Evaluating prediction performance of longitudinal biomarkers under cohort and two-phase study designs

Marlena Maziarz

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Yingye Zheng, Chair

Patrick Heagerty

Ying Qing Chen

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Evaluating prediction performance of longitudinal biomarkers under cohort and two-phase study designs

Marlena Maziarz

Chair of the Supervisory Committee:
Affiliate Professor Yingye Zheng
Department of Biostatistics

Risk prediction and evaluation of predictions based on longitudinal biomarkers are of interest in treatment selection, preventive medicine and management of chronic diseases. Methods to evaluate risk predictions in a longitudinal setting are limited to the area under the receiver operating characteristic curves and prediction error.

In this dissertation, we evaluate two approaches to risk prediction in the longitudinal setting: joint modeling and partly conditional modeling. We develop estimation procedures for more flexible and robust partly conditional models, demonstrate their adaptability and applicability, and provide a smoothing technique to account for measurement error in marker data. We develop nonparametric estimators of clinically relevant measures of prediction quality in the longitudinal setting under cohort, case-cohort, stratified case-cohort and nested case-control study designs. We provide resampling-based inference procedures for all estimators under the four study designs.

We evaluate our methods using simulation studies and illustrate them on the End Stage Renal Disease Study dataset and a nested case-control study within the HALT-C clinical trial.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I wish to thank my advisor, Yingye Zheng, for sharing her ideas, expertise and time with me. I am immensely grateful for your mentorship, honesty and your ongoing support throughout this journey.

The many excellent questions asked by committee members – Patrick Heagerty, Norm Breslow, Ying Qing Chen and Christopher Li – have not only kept me on my toes, but served to improve this dissertation and many other aspects of my work. A special thank you goes to Patrick, whose ideas were invaluable to this work. Your clarity of thought and experience in leading the way have given me direction at crucial moments in the process.

My first RA advisors at UW, Åke Lernmark and Norm Breslow have provided me with mentorship, guidance and support over the years. Åke, thank you for sharing your infectious enthusiasm for diabetes research with me. Norm, your passion for biostatistics, both theoretical and applied, was something I appreciated from day one and my appreciation only grew over the years. It was a pleasure to work with and learn from each of you. Both of you will always be an inspiration to me.

David Yanez's kindness, guidance and support throughout the years has meant a great deal to me. You always "had my back" and proved it time and time again. I am grateful for that.

And last, but not least, I would like to thank Yoshio Hall for sharing his knowledge and ideas with me, and for allowing me to use his fantastic data in my dissertation. Working with you was sheer joy.

My sister Adriana, you are my best and most trusted friend. Thank you for your love and support, and for dropping everything and jumping on a plane every time I break a leg. You and your beautiful family – your husband Piotr and your daughters Maya, Isabelle and Elise – make my world just that much happier and brighter.

Most of all, I would like to thank my parents. Mamuś and Tatuś, thank you for your love, your countless sacrifices, devotion and constant support. You always come up with the right words to say exactly when I need to hear them. Ever since I can remember you have been the kindest, truest and most unwavering supporters of mine in all my endeavors. That has meant the world to me.

DEDICATION

To my parents Zenobia and Roman, and my sister Adriana

Chapter 1
INTRODUCTION

In this dissertation we develop methods for risk prediction using longitudinal biomarker data with time-to-event outcomes, we develop an array of methods for evaluation of such risk predictions in a cohort study design, as well as case-cohort and nested case-control study designs. We evaluate our proposed methods with extensive simulation studies, and illustrate them on a cohort study of end-stage renal disease and a nested case-control study of hepatocellular carcinoma.

1.1 Motivation

This work was originally motivated by a problem in a type 1 diabetes (T1D) study. The data consisted of measurements of levels of anti-idiotypic antibodies (anti-Ids) targeting antibodies to glutamic acid decarboxylase 65kDa (GAD65), a protein found in the insulin-producing β-cells of the pancreas. Blood samples of Swedish children at high risk of T1D were collected every three months from birth to the time of diagnosis with T1D. Since T1D is a relatively rare disease even among those at high risk, a case-control study was conducted and anti-Ids were measured only in the blood samples of the individuals in the case-control study. The question was whether we could accurately predict the timeframe of the onset of T1D given the longitudinal anti-Id measurements.

We could not answer that question at the time, as the methods to address that problem did not yet exist. However, we realized that the problem had the necessary components for a dissertation topic: it was clinically relevant, open and, thanks to several methodological developments in recent years, feasible to tackle. We divided the problem into three aspects,

which are addressed in separate chapters of this dissertation:

- risk prediction in a longitudinal setting with time-to-event outcomes
- evaluation of risk predictions in a cohort study design setting
- evaluation of risk predictions in a two-phase study design setting

Over the course of this work we became aware that problems similar to our motivating problem arise quite often in clinical research studies, or the data is already collected and it is just a matter of querying an existing database. Thus, methods for risk prediction using longitudinal data with time-to-event outcomes are needed, as are methods to evaluate such predictions. The need for such methods extends to more efficient sampling designs than a cohort study design, since measuring longitudinal markers on everyone in the cohort is not always feasible, and is definitely not cost-effective especially when the outcome is rare, leading researchers to consider various two-phase sampling designs.

Next, we describe the two datasets that will be used to demonstrate our methods in the following chapters.

1.2 Illustrative datasets

1.2.1 End-stage renal disease (ESRD)

In the United States (US), chronic kidney disease (CKD) has been estimated to affect nearly 26 million Americans, is the ninth leading cause of death, and costs the federal government over $56 billion annually. Currently, over one-third of incident patients with ESRD are enrolled in Medicaid (the US joint federal and state health insurance program for the poor) or are uninsured at ESRD onset. In many instances, CKD can be slowed or prevented. Interventions such as blood pressure lowering, use of renin-angiotensin system inhibitors, and avoidance of nephrotoxins are effective in slowing CKD progression if initiated in earlier

stages of the disease process [Maziarz et al., 2014]. Thus, there is much interest in identifying biomarkers that can be used to accurately predict the risk of ESRD in a given timeframe. One such candidate biomarker is the estimated glomerular filtration rate (eGFR), which is currently the best test for measuring the level of kidney function.

ESRD Study dataset

The dataset we use to demonstrate the performance of our methods in Chapter 3 is the ESRD study (ESRDS) dataset. The ESRDS is a retrospective cohort of 698 individuals aged 18-60 years with severe non-dialysis requiring chronic kidney disease (CKD) who received ambulatory care in the Community Health Network (CHN) in San Francisco from January 1, 1996 to February 29, 2008. CKD was defined based on at least two outpatient estimated glomerular filtration rate (eGFR) measurements $\leq 60\, mL/min/1.73^2$ that were separated by at least three months, of these subjects we identified those with severe CKD defined based on the baseline eGFR measurement of $\leq 30\, mL/min/1.73^2$. The outcome was defined as time to ESRD (initiation of dialysis or kidney transplantation) or death. The longitudinal marker of interest was eGFR, a marker of kidney function, estimated using the Modification of Diet in Renal Disease (MDRD) formula. The MDRD formula estimates kidney function using a measurement of serum creatinine, adjusting for age, sex and race. The question of interest was to quantify the ability of eGFR to predict the risk of an event (ESRD or death) in a given timeframe τ_0 conditional on information available up to a given time point s, $s < \tau_0$.

1.2.2 Hepatocellular carcinoma

Patients with hepatocellular carcinoma (HCC) often have poor prognosis due to late diagnosis. Since cirrhosis of any cause and chronic infection with hepatitis B virus or hepatitis C virus are the most common risk factors for HCC, surveillance of high-risk population may

detect tumors at an early stage when curative interventions can be implemented. Results of several studies suggest that HCC surveillance improves survival among patients with HCC, and guidelines from professional organizations recommend HCC surveillance for at-risk populations [Lok et al., 2010].

