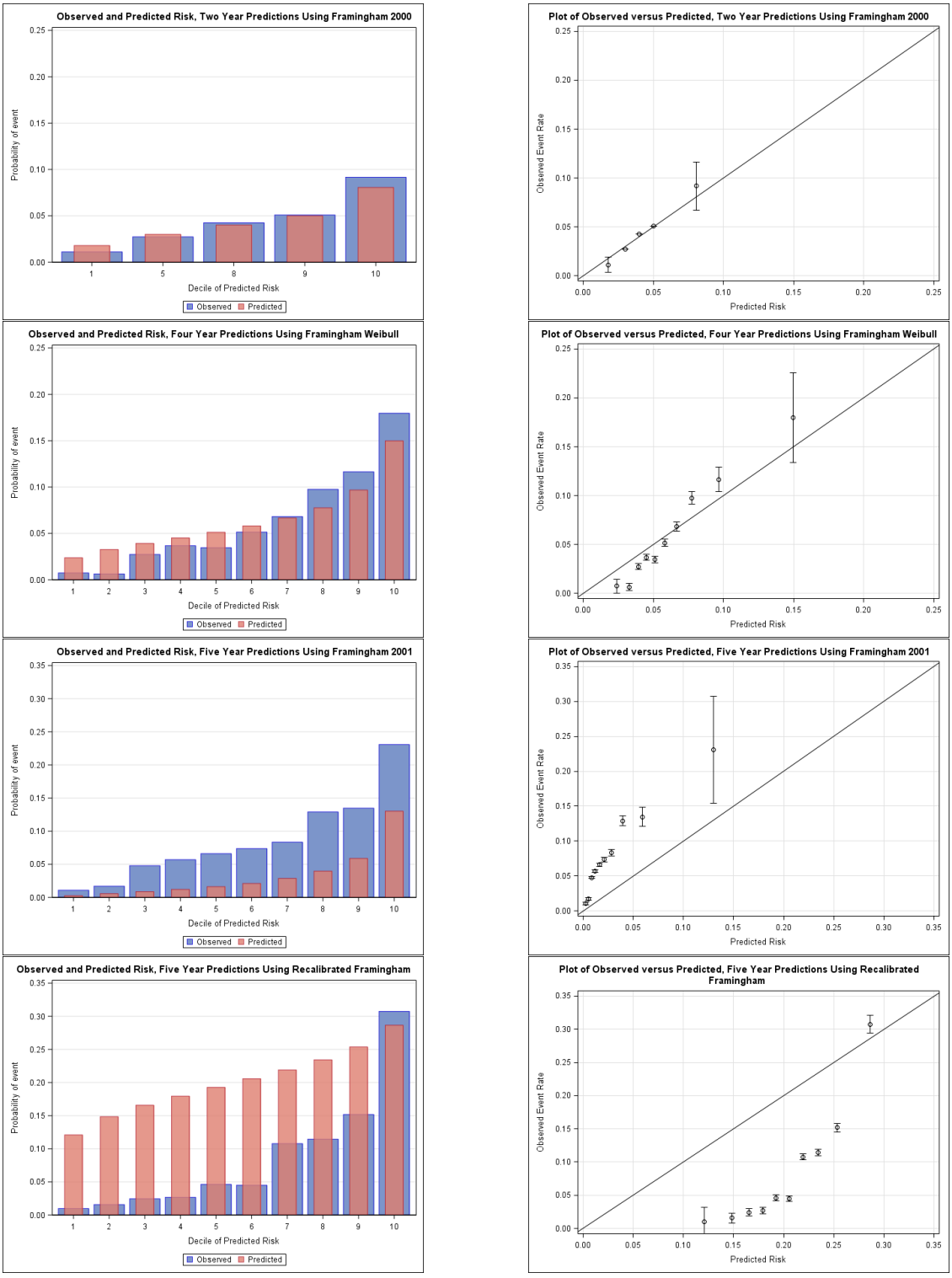


Figure 6.2: Calibration Plots of Framingham Models

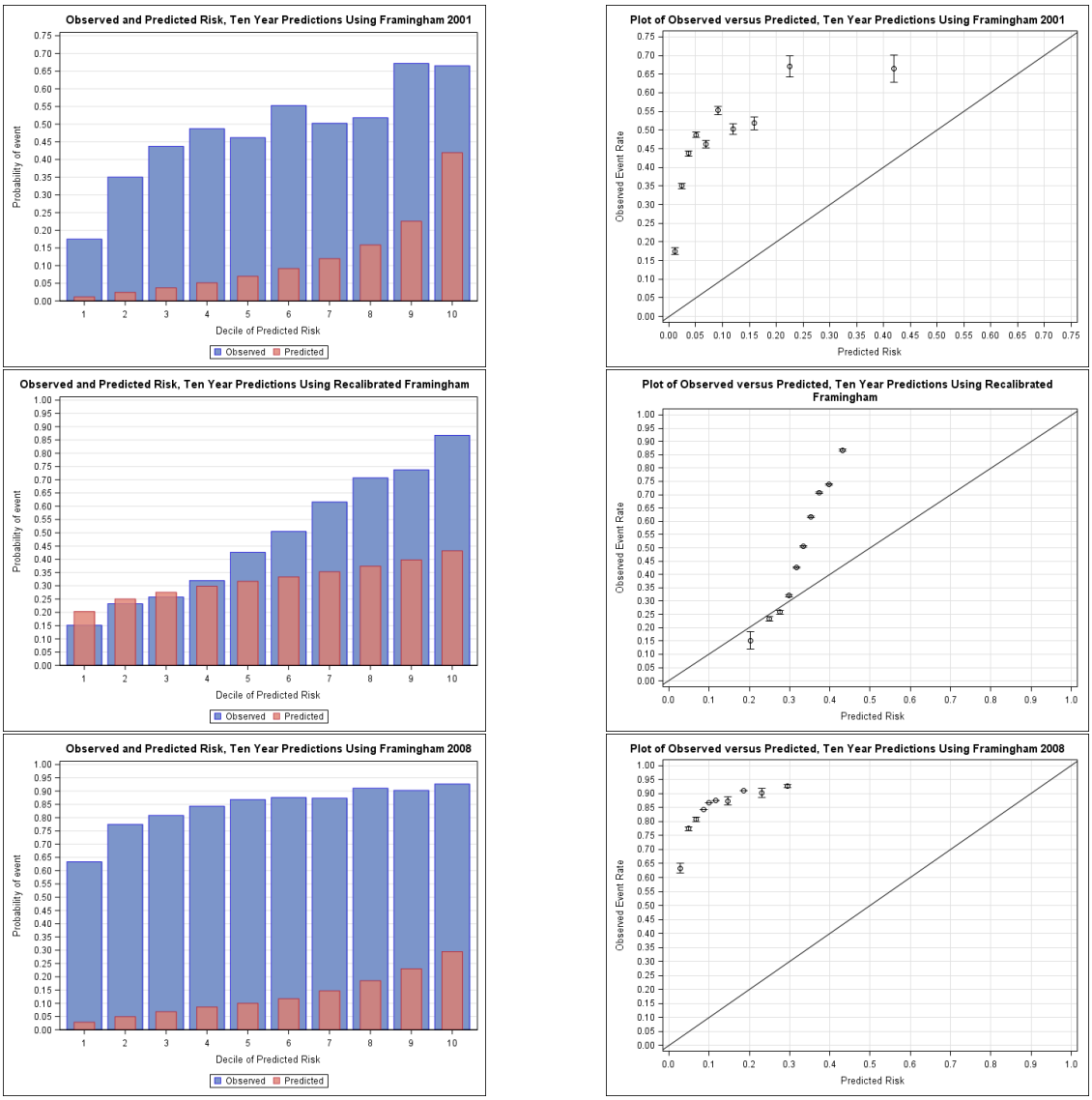


Figure 6.3: Additional Calibration Plots of Framingham Models

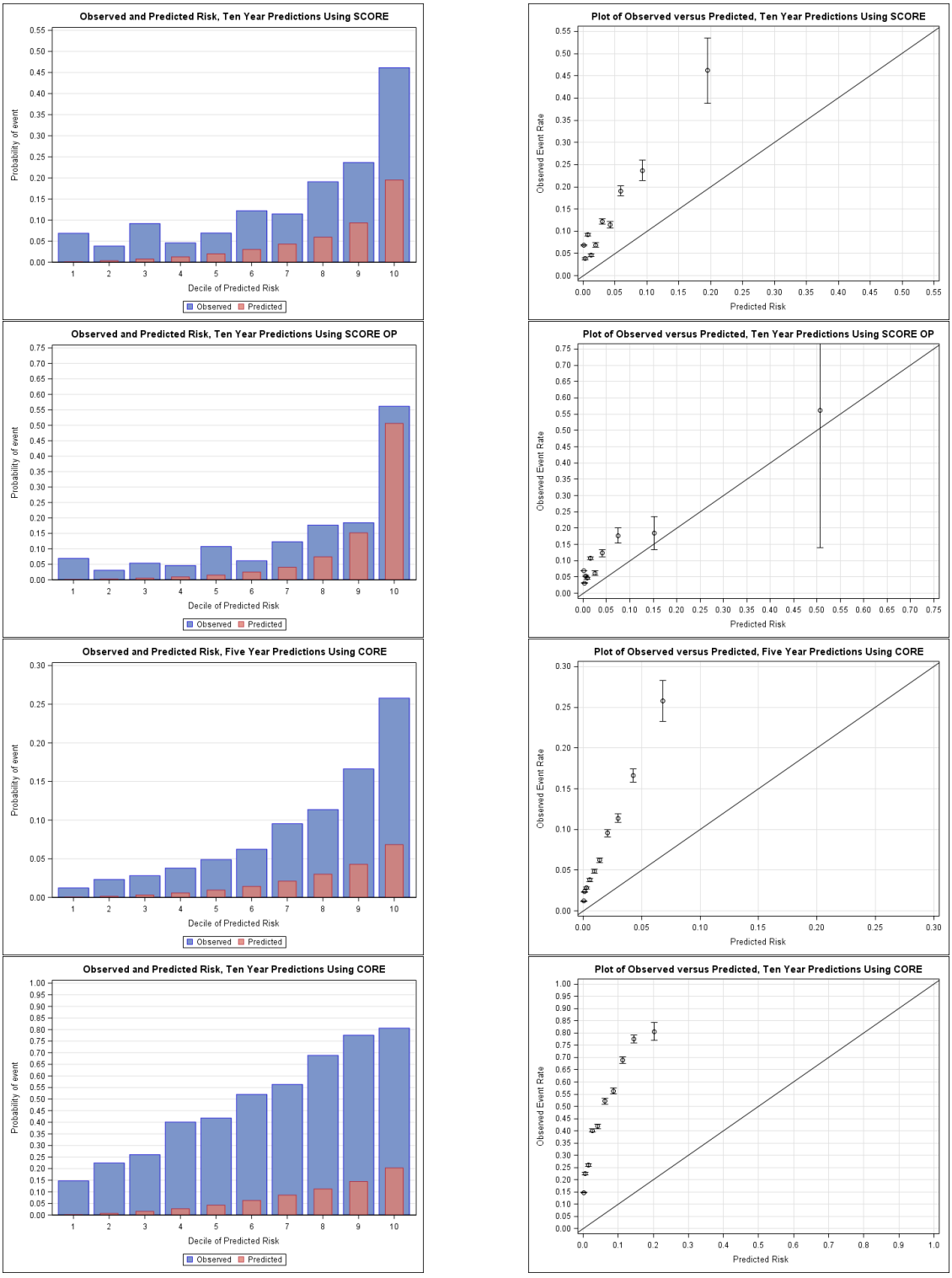


Figure 6.4: Calibration Plots of CORE and SCORE Models

Discussion: Aim I

While most of the cardiovascular disease risk models were established on a cohort that was younger and without a history of CVD events or comorbid complications, they still showed moderate discrimination in our cohort of breast cancer survivors. The model performance was widely influenced by the horizon for which it was predicting, as the more narrow prediction intervals (2 and 4 years) had better calibration to our data set. This was likely due to the shortened follow-up for a proportion of the data, as well as failures from competing causes being treated as censored observations at their time of failure. The Framingham model that could predict the first of any event, including strokes, PAD, etc., was not well suited to our data set, likely due to the fact that 14% of these women already had a documented event prior to their breast cancer diagnosis. This number could possibly be even higher if any of these women had an event prior to their enrollment in KP, which was not captured in the database. The SCORE models which only used cardiovascular mortality as an endpoint, performed slightly better at 10 years than the Framingham models, which also include myocardial infarction, however, the narrow definition for CVD or CHD death used by the model, greatly reduced the sample size upon which it was evaluated and may not be as clinically useful when the goal is to identify women for potential interventions.

Due to the limitations of the current sample, the next set of analyses will use the traditional cardiovascular and breast cancer recurrence risk factors, to create a new model that can potentially be used to identify breast cancer survivors who may be at an increased risk of cardiovascular death. By also modeling risk of breast cancer mortality, it can help clinicians identify which subsets of patients are at greater risk for breast cancer mortality versus cardiovascular mortality, to focus health and intervention strategies once a woman has completed her primary treatment for breast cancer and entered the remission phase.

6.3 Modeling the Risk of Breast and Cardiovascular Mortality

The next set of analyses examines the risk factors associated with cardiovascular and breast cancer mortality, using the competing risk methodology of CSH models, SH models, and Multi-state models. The endpoint of breast cancer death was chosen over recurrence due to the limitations of administrative data in reliably capturing breast cancer disease recurrence. Since the cardiovascular disease model performance was slightly better in our cohort when using mortality as the endpoint, and 14% of the sample had a history of a CVD event prior to cancer diagnosis, the mortality endpoint was chosen over non-fatal events for the cardiovascular disease process as well. Using a multi-state design, it would also have been possible to model the process from cancer diagnosis to intermediate non-fatal CVD events (such as peripheral artery disease or stroke), and then death, however, due to the variations in CVD event histories prior to breast cancer diagnosis, it was not pursued.

Table 6.7 describes the bivariate relationship between each risk factor and the outcome of interest in both a cause specific hazards (CSH) model (Cox) and a subdistribution hazards (SH) model (Fine and Gray). For the CSH models, the other events are censored at their time of occurrence, such that in the CSH model for cardiovascular death, the primary outcome of interest is cardiovascular death and deaths from breast cancer or other causes are treated as censored. In the Fine and Gray model for cardiovascular disease death, breast cancer and other deaths continue to be included in the risk set, so as to not overestimate the true risk of death from any one cause.

In comparing the results between the CSH and the SH models on the same cause of failure, we see that the estimates are very close between the two. This is comparable with the simulation results that showed when there is a large degree of censoring, the two models give nearly identical results. Many of the known risk factors for cardiovascular disease show statistically significant relationships with CVD death, but a handful of the breast cancer

disease characteristics do as well. This could be due to the large sample size and will be further examined in the multivariate models.

Table 6.7: Bivariate Results for each outcome of interest

	Cardiovascular Death		**Breast Cancer Death**	
	CSH Model HR (95%CI)	SH Model HR (95%CI)	CSH Model HR (95%CI)	SH Model HR (95%CI)
Age (at diagnosis)	1.09(1.08, 1.10)	1.08 (1.08, 1.09)	1.01 (1.01,1.02)	1.01(1.00, 1.01)
Age categories				
21-39	Reference	Reference	Reference	Reference
40-49	0.68(0.34, 1.34)	0.69(0.35, 1.38)	0.68(0.49, 0.94)	0.69 (0.50, 0.95)
50-54	0.89(0.45, 1.75)	0.92(0.46, 1.81)	0.55(0.39, 0.78)	0.56(0.40, 0.79)
55-59	0.90(0.46, 1.76)	0.93(0.47, 1.81)	0.55(0.39, 0.77)	0.55(0.39, 0.77)
60-64	1.27(0.66, 2.43)	1.30(0.68, 2.49)	0.59(0.42, 0.82)	0.59(0.42, 0.83)
65-69	1.94 (1.03, 3.65)	1.98(1.05, 3.73)	0.48(0.34, 0.69)	0.48(0.34, 0.69)
70-74	3.83(2.06, 7.11)	3.79(2.04, 7.05)	0.65(0.45, 0.92)	0.62(0.44, 0.88)
75+	9.06(7.97, 16.5)	8.29(4.54, 15.1)	1.09(0.80, 1.49)	0.96(0.70, 1.32)
Race				
White	2.44 (1.77, 3.36)	2.37 (1.72, 3.27)	1.46 (1.16, 1.84)	1.42 (1.13, 1.79)
Black	4.23 (2.90, 6.16)	4.01 (2.76, 5.84)	2.01(1.48, 2.74)	1.90 (1.40, 2.60)
Asian/Pacific Islander	Reference	Reference	Reference	Reference
Other/Unknown	2.40 (0.58, 9.94)	2.29 (0.56, 9.44)	1.78 (0.56, 5.65)	1.72 (0.54, 5.47)
Non- Hispanic	1.82 (1.31, 2.52)	1.80 (1.30, 2.50)	1.05 (0.83, 1.32)	1.03(0.82, 1.30)
Smoking status				
Current	2.29 (1.76, 2.98)	2.22 (1.70, 2.88)	2.07 (1.62, 2.64)	2.00 (1.57, 2.55)
Former	1.77 (1.36, 2.31)	1.74 (1.34, 2.27)	1.45 (1.13, 1.86)	1.42 (1.11, 1.82)
Non-smoker	Reference	Reference	Reference	Reference
Unknown	2.36 (1.92, 2.90)	2.32 (1.89, 2.83)	2.48 (2.07, 2.97)	2.43(2.04, 2.90)
Diabetes	2.12 (1.77, 2.52)	2.06 (1.72, 2.45)	0.98(0.80, 1.21)	0.95 (0.77, 1.16)
Charlson comorbidity				
0	Reference	Reference	Reference	reference
1-2	2.43(2.06, 2.86)	2.35 (1.99, 2.77)	1.01 (0.87, 1.18)	0.97 (0.84, 1.13)
3+	7.33(5.94, 9.04)	6.30 (5.11, 7.76)	1.53 (1.16, 2.01)	1.27 (0.96, 1.67)
unknown	1.52 (0.21, 10.8)	1.51 (0.21, 10.9)	1.51 (0.38, 6.04)	1.52 (0.38, 6.09)
Bilateral	1.82 (0.90, 3.64)	1.86 (0.91, 3.77)	0.74 (0.28, 1.97)	0.73 (0.27, 1.93)
Surgery				
None	3.20 (2.30, 4.46)	2.79 (1.99, 3.90)	5.75 (4.33, 7.61)	5.12 (3.86, 6.81)
Lumpectomy	Reference	Reference	Reference	Reference
Mastectomy	1.52 (1.31, 1.77)	1.46 (1.25, 1.70)	2.78 (2.40, 3.22)	2.71 (2.33, 3.14)
Unknown	6.82 (1.70, 27.4)	6.85 (1.87, 25.1)	3.95 (0.56, 28.2)	3.91 (0.59, 25.7)
Grade				
Well differentiated	Reference	Reference	Reference	Reference
Moderately differentiated	1.22 (0.99, 1.50)	1.19(0.97, 1.47)	4.52 (3.21, 6.36)	4.48 (3.18, 6.30)
Poorly differentiated	1.45 (1.16, 1.80)	1.34 (1.08, 1.68)	10.9 (7.77, 15.2)	10.5(7.52, 14.7)
Diffuse	0.95 (0.45, 2.04)	0.92 (0.43, 1.96)	9.75(5.66, 16.8)	9.80 (5.67, 16.9)
Unknown	1.96 (1.49, 2.56)	1.90 (1.45, 2.49)	5.77 (3.89, 8.56)	5.63 (3.80, 8.35)
Stage				
1	Reference	Reference	Reference	Reference
2	1.56 (1.33, 1.83)	1.51 (1.28, 1.76)	5.09(4.18, 6.20)	4.99 (4.09, 6.07)
3	2.22 (1.74, 2.82)	1.90 (1.49, 2.42)	15.9 (12.9, 19.7)	14.9 (12.0, 18.3)
Tumor size $\leq$2 cm	Reference	Reference	Reference	Reference
(2,5] cm	1.63 (1.39, 1.92)	1.54 (1.31, 1.81)	4.00 (3.43, 4.66)	3.86 (3.31, 4.51)
>5 cm	2.23 (1.63, 3.04)	1.93 (1.41, 2.64)	8.59 (6.91, 10.7)	7.94 (6.39, 9.87)
Diffuse or Unknown	4.32 (2.58, 7.24)	3.29 (1.95, 5.55)	18.7 (13.6, 25.6)	16.1(11.7, 22.2)
ER/PR				
positive	0.85(0.68, 1.05)	0.91 (0.74, 1.13)	0.31 (0.27, 0.37)	0.32 (0.28, 0.38)
Negative	Reference	Reference	Reference	Reference
Unknown/not done	0.63 (0.49, 0.81)	0.67 (0.53, 0.87)	0.34 (0.28, 0.41)	0.35 (0.29, 0.42)
HER2				
Positive	Reference	Reference	Reference	Reference
Negative	1.32 (0.98, 1.78)	1.34 (1.00, 1.81)	0.60 (0.48, 0.73)	0.59(0.48, 0.73)
Unknown/not done	1.21(0.89, 1.65)	1.23(0.90, 1.67)	0.74 (0.60, 0.92)	0. 74(0.60, 0.92)
Positive Lymph Nodes				
0	Reference	Reference	Reference	Reference
1-3	1.14 (0.95, 1.37)	1.11 (0.93, 1.34)	2.35 (1.98, 2.77)	2.33 (1.97, 2.76)

Table 6.7: Bivariate Results for each outcome of interest

	Cardiovascular Death		**Breast Cancer Death**	
	CSH Model HR (95%CI)	SH Model HR (95%CI)	CSH Model HR (95%CI)	SH Model HR (95%CI)
>3	1.71 (1.36, 2.16)	1.52 (1.21, 1.92)	7.51 (6.40, 8.81)	7.17 (6.12, 8.42)
None examined/unknown	4.69 (1.75, 12.6)	4.53 (1.66, 12.4)	3.07 (0.76, 12.3)	2.85 (0.70, 11.6)
Adjuvant Online Node Categories				
0	Reference	Reference	Reference	Reference
1-3	1.14 (0.95, 1.37)	1.11 (0.93, 1.34)	2.35(1.98, 2.78)	2.33 (1.97, 2.76)
4-9	1.49 (1.13, 1.96)	1.37(1.04, 1.81)	5.45(4.49, 6.60)	5.28 (4.35, 6.40)
$\geq$ 10	2.40 (1.66, 3.47)	1.93 (1.33, 2.80)	13.8 (11.2, 16.9)	12.6 (10.3, 15.5)
None examined/unknown	4.69 (1.75, 12.6)	4.54 (1.66, 12.4)	3.07 (0.77, 12.3)	2.85 (0.70, 11.6)
Treatments received				
Any chemotherapy	0.48 (0.41, 0.57)	0.48 (0.41, 0.57)	2.25 (1.95, 2.58)	2.30(2.00, 2.65)
Any hormonal therapy	0.92 (0.79, 1.07)	0.95 (0.81, 1.10)	0.55 (0.48, 0.64)	0.56 (0.48, 0.65)
Any radiation	0.69(0.58, 0.82)	0.70 (0.60, 0.83)	0.81 (0.70, 0.94)	0.83 (0.71, 0.96)
Blood Pressure				
<120 (optimal)	Reference	Reference	Reference	Reference
120-129	1.14 (0.88, 1.48)	1.14 (0.88, 1.48)	0.99 (0.81, 1.22)	0.99 (0.81, 1.22)
130-139	1.40 (1.09, 1.79)	1.40 (1.09, 1.80)	1.04(0.84, 1.28)	1.04 (0.84, 1.28)
140-149	1.66 (1.24, 2.21)	1.64 (1.23, 2.19)	1.32 (1.04, 1.68)	1.31 (1.03, 1.67)
150-159	2.20 (1.64, 2.94)	2.17 (1.62, 2.91)	1.01 (0.75, 1.36)	0.99 (0.73, 1.33)
160+	3.33 (2.46, 4.50)	3.20 (2.37, 4.32)	1.63 (1.20, 2.20)	1.55 (1.15, 2.10)
Total Cholesterol				
<160	Reference	Reference	Reference	Reference
160-199	0.89(0.67, 1.17)	0.91 (0.69, 1.20)	1.00 (0.79, 1.27)	1.02 (0.80, 1.30)
200-239	0.85 (0.65, 1.13)	0.88 (0.66, 1.16)	0.93 (0.72, 1.18)	0.95 (0.74, 1.21)
240-279	0.79 (0.57, 1.11)	0.81 (0.58, 1.13)	1.07 (0.81, 1.41)	1.09 (0.82, 1.44)
280+	0.83 (0.51, 1.37)	0.84 (0.51, 1.39)	1.17 (0.79, 1.73)	1.19 (0.80, 1.76)
HDL Cholesterol				
<35	2.44 (1.64, 3.64)	2.24 (1.50, 3.34)	2.81 (2.03, 3.88)	2.61 (1.89, 3.61)
35-44	1.38 (1.08, 1.76)	1.35 (1.06, 1.73)	1.44 (1.17, 1.78)	1.42 (1.15, 1.74)
45-49	1.00 (0.75, 1.35)	1.00 (0.75, 1.34)	1.04 (0.81, 1.33)	1.03 (0.80, 1.32)
50-59	0.91 (0.72, 1.15)	0.91 (0.72, 1.15)	1.08 (0.89, 1.31)	1.08 (0.89, 1.31)
60+	Reference	Reference	Reference	Reference
Log cholesterol ratio	1.41 (1.03, 1.93)	1.37 (0.99, 1.90)	2.04 (1.57, 2.64)	1.98 (1.53, 2.56)
History of CVD	6.44 (5.50, 7.54)	5.91 (5.05, 6.91)	1.40 (1.16, 1.69)	1.24(1.02, 1.50)

The first set of models, Tables 6.8 and 6.9 use cardiovascular death as the primary outcome of interest. We investigate both the full dataset, where missing covariates are included as a category labeled "missing", and then investigate only the complete case analysis, with the exception of the smoking variable, which is missing for a large percentage of the data set. Additionally, models that stratified by history of CVD prior to breast cancer diagnosis, were examined, but the sample size was greatly reduced, and the parameter estimates were similar to the complete case analysis (not shown).

Age, race, smoking status, history of CVD, and Charlson comorbidity score were all significantly associated with the risk of death from cardiovascular disease. Interestingly, several of the breast cancer risk factors were also associated with risk of death from CVD: grade, tumor size, number of positive lymph nodes, radiation (full data set only), chemotherapy

and HER2NEU (CSH models only). Only ages greater than 70 had an increased risk in CVD death, and the hazard doubled after age 75. Compared to the reference group of Asian/Pacific Islander, the black subgroup was also at a nearly two-fold increased risk of CVD death (2.6 Cause Specific Hazard Ratio, 2.5 Subdistribution Hazard Ratio). Those who were current smokers were also at an increased risk (2.7 CSHR, 2.5 SHR), while those who had quit were at a slightly increased risk (1.4 CSHR and SHR), though this was not statistically significant in the complete case model. Interestingly, the group that was missing information on smoking habits, yielded hazard ratios that were very similar to the current smokers in terms of their risk of death due to cardiovascular disease (2.5 CSHR, 2.4 SHR). In terms of a history of CVD, those with a prior event had twice the risk of those with no prior history (2.1 CSHR, 2.0 SHR). The only traditional CVD risk factor that remained statistically significant in the model was HDL cholesterol, with only the lowest group (HDL<35) having an increased risk (1.5 CSHR and SHR). Blood pressure and the use of blood pressure lowering medication, as well as diabetes, were not significantly associated with risk of CVD mortality in this population.

All three models gave very similar estimates for the risk factors, and had nearly the same subset of risk factors that were statistically significant. The estimates from the CSH and MS models are expected to be identical, however, the two models included a slightly different subset of variables, which is responsible for the slight differences seen in the parameter estimates. Tables 6.14 and 6.15 detail which variables remained in each of the multivariate models. For the endpoint of CVD death, the CSH and multi-state models contained the same subset of risk factors with the exception of HER2 (CSH only) and ERPR (Mstate only). The SH model included HER2 but did not find ERPR, chemotherapy or type of breast cancer surgery to be significantly associated with the cumulative incidence of CVD death. This may be due to the relationship between age, stage of breast cancer at diagnosis, and use of chemotherapy. In this cohort, the youngest group of patients were diagnosed, on average, with higher stages of disease, and thus had the highest rates of mastectomy and

chemotherapy usage, as well as the lowest rates of CVD death.

Table 6.8: Models for Cardiovascular Mortality, 696 events

	CSH Model HR (95%CI)	SH Model HR (95%CI)	MS Model HR (95%CI)
Age Categories			
21-39	Reference	Reference	Reference
40-49	0.81 (0.41,1.60)	0.81(0.41,1.60)	0.81 (0.13,1.49)
50-54	0.98 (0.49,1.95)	1.02 (0.51,2.06)	1.03 (0.34,1.72)
55-59	1.03 (0.52,2.03)	1.09(0.56,2.13)	1.04 (0.36,1.72)
60-64	1.33(0.68,2.55)	1.36 (0.70,2.63)	1.34 (0.68,2.00)
65-69	1.87 (0.98,3.58)	1.97 (1.04,3.75)	1.88 (1.23,2.53)
70-74	3.37 (1.78,6.37)	3.46 (1.85,6.48)	3.40 (2.76,4.04)
75+	6.28 (3.36,11.8)	6.09(3.30,11.2)	6.27 (5.64,6.90)
Race			
White	1.42 (1.02,1.97)	1.38(0.99,1.90)	1.42 (1.09,1.75)
Black	2.57 (1.76,3.77)	2.47 (1.68,3.63)	2.53 (2.15,2.91)
Asian/Pacific Islander	Reference	Reference	Reference
Other/Unknown	1.41 (0.34,5.84)	1.26 (0.31,5.14)	1.48 (0.06,2.90)
Smoking status			
Current	2.70 (2.06,3.54)	2.49 (1.89,3.28)	2.59 (2.32,2.86)
Former	1.38 (1.06,1.80)	1.36(1.03,1.78)	1.36 (1.09,1.63)
Non-smoker	Reference	Reference	Reference
Unknown	2.46(1.99,3.03)	2.38(1.92,2.94)	2.28 (2.07,2.49)
History of CVD			
No prior history	Reference	Reference	Reference
History of event	2.09 (1.70, 2.58)	2.01 (1.63,2.49)	2.10 (1.89,2.31)
Unknown	1.57(1.11, 2.20)	1.42(0.99,2.02)	1.53 (1.19,1.87)
HDL Cholesterol			
<35	1.52 (1.02,2.28)	1.47 (0.98,2.22)	1.54 (1.14,1.94)
35-44	1.02(0.80,1.31)	1.04 (0.81,1.34)	1.04 (0.79,1.29)
45-49	0.84 (0.62,1.13)	0.88 (0.65,1.17)	0.84 (0.54,1.14)
50-59	0.81(0.64,1.02)	0.81 (0.64,1.03)	0.81 (0.58,1.04)
60+	Reference	Reference	Reference
unknown	1.30 (1.00, 1.68)	1.35 (1.03,1.77)	1.32 (1.06, 1.58)
Charlson comorbidity			
0	Reference	Reference	Reference
1-2	1.44(1.20, 1.73)	1.38 (1.15,1.66)	1.44 (1.26,1.62)
3+	2.29(1.77,2.95)	1.95(1.51,2.53)	2.28 (2.03,2.53)
Grade			
Well differentiated	Reference	Reference	Reference
Moderately differentiated	1.13 (0.92, 1.40)	1.12(0.91,1.39)	1.11 (0.90,1.32)
Poorly differentiated	1.56 (1.28,1.98)	1.36(1.07,1.73)	1.45 (1.20,1.70)
Diffuse	1.16(0.54,2.50)	1.03(0.46,2.31)	1.05 (0.28,1.82)
Unknown	1.57 (1.19,2.06)	1.58(1.19,2.09)	1.52 (1.24, 1.80)
Tumor size			
≤2 cm	Reference	Reference	Reference
(2,5] cm	1.43(1.20, 1.71)	1.33(1.11, 1.59)	1.43 (1.25,1.61)
>5 cm	1.87(1.33,2.63)	1.51(1.06, 2.15)	1.91(1.57, 2.25)
Diffuse or Unknown	4.25(2.46,7.34)	3.12(1.76, 5.46)	3.74(3.20,4.28)
HER2			
Positive	reference	-	-
Negative	1.21(0.89,1.64)	-	-
Unknown/not done	0.97(0.71,1.34)	-	-
Adjuvant Online Node Categories			
0	Reference	Reference	Reference
1-3	1.24 (1.02, 1.51)	1.17(0.97, 1.41)	1.23 (1.03,1.43)
4-9	1.55(1.14, 2.10)	1.28(0.96, 1.72)	1.51 (1.21,1.81)
10+	2.29 (1.53,3.44)	1.59(1.05, 2.42)	2.35 (1.95,2.75)
None examined/unknown	2.29(0.83,6.32)	2.86(0.92, 8.94)	2.52 (1.51,3.53)
Radiation	0.82 (0.68,0.98)	0.82 (0.69,0.97)	0.82 (0.64,0.99)
Chemotherapy	0.80(0.64,0.99)	0.77 (0.55,0.99)	
Surgery			
None	1.90 (1.33,2.73)	-	1.95 (1.59,2.31
Lumpectomy	Reference	-	Reference
Mastectomy	1.05(0.88,1.25)	-	1.05 (0.88,1.22)

Table 6.8: Models for Cardiovascular Mortality, 696 events

	CSH Model HR (95%CI)	SH Model HR (95%CI)	MS Model HR (95%CI)
Unknown	3.05(0.74,12.6)	-	3.35 (1.93,4.77)
ER/PR			
Positive	-	-	0.82 (0.59,1.05)
Negative	-	-	Reference
Unknown/not done	-	-	0.67 (0.41,0.93)

Table 6.9: Models for Cardiovascular Mortality, complete data (N=11,152, 267 events)

	CSH Model HR (95%CI)	SH Model HR (95%CI)	MS Model HR (95%CI)
Age Categories			
21-39	Reference	Reference	Reference
40-49	1.25(0.28,5.64)	1.21(0.26,5.50)	1.27 (0.24, 2.78)
50-54	2.00(0.46,8.80)	1.99(0.45,8.80)	2.06 (0.58,3.54)
55-59	2.26(0.52,9.77)	2.17(0.50,9.33)	2.33 (0.87,3.79)
60-64	2.88(0.67,12.3)	2.65(0.62,11.3)	2.93 (1.47,4.39)
65-69	4.10(0.97,17.3)	3.93(0.93,16.6)	4.23 (2.79, 5.67)
70-74	7.41(1.77,31.0)	6.79(1.64,28.1)	7.61 (6.18, 9.04)
75+	15.45(3.76,63.6)	12.7(3.14,51.7)	16.1 (14.7, 17.6)
Race			
White	1.59(0.90,2.81)	1.56(0.89,2.74)	1.51 (0.96, 2.06)
Black	3.53(1.84,6.76)	3.38(1.75,6.51)	3.35 (2.72, 3.98)
Asian/Pacific Islander	Reference	Reference	Reference
Other/Unknown	2.88(0.37,22.5)	3.73(0.43,32.2)	n/a
Smoking status			
Current	2.71(1.76,4.17)	2.60(1.67,4.03)	2.71 (2.28, 3.14)
Former	1.38(0.90,2.11)	1.39(0.90,2.14)	1.39 (0.97, 1.81)
Non-smoker	Reference	Reference	Reference
Unknown	2.73(1.96,3.79)	2.80(2.02,3.88)	2.69 (2.36,3.02)
History of CVD	2.33(1.74,3.11)	2.22(1.65,2.98)	2.31 (2.02,2.60)
HDL Cholesterol			
<35	1.77(1.06,2.97)	1.86(1.11,3.12)	1.82 (1.30,2.34)
35-44	1.17(0.85,1.62)	1.20(0.87,1.68)	1.19 (0.87,1.51)
45-49	0.89(0.60,1.32)	0.96(0.64,1.42)	0.89 (0.49,1.29)
50-59	0.70(0.50,0.98)	0.72(0.51,1.00)	0.70 (0.36,1.04)
60+	Reference	Reference	Reference
Charlson comorbidity			
0	Reference	Reference	Reference
1-2	1.45(1.09,1.91)	1.44(1.08,1.90)	1.45 (1.17, 1.73)
3+	2.35(1.55,3.57)	1.85(1.20,2.84)	2.36 (1.94,2.78)
Grade			
Well differentiated	Reference		Reference
Moderately differentiated	1.15 (0.87, 1.52)		1.21 (0.89, 1.53)
Poorly differentiated	1.39 (1.01,1.89)		1.60 (1.25,1.95)
Diffuse	1.44 (0.62,3.32)		1.94 (1.01,2.87)
Tumor size			
≤2 cm	Reference	Reference	Reference
(2,5] cm	1.76(1.34,2.32)	1.73(1.32,2.27)	1.76 (1.49,2.03)
>5 cm	1.73(0.97,3.09)	1.66(0.90,3.04)	1.68 (1.10, 2.26)
HER2			
Positive	Reference		
Negative	1.56(1.04,2.37)		
Adjuvant Online Node Categories			
0	Reference	Reference	Reference
1-3	1.04(0.76,1.42)	1.03(0.75,1.41)	1.02 (0.71, 1.33)
4-9	1.83(1.20,2.79)	1.78(1.16,2.73)	1.77 (1.35, 2.19)
10+	2.85 (1.55,5.25)	2.07(1.08,3.96)	2.72 (2.11,3.33)

In evaluating risk factors for breast cancer (BC) death, models on the full data set and the subset with only complete risk factor data were examined and compared, with results in Tables 6.10 and 6.11. Once again, aside from the reduction in sample size leading to a loss of statistical power, the two models were quite similar. The majority of the risk factors found in the model on the full cohort, remained in the complete case model, and had very similar point estimates.

In both analyses, age greater than 70 was significantly associated with an increased risk of BC death, as well as smoking (both former and current smokers). Only the CSH models found an effect for race, with both the black and white groups having slightly higher risks than the Asian population. The cardiovascular risk factor of HDL less than 40 was also associated with an increased risk, relative to the reference category of HDL greater than 60. This was unexpected and may be either a spurious result or perhaps it is a marker for healthier lifestyle that could not be captured under the current limited set of covariates.

The known risk factors for breast cancer recurrence, namely grade, tumor size, and lymph node involvement, were all also significantly associated with risk of death from disease. ER/PR positive disease was also at a decreased risk relative to receptor negative cases. Type of surgery was significantly associated with risk of breast cancer death. Compared to the group who received a lumpectomy, the groups receiving both a mastectomy and no surgery, both were at an increased risk of death. However, it is likely that the mastectomy group presented with higher staged disease, whereas the group who did not undergo surgery could have done so due to comorbid conditions or older age. Lastly, in the model on the full cohort, use of radiation and chemotherapy were also associated with a decrease in the hazard of breast cancer death. However, as treatments received were not randomly assigned as in a clinical trial, there is likely confounding between patient age, comorbidity and use of chemotherapy. In fact, regardless of stage at diagnosis, the youngest group of women had the highest rates of chemotherapy, and a trend of decreasing chemotherapy use is seen with increasing age, across all stages of disease.