A major problem with HCC surveillance is the lack of reliable biomarkers. α-Fetoprotein (AFP) is the most commonly used biomarker for HCC surveillance, however not all HCC secrete AFP and often the AFP levels are elevated in individuals without HCC. Low sensitivity and specificity of AFP for detecting early HCC led the American Association for the Study of Liver Diseases (AASLD) Practice Guideline Committee to recommend that ultrasound alone be used for HCC surveillance. This is not optimal, since interpretation of the ultrasound is operator dependent and can be difficult in obese subjects or those with underlying cirrhosis [Lok et al., 2010].

Thus, reliable biomarkers for HCC surveillance and early detection are sought in order to detect the cancer earlier, when more treatment options are available. One candidate is des-γ carboxyprothrombin (DCP) which has been widely used in Japan for HCC diagnosis and surveillance [Lok et al., 2010] and has been shown to have better performance characteristics compared to AFP [Volk et al., 2007].

Nested case-control study within the HALT-C clinical trial

The Hepatitis C Antiviral Long-Term Treatment against Cirrhosis (HALT-C) Trial consisted of 1,002 patients with chronic hepatitis C and bridging fibrosis or cirrhosis, who failed to respond or to achieve a sustained virologic response to combination therapy. Patients were followed every 3 months for 3.5 years after randomization. Blood samples were collected at each visit for subsequent research testing including assays for HCC biomarkers. Ultrasound examinations were performed 6 months after enrollment and every 12 months thereafter. Patients with an elevated or rising AFP and those with new lesions on ultrasound were

evaluated further by CT or MRI. One of the goals of the HALT-C Trial was to identify and validate markers for the surveillance and early diagnosis of HCC. One of the markers of interest was des-γ-carboxyprothrombin (DCP).

A nested case-control study was used to evaluate the accuracy of DCP in the detection of HCC. For this study, 39 HCC cases diagnosed between randomization and 3.8 years after randomization were included in the study. For each case, 2 controls without HCC at the time of diagnosis of the case were selected matching on treatment assignment, presence of cirrhosis on baseline biopsy and length of followup. One control was later excluded because of high DCP values due to caumadin (anticoagulant) use, leaving 77 controls. DCP values played no role in diagnosis of HCC.

1.3 Purpose of this work

The work described in this dissertation aims to provide tools and practical guidance as to their use for risk prediction in a longitudinal setting, as well as tools for evaluating such risk predictions in a cohort, case-cohort and nested case-control study designs.

1.4 Scope of the dissertation

In Chapter 2 we review existing methods for risk prediction based on longitudinal data with time-to-event outcomes, methods for evaluation of predictions using markers measured at baseline in cohort and case-cohort studies. In Chapter 3 we extend the partly conditional (PC) model to provide it with more flexibility and accuracy, as well as to demonstrate the various possible approaches to estimating risk of an event in the next τ_0 time interval conditional on information up to time s, $R_i(\tau_0 \mid s)$, where $s < \tau_0$. We compare the PC models with the joint model, the current gold standard of risk prediction in a longitudinal setting. We evaluate and compare the models using simulation studies and we illustrate their performance using the ESRDS dataset. In Chapter 4 we add to the set of existing measures to evaluate risk predictions in a longitudinal setting under a cohort study design setting. We

develop resampling-based variance estimators for all of the measures discussed in Chapter 4. We evaluate our methods using simulation studies and we apply them in an analysis of the ESRDS dataset. In Chapter 5 we develop estimators and resampling-based estimators for all measures described in Chapter 4 under the case-cohort and nested case-control study designs. We evaluate our methods using simulation studies and illustrate our methods on the nested case-control study within the HALT-C clinical trial. Concluding remarks and ideas generated by this work to be addressed in the future are summarized in Chapter 6.

Chapter 2

LITERATURE REVIEW

In this chapter we recount the notable advances that have led to the current approaches to risk prediction based on longitudinal biomarkers with time-to-event outcomes and survey the existing methods for evaluation of predictions in cohort, case-cohort and nested case-control study designs.

Interest in survival models utilizing longitudinal observations, as well as methods to evaluate and compare such models has increased in recent years, and there are several reasons for that. First, methods for risk prediction based on longitudinal data with time-to-event outcomes have many public health applications. Development of methods to model longitudinal covariates and time-to-event data was motivated by the need to better understand the progression of diseases such as AIDS and prostate cancer. In the case of AIDS, such models were used to characterize the association between longitudinal CD4 count profiles and hazard of AIDS [Wang and Taylor, 2001] or time to progression to AIDS among HIV-positive subjects [Tsiatis and Davidian, 2004]. In the case of prostate cancer, longitudinal measurements of prostate-specific antigen (PSA) were used to predict prostate cancer [Lin et al., 2002] or recurrence of prostate cancer [Pauler and Finkelstein, 2002, Law et al., 2002, Yu et al., 2004, Taylor et al., 2005]. Other applications include analysis of data from a clinical trial evaluating the response to a drug to treat schizophrenia [Xu and Zeger, 2001]. Second, the necessary data are now more accessible with the increase in the use of digitized patient charts and databases to store clinical information. Lastly, the methods needed for fitting survival models with longitudinal observations are often computationally intensive and involve steps such as numerical integration, resampling, maximization by iterating through

complex expectation-maximization steps and so on. Many of these calculations were too time consuming a decade ago, but have now become feasible.

Methods to analyze longitudinal data with time-to-event outcomes, as well as methods to predict risk based on that data, are available in some software packages. It is tempting to think that fitting such models has become a relatively straightforward process from the perspective of a user. However, the methods behind the 'black-box' functions available in today's computational packages are far from simple. They are computationally intensive, but their main drawback is their lack of flexibility in case they need to be adapted to a new problem. Lastly, methods to evaluate predictions obtained from predictive models utilizing longitudinal data are lacking and most of such methods have focused on models utilizing baseline biomarkers [Taylor et al., 2013].

We have divided this review into three main sections. In Section 2.1 we review methods for risk prediction using longitudinal biomarkers and time-to-event outcomes focusing on the the joint and the partly conditional modeling approaches. In Section 2.2 we review existing methods for evaluating risk prediction models in a cohort study design setting. In the last section we review approaches to evaluate risk predictions in a case-cohort study design setting. After each section we remark on the shortcomings of these methods and comment on what methods are lacking, yet needed and comment on how the work on the three goals of this dissertation is aimed at filling some of those gaps.

2.1 Methods for risk prediction based on longitudinal data with time-to-event outcomes

In this section, we review two approaches to risk prediction using longitudinal biomarkers and time-to-event outcomes – joint models and partly conditional models. For each of these approaches we briefly review the relevant methods comprising preliminary work leading up to their development, we give examples of their application and some variations that have been proposed for specific applications. We then review joint and partly conditional

models in detail: their construction, assumptions, estimation of model parameters, as well as obtaining risk predictions using each approach. We now introduce some notation that will be used throughout this chapter; additional notation will be introduced as needed.

2.1.1 Notation

Let T_i and C_i be random variables with values in $\mathbb{R}_+$ measuring the time to failure and time to censoring, respectively, for subject i. We may not observe T_i for all subjects, but rather we observe $X_i = \min(T_i, C_i)$ and $\delta_i = \mathrm{I}(T_i \leq C_i)$. We assume that the survival time is independent of censoring time. We use $\mathbf{Z}_i$ denote time-constant covariates such as gender. The observed longitudinal biomarker on subject i is denoted by $\mathbf{Y}_i = \{Y_i(s_{i1}), \ldots, Y_i(s_{im_i})\}$ measured at times $\mathbf{s}_i = \{s_{i1}, \ldots, s_{im_i};\, s_{im_i} < X_i\},\ i = 1, \ldots, n,\ j = 1, \ldots, m_i$, where $Y_i(s_{ij})$ is the j^{th} measurement for subject i observed at time s_{ij}. The observation times are assumed to be specified by a study protocol, although deviations from protocol are not unusual. The observed covariate information for subject i consists of $\mathbf{H}_i = (\mathbf{Z}_i, \mathbf{Y}_i, \mathbf{s}_i)$. At time $u \geq 0$, the history of the covariate process for subject i is known and equals to $\mathbf{H}_i(u) = \{\mathbf{Z}_i, \mathbf{Y}_i(u), \mathbf{s}_i(u)\}$, where $\mathbf{Y}_i(u) = \{Y_i(s_{ij}) : 0 \leq s_{ij} \leq u,\ j = 1, \ldots, m_i,\ u < X_i\}$ and $\mathbf{s}_i(u) = \{s_{ij} : 0 \leq s_{ij} \leq u,\ j = 1, \ldots, m_i,\ u < X_i\}$. The full data sample is denoted by $\mathcal{D}_n = \{X_i, \delta_i, \mathbf{H}_i, i = 1, \ldots, n\}$. A subscript $\cdot_\circ$ is used to denote data for a new, or future, individual and any relevant notation applied to such an individual will have $\circ$ in place of i.

Throughout this document we use a generic term *risk* to mean the probability of an event for an individual in a given timeframe, given their survival and covariate information up to some earlier time, s. We define risk more precisely as the probability of an event for a specific individual i in a given timeframe τ_0, given biomarker data up to time s, $\mathbf{Y}_i(s)$, and possibly baseline covariate information, $\mathbf{Z}_i$, and denote that risk by $R_i(\tau_0 \mid s) = P(s < T_i \leq s + \tau_0 \mid T_i > s,\, \mathbf{Y}_i(s),\, s,\, \mathbf{Z}_i)$.

2.1.2 Joint models

Early methods to analyze longitudinal data with time-to-event outcomes focused on estimating associations between longitudinal data and survival data. Building on the Cox model [Cox, 1972], several models to analyze survival data with longitudinal covariates have been developed, such as a time-varying covariate Cox model [Crowley and Hu, 1977, Andersen and Gill, 1982]. These models require that covariate measurements be available at every event time, which is often not the case in practice. One approach to address that is to carry forward the last available value of the covariate to the current time, but this introduces bias in the estimation of β [Prentice, 1982, Tsiatis and Davidian, 2001]. More recently, joint models to jointly model the longitudinal covariate with time-to-event outcomes were developed [Tsiatis et al., 1995, Faucett and Thomas, 1996, Wulfsohn and Tsiatis, 1997]. By modeling the longitudinal process, joint models provide a fitted covariate measurement at every failure time, thus are able to deal with missing covariate data and can potentially account for random measurement fluctuations due to measurement error [Tsiatis and Davidian, 2004, Ye et al., 2008]. If the model is correctly specified, joint modeling is also able to account for informative censoring and missingness in longitudinal data that may depend on observed, but not on missing, data (missing at random) [Tsiatis and Davidian, 2004, Rizopoulos et al., 2013].

Two major approaches to estimation of parameters of a joint model are a two-stage approach [Pawitan, 1993, Tsiatis et al., 1995, Dafni, 1998] and a likelihood-based approach [Faucett and Thomas, 1996, Henderson et al., 2000, Wang and Taylor, 2001]. The two-stage approach is comprised of two stages. First, a model is assumed for the marker process and estimates of missing marker values are imputed for all subjects in each risk set. Next, these imputed marker values are used to fit a time-dependent Cox proportional hazards model [Tsiatis et al., 1995, Yu et al., 2004]. This approach is relatively simple, but it has a few drawbacks. The survival information is not used in modeling the marker process, which may

result in bias and loss of efficiency [Faucett and Thomas, 1996]. In stage two, the marker data is treated as fixed, thus uncertainty in the modeling of the marker process stage is not propagated to stage two [Yu et al., 2004]. In the likelihood-based approach, the marker data and survival data is modeled jointly and estimation and inference is based on the joint likelihood [Faucett and Thomas, 1996, Tsiatis and Davidian, 2004, Yu et al., 2004]. Joint analysis tends to produce estimates that are less biased and more efficient [Faucett and Thomas, 1996]. We focus our discussion on the likelihood based approach to estimation of parameters of the joint model.

Joint modeling framework for longitudinal and time-to-event data allows the full representation of the joint likelihood given the observed data, $\mathcal{D}_n$ [Tsiatis et al., 1995, Faucett and Thomas, 1996, Wulfsohn and Tsiatis, 1997, Henderson et al., 2000]. It comprises two linked sub-models, one for the underlying 'true' biomarker process as a function of time, and one linking the failure time T_i to the underlying 'true' biomarker process. To model the biomarker process, a linear mixed effects model [Laird and Ware, 1982] with random intercepts and slopes is often considered: $Y_i(s_{ij}) = W_i(s_{ij}) + e_i(s_{ij})$, $W_i(s_{ij}) = \theta_{0i} + \theta_{1i} f(s_{ij})$, where $W_i(s_{ij})$ represents the underlying 'true' biomarker for subject i at time s_{ij}, $(\theta_{i0}, \theta_{i1} \,|\, \mathbf{Z}_i)$ is assumed to be normally distributed with mean (θ_0, θ_1) and variance $\mathbf{\Sigma}$. The measurement error $e_i(s_{ij})$ is assumed to be distributed $N(0, \sigma_e^2)$, $\text{cov}(e_i(s_{ij}), e_i(s_{ij'})) = 0$ for $j \neq j'$.

The hazard function

$$\lambda_i(t|W_i(t), \mathbf{Z}_i) = \frac{1}{\delta t} \lim_{\delta t \to 0} P(t \leq T_i < t + \delta t \mid T_i \geq t,\, W_i(t), \mathbf{Z}_i)$$

is typically modeled using a time-varying covariate proportional hazards regression model $\lambda_i(t) = \lambda_0(t) \exp(\beta_w W_i(t) + \boldsymbol{\beta}_z^T \mathbf{Z}_i)$ [Crowley and Hu, 1977, Wulfsohn and Tsiatis, 1997, Tsiatis and Davidian, 2004].

The resulting joint likelihood is difficult to maximize, since it involves unobserved quantities and integrals which do not have analytical solutions. The EM algorithm can be used

to obtain maximum likelihood estimates of the joint model parameters [Wulfsohn and Tsiatis, 1997, Tsiatis and Davidian, 2004] or they can be obtained by direct maximization [Wu et al., 2012]. To approximate integrals numerical integration techniques, such as Monte Carlo or Gauss-Hermite quadrature, have been used in the joint modeling literature [Henderson et al., 2000, Rizopoulos, 2011], or for high-dimensional random effects problems Laplace approximations can be used [Ye et al., 2008, Rizopoulos et al., 2009].

2.1.3 Prediction based on a joint model

A natural extension of methods for modeling longitudinal and time-to-event data is to attempt to look into the 'future' given the information we have so far. Some examples of early applications of predictive methods using joint models were applied to prostate cancer. Pauler and Finkelstein [Pauler and Finkelstein, 2002] used joint modeling to obtain posterior predictive distributions for time to relapse of prostate cancer. Yu *et al.* estimated conditional recurrence probability of prostate cancer using a joint modeling approach via two estimation methods, a Monte Carlo EM algorithm and a Markov chain Monte Carlo [Yu et al., 2004].