Table 6.10: Models for Breast Cancer Mortality, 842 events)

	CSH Model HR (95%CI)	SH Model HR (95%CI)	MS Model HR (95%CI)
Age Categories			
21-39	Reference	Reference	Reference
40-49	1.08(0.78,1.49)	1.10 (0.79, 1.54)	1.09 (0.76,1.42)
50-54	0.97(0.68,1.38)	1.00 (0.70,1.43)	0.97 (0.62,1.32)
55-59	1.08 (0.76,1.52)	1.11(0.79,1.58)	1.11 (0.77,1.45)
60-64	1.18(0.82,1.70)	1.34(0.94,1.90)	1.30 (0.96,1.64)
65-69	1.19(0.83,1.71)	1.14 (0.79,1.65)	1.19 (0.83,1.55)
70-74	1.63(1.13,2.33)	1.50(1.03,2.19)	1.65 (1.29,2.01)
75+	2.45(1.77,3.38)	2.07(1.48,2.90)	2.47 (2.15,2.79)
Race			
White	1.31(1.03,1.66)		1.30 (1.06,1.54)
Black	1.66(1.21,2.28)		1.64 (1.33,1.95)
Asian/Pacific Islander	Reference		Reference
Other/Unknown	1.73(0.54,5.50)		1.65 (0.49,2.81)
Smoking status			
Current	1.97(1.54,2.52)	1.96 (1.53,2.51)	1.96 (1.71,2.21)
Former	1.32(1.03,1.70)	1.38(1.07,1.78)	1.33 (1.08,1.58)
Non-smoker	Reference	Reference	Reference
Unknown	2.33 (1.94,2.80)	2.22(1.85,2.65)	2.29 (2.11,2.47)
HDL Cholesterol			
<35	1.91 (1.38,2.65)	1.72 (1.20,2.45)	1.92 (1.59, 2.25)
35-44	1.27 (1.03,1.57)	1.24 (0.99,1.53)	1.28 (1.07, 1.49)
45-49	0.88 (0.68,1.14)	0.89 (0.69,1.14)	0.90 (0.65,1.15)
50-59	0.99(0.82,1.20)	0.98(0.81, 1.19)	0.99 (0.80,1.18)
60+	Reference	Reference	Reference
Unknown	1.59 (1.29,1.96)	1.39(1.11,1.73)	1.57 (1.36,1.78)
Grade			
Well differentiated	Reference	Reference	Reference
Moderately differentiated	2.87(2.03,4.06)	2.90(2.05, 4.09)	2.86 (2.51,3.21)
Poorly differentiated	4.34 (3.05,6.17)	4.19(2.92, 6.01)	4.34 (3.99,4.69)
Diffuse	4.59 (2.63, 8.02)	4.86(2.74,8.62)	4.59 (4.03, 5.15)
Unknown	2.65 (1.77,3.97)	2.77 (1.87,4.13)	2.63 (2.23,3.03)
Tumor size			
≤2 cm	Reference	Reference	Reference
(2,5] cm	2.13(1.81,2.52)	2.10(1.76,2.50)	2.13 (1.96,2.30)
>5 cm	3.20(2.51,4.08)	2.87(2.21, 3.73)	3.19 (2.95,3.43)
Diffuse or Unknown	5.49(3.90,7.73)	4.26(2.88, 6.31)	5.26 (4.92,5.60)
Adjuvant Online Node Categories			
0	Reference	Reference	Reference
1-3	1.71 (1.44,2.04)	1.69 (1.41,2.02)	1.71 (1.5,4 1.88)
4-9	3.07(2.49,3.79)	3.00(2.41,3.73)	3.04 (2.83,3.25)
10+	6.49(5.17,8.15)	5.80(4.51,7.46)	6.48 (6.25,6.71)
None examined/unknown	2.13(0.52,8.70)	1.81 (0.41,8.02)	2.13 (0.72,3.54)
ER/PR			
Positive	0.48 (0.40, .57)	0.52 (0.43,0.62)	0.48 (0.31,0.65)
Negative	Reference	Reference	Reference
Unknown/not done	0.54 (0.45,0.66)	0.56(0.46,0.69)	0.53 (0.34,0.72)
Surgery			
None	3.20(2.38,4.30)	2.78 (2.05,3.77)	3.25 (2.95,3.55)
Lumpectomy	Reference	Reference	Reference
Mastectomy	1.32(1.12,1.55)	1.29 (1.10,1.51)	1.33 (1.17, 1.49)
Unknown	1.94 (0.27,14.2)	2.17 (0.56,8.37)	1.96 (0.03,3.95)

Table 6.11: Models for Breast Cancer Mortality, complete cases (N=11,783, 469 events)

	CSH Model HR (95%CI)	SH Model HR (95%CI)	MS Model HR (95%CI)
Age Categories			
21-39	Reference	Reference	Reference
40-49	1.06(0.67,1.66)	1.07(0.67,1.71)	1.06(0.61,1.51)
50-54	0.96(0.60,1.55)	0.96(0.59,1.56)	0.96(0.48,1.44)
55-59	0.99(0.62,1.60)	0.95(0.58,1.56)	0.99 (0.51,1.47)
60-64	1.33(0.83,2.12)	1.29(0.79,2.10)	1.33 (0.86,1.80)
65-69	1.15(0.71,1.88)	1.08(0.65,1.79)	1.15 (0.66,1.64)
70-74	2.05(1.27,3.31)	1.86(1.14,3.05)	2.05 (1.57,2.53)
75+	3.15(2.02,4.90)	2.58(1.63,4.08)	3.15 (2.71,3.59)
Smoking status			
Current	1.83(1.30,2.57)	1.77(1.25,2.50)	1.82 (1.48,2.16)
Former	1.47(1.05,2.07)	1.48(1.05,2.10)	1.47 (1.13, 1.81)
Non-smoker	Reference	Reference	Reference
Unknown	2.12(1.65,2.71)	2.11(1.65,2.70)	2.12 (1.87, 2.37)
HDL Cholesterol			
<35	1.65(1.07,2.54)	1.52(0.98,2.36)	1.65 (1.22, 2.08)
35-44	1.50(1.18,1.92)	1.49(1.16,1.92)	1.50 (1.25,1.75)
45-49	0.89(0.66,1.21)	0.89(0.65,1.22)	0.89 (0.58,1.20)
50-59	0.95(0.75,1.20)	0.96(0.75,1.21)	0.95 (0.71,1.19)
60+	Reference	Reference	Reference
Grade			
Well differentiated	Reference	Reference	Reference
Moderately differentiated	2.54(1.69,3.83)	2.53(1.68,3.81)	2.54(2.13,2.95)
Poorly differentiated	3.35 (2.19,5.13)	3.24(2.08,5.05)	3.35(2.92,3.78)
Diffuse	4.12(2.15,7.90)	4.42(2.26,8.64)	4.12 (3.4, 4.77)
Tumor size			
≤2 cm	Reference	Reference	Reference
(2,5] cm	2.07(1.67,2.55)	1.98(1.58, 2.48)	2.07 (1.86, 2.28)
>5 cm	2.75(1.96, 3.87)	2.52(1.75,3.65)	2.75 (2.41,3.09)
Adjuvant Online Node Categories			
0	Reference	Reference	Reference
1-3	2.02(1.60,2.53)	1.99(1.57,2.51)	2.02 (1.79,2.25)
4-9	3.67(2.78,4.83)	3.58(2.69,4.78)	3.66 (3.38,3.94)
10+	7.67 (5.58,10.5)	7.04(5.01,9.88)	7.67 (7.35,7.99)
ER/PR			
Positive	0.41(0.33,0.51)	0.44(0.35,0.55)	0.41(0.20,0.62)
Negative	Reference	Reference	Reference
Surgery			
None	2.12(1.24,3.62)	1.86(1.07,3.25)	2.12 (1.58,2.66)
Lumpectomy	Reference	Reference	Reference
Mastectomy	1.24(1.01,1.52)	1.23(1.00,4.50)	1.24 (1.04,1.44)

Reweighting Approach for CSH model

To examine the effect of missingness, we used the methods of Xu et al. [Xu et al., 2009, Xu et al., 2011], by applying weights to the subset of the cohort that had complete information on smoking status, and re-running the cause specific hazards models. This method has not been adapted or evaluated for use in a SH or MS model. The weighting is performed in two steps. The first step weights each complete subject by the inverse of the observation probability, $\pi(w_i)$, to correct for the selection bias introduced by running a complete case

analysis. $\pi(w_i)$ can be estimated from a logistic regression model on R_i, the indicator for whether the covariate is observed or missing. The second step is to impose a time-variant selection probability on the "pseudo-unbiased" samples from the first step (to improve estimation efficiency), using the empirical estimator of the marginal observation probability given a risk set at time t;

$$\hat{\pi^*}(t) = \frac{\sum_{i=1}^{n} R_i Y_i(t)}{\sum_{i=1}^{n} Y_i(t)}$$

This can be done provided that the data are entered in counting process style, with one row per each patient for each observed failure time interval, $(t_a, t_b]$. The final weight then becomes: $\frac{\hat{\pi^*}(t)}{\pi(w_i)}$, and can be implemented in standard software using a weight statement in a Cox model.

Results using this reweighting approach for the endpoint of CVD mortality appear in Table 6.12. Parameter estimates for the relevant covariates are not appreciably different than the results on the full cohort using an unweighted approach. There is, however, a slight reduction in power due to the smaller sample size. If the reweighted estimates are instead compared to the unweighted estimates of the complete case analysis, shown in Table 6.9, the results are also quite similar, with overlapping confidence intervals between the two models for the majority of the covariates.

Re-weighted results for breast cancer mortality also yielded similar results to the cause specific hazards model on the full cohort (not shown), thus the model created using the full cohort was selected for evaluation of predictive accuracy.

Table 6.12: Reweighting Approach for CSH model of CVD (N=12,914)

	CSH Model (696 events) HR (95%CI)	**RWE Model (320 events) HR (95%CI)**
Age Categories		
21-39	Reference	Reference
40-49	0.81 (0.41, 1.60)	0.85 (0.32, 2.27)
50-54	0.98 (0.49 ,1.95)	0.77 (0.28, 2.15)
55-59	1.03 (0.52, 2.03)	1.02 (0.38, 2.70)
60-64	1.33 (0.68, 2.55)	1.47 (0.57, 3.79)
65-69	1.87 (0.98, 3.58)	1.75 (0.69, 4.46)
70-74	3.37 (1.78, 6.37)	3.21 (1.28, 8.06)
75+	6.28 (3.36, 11.8)	4.76 (1.92, 11.8)
Race		
White	1.42 (1.02, 1.97)	1.75 (0.99, 3.08)
Black	2.57 (1.76, 3.77)	2.85 (1.51, 5.40)
Asian/Pac Isl	Reference	Reference
Smoking status		
Current	2.70 (2.06, 3.54)	2.03 (1.50, 2.75)
Former	1.38 (1.06, 1.80)	1.19 (0.90, 1.58)
Non-smoker	Reference	Reference
History of CVD		
No prior history	Reference	Reference
History of event	2.09 (1.70, 2.58)	1.93 (1.41, 2.64)
Unknown	1.57 (1.11, 2.20)	1.34 (0.79, 2.28)
HDL Cholesterol		
<35	1.52 (1.02, 2.28)	1.89 (1.07, 3.34)
35-44	1.02 (0.80, 1.31)	1.03 (0.71, 1.49)
45-49	0.84 (0.62, 1.13)	0.87 (0.56, 1.35)
50-59	0.81 (0.64, 1.02)	0.84 (0.59, 1.20)
60+	Reference	Reference
Unknown	1.30 (1.00, 1.68)	1.30 (0.88, 1.90)
Charlson		
0	Reference	Reference
1-2	1.44 (1.20, 1.73)	1.28 (0.97, 1.68)
3+	2.29 (1.77, 2.95)	2.03 (1.41, 2.91)
Grade		
1	Reference	Reference
2	1.13 (0.92, 1.40)	1.18 (0.85, 1.65)
3	1.56 (1.28, 1.98)	1.77 (1.23, 2.55)
4	1.16 (0.54, 2.50)	0.52 (0.11, 2.59)
Unknown	1.57 (1.19, 2.06)	1.76 (1.17, 2.66)
Tumor size		
≤2 cm	Reference	Reference
(2,5] cm	1.43 (1.20, 1.71)	1.71 (1.32, 2.24)
>5 cm	1.87 (1.33, 2.63)	2.55 (1.58, 4.11)
Diffuse/Unk	4.25 (2.46, 7.34)	5.38 (2.35, 12.3)
Node Categories		
0	Reference	Reference
1-3	1.24 (1.02, 1.51)	1.26 (0.94, 1.70)
4-9	1.55 (1.14, 2.10)	1.25 (0.78, 2.00)
10+	2.29 (1.53, 3.44)	1.69 (0.90, 3.18)
Radiation	0.82 (0.68, 0.98)	0.69 (0.52, 0.91)
Chemotherapy	0.80 (0.64, 0.99)	0.56 (0.40, 0.78)
HER2		
Positive	1.21 (0.89, 1.64)	0.65 (0.25, 1.68)
n/a	0.97 (0.71, 1.34)	0.45 (0.32, 0.63)
Surgery		
None	1.90 (1.33, 2.73)	2.43 (1.50, 3.92)
Lumpectomy	Reference	
Mastectomy	1.05(0.88, 1.25)	1.00 (0.77, 1.31)

Models Stratified by Stage and Age

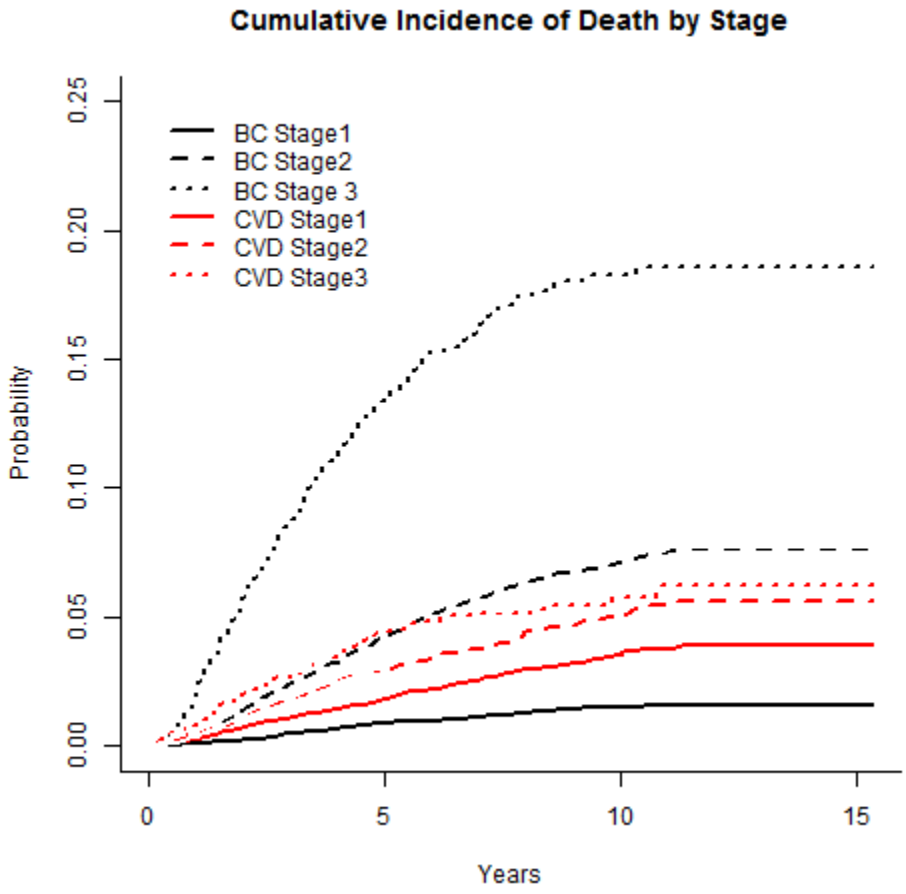
Due to the relationship between age and CVD mortality, and stage of disease and breast cancer mortality, models stratified by stage and age were also examined.

The cumulative incidence curves by stage of breast cancer (Figure 6.5) indicate a clear pattern of higher risk of breast cancer death for higher staged disease. There is a slight trend of increasing risk of CVD death for higher stage cancer as well, though all stages have a less than 5% occurrence of CVD death by 10 years, therefore, the apparent stacking of the curves may not translate into a clinically meaningful difference.

Table 6.13: Outcomes by Stage and Age

	N	**BC death**	**CVD death**	**Other death**	**Censored**
Stage 1	10,843	127 (1.6%)	294(3.9%)	551 (7.5%)	9871 (87.0%)
Stage 2	7,806	444 (7.6%)	316 (5.6%)	497 (8.6%)	6549 (78.2%)
Stage 3	1,813	271 (18.6%)	86 (6.2%)	143 (10.2%)	1313 (65.0%)
Total	**20,462**	**842 (5.3%)**	**696 (4.7%)**	**1191 (8.2%)**	**17733 (81.8%)**
Age<40	873	49 (9.5%)	11 (1.5%)	31 (5.1%)	782 (83.9%)
Age 40-50	4,003	163 (5.6%)	40 (1.3%)	85 (2.8%)	3715 (90.3%)
Age 50-65	7,936	283 (4.7%)	117 (1.9%)	250 (4.2%)	7286 (89.2%)
Age 65+	7,650	347 (5.5%)	528 (9.5%)	825 (14.9%)	5950 (70.1%)

Figure 6.5: Cumulative Incidence of Mortality by Stage



In running models stratified by stage of breast cancer, for the CVD endpoint there are no major differences seen across the 3 stages, and the stratification leads to a great deal of power loss, especially for the stage 3 group that experiences only 86 CVD deaths across the entire follow-up period. Therefore, we refer to the previous, unstratified model for the prediction of CVD mortality.