Since then, the interest in prediction based on longitudinal and time-to-event data has grown [Garre et al., 2008, Song and Wang, 2008, Yu et al., 2008, Proust-Lima and Taylor, 2009, Rizopoulos, 2011, Taylor et al., 2013]. In 2010, Rizopoulos developed an `R` package `JM` for fitting joint models and for obtaining predictions using them [Rizopoulos, 2010]. We now describe the methods used to obtain risk predictions using joint models.

We are interested in predicting risk of an event in an interval τ_0 of time from s, $R_i(\tau_0 \,|\, s)$, for a new subject i given their longitudinal measurements up to time s, $\mathbf{Y}_i(s)$, using the joint modeling framework: $R_i(\tau_0|s) = \mathrm{P}(s < T_i \leq s + \tau_0 \,|\, T_i > s, \mathbf{Y}_i(s), \mathbf{Z}_i, \mathcal{D}_n, \theta)$, where θ denotes the parameter vector of the joint model. The estimation of this quantity takes advantage of the conditional independence assumptions used to define the joint model. Rizopoulos [Rizopoulos, 2011] showed that given the condition $p(X_i, \delta_i, Y_i|b_i, \theta) = p(X_i, \delta_i|b_i, \theta)p(Y_i|b_i, \theta)$

where b_i are the random effects, $R_i(\tau_0|s)$ can be rewritten as:

$$\begin{aligned} R_i(\tau_0|s) &= 1 - P(T_i > s + \tau_0 \mid T_i > s, \mathbf{Y}_i(s), \mathbf{Z}_i, \mathcal{D}_n, \theta) \\ &= 1 - \int \frac{S_i(s + \tau_0 \mid \mathbf{W}_i(s + \tau_0, b_i, \theta), \theta)}{S_i(s \mid \mathbf{W}_i(s, b_i, \theta), \theta)} p(b \mid T_i > s, \mathbf{Y}_i(s), \theta) db \end{aligned}$$

where $S_i(\cdot)$ denotes the survival function, $\mathbf{W}_i(\cdot)$ is the longitudinal history as approximated by the LMEM and is a function of the random effects and the parameters. The first order estimate of $R_i(\tau_0|s)$ was derived by [Rizopoulos, 2011] using the empirical Bayes estimates for the random effects:

$$\widehat{R}_i(\tau_0|s) = 1 - \frac{\widehat{S}_i(s + \tau_0 \mid \mathbf{W}_i(s + \tau_0, \widehat{b}_i^{(s)}, \widehat{\theta}), \widehat{\theta})}{\widehat{S}_i(s \mid \mathbf{W}_i(s, \widehat{b}_i^{(s)}, \widehat{\theta}), \widehat{\theta})} + O\left(\frac{1}{n_i(s)}\right)$$

where $\widehat{\theta}$ are the maximum likelihood estimates, $\widehat{b}_i^{(s)}$ is the mode of the conditional distribution $log\{p(b_i \mid T_i > s, \mathbf{Y}_i(s), \widehat{\theta})\}$ and $n_i(s)$ is the number of longitudinal measurements for subject i at time s. Rizopoulos [Rizopoulos, 2011] showed that this estimator works well in practice. Standard errors and confidence intervals can be obtained using Monte Carlo simulation schemes [Proust-Lima and Taylor, 2009, Rizopoulos, 2011].

2.1.4 Partly conditional and landmark models

In the early 1980's it became apparent that analysis of survival time in responders vs. non-responders to a treatment, where the responder group was identified sometime after the start of the study, was flawed. The non-responder group had advantages, for example to be in the responder group one had to survive until the time when the response was evaluated. Thus, there was a survival bias in favor of the responder group which could lead to a significant difference in survival between the two groups even if there was no such difference in reality [Anderson et al., 1983]. Landmarking was one of the methods addressing this issue. At the start of the study a specific time, a *landmark time*, would be set for evaluation of response.

The analysis of survival time in responders and non-responders would then proceed with survival time reset to zero at that landmark time [Ettinger and Lagakos, 1982, Anderson et al., 1983].

It turns out that a similar approach can be taken to work around a problem in prediction using longitudinal data. The problem only exists for covariates that are *internal*, meaning they can only be observed for individuals that are alive and uncensored. Blood pressure is an example of an internal covariate. Internal covariates may carry information about the failure time [Kalbfleisch and Prentice, 2002]. Thus, using an internal covariate measured at time t^- to predict survival beyond time t is not interesting, since $\mathrm{P}(T \geq t \mid Y(t^-)) = 1$, where $Y(\cdot)$ denotes an internal covariate [Fisher and Lin, 1999, Kalbfleisch and Prentice, 2002]. Landmarking (LM) gets around that problem by conditioning on a part of the covariate trajectory up to a landmark time t^L for each individual in the dataset [van Houwelingen, 2007]. Everyone at risk at t^L is included in the analysis. Landmark times are chosen in advance and there can be multiple landmark times. If more than one landmark time is chosen, individuals enter the analysis multiple times, once per landmark time as long as they are at risk at that landmark time, and they remain in the analysis until their event or censoring time. This results in multiple records per individual with tied survival times, requiring methods accounting for tied events to estimate model parameters [van Houwelingen, 2007].

In 2005, Zheng and Heagerty proposed a class of semiparametric models, called *partly conditional survival models* (PC), where the conditional hazard is specified by conditioning on a part of the marker trajectory [Zheng and Heagerty, 2005]. This approach is related to landmarking, but there are some notable differences between the two approaches. In the PC approach the observations are also stacked, but they are stacked with all measurement times aligned at time zero as opposed to all survival times for a given individual aligned at their event time in the LM approach. In other words, in the PC approach, the trajectory of a given individual is *cloned* and reset to zero at every measurement time for that individual, with the

original trajectory allowed to continue (Figure 2.1). This results in as many trajectories for a given individual as they have measurements, each shorter than the previous but all starting at analysis time equal to zero. The event times for that individual are not tied, as is the case in the LM approach. Also, the times at which the records are cloned is determined by times at which measurements were taken in the PC approach, and in the LM approach those times are arbitrary and decided on *a priori* [Zheng and Heagerty, 2005, van Houwelingen, 2007].

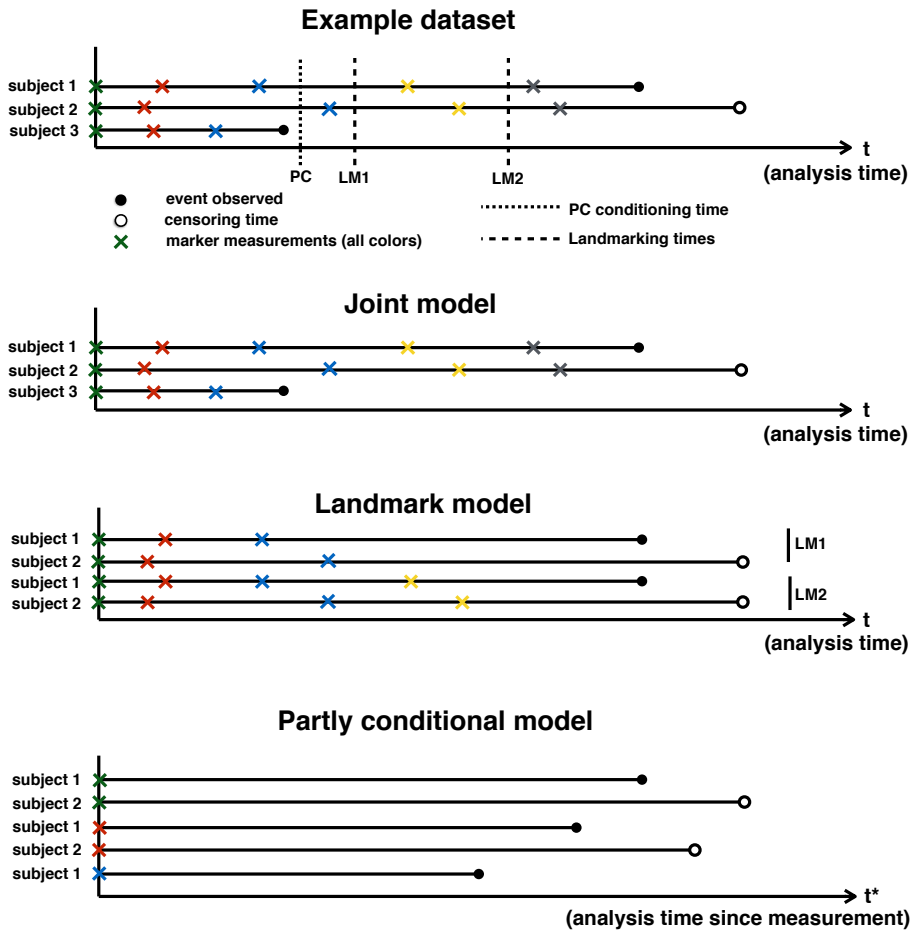

Figure 2.1: Example of analysis setup for analysis using a joint model, a landmark model (LM) and a partly conditional (PC) model given the same dataset containing three subjects.

Due to similarities in these two approaches we will refer to all models conditioning on a portion of the marker trajectory as *partly conditional survival models*. Partly conditional modeling does not require parametric assumptions about the marker process. The PC approach decouples the analysis time from the observation time. This has interesting implications for the relationship between the longitudinal measurements and the hazard function. The main idea is that the marker measurement for a given visit becomes 'frozen' in time and is treated as a baseline measurement in the analysis [Zheng and Heagerty, 2005]. Standard statistical software can be used to fit some simpler partly conditional models, though some additional work is required for more elaborate models [Zheng and Heagerty, 2005, van Houwelingen, 2007].

In the partly conditional survival model the analysis time is the time since measurement and is denoted by t^*. Let $T^*_{ij} = T_i - s_{ij}$ ($C^*_{ij} = C_i - s_{ij}$) be a random variable with values in $\mathbb{R}_+$ measuring the time from measurement j to failure (censoring) for subject i, $j = 1, \ldots, m_i$. Thus, each measurement time j for subject i is associated with an observed vector $\mathbf{O}_{ij} = (X^*_{ij}, \delta_i, Y_{ij}, s_{ij})$ where $X^*_{ij} = \min(T^*_{ij}, C^*_{ij})$ and we assume that all measurements of the marker for subject i occur before X_i.

The partly conditional hazard function $\lambda_{ij}(t^*)$ for the derived survival outcomes T^*_{ij} is

$$\lambda_{ij}(t^* \mid Y_i(s_{ij}),\, T^*_{ij} \geq 0) = \lim_{\delta t \to 0} \frac{1}{\delta t} \mathrm{P}(t^* \leq T^*_{ij} < t^* + \delta t \mid T^*_{ij} \geq t^*,\, Y_i(s_{ij}),\, T^*_{ij} > 0)$$

A regression model for the hazard can take the general form:
$\lambda_{ij}(t^* \mid Y_i(s_{ij}),\, 0 \leq s_{ij} < T_i) = g(\lambda_0(t^*, s),\, \beta(t^*, s) Y_i(s_{ij}))$ where $g(\lambda, \nu)$ is a link function. The authors showed examples of models fitting into this framework and how the dependence of $\beta(t^*, s) Y_i(s_{ij})$ on s can be handled [Zheng and Heagerty, 2005].

2.1.5 Prediction based on a partly conditional survival model

Methods for obtaining risk predictions based on a longitudinal biomarker using a partly conditional model stem from methods for obtaining predictions based on a baseline biomarker using a Cox model. For a simple Cox model, for example $\lambda_i(t) = \lambda_0(t)\exp(\beta Y_i)$ where Y_i is a marker measurement at baseline for subject i, the estimate of risk of an event in the time interval τ_0 from baseline given Y_i is $\widehat{R}_i(\tau_0) = 1 - \widehat{S}_i(\tau_0)$, where $\widehat{S}_i(\tau_0) = \widehat{S}_0(\tau_0)^{\exp(\widehat{\beta}Y_i)}$, $\widehat{S}_0(\tau_0) = \exp(-\widehat{\Lambda}_0(\tau_0))$, $\widehat{\Lambda}_0(\tau_0)$ is the Breslow estimator of the cumulative baseline hazard and $\widehat{\beta}$ is estimated by maximizing the partial likelihood.

In the case of a longitudinal biomarker, given a partly conditional working model for the hazard, we can estimate the risk of an event in the time interval τ_0 past time s, given data up to time s by: $\widehat{R}_i(\tau_0|s) = 1-\widehat{S}_i(\tau_0|s) = 1-\exp(-\widehat{\Lambda}(\tau_0|s))$ where $\widehat{\Lambda}(\tau_0|s)$ denotes the estimate of the cumulative hazard at time τ_0 on a 'time since measurement' scale given marker data up to time s [Zheng and Heagerty, 2005]. The details of the estimation of $\Lambda_i(\tau_0|s)$ follow.

Zheng and Heagerty [Zheng and Heagerty, 2005] described three PC working models for the hazard: (1) a PC proportional hazards model stratified on measurement time; (2) a PC varying coefficient hazards model and (3) a PC proportional and varying coefficient hazards model. However, other models in the general class of PC models could be used as well. For the moment, let us assume that our working regression model for the hazard is:

$$\lambda_{ij}(t^* \mid Y_i(s_{ij}), T^*_{ij} > 0) = \lambda_0(t^*)\exp\big(\beta(t^*)Y_i(s_{ij}) + \boldsymbol{\alpha}^T\{f_k(s_{ij})\}_{k=1}^p\big) \tag{2.1}$$

where $\{f_k(s)\}_{k=1}^p = \{f_1(s), \ldots, f_p(s)\}$ and $f_k(s_{ij})$ represents basis functions for some parametric but flexible function of s_{ij}.

Using standard counting process notation for the time-dependent setting, we have $N_{ij}(u^*) = \mathrm{I}(X^*_{ij} \leq u^*, \delta_i = 1)$, $R_{ij}(u^*) = \mathrm{I}(X^*_{ij} \geq u^*)$ denotes the at-risk process and $\bar{N} = \sum_{i=1}^n \sum_{j=1}^{m_i} N_{ij}(u^*)$.

Then the cumulative baseline hazard can be estimated using a Breslow-type estimator

$$\widehat{\Lambda}_0(t^*) = \int_0^{t^*} \frac{d\bar{N}(u^*)}{S^{(0)}(\widehat{\beta}, u^*)}$$

where

$$S^{(0)}(\widehat{\beta}, u^*) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} R_{ij}(u^*) \exp\big(\beta(t^*) Y_i(s_{ij}) + \boldsymbol{\alpha}^T \{f_k(s_{ij})\}_{k=1}^{p}\big) \tag{2.2}$$

Then

$$\widehat{\Lambda}(t^* \mid Y_i(s), s) = \int_0^{t^*} \exp\big(\beta(t^*) Y_i(s) + \boldsymbol{\alpha}^T \{f_k(s)\}_{k=1}^{p}\big) d\widehat{\Lambda}_0(u^*) \tag{2.3}$$

and $\widehat{\Lambda}(\tau_0 \mid s) = \widehat{\Lambda}(t^* \mid Y_i(s), s)$ when $t^* = \tau_0$. The estimate of risk $R_i(\tau_0|s)$ follows. The estimation proceeds similarly for different working models for the hazard, in which case equations 2.1, 2.2 and 2.3 would need to be modified accordingly.