For the models predicting breast cancer death, stratified by stage, we again see a reductionin statistical power, especitally for the stage 1 patients, who experience only 127 breast cancer deaths over the course of the study. For this stage, continuing to smoke, higher grade of disease, and ER/PR negativity remain the only risk factors associated with an increased risk of breast cancer death. For the stage 2 women, of whom 444 died from their disease, age

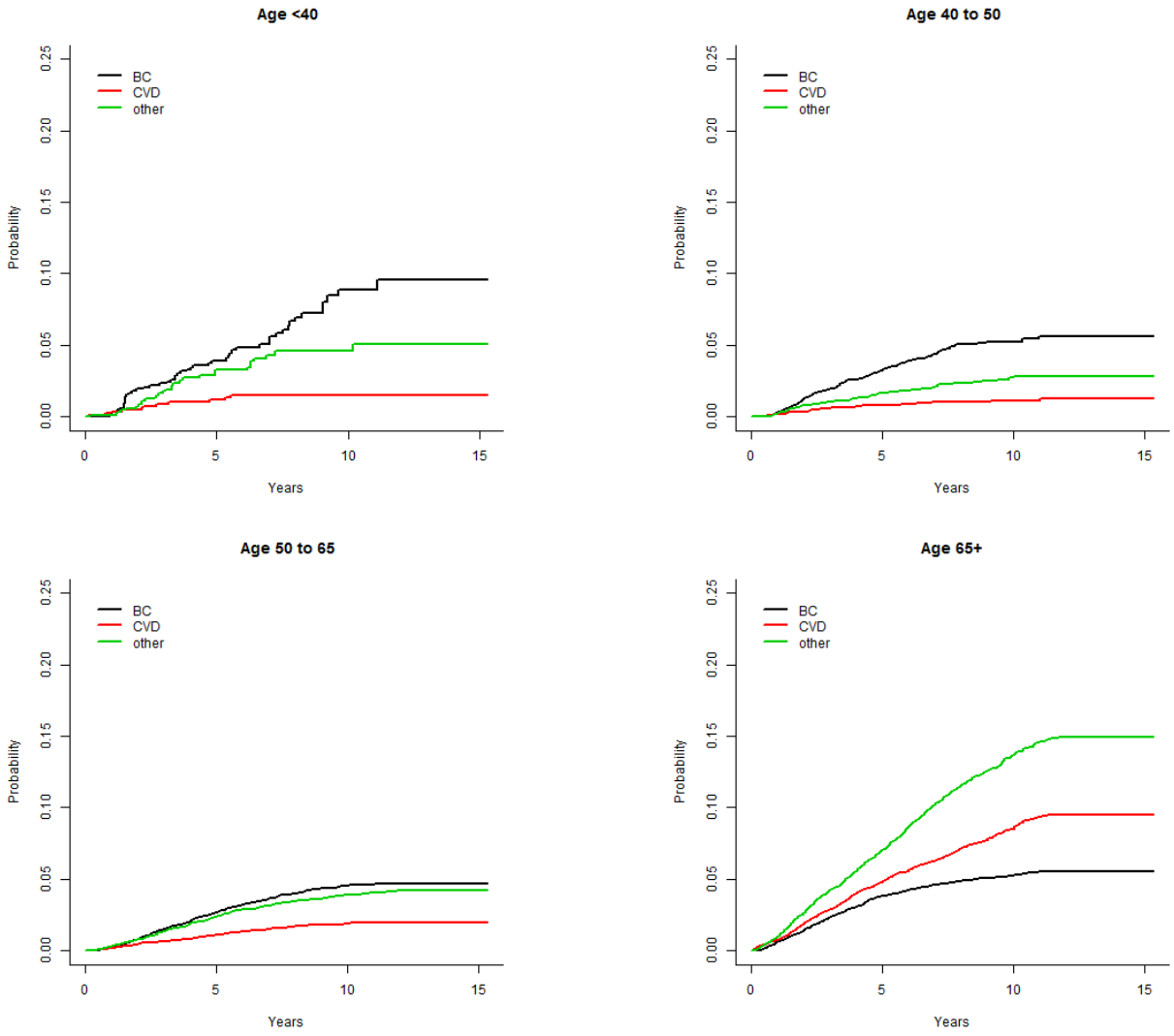


Figure 6.6: Cumulative Incidence Curves by Age

75 or older, smoking, higher grade, larger tumor size, increasing number of positive nodes, having no surgery, and ER negativity all remained significant predictors of an increased risk of breast cancer death. For the 1,813 women with stage 3 disease, age 75 and higher, smoking, poorly differentiated tumor, ER/PR negativity, HER2 negativity and having no surgery, were all associated with an increased risk.

In the cumulative incidence curves stratified by age, we see a clear increase in the cumulative incidence of CVD death for older ages, as well as deaths due to other causes. In fact, for the 75+ group, the lowest death risk seems to be from breast cancer. Contrarily, for the younger ages, the greatest mortality risk is death from breast cancer, and the lowest cunulative incidence of death is due to CVD. While differences in death risk are seen by age, stratified models (age <50 vs age 50+) did not yield major differences in the hazard ratios for the remainder of the risk factors examined, and thus only the unstratified results are presented.

6.3.1 Comparison of Model Estimation and Risk Factor Inclusion

Tables 6.14 and 6.15 list the risk factors that were included in each of the models. As expected, when the analysis was performed on only the complete case data, the sample size and power were reduced, and fewer variables remained statistically significant compared to the models on the full data set. For the variables that remained in both models, the parameter estimates were fairly close between the full data set and the complete case analysis, and the levels of a risk factor that were statistically significant remained consistent from one model to the next.

As was seen in the simulation analysis, as the degree of censoring increases, the parameter estimates from the different model types (CSH, SH, and MS) become nearly identical, especially in the situation where risk factors are only directly related to a single cause of failure. Therefore, for this analysis, any model choice is appropriate for risk factor estimation. However, we can expect to see differences in the individual risk predictions.

Table 6.14: Variables Included in Multivariate Models

	CVD death, all data (N=20,462 : 695 events)			Breast Cancer death, all data (N=20,462 : 842 events)		
	CSH	SH	Mstate	CSH	SH	Mstate
Age	X	X	X	X	X	X
Race	X	X	X	X		X
Smoking	X	X	X	X	X	X
History CVD	X	X	X			
HDL	X	X	X	X	X	X
Charlson	X	X	X			
Grade	X	X	X	X	X	X
Tumor size	X	X	X	X	X	X
Her2Neu	X	X				
Lymph Nodes	X	X	X	X	X	X
Radiation	X	X	X			
Chemo	X		X			
Surgery	X		X	X	X	X
ERPR			X	X	X	X

Table 6.15: Variables Included in Complete Case Multivariate Models

	CVD death, complete data (N=11,152 : 267 events)			Breast Cancer death, complete data (N=11,783 :469 events)		
	CSH	SH	Mstate	CSH	SH	Mstate
Age	X	X	X	X	X	X
Race	X	X	X			
Smoking	X	X	X	X	X	X
History CVD	X	X	X			
HDL	X	X	X	X	X	X
Charlson	X	X	X			
Grade	X		X	X	X	X
Tumor size	X	X	X	X	X	X
Her2Neu	X					
Lymph Nodes	X	X	X	X	X	X
Surgery				X	X	X
ERPR				X	X	X

6.3.2 Model Performance

To further evaluate the choice of statistical model, we examine how the model performs when used for prediction. Due to the low percentage of events and high degree of censoring, no data splitting (into a training and test set) was done in this primary analysis. While the best method of model validation would include examination on an external data set, none was available at current time, and remains a limitation to evaluating the model generalizability. Therefore, the results of model performance should be interpreted with caution, as they will likely overstate the model's capabilities.

Calculating Individual Predicted Values

To assess the models' performances, predicted values were calculated at 3, 5 and 10 years after breast cancer diagnosis.

For the cause specific hazards, the model based prediction can be performed using the method described in [D'Agostino Sr et al., 2008] and others, by evaluating the below formula for the probability of the event by the time-point of interest:

$$\hat{p} = 1 - S_0(t)^{\exp\left[\sum_{i=1}^{p} \beta_i X_i - \sum_{i=1}^{p} \beta_i \bar{X}_i\right]},$$

where $S_{0(t)}$ is the baseline survival at follow-up time t (t = 3, 5 or 10 years), provided in Table 6.16, β_i are the estimated regression coefficients from Tables 6.8 and 6.10), X_i is the individual's value of the risk factor, $\bar{X}_i$ is the corresponding sample mean, and p denotes the number of risk factors.

For the subdistribution hazards model, the probability of an event by a given time, t, depends on the covariates and the baseline cumulative subdistribution hazard and can be calculated just as straightforwardly, using [Wolbers et al., 2009] :

$$\hat{p*} = 1 - \exp\Big(- \exp\Big(\sum_{k=1}^{p} \beta_k X_{ik} \Big) \cdot \int_0^t \bar{\lambda}_{1,0}(s) ds \Big),$$

where $\int_0^t \bar{\lambda}_{1,0}(s)ds$ is the baseline cumulative subdistribution hazard at time t and $\sum_{k=1}^{p} \beta_k X_{ik}$ is the linear predictor for a patient based on their risk factors and the coefficients from the subdistribution hazards models in Tables 6.10 and 6.8. To obtain the baseline cumulative subdistribution hazards, the SAS macro %DACIF [Zhang and Zhang, 2011] was used.

Calculation of the patient-specific transition probabilities from the multi-state model is done in two steps [Putter, 2014]. The first step uses the estimated parameters and baseline transition hazards along with the patient-specific covariate values to obtain the matrix of patient-specific transition hazards, including the hazard of remaining in the initial state at time t. The second step uses the patient specific transition hazards, along with the variance-covariance matrix, to obtain the patient specific transition probabilities using the Aalen-Johansen estimator [Aalen and Johansen, 1978]:

$$\hat{\mathbf{P}}(s,t) = \prod_{s<u\leq t} (\mathbf{I} + d\hat{\mathbf{\Lambda}}(u)),$$

where $\hat{\mathbf{\Lambda}}(t))$ is the matrix of patient specific transition hazards at time t, and for our specific model is as follows:

$$\begin{bmatrix} -\sum_{h\neq g} \Lambda_{gh}(t) & \hat{\alpha_{01}}(t) & \hat{\alpha_{02}}(t) & \hat{\alpha_{03}}(t) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

and $\mathbf{I}$ is the (SxS) identity matrix where S denotes the number of possible states. Calculations can be done using the *mstate* package in R.

Prediction Results

Table 6.17 summarizes the average risk (and standard deviation) for each failure type at 3, 5 and 10 years. Across all thre time points examined, the risk of breast cancer death was higher than the risk of cardiovascular death in this cohort.

Table 6.16: Survival and Cumulative Subdistribution Baseline Hazard estimates for making predictions

$\mathbf{S_{0(t)}}$	**3 years**	**5 years**	**10 years**
Cardiovascular Survival	0.9848	0.9745	0.9541
Breast Cancer Survival	0.9803	0.9661	0.9460
Cumulative Subdistribution Baseline Hazard			
CVD	0.0006056	0.0010126	0.001768
Breast Cancer	0.00355	0.00614	0.00952

Table 6.17: Average Predicted Risks at Selected Timepoints

	3 years	**5 years**	**10 years**
BC deaths (observed)	**1.9%**(0.001)	**3.3%** (0.001)	**5.1%**(0.002)
CSH	4.2% (0.08)	6.9% (0.11)	10.2% (0.14)
SH	3.8% (0.06)	6.3% (0.09)	9.2% (0.12)
MS	1.9% (0.03)	3.3% (0.04)	4.9% (0.05)
CVD deaths (observed)	**1.5%**(0.001)	**2.5%**(0.001)	**4.3%**(0.002)
CSH	3.4% (0.06)	5.5% (0.09)	9.1% (0.14)
SH	0.7% (0.01)	1.2% (0.02)	2.1% (0.03)
MS	1.5% (0.02)	3.2% (0.03)	4.1% (0.04)

Though the CSH model is known to over-estimate the true risk, with the high degree of censoring in this cohort, the two models give very similar results for breast cancer mortality. For the endpoint of CVD death, there is a greater discrepancy in predicted risk between the two models, which increases with increasing time. A possible explanation for this is that a large proportion of the breast cancer deaths occur within 5 years of diagnosis, which can especially be seen in Figure 6.5, the cumulative incidence curves for each cause of failure, whereas the cumulative incidence of CVD death increases at a fairly constant rate over the time period. Thus, the endpoint of CVD death is more affected by the competing cause of BC death, and how the risk set is defined, which varies between the two models, leads to big differences in predicted risk.

Looking in greater detail at the predicted risks stratified by stage, Table 6.18, we can see the wide differences in predicted risk of breast cancer death by stage across all 3 timepoints, with 5 year risk greater than 20% for stage 3 diesase. The corresponding stratification for CVD mortality risk shows little to no difference based on stage, with generally low risk across the entire length of follow-up.

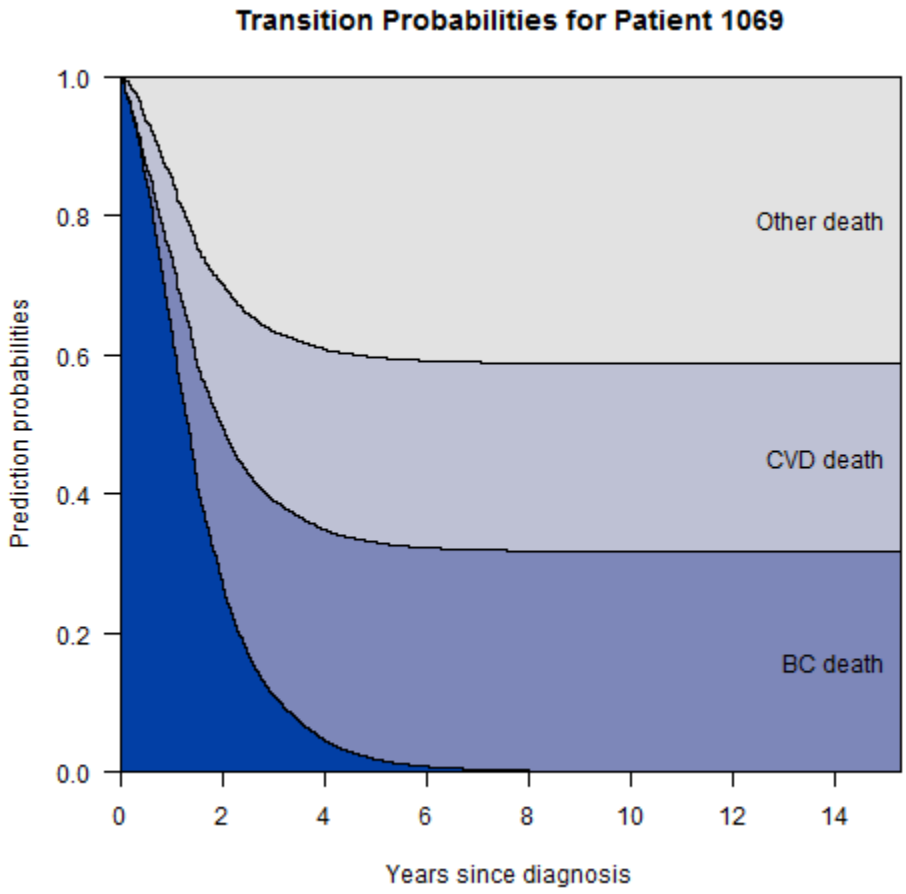
Table 6.18: Average Predicted Risk at Selected Timepoints, by Stage (SH Model)

Predicted BC mortality risk	3 years	5 years	10 years
Stage 1	1.4% (0.012)	2.4% (0.02)	3.6% (0.03)
Stage 2	5.1% (0.06)	8.4%(0.09)	12.5% (0.12)
Stage 3	13.1% (0.12)	21.0% (0.17)	28.6%(0.20)
Predicted CVD mortality risk	3 years	5 years	10 years
Stage 1	0.6% (0.01)	0.9% (0.02)	1.7% (0.03)
Stage 2	0.8% (0.014)	1.4% (0.02)	2.4% (0.04)
Stage 3	1.1% (0.02)	1.8% (0.03)	3.0%(0.05)

Multi-state Model Output

In addition to the predicted risk at selected timepoints, a user-friendly feature of the multi-state model is the graphical representation of a person's risk over time. In the sample figure below we see the patient's risk of transitioning to any of the three states at any point in time.

Figure 6.7: Sample Multi-state Prediction Curve



This patient was a 75 year old, white female with stage 3, grade 3 breast cancer, a history of CVD, Charlson score of 2, highl total cholesterol, poor HDL and smoking missing. Her 5 year predicted risk of breast cancer mortality was 31% from the MS model, 90% from the CSH model, and 74% from the SH model. Her 5 year predicted risks of CVD mortality were 27% from the MS model, 85% from the CSH model, and 23% from the SH model. Her actual outcome was death due to breast cancer at 3.5 years post-diagnosis.

Discrimination and Calibration

To summarize model discriminative ability, we again examine the AUCs at 3, 5 and 10 years. Results are summarized in Table 6.19. While the AUCs indicate good discrimination,

they are inflated due to evaluating the models on the same data set upon which they were created. No internal cross-validation was performed, as the best method of model validation is truly on an external data source [Steyerberg et al., 2010].

Table 6.19: Summary of AUC for selected models and timepoints

	3 years	5 years	10 years
CVD - CSH	0.85 (0.83, 0.87)	0.85 (0.84,0.87)	0.85 (0.83,0.86)
CVD - SH	0.84 (0.82,0.86)	0.85 (0.83,0.86)	0.84 (0.82,0.85)
CVD - MS	0.84 (0.82,0.86)	0.85 (0.83,0.86)	0.82 (0.81,0.84)
BC - CSH	0.87 (0.86,0.89)	0.85 (0.84,0.87)	0.82 (0.80,0.83)
BC - SH	0.87 (0.85,0.88)	0.85 (0.84,0.86)	0.81 (0.80,0.83)
BC - MS	0.82 (0.80,0.84)	0.80 (0.78,0.87)	0.77 (0.75,0.79)

Plots of observed versus predicted appear as both bar charts and traditional calibration plots in the following pages. For cardiovascular mortality, the plots indicate that using CSH models, the predicted risk is slightly higher than the observed event rate, with the exception of the 10 year mark, when a large portion of our data set has been censored. Conversely, the MS and SH models have similar observed and predicted risks for all except the highest decile of risk, for which it largely underestimated the event rate. A possible reason could be due to the low CVD event rate in our population. Regardless of decile of predicted risk, at 3 years, there is a less than 5% event rate for every decile except the highest, which still has a less than 10% event rate. This is still the case at 5 years, with the exception of the 2 highest deciles having observed rates of slightly higher than 5% and 15%, respectively.

Interestingly, the same pattern does not hold for the breast cancer deaths, where the plots indicate only slight over-estimation of risk by the CSH and SH models at both 3 and 5 years, but otherwise good calibration. The MS model, however, shows excellent calibration at 3 and 5 years. At 10 years, the CSH and SH models slightly underestimate the event rate, while the MS model has a greater degree of under-estimation. However, this time-point is

subject to a high degree of overall censoring and reduced sample size, thus evaluation at 3 and 5 years remains more meaningful in this cohort.

6.4 Summary

In summary, while the previously established models of CVD risk have adequate discrimination in our cohort, the calibration is poor. This is likely due to a number of reasons, including the high proportion of competing events seen in our data set, but not accounted for under the traditional survival methods used to create the original models. It could also be due to the fact that our data set does not have 10 years of follow-up for a large percentage of the cohort, and removing those with less follow-up and those failing from competing causes, heavily biases the sample used for model evaluation. However, even for the shorter prediction horizons, such as 2 - 5 years, poor calibration is still seen.

In our analysis of CVD death using competing risk methodology, we discovered that many of the traditional CVD risk factors were not as prognostic in this data set. This could be influenced by the nontrivial amounts of missing data, such as for smoking status, or timeframe for which risk factors such as cholesterol measurements were collected. The additional risk of breast cancer mortality in this population led to the finding that some of the breast cancer risk factors, such as lymph nodes, grade and tumor size, were also associated with an increased risk of CVD death, as well as patient characteristics including Charlson comorbidity, smoking, and history of a CVD event.

Lastly, it is plausible that in this data set, the CVD event rate happens to be much lower than the population upon which the original models were created. We found that in this cohort, the occurrence of CVD death is higher than 5% only for those aged 70 or more, and is an extremely rare event for the remainder of the cohort. This could be due to the differences in distribution of risk factors between the two populations, as well as the high risk of competing causes of death in our study.

While we were not able to directly evaluate the established breast cancer recurrence risk models because recurrence was not routinely collected in our data set, we found that the risk factors included in the traditional models such as tumor size, lymph node involvement and grade, remained highly predictive of breast cancer mortality in this patient population.