These risk predictions can be estimated using standard software for fitting Cox regression models for most PC working models. Modifications to the software would be required for fitting more complex PC models [Zheng and Heagerty, 2005].

2.1.6 Remarks

Joint models (JM) and partly conditional models (PC) can be used to obtain risk predictions based on longitudinal biomarker with time-to-event outcomes. The two approaches differ substantially in the assumptions required, computational complexity, robustness and flexibility. If the JM is correctly specified, it can appropriately handle data missing at random, and situations where visit times and censoring depends on the observed longitudinal data. The JM assumes that the underlying, or true, marker process is associated with the hazard of an event, and that the observed marker data is likely to be contaminated with measurement error. Thus by modeling the relationship of the underlying process with the time-to-event outcome, it aims to reduce the impact of measurement error on the model parameter estimates.

The PC model does not make any parametric assumptions on the marker process. It does, however, assume that any missing data is missing completely at random, that visit times are independent of the longitudinal data and survival time (or are specified in advance by the study protocol), and that censoring is independent of the observed longitudinal data.

The JM is computationally complex, and is difficult and time consuming to fit. The PC model can be fit using standard statistical software with possibly small modifications. It is also more robust compared to the JM. Lastly, the PC approach is very flexible, lending itself to various types of model modifications, such as incorporation of the measurement time as a covariate or treating the marker as a time-varying covariate, as well as various estimation techniques such as kernel smoothing or penalizing the contribution to the partial likelihood.

The predictive performance of JM and PC approaches is likely to differ, but so far little attention has been given to comparing of the two approaches in terms of their predictive performance. It would be interesting to compare the predictive performance of those two approaches in several settings: misspecification to model assumptions about the longitudinal and event time processes, varying amount of longitudinal information per subject, different marker trajectories, varying the ratio of measurement error variance to between subject variability. Rizopoulos *et al.* made an attempt to address some of these issues in a recent technical report, but they focused on the comparison between JM and landmarking where only the last available measurement per subject is used [Rizopoulos et al., 2013]. Thus, questions of which approach would be more appropriate to use in practice in these different situations remain.

We address these questions in Chapter 3 of this dissertation. Specifically, we discuss the modeling choices and evaluate the trade-offs between JM and several extensions of the PC model. We emphasize the practical importance of such a comparison. Using longitudinal data with time-to-event outcomes to predict risk of an event in a given timeframe is a difficult problem, both from a statistical and computational standpoint. However, the ability to make such predictions accurately has the potential to impact public health decisions on an

individual and policy levels.

2.2 Methods for evaluation of predictions in cohort studies

Despite many theoretical developments in methods for risk prediction using longitudinal biomarkers, development of methods to evaluate the performance of such predictive models is lagging [Schoop et al., 2008, Taylor et al., 2013]. There has been a lot of work done on methods to evaluate tests for classification and prediction based on baseline biomarkers [Pepe, 2003]. More recently, some of these methods have been extended to a time-dependent setting [Leisenring et al., 1997, Etzioni et al., 1999, Heagerty et al., 2000]. We start our review with measures to evaluate diagnostic tests, or tests used to classify subjects into those with and without the outcome of interest. These measures can also be used to evaluate tests for prediction of disease based on biomarkers measured at baseline. We then review extensions of some of these measures to a time-dependent setting.

Consider a situation where a binary test is used in screening for a disease. Let D denote the disease status, with $D = 1$ for a subject with an outcome of interest (case) and $\bar{D} \equiv D = 0$ otherwise (control). Let M_b denote the result of a binary medical test, where $M_b = 1$ denotes a positive test result and $M_b = 0$ a negative one. Then the true positive fraction (TPF) and the false positive fraction (FPF) are defined as $\text{TPF} = \text{P}(M_b = 1 \mid D = 1)$ and $\text{FPF} = \text{P}(M_b = 1 \mid D = 0)$, respectively. These measures are also used to define the sensitivity and specificity of a test: Sensitivity = TPF and Specificity = $1 - \text{FPF}$.

The purpose of a medical test is to aid in clinical decision-making and such decisions are almost always binary. Medical providers must decide whether their patient has a given disease or not, whether they should perform a medical procedure or not, whether the patient is responding to treatment or not. Thus, tests on a continuous scale are usually dichotomized given some threshold value. The test can be based on a marker value for some continuous marker Y, it can also be based on risk predicted using a risk prediction model (using marker measured at baseline), R. Such a prediction model is essentially a mapping from the range

of Y to $(0, 1)$. In order to emphasize the fact that tools to evaluate medical tests can be applied to either, the raw data or some function of it, we use a generic M to denote the result of a medical test, where M could represent the marker value or predicted risk, or possibly some other continuous quantity summarizing the test result. The threshold, m, used to dichotomize the continuous test result is assumed to have the same range as M. We also assume that a larger M is more indicative of disease.

A binary test based on a continuous test result is defined as positive if $M \geq m$ and negative if $M < m$. The true and false positive fractions depend on the threshold m and are defined as $\text{TPF}(m) = \text{P}(M \geq m \mid D = 1)$ and $\text{FPF}(m) = \text{P}(M \geq m \mid D = 0)$. Definitions of sensitivity and specificity of continuous tests follow [Pepe, 2003].

Comparison of true and false positive fractions for a binary test is simple, since each test has one TPF and one FPF associated with it. When the test is based on a continuous test result, there are as many TPF and FPF as there are thresholds, making comparisons between tests using TPF and FPF difficult and threshold-dependent. The receiver operating characteristic (ROC) curve represents the entire set of true and false positive fractions obtained by dichotomizing a continuous test at all possible thresholds [Pepe, 1998]. The ROC curve was introduced to the field of medical diagnostic testing in the 1960's and since then it has become the best-developed statistical tool for describing the performance of tests with results on a continuous scale [Pepe, 2003]. The ROC is defined as $\text{ROC}(\cdot) = \{\text{TPF}(\text{FPF}^{-1}(u)), \quad u \in (-\infty, \infty)\}$. One of the advantages of the ROC curve is that it can be used to compare the discriminatory capacity of different markers and provides a valid approach to comparing markers even when they are measured on different scales [Heagerty et al., 2000].

The most widely used single-number summary of the ability of a diagnostic tool to discriminate between cases and controls is the area under the ROC curve (AUC). The AUC is defined as $\text{AUC} = \int_0^1 \text{ROC}(u)du$. For a perfect test the AUC = 1.0 and it is 0.5 for an uninformative test. The AUC measure can be interpreted as the probability that a randomly

chosen case subject will have a higher test result than a randomly chosen control subject: $P(M_D > M_{\bar{D}})$. The ROC, and consequently the AUC, is likely to be most useful in the development phase of a diagnostic test. Issues of cost, disease prevalence, as well as consequences of misdiagnosis will likely be considered when evaluating the usefulness of the test in practice [Pepe, 2000].

The AUC has been criticized for having little relevance to clinical practice, since clinicians are not usually attempting to determine which of two patients in a case-control pair is the case [Kerr et al., 2011, Pepe et al., 2013]. The use of the AUC to evaluate models that predict future risk or stratify individuals into risk categories has also been criticized. In this setting, calibration of the model plays an equally important role in the accurate assessment of risk [Cook, 2007]. Lastly, the AUC may be dominated by differences in risk distributions that are clinically irrelevant and thus may be insensitive to differences in risk distributions over more clinically relevant ranges [Pepe and Janes, 2013]. Pepe points out that practical irrelevance and insensitivity to clinically important differences in distributions apply not only to the AUC but to any measure of distance between case and control risk distributions [Pepe and Janes, 2013].

In the early 2000's, papers incorporating the time dimension into the methods to evaluate predictive biomarkers using ROC first appeared in the literature. Etzioni *et al.* [Etzioni et al., 1999] proposed two methods to model time-specific ROC curves using a biomarker measured longitudinally. The disease status was fixed, since these methods were motivated by a matched case-control study to evaluate the ability of PSA to predict prostate cancer at various times prior to diagnosis of the cases. Heagerty, Lumley and Pepe [Heagerty et al., 2000] tackled the complementary aspect of time-dependent ROC estimation. They proposed two estimators of time-dependent ROC for a continuous marker measured at baseline, but allowing the disease status to change over time. Sensitivity and specificity were defined in

terms of the marker threshold, c, and time at which disease status was ascertained, t:

$$\text{sens}(c,t) = \text{P}(Y_i > c \mid D_i(t) = 1) \qquad \text{spec}(c,t) = \text{P}(Y_i \leq c \mid D_i(t) = 0)$$

where Y_i is the marker value at baseline and $D_i(t)$ denotes disease status of subject i at time t. Using these definitions of sensitivity and specificity, they were able to define the corresponding ROC curve for any time t, ROC(t) [Heagerty et al., 2000]. The concept of sensitivity and specificity in this study was further explored by Heagerty and Zheng [Heagerty and Zheng, 2005]. They defined cumulative and incident sensitivity, and static and dynamic specificity as:

$$\begin{aligned} \text{sens}^C(c,t) &= \text{P}(Y_i > c \mid T_i < t) & \text{spec}^S(c,t) &= \text{P}(Y_i \leq c \mid T_i > t^\star) \\ \text{sens}^I(c,t) &= \text{P}(Y_i > c \mid T_i = t) & \text{spec}^D(c,t) &= \text{P}(Y_i \leq c \mid T_i > t) \end{aligned}$$

where T_i denotes the event time of subject i and $t^\star$ is a fixed follow-up time, so $T_i > t^\star$ denotes a long-term survivor. Focusing on incident/dynamic (I/D) ROC curves, they developed estimators under proportional and non-proportional hazards, and showed that these can be obtained from standard Cox regression output [Heagerty and Zheng, 2005].

Time-dependent ROC estimation for a longitudinal marker and time-dependent disease status was developed by Zheng and Heagerty [Zheng and Heagerty, 2004]. They focused on incident ROC curves using sensitivity and specificity extended to recognize explicitly the timing of the marker measurement and the timing of the disease. Sensitivity and specificity were defined as:

$$\begin{aligned} \text{sens}(s, c \mid t, \mathbf{Z}_i) &= \text{P}(X_i(s) > c \mid T_i = t, T_i > s, \mathbf{Z}_i) \\ \text{spec}(s, c \mid t, \mathbf{Z}_i) &= \text{P}(X_i(s) \leq c \mid T_i > t^\star, T_i > s, \mathbf{Z}_i) \end{aligned}$$

where s is the marker measurement time and $\mathbf{Z}_i$ denotes a vector of baseline covariates. In 2007 they extended the estimation procedure of the time-dependent ROC to account for censoring [Zheng and Heagerty, 2007].

Another interesting measure for evaluating the calibration of predictive models has its beginnings in weather prediction in the 1950's. The Brier score [Brier, 1950] measures the deviation of the predicted event probability from the observed event status and it is defined as the average of squared differences between the prediction and the observed outcome. The Brier score utilizes a square loss function, though other loss functions could potentially be used. Four decades later, the Brier score was adapted to the survival setting [Korn and Simon, 1990, Graf et al., 1999, Gerds and Schumacher, 2006] and it became known as the *prediction error*. Soon after, it was adapted to a longitudinal setting where its definition was now conditional on being at risk at the time of prediction, thus it was referred to as *conditional prediction error* [Schoop et al., 2008, Schoop et al., 2011]. An average of the prediction error over the time period for which predictions are made was proposed as a summary measure of prediction error over time [Schoop et al., 2008].

In 2011, Pfeiffer and Gail proposed two measures for evaluating the ability of a predictive model to discriminate between cases and controls that are geared towards use in a clinical setting, providing intuition and clinically relevant interpretation [Pfeiffer and Gail, 2011]. The two measures proposed were: the proportion of cases followed (PCF) and the proportion of the population needed to be followed (PNF). The PCF(q) represents the estimated proportion of cases that would be captured if we followed proportion q of the population at highest risk. Larger values of PCF(q) indicate better performance. PNF(p) represents the estimated proportion of the population at highest risk that would need to be followed in order to capture proportion p of the cases. Smaller values of PNF(p) indicate better performance. To define PCF, let ψ_q denote a risk threshold such that $\mathrm{P}(R > \psi_q) = q$, where R denotes predicted risk. Then $\mathrm{PCF}(q) = \mathrm{P}(R > \psi_q \mid T \leq t_p)$, where t_p denotes the time for which the risk prediction is made. To define PNF, let ψ_p denote a risk threshold such

that $\mathrm{P}(R > \psi_p \mid T \leq t_p) = p$, then $\mathrm{PNF}(q) = \mathrm{P}(R > \psi_p)$. The advantages of these measures include their ease of interpretation and their relevance in a clinical setting. They provide information that is relevant not only for comparing biomarkers or predictive models, but also in study design and feasibility studies [Pfeiffer and Gail, 2011]. Methods for estimation of these measures in a longitudinal setting have not yet been developed.

2.2.1 Remarks

Predictive biomarkers differ from diagnostic biomarker in that they are associated with the disease in the early stages of its progression. Often, as is the case with prostate-specific antigen (PSA), a single biomarker can serve as a predictive and diagnostic biomarker, though the analysis differs between the two applications.

We focus on methods to evaluate biomarkers that can be used to predict the onset of disease. There are many applications for such methods. Predictive biomarkers have great potential for guiding treatment decisions for slowly progressing diseases, such as AIDS, heart disease or some types of cancer. They can also be used in guiding preventive treatment for diseases such as end stage renal disease (ESRD), diabetes or heart disease. For many of these diseases, it is the longitudinal trend of a marker associated with the disease process that is most informative about where in the disease process the subject currently is.

The most studied measures to evaluate the predictive capacity of a biomarker are the ROC and its related measures: the sensitivity and specificity of a test, as well as the area under the ROC curve (AUC). Methods to estimate these measures in a typical setting (baseline biomarker and fixed disease status), a time-dependent setting (baseline biomarker and time-dependent disease status) and the longitudinal setting (longitudinal biomarker and time-dependent disease status) have been developed. Though these measures are immensely useful, they provide information about discrimination, or how accurately does the biomarker (or a predictive model based on it) classify subjects into future cases and controls. Though

discrimination ability is a greatly important attribute of a biomarker (predictive model), there are other aspects of the performance of a predictive biomarker (predictive model) that we would like to be able to quantify.

Two measures that we believe have enormous potential for public health applications are the *proportion of cases followed* (PCF) and the *proportion of the population needed to be followed* (PNF). These measures are very useful in feasibility studies and are directly relevant to public health decisions. In short, the PNF estimates the proportion of the population that would need to be followed in order to capture a given proportion of the cases. Alternatively, the PCF estimates the proportion of cases that would be captured if a given proportion of the population at highest risk for a given disease were to be followed. From a public health standpoint, these are very useful when making decisions about resource allocation or feasibility of large scale preventive interventions.