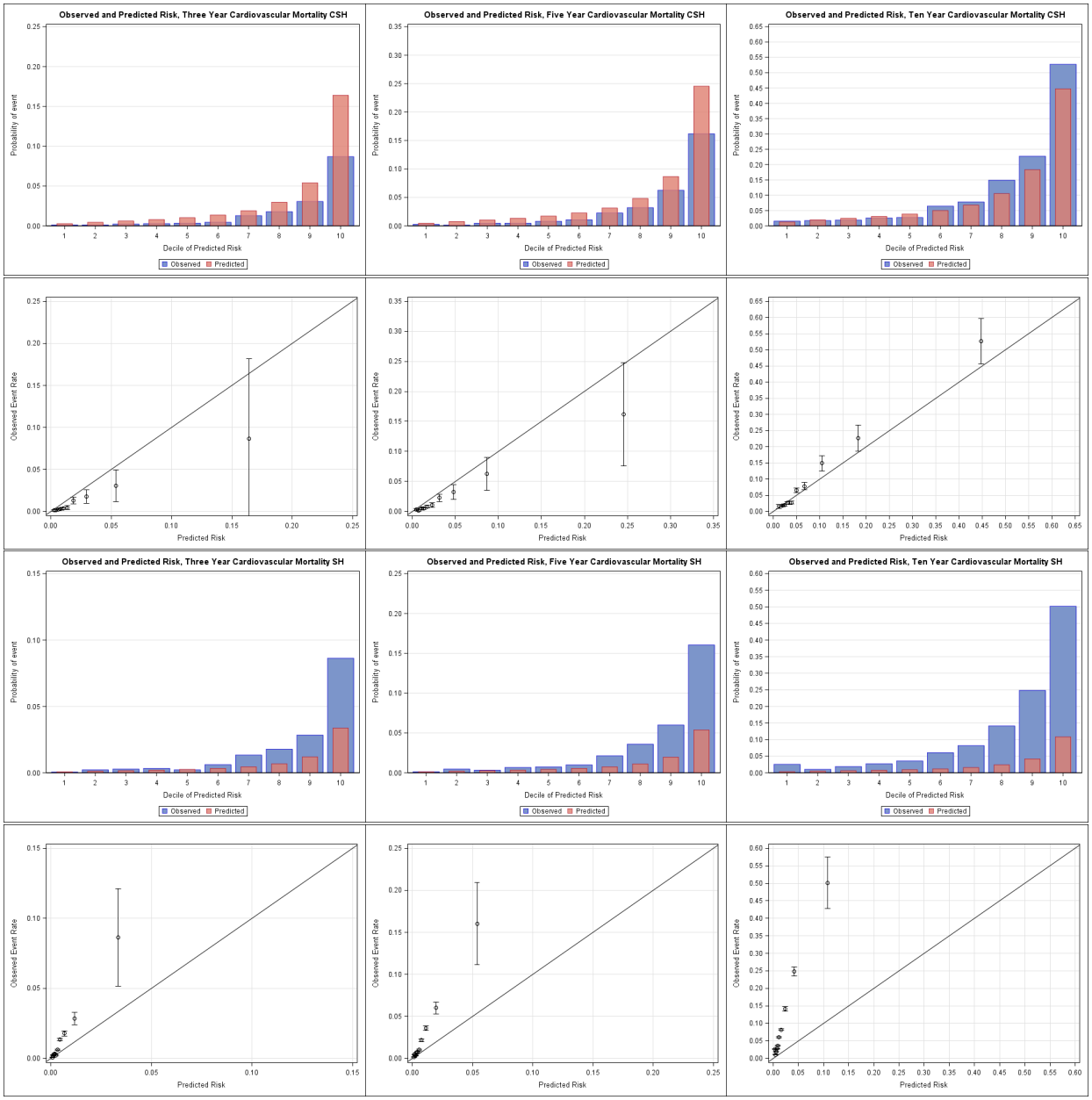


Figure 6.8: Calibration Plots for CVD Endpoint

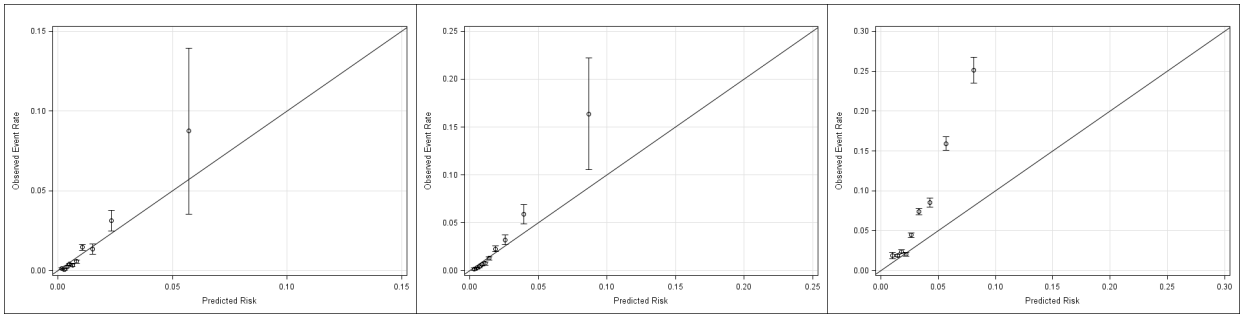


Figure 6.9: Calibration Plots for Multi-State Models: CVD Mortality

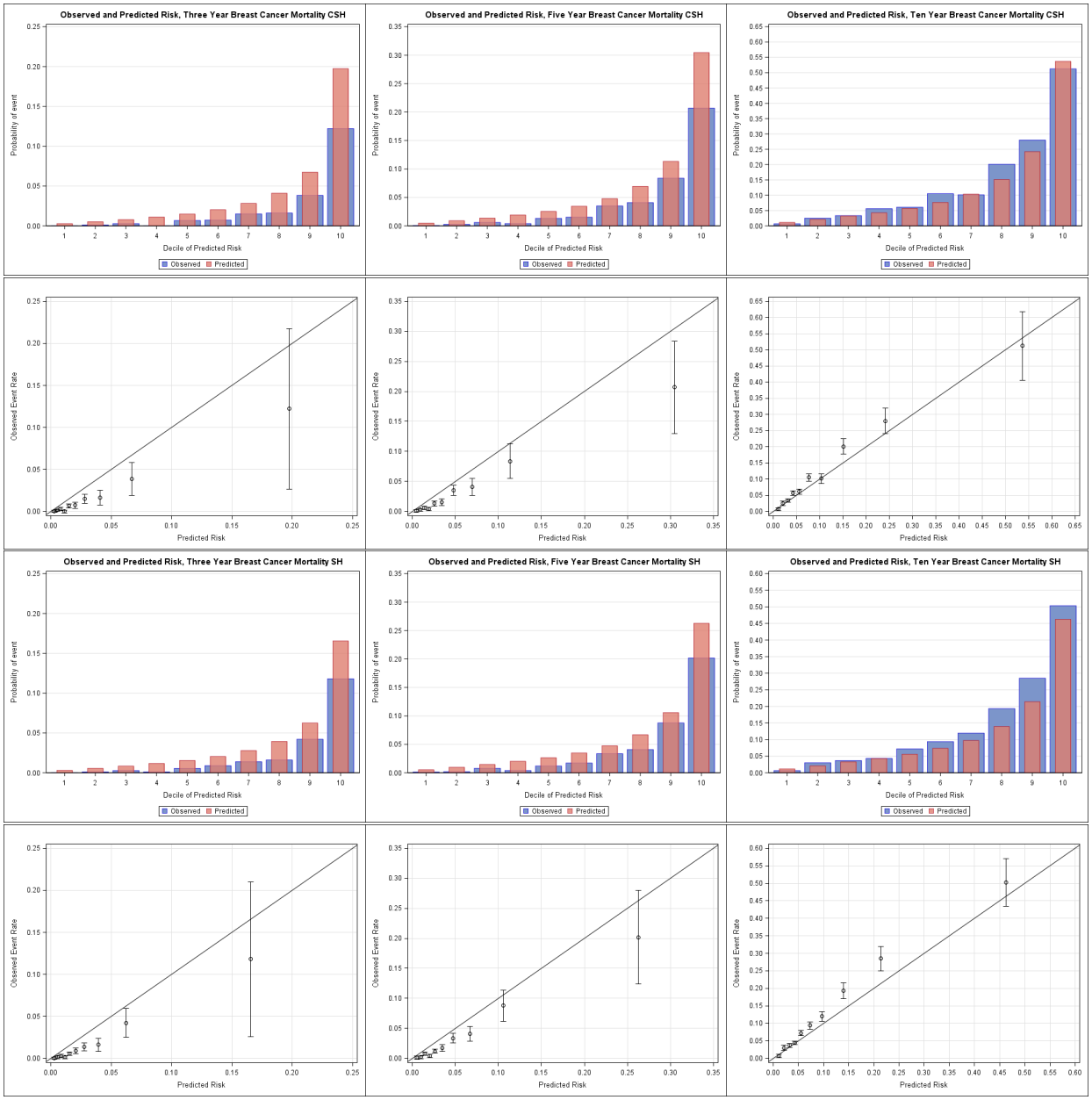


Figure 6.10: Calibration Plots for BC Endpoint

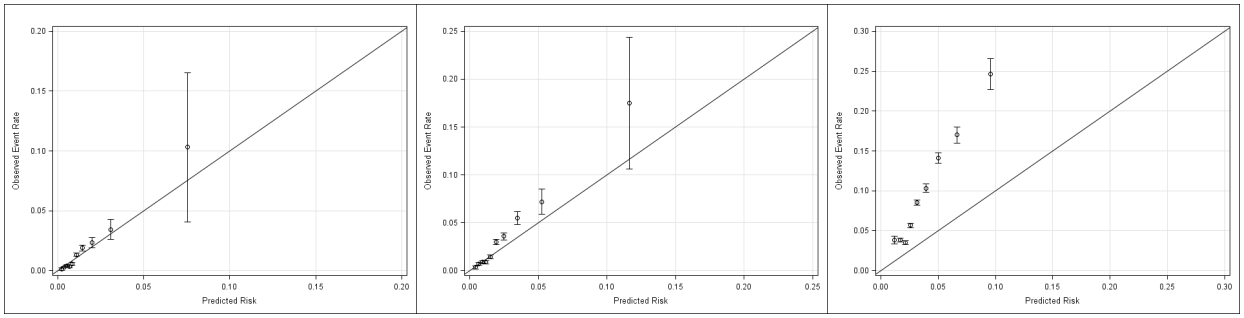


Figure 6.11: Calibration Plots for Multi-State Models: Breast Cancer Mortality

Additional covariates of ERPR status, surgery, and smoking were also significantly associated with breast cancer mortality. Stage, while not included in the models due to its potential co-linear effect with tumor size and nodes, was seen to be highly predictive in stratified cumulative incidence curves, though a specific stratified analysis did not reveal additional findings beyond the unstratified models.

The AUCs for the models were very high, possibly due to the fact that they were calculated on the same data set from which the models were built. However, despite the good discrimination, some were not well calibrated and seemed to over-estimate a woman's risk at 3 and 5 years. Perhaps setting a risk/treatment threshoold such as >5% at 3 years, or >10% risk at 5 years, and evaluating the models under decision analytic techniques would further promote the model's clinical utility, beyond traditional measures of discrimination and calibration.

Regardless, our findings indicate that following a breast cancer diagnosis, a woman's risk of breast cancer mortality is far greater than her risk of CVD mortality, and the primary focus remains breast cancer mitigation and treatment. Specific subgroups, such as those greater than 65 years old, or those with a prior history of CVD, are at the greatest risk for a CVD event. However, use of the model for women outside those distinct subgroups may help identify additional patients with specific combinations of risk factors that would establish them as high risk for a CVD event or mortality, for which beneficial interventions may be available.

Lastly, we have demonstrated the utility of a multi-state model when there is interest in predicting for more than one competing event. The multi-state model demonstrated the best calibration of the three models at 3 and 5 years, and the individual plot of patient risk over time provides a user-friendly graphical summary that can be easily understood by clinicians and patients.

Chapter 7

Concluding Remarks

To the best of our knowledge, this is the first study to validate any established CVD risk models in a cancer population, or to model the simultaneous risk of CVD versus breast cancer death via competing risk methodology. There are likely a number of reasons why it has not been previously persued. The first could be lack of an awareness of a public health burden. For many cancers, no curative therapy exists, and the primary focus for the patient and oncologist is on managing the cancer; other causes of death are trivial. However, for certain early stage cancers where there is potential for cure, such as in breast cancer, once a patient has entered the remission phase, they may become as at-risk for other causes of death as the general population. We have explored some of the major causes of death for a cohort of early stage breast cancer patients, but there are likely other cancers, such as early stage colorectal and low grade prostate, for which competing causes of death could equal or perhaps surpass the burden of cancer mortality, and patient awareness is beneficial.

Another difficulty arises in obtaining a large enough sample size, with all relevant covariates. Cancer overall is not a rare disease, but when studying a specific cancer diagnosis, each site must be thought of as its own disease and disease process. In focusing on a single site, such as breast cancer, with less than 300,000 cases diagnosed in the U.S. each year, it is a relatively rare disease. Thus any analysis of a cancer site would likely require a coop-

erative group arrangement in order to obtain a meaningful sample size, which can become very costly and logistically complex. In addition, many of the cooperative groups may focus solely on the specific cancer diagnosis and treatment, and thus may not collect data on risk factors for competing causes of failure. One solution to this is to examine data from clinical trials, where more extensive data are collected on side effects, toxicities, and patient vitals. However, once the trial is closed or patients have reached the primary trial endpoint, follow-up is not continued and there becomes a high degree of administrative censoring.

A further challenge may be due to the current healthcare delivery system. A large proportion of cancer patients likely receive their oncologic care from an onocology specialist or specialty institution that may focus exclusively on cancer care. This could lead to difficulties in data collection and availability. For instance, the current CVD risk models require data on serum levels such as cholesterol, which may not be monitored or stored electronically in the oncology setting. Oncologists not associated with the same institution as the patient's primary care physician will also not have access to detailed patient histories including prior non-cancer events or patient medications. Similarly, primary care physicians may not have detailed information on cancer characteristics for their patients who have had a cancer diagnosis, but receive their care elsewhere. An integrated healthcare delivery system with linked electronic medical records, would allow access to more comprehensive patient data for research questions such as these that may bridge more than one general area of patient health. Another solution could be to develop models that contain alternative risk factors, ones that would be readily available, such as BMI or questionnaire data on lifestyle habits and history of health issues, which is a potential area of future research.

Despite these limitations, we have reported on an extremely large cohort of breast cancer patients from Kaiser Permanente in Northern California. Our analyses indicate that while the previously published CVD risk models have moderate discrimination in this cohort, the calibration is poor. Some of the risk factors that have been traditionally associated with CVD risk, are not as prognostic in this population, and there could be several reasons for

this. While our cohort size was large, we chose to focus on cardiovascular mortality rather than intermediate, non-fatal events, and thus had a low number of events leading to reduced statistical power. We also elected to include all of the risk factors as categorical covariates, which increases the number of variables in the model.

Our cohort was also very different from the healthy cohorts upon which the models were developed. The current cohort was on average, much older, with a range of comorbid conditions, including their cancer diagnosis. We also did not exclude patients with a prior history of CVD, in order to present a more "real-world" setting, though it is likely that patients with a history of CVD are already well aware of their mortality risks from heart disease, despite their new cancer diagnosis. In fact, those with a history of CVD were at an increased risk for death from CVD but not from breast cancer. Our data set was also affected by missing information, especially for smoking and cholesterol. Rather than exclude patients with missing information, we elected to broaden the time frame for which cholesterol data was obtained, but it is possible that data not collected within a short time of the breast cancer diagnosis, may not be a good indication of patient health surrounding the time of their primary breast cancer treatment. However, an analysis using only complete cases revealed similar results to the analysis that included missing information as its own "missing" category, and only those with very low HDL cholesterol remained at an increased risk.

Another reason for the poor calibration could be due to the methodology upon which the original models were developed. None of the previously published CVD models used methods to account for competing risk of death, as it was a trivial issue for their young, healthy cohorts. Therefore, we would expect these models to over-estimate the risk of CVD death in our patient population, and this is reflected in several of the calibration plots, especially those predicting for 5 years or less. However, some of the calibration plots indicate the opposite effect, that the model is understimating the observed event rate. This is likely due to the discrepancy in time frame for which these models predicts (generally 10 years) and the fact

that we do not have 10 years of follow-up for a large proportion of our cohort. In excluding patients who are still alive but have not yet been followed for 10 years, we are reducing our overall sample size, while keeping the same number of failures (as most failures are observed before 10 years) and thus increasing the observed event rate in our sample. As the cumulative incidence of both BC and CVD death continues to increase between years 5 and 10 post-diagnosis, it would be interesting to repeat the analysis once 10 years of follow-up have been achieved for the majority of survivors. Alternatively, perhaps predicting risk at 10 years is not useful for a woman undergoing treatment for her primary breast cancer diagnosis, and understanding her risk of competing events, such as CVD, in the next 5 years would be more beneficial, especially while weighing treatment decisions. The majority of the established CVD models chose 10 years as their timepoint for predicition, likely due to the low number of observed events. For example, in the 1998 Framingham paper [Wilson et al., 1998],the sample of 2,856 women, yielded only 227 events, and no information was provided on the timing of these events. However, it is possible that for a breast cancer patient, her greatest risk may be in the next 5 years following her diagnosis, and established models would be more useful if calibrated to predict in that time horizon.

It also may be possible that the cancer disease process modifies the body in such a way that traditional CVD risk factors are no longer as prognostic. This is especially echoed in the multivariate models presented for CVD death, where a number of the breast cancer characteristics are significantly associated with increased risk of CVD death, but the traditional CVD risk factors are not. Unfortunately, we did not have specific treatment information on type and duration of chemotherapy and hormonal therapies, and were unable to determine whether this could be treatment related. There was also confounding with age, stage of diagnosis and aggressiveness of treatment. For example, younger women were more likely to have stage 3 disease, and regardless of stage at diagnosis, were also much more likely to receive chemotherapy. Older women (greater than 75) were less likely to receive any surgery or chemotherapy, but also remain the most at risk for death from other causes. However, an

analysis stratified by age did not reveal any major differences in the covariate effects of the remaining risk factors.

Another possibility is that we are missing other relevant covariates. Our data was limited to what is historically collected for cancer registries, and traditional risk factors included in the Framingham models. In doing so, we are missing information on exposures such as alcohol, diet, BMI, reproductive factors, or possible genetic information related to breast cancer or CVD, including family history and known genetic mutations. Lastly, our sample had a high proportion of other causes of mortality that were not directly analyzed as an endpoint, but rather accounted for under a general "death from other" category, the largest subgroup of which was "other cancer". Perhaps an additional transition from diagnosis to death from other cancer, or including it with breast cancer deaths, would be worthwhile in presenting a more complete picture of risks following a breast cancer diagnosis, especially if it is possible that the "other" cancers included a number of breast cancers that had metastasized.

Nevertheless, this is the first study to evaluate CVD risk models in a cancer population, as well as develop a prognostic model that evaluates risk of cancer death and CVD death. While the number of CVD deaths was small, it was similar to the number of BC deaths observed during the follow-up period, indicating that the risk for CVD death remains as high in breast cancer survivors as the risk for BC death. Based on the current data set, if clinicians elect to use the Framingham models, it is recommended to choose the model that predicts in a 2 or 4 year horizon. If longer prediction is of interest, or simultaneous estimates of BC and CVD mortality risk, our models showed good performance, especially the user-friendly multi-state model, though their generalizability to an outside patient population still remains to be assessed.

In terms of the methodology utilized in our analysis, we found no major differences in the results between the three model types. This is likely due to the low number of events relative to our sample size, and the high degree of censoring. Perhaps a better model choice

for this population would have even been a transformation model with a cure fraction as detailed by Zeng et al. [Zeng et al., 2006], however, it has not yet been extended to include competing events, and software availability remains limited.

One further area of methodological debate is the choice of the time scale for the model. While the greatest predictor of mortality is clearly patient age, we chose not to use age as the time-scale, but rather include it as a covariate in the model, as well as a stratification variable in sensitivity analyses. The debate over choice of time scale is ongoing, with no general consensus over which should be used[Korn et al., 1997, Cheung et al., 2003], as the two methods have been shown to produce differing results on the same set of data. Additionally, we elected to use age at breast cancer diagnosis as the starting point, but this is a left-truncated entry time; by the time a woman has been diagnosed, the cancer has already been present for an unknown amount of time. Currently it is not possible to analyze left truncated data using a subdistribution hazards model, and thus was not pursued in the current work. However, due to the different effects of age on each competing endpoint in the analysis, it remains an interesting area of future research, especially under the multi-state modeling approach.

In simulation studies of available methods for analyzing competing risk data, only the CSH and SH models have been evaluated. The major findings of these studies have been that the covariate effect on the CSH for a specific cause may be different from its effect on the cumulative incidence of the cause, and that if one of the model choices holds, the other will be misspecified. This means that if the proportionality assumption of the CSH model is met, then the corresponding SH model may be violated. However, it has become widely accepted that even when the model is mis-specified, it may still be useful if the covariate effect is interpreted as a time-average effect [Latouche et al., 2013].

Our simulation did not aim to address the behavior of a covariate that is known to be associated with more than one cause of failure, as that has been reported previously [Dignam and Kocherginsky, 2008, Allignol et al., 2011]. We rather turned to the goal of

prediction and deciding which methodl to use when creating a disease prognostic model. Based on the previously published risk models for CVD and breast cancer recurrence, we see no overlap in covariates other than age, and possibly smoking. Therefore, our simulations focus on the accuracy of covariate effect under each model specification, when each endpoint has its own independent set of risk factors. Regardless of the method of data generation or degree of censoring, we found that the CSH and MS models give accurate parameter estimates for both the binary and standard normal covariates that were generated. Given the excellent performance of the CSH model in simulations, one could infer that the SH model would possibly be mis-specified, and this is apparent given the poor estimation for the covariate effect. However, as the censoring percentage increases, we see the SH model results asymptote to the CSH and MS model results. Even when the parameter estimates are identical, the one difference that remains between the two models is in the inclusion of covariates. Since the SH model models not the specific event of interest, but the cumulative incidence of events, covariates that are associated with competing causes of failure will have a non-trivial hazard ratio for the subdistribution hazard. This was shown both in the simulation results and in the real data analysis. Therefore, while the estimates for the primary risk factors will be similar for the three models when there is a high degree of censoring, the SH model will likely include covariates that are strongly related to all causes of failure, while a parsimonious CSH model should only include covariates that act upon the failure type being modeled. The multi-state model should include covariates related to all failure types or transitions.

While it has been described in the literature that the model choice depends on the question of interest, and CSH models should be used when interest lies in the effect of a covariate on a specific cause of failure, whereas SH models should be reserved for absolute risk prediction, we have shown through simulation and real data analysis, that when the degree of censoring is high, as is often the case in a retrospective study with short follow-up, the models give nearly identical results. However, even though we evaluated our models on

the same data set that was used to create them, there is room for improvement in model calibration. Perhaps neither one of our models is the optimum choice, and one that can be adapted to change after a cure threshold has been reached, or even Bayesian methods incorporating a prior distribution based on population data, would be better suited for this type of competing risk analysis, and are potential areas of future exploration. Additionally, new evaluation methods for the simultaneous prediction of multiple endpoints, have yet to be developed and are another area of future research.

Appendices

Appendix A: Cardiovascular Disease Risk Models Worksheets

1. Framingham 1991 worksheet
2. Framingham 1998 worksheet
3. DAgostino2008worksheet
4. SCORE worksheet

TABLE 5. Framingham Heart Study Coronary Heart Disease Risk Prediction Chart

1. Find points for each risk factor

Age (if female) (yr)				Age (if male) (yr)				HDL cholesterol			
Age	Points	Age	Points	Age	Points	Age	Points	HDL	Points	HDL	Points
30	−12	41	1	30	−2	48–49	9	25–26	7	67–73	−4
31	−11	42–43	2	31	−1	50–51	10	27–29	6	74–80	−5
32	−9	44	3	32–33	0	52–54	11	30–32	5	81–87	−6
33	−8	45–46	4	34	1	55–56	12	33–35	4	88–96	−7
34	−6	47–48	5	35–36	2	57–59	13	36–38	3		
35	−5	49–50	6	37–38	3	60–61	14	39–42	2		
36	−4	51–52	7	39	4	62–64	15	43–46	1		
37	−3	53–55	8	40–41	5	65–67	16	47–50	0		
38	−2	56–60	9	42–43	6	68–70	17	51–55	−1		
39	−1	61–67	10	44–45	7	71–73	18	56–60	−2		
40	0	68–74	11	46–47	8	74	19	61–66	−3		

Total cholesterol (mg/dl)				Systolic blood pressure (mm Hg)					Points	
Chol	Points	Chol	Points	SBP	Points	SBP	Points	Other factors	Yes	No
139–151	−3	220–239	2	98–104	−2	150–160	4	Cigarette smoking	4	0
152–166	−2	240–262	3	105–112	−1	161–172	5	Diabetes		
167–182	−1	263–288	4	113–120	0	173–185	6	Male	3	0
183–199	0	289–315	5	121–129	1			Female	6	0
200–219	1	316–330	6	130–139	2			ECG-LVH	9	0
				140–149	3					

2. Add points for all risk factors

____	+	____	+	____	+	____	+	____	+	____	+	____	=	____
(Age)		(Total chol)		(HDL)		(SBP)		(Smoking)		(Diabetes)		(ECG-LVH)		(Total)

Note: *Minus points subtract from total.*

3. Look up risk corresponding to point total

	Probability (%)			Probability (%)			Probability (%)			Probability (%)	
Points	5 yr	10 yr	Points	5 yr	10 yr	Points	5 yr	10 yr	Points	5 yr	10 yr
≤1	<1	<2	9	2	5	17	6	13	25	14	27
2	1	2	10	2	6	18	7	14	26	16	29
3	1	2	11	3	6	19	8	16	27	17	31
4	1	2	12	3	7	20	8	18	28	19	33
5	1	3	13	3	8	21	9	19	29	20	36
6	1	3	14	4	9	22	11	21	30	22	38
7	1	4	15	5	10	23	12	23	31	24	40
8	2	4	16	5	12	24	13	25	32	25	42

4. Compare with average 10-year risk

Age	Probability (%)		Age	Probability (%)		Age	Probability (%)	
(yr)	Women	Men	(yr)	Women	Men	(yr)	Women	Men
30–34	<1	3	45–49	5	10	60–64	13	21
35–39	<1	5	50–54	8	14	65–69	9	30
40–44	2	6	55–59	12	16	70–74	12	24

HDL, high density lipoprotein; SBP, systolic blood pressure; ECG-LVH, left ventricular hypertrophy by electrocardiography.

Step 1

Age		
Years	LDL Pts	Chol Pts
30-34	-9	[-9]
35-39	-4	[-4]
40-44	0	[0]
45-49	3	[3]
50-54	6	[6]
55-59	7	[7]
60-64	8	[8]
65-69	8	[8]
70-74	8	[8]

Step 2

LDL - C		
(mg/dl)	(mmol/L)	LDL Pts
<100	<2.59	-2
100-129	2.60-3.36	0
130-159	3.37-4.14	0
160-190	4.15-4.92	2
≥190	≥4.92	2

Cholesterol		
(mg/dl)	(mmol/L)	Chol Pts
<160	<4.14	[-2]
160-199	4.15-5.17	[0]
200-239	5.18-6.21	[1]
240-279	6.22-7.24	[1]
≥280	≥7.25	[3]

Step 3

HDL - C			
(mg/dl)	(mmol/L)	LDL Pts	Chol Pts
<35	<0.90	5	[5]
35-44	0.91-1.16	2	[2]
45-49	1.17-1.29	1	[1]
50-59	1.30-1.55	0	[0]
≥60	≥1.56	-2	[-3]

Step 4

Blood Pressure					
Systolic (mm Hg)	Diastolic (mm Hg) <80	80-84	85-89	90-99	≥100
<120	-3 [-3] pts				
120-129		0 [0] pts			
130-139			0 [0] pts		
140-159				2 [2] pts	
≥160					3 [3] pts

+ Note: When systolic and diastolic pressures provide different estimates for point scores, use the higher number

Step 5

Diabetes		
	LDL Pts	Chol Pts
No	0	[0]
Yes	4	[4]

Step 6

Smoker		
	LDL Pts	Chol Pts
No	0	[0]
Yes	2	[2]

(sum from steps 1-6)

Step 7

Adding up the points	
Age	_____
LDL-C or Chol	_____
HDL - C	
Blood Pressure	_____
Diabetes	_____
Smoker	_____
Point total	_____

Key	
Color	Relative Risk
green	Very low
white	Low
yellow	Moderate
rose	High
red	Very high

(determine CHD risk from point total)

Step 8

CHD Risk			
LDL Pts Total	10 Yr CHD Risk	Chol Pts Total	10 Yr CHD Risk
≤-2	1%	[≤-2]	[1%]
-1	2%	[-1]	[2%]
0	2%	[0]	[2%]
1	2%	[1]	[2%]
2	3%	[2]	[3%]
3	3%	[3]	[3%]
4	4%	[4]	[4%]
5	5%	[5]	[4%]
6	6%	[6]	[5%]
7	7%	[7]	[6%]
8	8%	[8]	[7%]
9	9%	[9]	[8%]
10	11%	[10]	[10%]
11	13%	[11]	[11%]
12	15%	[12]	[13%]
13	17%	[13]	[15%]
14	20%	[14]	[18%]
15	24%	[15]	[20%]
16	27%	[16]	[24%]
≥17	≥32%	[≥17]	[≥27%]

(compare to average person your age)

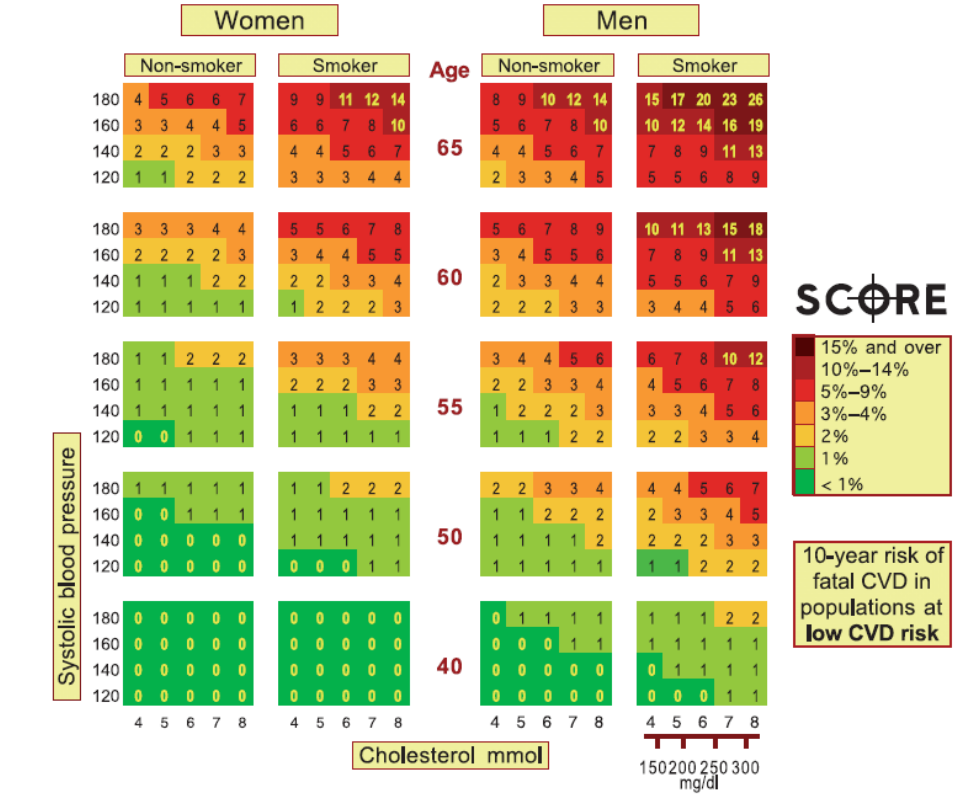
Step 9

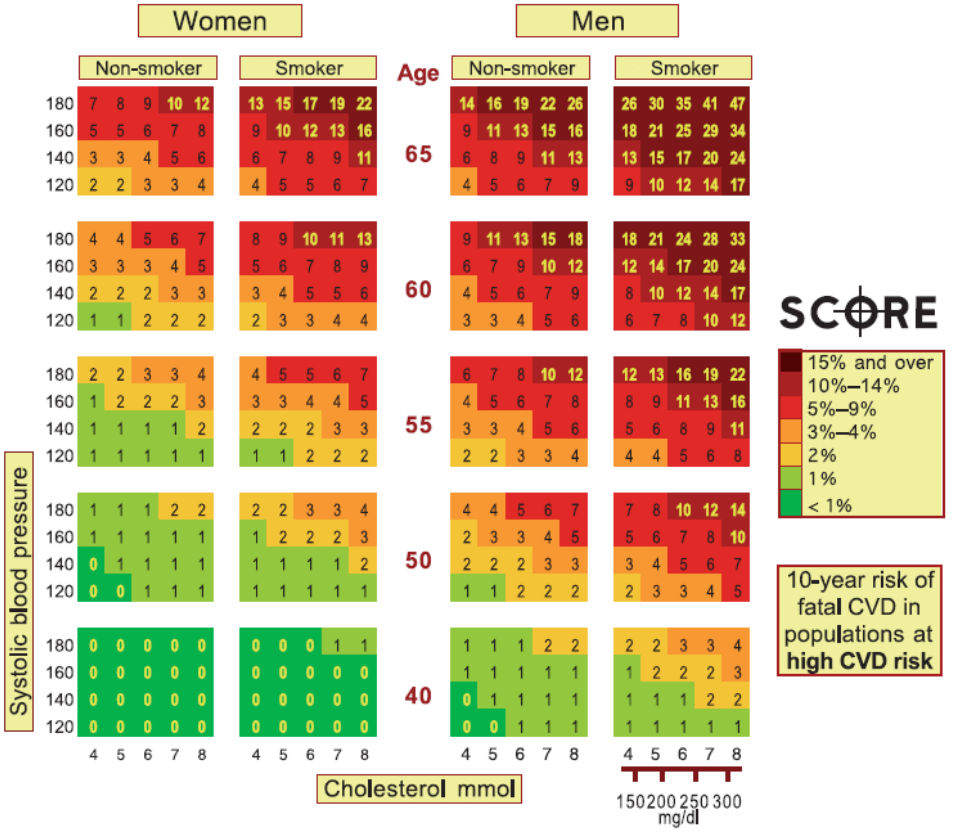
Comparative Risk			
Age (years)	Average 10 Yr CHD Risk	Average 10 Yr Hard* CHD Risk	Low** 10 Yr CHD Risk
30-34	<1%	<1%	<1%
35-39	<1%	<1%	1%
40-44	2%	1%	2%
45-49	5%	2%	3%
50-54	8%	3%	5%
55-59	12%	7%	7%
60-64	12%	8%	8%
65-69	13%	8%	8%
70-74	14%	11%	8%

* Hard CHD events exclude angina pectoris

** Low risk was calculated for a person the same age, optimal blood pressure, LDL-C 100-129 mg/dL or cholesterol 160-199 mg/dl, HDL-C 45 mg/dL for men or 55 mg/dL for women, non-smoker, no diabetes

Risk estimates were derived from the experience of the Framingham Heart Study, a predominantly Caucasian population in Massachusetts, USA


Women
Men
Non-smoker
Smoker
Age
65
60
55
50
40
Systolic blood pressure
Cholesterol mmol
150 200 250 300
mg/dl
SCORE
15% and over
10%–14%
5%–9%
3%–4%
2%
1%
< 1%
10-year risk of fatal CVD in populations at low CVD risk


Women
Men
Non-smoker
Smoker
Age
65
60
55
50
40
Systolic blood pressure
Cholesterol mmol
150 200 250 300
mg/dl
SCORE
15% and over
10%–14%
5%–9%
3%–4%
2%
1%
< 1%
10-year risk of fatal CVD in populations at high CVD risk

Appendix B: Breast Cancer Recurrence Models

1. Nottingham Prognostic Index
2. Kattan Nomogram
3. Adjuvant!Online
4. CancerMath
5. Predict
6. Oxford Model

TABLE I.—*The coding for the various prognostic factors*

Prognostic factor	Codes used in Cox analysis
Age	In years
Menopausal state	0 = premenopausal 1 = postmenopausal
Size	In cm
Lymph-node stage	1 = A 2 = B 3 = C
Tumour grade	1 = I 2 = II 3 = III
Cellular reaction	1 = marked 2 = moderate 3 = slight 4 = none
Sinus histiocytosis	0 = nodes completely replaced by tumour 1 = absent 2 = present
Oestrogen receptor (RE)	0 = negative 1 = positive
Adjuvant therapy	0 = none 1 = therapy given

$$\mathrm{I} = 0{\cdot}2 \times \mathrm{size} + \mathrm{stage} + \mathrm{grade}$$

Index value
High ($> 4{\cdot}4$)
Medium ($2{\cdot}8$–$4{\cdot}4$)
Low ($< 2{\cdot}8$)

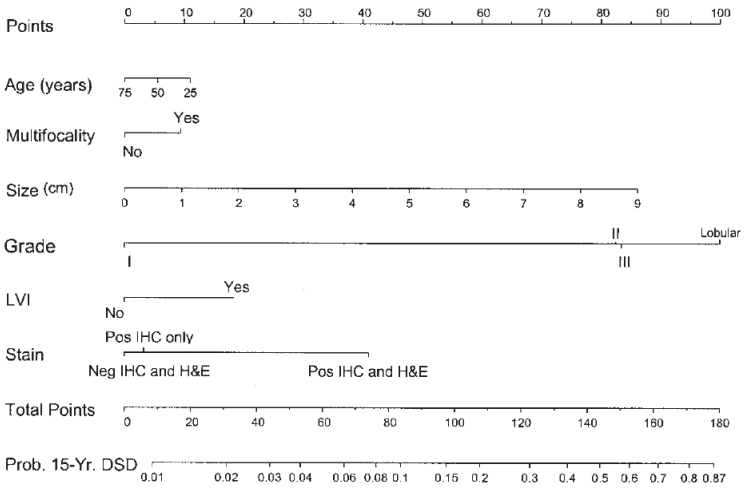
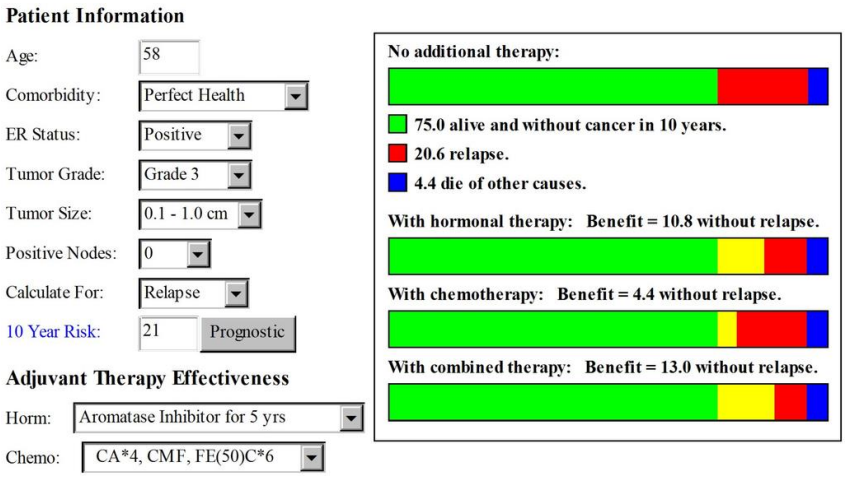
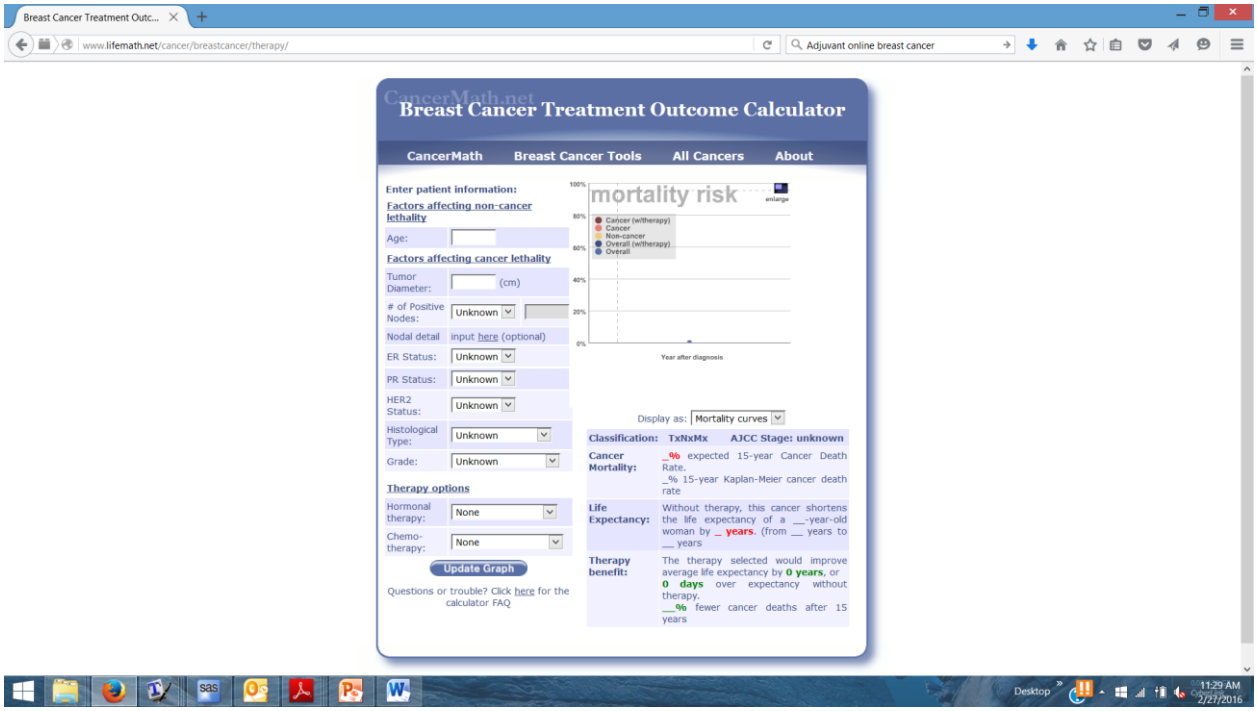
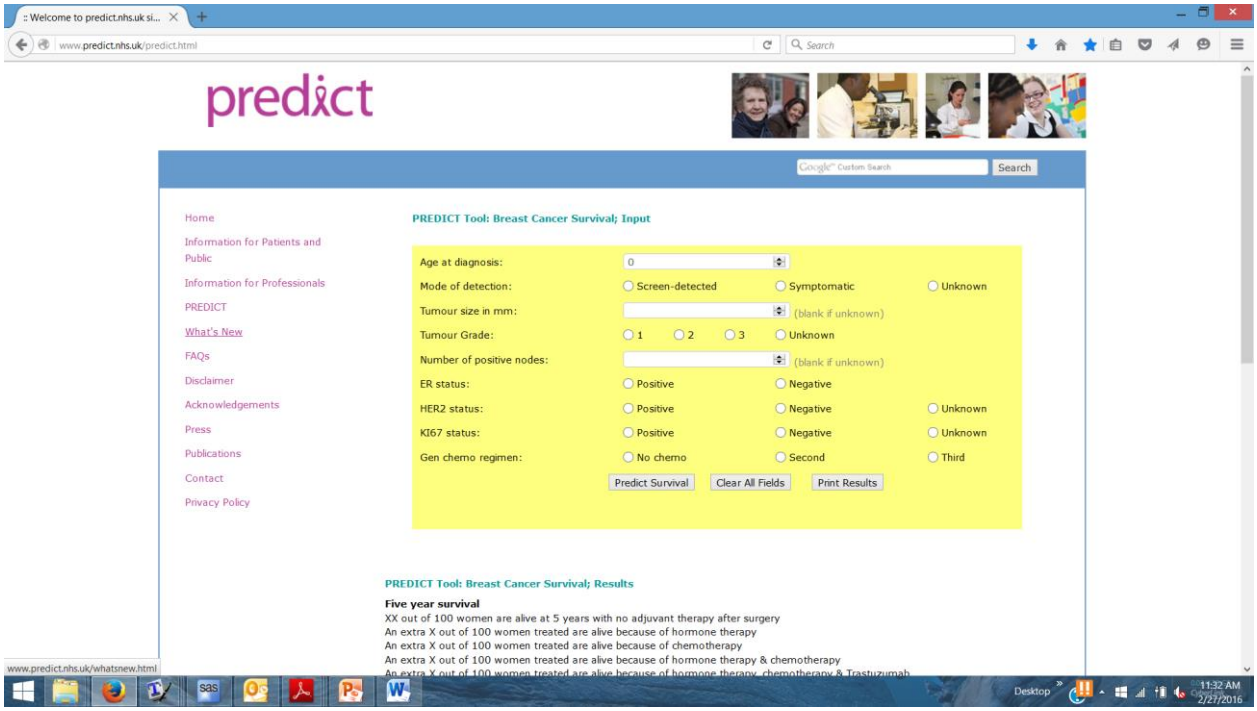


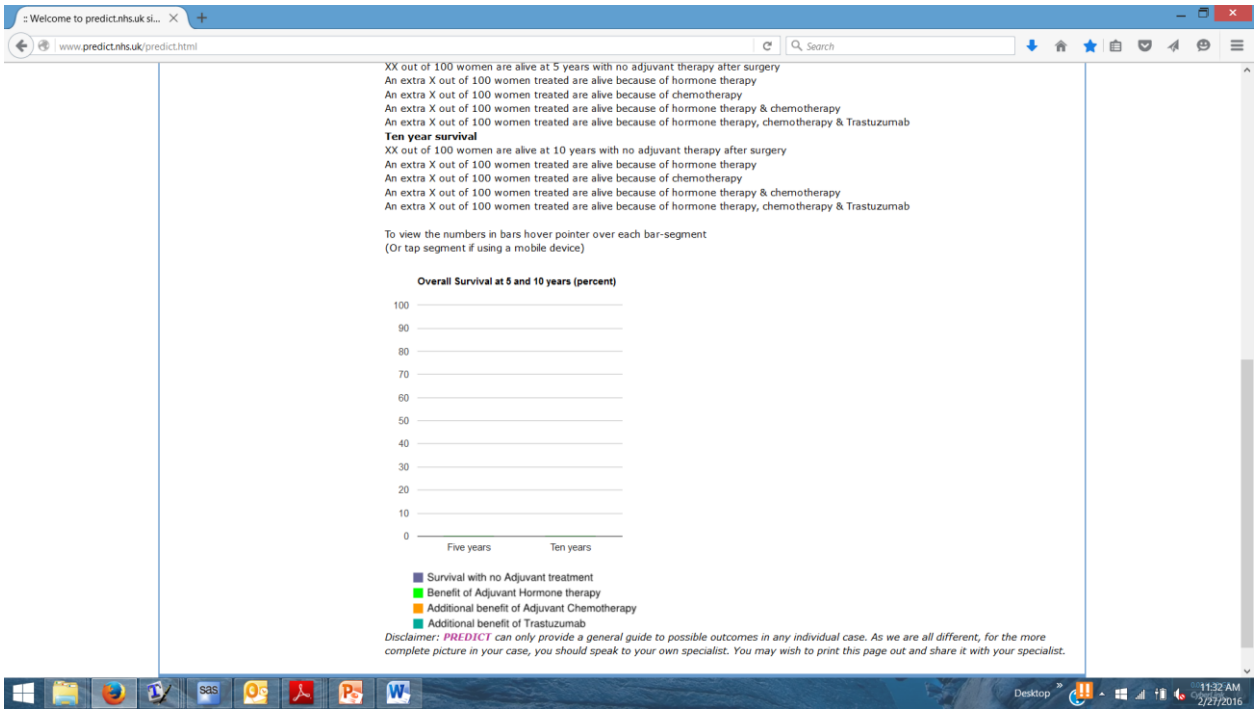
FIGURE 4. Nomogram for predicting 15-year breast carcinoma–specific mortality. LVI: lymphovascular invasion; Pos: positive; Neg: negative; IHC: immunohistochemistry; H&E: hematoxylin and eosin; Prob.: probability; DSD: disease-specific death.

Adjuvant! Online sample screenshot




Breast Cancer Treatment Outc...
www.lifemath.net/cancer/breastcancer/therapy/
Adjuvant online breast cancer
CancerMath.net
Breast Cancer Treatment Outcome Calculator
CancerMath
Breast Cancer Tools
All Cancers
About
Enter patient information:
Factors affecting non-cancer lethality
Age:
Factors affecting cancer lethality
Tumor Diameter:
(cm)
# of Positive Nodes:
Unknown
Nodal detail
input here (optional)
ER Status:
Unknown
PR Status:
Unknown
HER2 Status:
Unknown
Histological Type:
Unknown
Grade:
Unknown
Therapy options
Hormonal therapy:
None
Chemo-therapy:
None
Update Graph
Questions or trouble? Click here for the calculator FAQ
mortality risk
enlarge
Cancer (w/therapy)
Cancer
Non-cancer
Overall (w/therapy)
Overall
100%
80%
60%
40%
20%
0%
Year after diagnosis
Display as:
Mortality curves
Classification:
TxNxMx
AJCC Stage: unknown
Cancer Mortality:
_% expected 15-year Cancer Death Rate.
_% 15-year Kaplan-Meier cancer death rate
Life Expectancy:
Without therapy, this cancer shortens the life expectancy of a __-year-old woman by _ years. (from __ years to __ years
Therapy benefit:
The therapy selected would improve average life expectancy by 0 years, or 0 days over expectancy without therapy.
__% fewer cancer deaths after 15 years
Desktop
11:29 AM
2/27/2016


:: Welcome to predict.nhs.uk si...
www.predict.nhs.uk/predict.html
predict
Search
Home
Information for Patients and Public
Information for Professionals
PREDICT
What's New
FAQs
Disclaimer
Acknowledgements
Press
Publications
Contact
Privacy Policy
PREDICT Tool: Breast Cancer Survival; Input
Age at diagnosis:
Mode of detection:
Screen-detected
Symptomatic
Unknown
Tumour size in mm:
(blank if unknown)
Tumour Grade:
1
2
3
Unknown
Number of positive nodes:
(blank if unknown)
ER status:
Positive
Negative
HER2 status:
Positive
Negative
Unknown
KI67 status:
Positive
Negative
Unknown
Gen chemo regimen:
No chemo
Second
Third
Predict Survival
Clear All Fields
Print Results
PREDICT Tool: Breast Cancer Survival; Results
Five year survival
XX out of 100 women are alive at 5 years with no adjuvant therapy after surgery
An extra X out of 100 women treated are alive because of hormone therapy
An extra X out of 100 women treated are alive because of chemotherapy
An extra X out of 100 women treated are alive because of hormone therapy & chemotherapy
www.predict.nhs.uk/whatsnew.html
Desktop
11:32 AM
2/27/2016


:: Welcome to predict.nhs.uk si...
www.predict.nhs.uk/predict.html
XX out of 100 women are alive at 5 years with no adjuvant therapy after surgery
An extra X out of 100 women treated are alive because of hormone therapy
An extra X out of 100 women treated are alive because of chemotherapy
An extra X out of 100 women treated are alive because of hormone therapy & chemotherapy
An extra X out of 100 women treated are alive because of hormone therapy, chemotherapy & Trastuzumab
Ten year survival
XX out of 100 women are alive at 10 years with no adjuvant therapy after surgery
An extra X out of 100 women treated are alive because of hormone therapy
An extra X out of 100 women treated are alive because of chemotherapy
An extra X out of 100 women treated are alive because of hormone therapy & chemotherapy
An extra X out of 100 women treated are alive because of hormone therapy, chemotherapy & Trastuzumab
To view the numbers in bars hover pointer over each bar-segment
(Or tap segment if using a mobile device)
Overall Survival at 5 and 10 years (percent)
100
90
80
70
60
50
40
30
20
10
0
Five years
Ten years
Survival with no Adjuvant treatment
Benefit of Adjuvant Hormone therapy
Additional benefit of Adjuvant Chemotherapy
Additional benefit of Trastuzumab
Disclaimer: PREDICT can only provide a general guide to possible outcomes in any individual case. As we are all different, for the more complete picture in your case, you should speak to your own specialist. You may wish to print this page out and share it with your specialist.
Desktop
11:32 AM
2/27/2016

Oxford Model

Table 2 Exponentiated coefficients (time ratios) from the final aggregated version of the prognostic model (estimated assuming survival times follow a gamma distribution)

	Coefficient	*P*	95% CI
ln (positive nodes)	0.402	<0.001	0.333–0.485
Tumour size^2	0.898	<0.001	0.854–0.944
Tumour size^2 × ln (tumour size)	1.045	<0.001	1.021–1.070
Tumour grade	0.647	<0.001	0.557–0.751
Age, years	1.015	0.005	1.004–1.026
Ercat	1.209	0.404	0.774–1.888
Adjrt	1.546	0.001	1.192–2.006
Adjhormones	1.230	0.234	0.875–1.730
Adjchemo	0.357	0.047	0.129–0.985
Ercat × adjhormones	1.226	0.481	0.696–2.160
Adjchemo × age	1.023	0.029	1.002–1.044
Ln (positive nodes) × adjchemo	1.418	0.015	1.070–1.878
Ancillary 1[a]	1.698	<0.001	1.572–1.835
Ancillary 2[a]	0.567	<0.001	0.411–0.782

Abbreviations: Adjchemo = adjuvant chemotherapy; Adjhormones = adjuvant hormone therapy; Adjrt = adjuvant radiotherapy; CI = confidence intervals; ER = oestrogen receptor; Ercat = ER status; ln = natural logarithm. [a]Ancillary parameters 1 and 2 determine the shape and scale of the hazard function of the generalized gamma distribution. A literal interpretation of these parameter values is difficult however, on their original scales, they are useful for ruling out models with other functional forms nested within the gamma model. For example, if ancillary 1 = ancillary 2 = 1, survival times follow an exponential distribution, and if ancillary 2 = 0 a log-normal model is appropriate.

Bibliography

[Aalen and Johansen, 1978] Aalen, O. and Johansen, S. (1978). An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5:141150.

[ACS, 2014] ACS (2014). American cancer society facts and figures.

[Adult Treatment Panel (ATP), 2001] Adult Treatment Panel (ATP), I. (2001). Expert panel on detection, evaluation, and treatment of high blood cholesterol in adults. executive summary of the third report of the national cholesterol education program (ncep) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults. *JAMA*, May 16;285(19):2486–2497.

[Agresti, 2002] Agresti, A. (2002). *Categorical Data Analysis Second Edition*. Wiley Interscience, Hoboken, NJ.

[Allignol et al., 2011] Allignol, A., Schumacher, M., Wanner, C., Drechsler, C., and Beyersmann, J. (2011). Understanding competing risks: a simulation point of view. *BMC Med Res Methodol*, Jun 3;(11):86.

[Anderson et al., 1991a] Anderson, K., Odell, P., Wilson, P., and Kannel, W. (1991a). Cardiovascular disease risk profiles. *American Heart Journal*, Jan;121(1 Pt 2):293–8.

[Anderson et al., 1991b] Anderson, K., Wilson, P., Odell, P., and Kannel, W. (1991b). An updated coronary risk profile. a statement for health professionals. *Circulation*, Jan;83(1):356–62.

[Azim Jr et al., 2011] Azim Jr, H., de Azambuja, E., Colozza, M., Bines, J., and Piccart, M. (2011). Long-term toxic effects of adjuvant chemotherapy in breast cancer. *Annals of Oncology*, Sep;22(9):1939–47.

[Bardia et al., 2012] Bardia, A., Arieas, E., Zhang, Z., Defilippis, A., Tarpinian, K., Jeter, S., Nguyen, A., Henry, N., Flockhart, D., Hayes, D., Hayden, J., Storniolo, A., Armstrong, D., Davidson, N., Fetting, J., Ouyang, P., Wolff, A., Blumenthal, R., Ashen, M., and Stearns, V. (2012). Comparison of breast cancer recurrence risk and cardiovascular disease incidence risk among postmenopausal women with breast cancer. *Breast Cancer Res Treat*, Feb;131(3):907–914.

[Bender et al., 2005] Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, Jun 15;24(11):1713–1723.

[Beyersmann et al., 2012] Beyersmann, J., Allignol, A., and Schumacher, M. (2012). *Competing Risks and Multistate Models with R*. Springer, New York, New York.

[Beyersmann et al., 2007] Beyersmann, J., Dettenkofer, M., Bertz, H., and Schumacher, M. (2007). A competing risks analysis of bloodstream infection after stem-cell transplantation using subdistribution hazards and cause-specific hazards. *Statistics in Medicine*, Dec 30;26(30):5360–5369.

[Beyersmann et al., 2009] Beyersmann, J., Latouche, A., Buchholz, A., and Schumacher, M. (2009). Simulating competing risks data in survival analysis. *Statistics in Medicine*, Mar 15;28(6):956–971.

[Beyersmann and Schumacher, 2007] Beyersmann, J. and Schumacher, M. (2007). Misspecified regression model for the subdistribution hazard of a competing risk. *Statistics in Medicine*, Mar 30;26(7):164916–51.

[Blamey et al., 2007a] Blamey, R., Ellis, I., Pinder, S., Lee, A., Macmillan, R., Morgan, D., Robertson, J., Mitchell, M., Ball, G., Haybittle, J., and Elston, C. (2007a). Survival of invasive breast cancer according to the nottingham prognostic index in cases diagnosed in 1990-1999. *European Journal of Cancer*, 43:1548–1555.

[Blamey et al., 2007b] Blamey, R., Pinder, S., Balla, G., Ellis, I., Elston, C., Mitchell, M., and Haybittle, J. (2007b). Reading the prognosis of the individual with breast cancer. *European Journal of Cancer*, 43:1545–1547.

[Brier, 1950] Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

[Campbell et al., 2010] Campbell, H., Gray, A., Harris, A., Briggs, A., and Taylor, M. (2010). Estimation and external validation of a new prognostic model for predicting recurrence-free survival for early breast cancer patients in the uk. *British Journal of Cancer*, 103:776–786.

[Campbell et al., 2009] Campbell, H., Taylor, M., Harris, A., and Gray, A. (2009). An investigation into the performance of the adjuvant! online prognostic programme in early breast cancer for a cohort of patients in the united kingdom. *British Journal of Cancer*, 101:1074–1084.

[Chambless and Diao, 2006] Chambless, L. and Diao, G. (2006). Estimation of time-dependent area under the roc curve for long-term risk prediction. *Statistics in Medicine*, 25:34743486.

[Chen et al., 2002] Chen, K., Jin, Z., and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, 89:659–668.

[Cheng et al., 1995] Cheng, S., Wei, L., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, 82:835–845.

[Cheung et al., 2003] Cheung, Y., Gao, F., and Khoo, K. (2003). Age at diagnosis and the choice of survival analysis methods in cancer epidemiology. *Journal of Clinical Epidemiology*, Jan;56(1):38–43.

[Conroy et al., 2003] Conroy, R., Pyrl, K., Fitzgerald, A., Sans, S., Menotti, A., DeBacker, G., De Bacquer, D., Ducimetire, P., Jousilahti, P., Keil, U., Njlstad, I., Olganov, R., Thomsen, T., Tunstall-Pedoe, H., Tverdal, A., Wedel, H., Whincup, P., Wilhelmsen, L., and Graham, I. (2003). Estimation of ten-year risk of fatal cardiovascular disease in europe: the score project. *European Heart Journal*, Jun;24(11):987–1003.

[Cook et al., 2006] Cook, N., Buring, J., and Ridker, P. (2006). The effect of including c-reactive protein in cardiovascular prediction models for women. *Ann Intern Med*, 145:21–29.

[Cook and Ridker, 2009] Cook, N. and Ridker, P. (2009). Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Annals of Internal Medicine*, Jun 2;150(11):795–802.

[Cooney et al., 2015] Cooney, M., Selmer, R., Lindman, A., Tverdal, A., Menotti, A., Thomsen, T., DeBacker, G., De Bacquer, D., Tell, G., Njolstad, I., and Graham, I. (2015). Cardiovascular risk estimation in older persons: Score o.p. *European Journal of Preventive Cardiology*, Jun:1–11.

[Cox, 1972] Cox, D. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society Series B*, 34:187–220.

[D'Agostino et al., 2000] D'Agostino, R., Russell, M., Huse, D., Ellison, R., Silbershatz, H., Wilson, P., and Hartz, S. (2000). Primary and subsequent coronary risk appraisal: new results from the framingham study. *American Journal of the Heart*, Feb;139(2 Pt 1):272–281.

[D'Agostino Sr et al., 2001] D'Agostino Sr, R., Grundy, S., Sullivan, L., Wilson, P., and Group., C. R. P. (2001). Validation of the framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA*, Jul 11;286(2):180–187.

[D'Agostino Sr et al., 2008] D'Agostino Sr, R., Vasan, R., Pencina, M., Wolf, P., Cobain, M., Massaro, J., and Kannel, W. (2008). General cardiovascular risk profile for use in primary care: the framingham heart study. *Circulation*, Feb 12;117(6):743–753.

[de Wreede et al., 2010] de Wreede, L., Fiocco, M., and Putter, H. (2010). The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Computer Methods and Programs in Biomedicine*, 99:261–274.

[DeLong et al., 1988] DeLong, E., DeLong, D., and Clarke-Pearson, D. (1988). Problems with risk reclassification methods for evaluating prediction modelscomparing the areas

under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, Sep;44(3):837–45.

[Dignam and Kocherginsky, 2008] Dignam, J. and Kocherginsky, M. (2008). Choice and interpretation of statistical tests used when competing risks are present. *Journal of Clinical Oncology*, 26:40274034.

[Dignam et al., 2012] Dignam, J., Zhang, Q., and Kocherginsky, M. (2012). The use and interpretation of competing risks regression models. *Clin Cancer Res*, Apr 15;18(8):2301 – 2308.

[Engelhardt et al., 2014] Engelhardt, E., Garvelink, M., de Haes, J., van der Hoeven, J., Smets, E., Pieterse, A., and Stiggelbout, A. (2014). Predicting and communicating the risk of recurrence and death in women with early-stage breast cancer: a systematic review of risk prediction models. *Journal of Clinical Oncology*, 32:238–250.

[Ferguson et al., 2012] Ferguson, N., Somnath, D., and Brock, G. (2012). mssurv: An r package for nonparametric estimation of multistate models. *Journal of Statistical Software*, 50:1–24.

[Fine and Gray, 1999] Fine, J. and Gray, R. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94 (446):496–509.

[Fine et al., 2001] Fine, J., Jiang, H., and Chappell, R. (2001). On semi-competing risks data. *Biometrika*, 88:907919.

[Gagliardi et al., 1998] Gagliardi, G., Lax, I., Sderstrm, S., Gyenes, G., and Rutqvist, L. (1998). Prediction of excess risk of long-term cardiac mortality after radiotherapy of stage i breast cancer. *Radiotherapy Oncology*, Jan;46(1):63–71.

[Ganz, 2009] Ganz, P. (2009). Survivorship: adult cancer survivors. *Primary Care*, Dec;36(4):721–41.

[Gaziano et al., 2008] Gaziano, T., Young, C., Fitzmaurice, G., Atwood, S., and Gaziano, J. (2008). Laboratory-based versus non-laboratory-based method for assessment of cardiovascular disease risk: the nhanes i follow-up study cohort. *Lancet*, Mar 15;371(9616):923–931.

[Gerds et al., 2012] Gerds, T., Scheike, T., and Andersen, P. (2012). Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in Medicine*, Dec 20;31(29):3921 – 3930.

[Gerds and Schumacher, 2006] Gerds, T. and Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040.

[Gönen and Heller, 2005] Gönen, M. and Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965970.

[Gordon et al., 1973] Gordon, T., Sorlie, P., and Kannel, W. (1973). Coronary heart disease, atherothrombotic brain infarction, intermittent claudication- a multivariate analysis of some factors related to their incidence: Framingham study, 16-year followup. In Kannel, W. and Gordon, T., editors, *The Framingham Study: An Epidemiological Investigation of Cardiovascular Disease*. US Government Printing Office No. 426-1301/1345.

[Grambauer et al., 2010] Grambauer, N., Schumacher, M., and Beyersmann, J. (2010). Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified. *Statistics in Medicine*, Mar 30;29(7-8):875–884.

[Hajage et al., 2011] Hajage, D., de Rycke, Y., Bollet, M., Savignoni, A., and Caly M, e. a. (2011). External validation of adjuvant! online breast cancer prognosis tool. prioritising recommendations for improvement. *PLoS ONE*, 6:e27446.

[Hanley and McNeil, 1982] Hanley, J. and McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36.

[Harrell Jr et al., 1996] Harrell Jr, F., Lee, K., and Mark, D. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361387.

[Haybittle et al., 1982] Haybittle, J., Blamey, R., Elston, C., Johnson, J., Doyle, P., Campbell, F., Nicholson, R., and Griffiths, K. (1982). A prognostic index in primary breast cancer. *British Journal of Cancer*, 45:361–366.

[Hense et al., 2003] Hense, H., Schulte, H., Lowel, H., Assmann, G., and Keil, U. (2003). Framingham risk function overestimates risk of coronary heart disease in men and women from germany. results from the monica augsburg cohort and the procam cohort. *European Heart Journal*, May;24(10):937–945.

[Hsieh and Huang, 2012] Hsieh, J. and Huang, Y. (2012). Regression analysis based on conditional likelihood approach under semi-competing risks data. *Lifetime Data Analysis*, 18:302–320.

[Hsieh et al., 2008] Hsieh, J., Wang, W., and Ding, A. (2008). Regression analysis based on semi-competing risks data. *Journal of the Royal Statistical Society B*, 70(Part 1):320.

[Huang and Jin, 2007] Huang, L. and Jin, Z. (2007). Lss: an s-plus/r program for the accelerated failure time model to right censored data based on least-squares principle. *Comput Methods Programs Biomed*, Apr;86(1):45–50.

[Ishwaran et al., 2014] Ishwaran, H., Gerds, T., Kogalur, U., Moore, R., Gange, S., and Lau, B. (2014). Random survival forests for competing risks. *Biostatistics*, Oct;15(4):757–73.

[Jackson et al., 2003] Jackson, C., Sharples, L., Thompson, S., Duffy, S., and Couto, E. (2003). Multi-state markov models for disease progression with classification error. *The Statistician*, 52:193209.

[Jin et al., 2003] Jin, Z., Lin, D., Wei, L., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, 90:341 – 353.

[Jin et al., 2006] Jin, Z., Lin, D., and Ying, Z. (2006). On least-squares regression with censored data. *Biometrika*, 93:147–161.

[Kalbfleisch and Prentice, 2002] Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York, USA.

[Kannel et al., 1976] Kannel, W., McGee, D., and Gordon, T. (1976). A general cardiovascular risk profile: the framingham study. *Am J Cardiol*, Jul;38(1):146-5:146–151.

[Kattan et al., 2004] Kattan, M., Giri, D., Panageas, K., Hummer, A., Cranor, M., Van Zee, K., Hudis, C., Norton, L., Borgen, P., and Tan, L. (2004). A tool for predicting breast carcinoma mortality in women who do not receive adjuvant therapy. *Cancer*, 101:2509–2515.

[Kerr et al., 2014] Kerr, K., Wang, Z., Janes, H., McClelland, R., Psaty, B., and Pepe, M. (2014). Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology*, Jan;25(1):114–121.

[Khot et al., 2003] Khot, U., Khot, M., Bajzer, C., Sapp, S., Ohman, E., Brener, S., Ellis, S., Lincoff, A., and Topol, E. (2003). Prevalence of conventional risk factors in patients with coronary heart disease. *JAMA*, 290:898–904.

[Klein, 2006] Klein, J. (2006). Modelling competing risks in cancer studies. *Statistics in Medicine*, Mar 30;25(6):1015 – 1034.

[Klein and Andersen, 2005] Klein, J. and Andersen, P. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*, Mar;61(1):223–229.

[Kohl et al., 2015] Kohl, M., Plischke, M., Leffondre, K., and Heinze, G. (2015). Pshreg: a sas macro for proportional and nonproportional subdistribution hazards regression. *Computer Methods and Programs in Biomedicine*, 2015 Feb;118(2):218–233.

[Koller et al., 2012] Koller, M., Leening, M., Wolbers, M., Steyerberg, E., Hunink, M., Schoop, R., Hofman, A., Bucher, H., Psaty, B., Lloyd-Jones, D., and Witteman, J. (2012). Development and validation of a coronary risk prediction model for older u.s. and european persons in the cardiovascular health study and the rotterdam study. *Annals of Internal Medicine*, 6:389–397.

[Korn et al., 1997] Korn, E., Graubard, B., and Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *American Journal of Epidemiology*, Jan 1;145(1):72–80.

[Kurian et al., 2013] Kurian, A., Lichtensztajn, D., Keegan, T. H., Leung, R., Shema, S., Hershman, D., Kushi, L., Habel, L., Kolevska, T., Caan, B., and Gomez, S. (2013). Patterns and predictors of breast cancer chemotherapy use in kaiser permanente northern california, 2004-2007. *Breast Cancer Research and Treatment*, 137 (1):247–260.

[Latouche et al., 2013] Latouche, A., Allignol, A., Beyersmann, J., Labopin, M., and Fine, J. (2013). A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *Journal of Clinical Epidemiology*, Jun;66(6):648–653.

[Latouche et al., 2007] Latouche, A., Boisson, V., Chevret, S., and Porcher, R. (2007). Misspecified regression model for the subdistribution hazard of a competing risk. *Statistics in Medicine*, Feb 28;26(5):965–974.

[Leemis, 1987] Leemis, L. (1987). Variate generation for accelerated life and proportional hazards models. *Operations Research*, 35:892–894.

[Liao et al., 1999] Liao, Y., McGee, D., and Cooper, R. (1999). Prediction of coronary heart disease mortality in blacks and whites: pooled data from two national cohorts. *Am J Cardiol*, Jul 1;84(1):31–36.

[Liu and Jin, 2009] Liu, X. and Jin, Z. (2009). A non-parametric approach to scale reduction for uni-dimensional screening scales. *International Journal of Biostatistics*, 15:1–7.

[Lumley et al., 2006] Lumley, T., Kronmal, R., and Ma, S. (2006). Relative risk regression in medical research: Models, contrasts, estimators, and algorithms. *UW Biostatistics Working Paper Series*, July 2006:Working Paper 293.

[Lundin, 2007] Lundin, J. (2007). The nottingham prognostic index - from relative to absolute risk prediction. *European Journal of Cancer*, 43:1498–1500.

[Michaelson et al., 2011] Michaelson, J., Chen, L., Bush, D., Fong, A., Smoth, B., and Younger, J. (2011). mproved web-based calculators for predicting breast carcinoma outcomes. *Breast Cancer Research and Treatment*, 128:827–835.

[Mook et al., 2009] Mook, S., Schmidt, M., E.J., R., van de Velde, A., Visser, O., Rutgers, S., Armstrong, N., van't Veer, L., and Ravdin, P. (2009). Calibration and discriminatory accuracy of prognosis calculation for breast cancer with the online adjuvant! program: a hospital-based retrospective cohort study. *Lancet Oncology*, 11:1070–1076.

[Nagelkerke, 1991] Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78:691–692.

[Nichols et al., 2013] Nichols, H., Trentham-Dietz, A., Newcomb, P., Egan, K., Titus, L., Hampton, J., and Visvanathan, K. (2013). Pre-diagnosis oophorectomy, estrogen therapy and mortality in a cohort of women diagnosed with breast cancer. *Breast Cancer Res*, 15(5):R99.

[Olivotto IA, 2005] Olivotto IA, Bajdik CD, R. P. S. C. C. A. N. B. D. G. C. S. G. K. (2005). Population-based validation of the prognostic model adjuvant! for early breast cancer. *Journal of Clinical Oncology*, 23:2716–2725.

[Orbe et al., 2002] Orbe, J., Ferreira, E., and Nez-Antn, V. (2002). Comparing proportional hazards and accelerated failure time models for survival analysis. *Statistics in Medicine*, Nov 30;21(22):3493–3510.

[Patnaik et al., 2011] Patnaik, J., Byers, T., DiGuiseppi, C., Dabelea, D., and Denberg, T. (2011). Cardiovascular disease competes with breast cancer as the leading cause of death for older females diagnosed with breast cancer: a retrospective cohort study. *Breast Cancer Res*, Jun 20;13(3):R64.

[Pencina and DAgostino Sr, 2004] Pencina, M. and DAgostino Sr, R. (2004). Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23:2109–2123.

[Pencina et al., 2008] Pencina, M., DAgostino Sr, R., DAgostino Jr, R., and Vasan, R. (2008). Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Statistics in Medicine*, 27:157–172.

[Pencina et al., 2012a] Pencina, M., DAgostino Sr, R., and Demler, O. (2012a). Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Statistics in Medicine*, Jan 30;31(2):101–113.

[Pencina et al., 2012b] Pencina, M., DAgostino Sr, R., and Song, L. (2012b). Quantifying discrimination of framingham risk functions with different survival c statistics. *Statistics in Medicine*, Jul 10;31(15):1543–1553.

[Peng and Fine, 2007] Peng, L. and Fine, J. (2007). Regression modeling of semi-competing risks data. *Biometrics*, 63:96108.

[Pepe, 2011] Pepe, M. (2011). Problems with risk reclassification methods for evaluating prediction models. *American Journal of Epidemiology*, Jun 1;173(11):1327–1335.

[Pepe et al., 2008] Pepe, M., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I., and Zheng, Y. (2008). Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology*, Feb 1;167(3):362–368.

[Pepe et al., 2014] Pepe, M., Janes, H., and Li, C. (2014). Net risk reclassification p values: valid or misleading? *Journal National Cancer Institute*, Apr;106(4):1–6.

[Pepe et al., 2013] Pepe, M., Kerr, K., Longton, G., and Wang, Z. (2013). Testing for improvement in prediction model performance. *Statistics in Medicine*, Apr 30;32(9):1467–1482.

[Pepe and Mori, 1993] Pepe, M. and Mori, M. (1993). Kaplan-meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine*, Apr 30;12(8):737–751.

[Portnoy, 2003] Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association*, 198:1001–1012.

[Prentice et al., 1978] Prentice, R., Kalbfleisch, J., Peterson, A. J., Flournoy, N., Farewell, V., and Breslow, N. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, Dec;34(4):541–554.

[Putter, 2014] Putter, H. (2014 (accessed October 18, 2014)). *Tutorial in Biostatistics: Competing Risks and Multi-State Models. Analysis Using the mstate Package*.

[Putter et al., 2007] Putter, H., Fiocco, M., and Geskus, R. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*, 26:2389–2430.

[Ravdin et al., 2001] Ravdin, P., Siminoff, L., Davis, G., Mercer, M., Hewlett, J., Gerson, N., and Parker, H. (2001). Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *Journal of Clinical Oncology*, 19:980–991.

[Ridker et al., 2007] Ridker, P., Buring, J., Rifai, N., and Cook, N. (2007). Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the reynolds risk score. *JAMA*, Feb 14;297(6):611–619.

[Royston and Mahesh, 2002] Royston, P. and Mahesh, K. (2002). Flexible parametric proportional hazards and proportional odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21:21752197.

[Schairer et al., 2004] Schairer, C., Mink, P., Carroll, L., and Devesa, S. (2004). Probabilities of death from breast cancer and other causes among female breast cancer patients. *J Natl Cancer Inst*, Sep 1;96(17):1311–21.

[Schonberg et al., 2011] Schonberg, M., Marcantonio, E., Ngo, L., Li, D., Silliman, R., and McCarthy, E. (2011). Causes of death and relative survival of older women after a breast cancer diagnosis. *J Clin Oncol*, Apr 20;29(12):1570–7.

[Singla et al., 2012] Singla, A., Kumar, G., and Bardia, A. (2012). Personalizing cardiovascular disease prevention among breast cancer survivors. *Curr Opin Cardiol*, Sep;27(5):515–24.

[Steyerberg et al., 2010] Steyerberg, E., Vickers, A., Cook, N., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M., and Kattan, M. (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, Jan;21(1):128–138.

[Thomsen et al., 2002] Thomsen, T., McGee, D., Davidsen, M., and Jorgensen, T. (2002). A cross-validation of risk-scores for coronary heart disease mortality based on data from the glostrup population studies and framingham heart study. *International Journal of Epidemiology*, 31:817822.

[Vickers et al., 2011] Vickers, A., Cronin, A., and Begg, C. (2011). One statistical test is sufficient for assessing new predictive markers. *BMC Med Res Methodology*, Jan 28:11–13.

[Vickers et al., 2008] Vickers, A., Cronin, A., Elkin, E., and Gonen, M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak.*, Nov 26;8:566–578.

[Vickers and Elkin, 2006] Vickers, A. and Elkin, E. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*, Nov-Dec;26(6):565–74.

[Wei, 1992] Wei, L. (1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15):1871–1879.

[Willekens and Putter, 2014] Willekens, F. and Putter, H. (2014). Software for multistate analysis. *Demographic Research*, 31:381420.

[Wilson et al., 1998] Wilson, P., D'Agostino, R., Levy, D., Belanger, A., Silbershatz, H., and Kannel, W. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, May 12;97(18):1837–47.

[Wishart et al., 2010] Wishart, G., Azzato, E., Greenberg, D., Rashbass, J., Kearins, O., Lawrence, G., Caldas, C., and Pharoah, P. (2010). Predict: a new uk prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Research*, 12:2–10.

[Wishart et al., 2012] Wishart, G., Bajdik, C., Dicks, E., Provenzano, E., Schmidt, M., Sherman, M., Greenberg, D., Pharoah, P., and et. al (2012). Predict plus: development and validation of a prognostic model for early breast cancer that includes her2. *British Journal of Cancer.*, 5:800–807.

[Wishart et al., 2014] Wishart, G., Rakha, E., Green, A., Ellis, I., Ali, H., Provenzano, E., Blows, F., Caldas, C., and Pharoah, P. (2014). Inclusion of ki67 significantly improves performance of the predict prognostication and prediction model for early breast cancer. *BMC Cancer.*, 5:908–914.

[Wolbers et al., 2014] Wolbers, M., Blanche, P., Koller, M., Witteman, J., and Gerds, T. (2014). Concordance for prognostic models with competing risks. *Biostatistics*, 15:526 – 539.

[Wolbers et al., 2009] Wolbers, M., Koller, M., Witteman, J., and Steyerberg, E. (2009). Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology*, 20:555–561.

[Xu et al., 2009] Xu, Q., Paik, M., Luo, X., and Tsai, W. (2009). Reweighting estimators for cox regression with missing covariates. *Journal of the American Statistical Association*, 104:1155–1167.

[Xu et al., 2011] Xu, Q., Paik, M., Rundek, T., Elkind, M., and Sacco, R. (2011). Reweighting estimators for cox regression with missing covariate data: analysis of insulin resistance and risk of stroke in the northern manhattan study. *Statistics in Medicine*, Dec 10;30(28):3328–3340.

[Zeng et al., 2006] Zeng, D., Yin, G., and Ibrahim, J. (2006). Semiparametric transformation models for survival data with a cure fraction. *Journal of the American Statistical Association*, 101:670–684.

[Zhang et al., 2008] Zhang, M., Zhang, X., and Scheike, T. (2008). Modeling cumulative incidence function for competing risks data. *Expert Rev Clin Pharmacology*, May 1;1(3):391–400.

[Zhang and Zhang, 2011] Zhang, X. and Zhang, M. (2011). Sas macros for estimation of direct adjusted cumulative incidence curves under proportional subdistribution hazards models. *Computer Methods and Programs in Biomedicine*, 101:87–93.