Methods to estimate PCF and PNF in a time-dependent setting (time-dependent disease status) or based on a longitudinal marker, or both, do not exist. Such methods are not only needed, but are highly applicable to public health. Though these measures are relatively new, proposed only in 2011, we anticipate their popularity growing as was the case with the ROC in the early 1980's.

In the second part of the dissertation (Chapter 4) we develop estimators of these measures in a longitudinal setting with a time-dependent disease status, as well as provide resampling-based procedures for making statistical inference. We evaluate their performance using extensive simulation studies and illustrate their performance on the ESRDS dataset.

2.3 Methods for prediction and evaluation of predictions in two-phase studies

The term case-cohort study design for survival analysis was coined by Ross Prentice in 1986 [Prentice, 1986]. It is a type of a two-phase study design, where the sampling of the study subjects is performed in two phases. In the first phase, a cohort is defined and inexpensive exposure information is collected for everyone in the cohort. In the second phase a subset of

the cohort is sampled and full covariate histories, such as expensive to measure biomarkers, are obtained only for the subjects in the subcohort. Another widely used two-phase study design is a nested case-control study [Thomas, 1977]. The main motivation behind two-phase study designs is their cost effectiveness. Large cohort studies are expensive and produce information that is largely redundant, especially when the outcome is rare [Self and Prentice, 1988]. By reducing the redundancy, the financial, computational and logistical burdens can be substantially reduced.

The sampling for the case-cohort (CCH) and the nested case-control (NCC) study designs are as follows. In the CCH study, a sub-cohort of a desired size is randomly selected from the full cohort, then the subjects who experienced the event of interest (cases) at any point during the study are added to the analysis group. In the NCC study, one or more control subjects are randomly selected from the risk set of each case at the time of failure of that case, thus the analysis group comprises all cases and their control sets.

The main advantage of the CCH over the NCC study design is that the same set of controls can be used in analysis of multiple outcomes and the analysis can be performed on different time scales. It also lends itself very well to prevention trials, since the sub-cohort is identified at the start of the study and the achievement of the intervention goals can be monitored in the sub-cohort [Self and Prentice, 1988, Barlow et al., 1999]. The disadvantage of CCH design is that the variance estimation is more complex compared to that for the NCC design. In the CCH design, the individual contributions to the pseudolikelihood are not independent due to sampling and therefore special procedures are needed for estimation and inference. In the NCC design, estimation and inference can be made based on partial likelihood or, equivalently, conditional logistic regression models [Prentice, 1986, Barlow et al., 1999].

Estimation procedures focus on estimation of the proportional hazards regression parameters, as well as their variance. Prentice [Prentice, 1986] proposed estimators of β as well as the cumulative baseline hazard and informally showed that they are consistent and

asymptotically normal. This was indeed the case and was proven shortly after by Self and Prentice in 1988, with a slight modification of the pseudolikelihood proposed by Prentice [Prentice, 1986]. Their proofs utilized martingale and finite sample convergence results [Self and Prentice, 1988]. The form of the Self and Prentice variance estimator was difficult to use in practice [Barlow et al., 1999, Therneau and Li, 1999] and in 1994 Barlow proposed a robust variance estimator for the CCH study design [Barlow, 1994]. Barlow's robust variance estimator used a jackknife estimate of the variance of the individual influence function and was shown to be equivalent to a robust variance estimator proposed by Lin and Wei [Lin and Wei, 1989] for the standard Cox model [Barlow, 1994].

With the availability of statistical software that supports offsets, or weighting of subjects, *dfbeta* residuals and the *start* and *stop* notation, obtaining parameter estimates and their variances in a CCH setting is as easy as it is in a NCC setting [Barlow et al., 1999, Therneau and Li, 1999]. With the theoretical and computational advances, the drawback related to the complex variance estimation in CCH studies has lost its significance over the years. However, one of the remaining major drawbacks of these methods was that they ignored the information on cohort members not samples as cases or controls [Breslow et al., 2009].

In 2000, Borgan *et al.* [Borgan et al., 2000] proposed a study design where the subcohort is selected by stratified random sampling, stratified on a correlate of exposure of interest that is available for all cohort members. Referred to as *exposure stratified case-cohort* design, this method provides improved efficiency compared to a case-cohort study design by sampling the cohort to over-represent more highly exposed subjects [Borgan et al., 2000], information that could potentially be used to improve the efficiency of estimation in CCH studies.

Over the next several years, considerable attention was given to improving the efficiency of estimators for two-phase designs by using the first-phase data [Kulich and Lin, 2004], relaxing assumptions that are likely to be unrealistic in practice [Chen, 2002, Nan, 2004] and using time-dependent sampling weights [Borgan et al., 2000, Kulich and Lin, 2004]. More general two-phase sampling was considered by Lin with focus on survey sampling

[Lin, 2000], and Breslow and Wellner for both finite and Bernoulli sampling [Breslow and Wellner, 2007, Breslow and Wellner, 2008]. The authors weighted each subject's contribution to the score function and the Breslow estimator of the baseline cumulative hazard by inverse probability of selection into the cohort, and derived consistent and asymptotically normal estimators of the regression coefficients and the cumulative baseline hazard for the Cox model [Lin, 2000, Breslow and Wellner, 2007, Lin, 2007, Breslow and Wellner, 2008]. Breslow and Wellner developed a theory of weighted likelihood estimation in semiparametric models that is general, encompasses previous results such as those of Lin [Lin, 2000] and Borgan *et al.* [Borgan et al., 2000], and provides a foundation for development of further applications [Breslow and Wellner, 2007].

To date, most CCH literature has focused on the estimation of the Cox model regression parameters, while estimation of summary measures to evaluate biomarkers or predictive models in the CCH setting remained an open problem. In 2012, Liu *et al.* [Liu et al., 2012] developed statistical methods to evaluate the accuracy and predictiveness of a risk prediction biomarker measured at baseline, with censored time-to-event outcome under stratified case-cohort sampling. The authors cast the problem of estimation of predictive summary measures into a missing data framework and used the inverse probability weighting (IPW) approach to account for the outcome dependent missingness induced by the CCH design. Liu *et al.* proposed a class of estimators based on IPW estimators for several measures commonly used in biomarker validation studies. They established consistency and asymptotic normality of their estimators by relying on the results of Breslow and Wellner [Breslow and Wellner, 2007, Breslow and Wellner, 2008] and standard empirical process theory [Pollard, 1990, Liu et al., 2012].

2.3.1 Remarks

Most studies utilizing the case-cohort study design have focused on estimating the association between exposure covariate and time-to-event outcomes and thus, on estimation of model parameters. The main advantage of a case-cohort design is its cost effectiveness with little, of any, loss of efficiency. This advantage extends beyond estimation of model parameters and into methods to evaluate risk prediction biomarkers and models. The need for such methods is substantial in practice. Many diseases that develop slowly, over the course of several years or decades, are rare. Examples include end-stage renal disease and type 1 diabetes.

Studying biomarkers that could be used to predict the risk of rare diseases is, if not infeasible, highly impractical using the cohort study design. In addition, it is often the longitudinal trajectory of the biomarker that is informative about where in the disease process the subject currently is. That, together with possibly other covariates, can be used to determine the subject's risk of the disease. Thus, methods to evaluate longitudinal predictive biomarkers in a case-cohort study design setting would be applicable in practice, but none exist at this time. Thanks to several methodological developments in recent years, development of such methods is now feasible.

In the third part of the dissertation (Chapter 5) we develop estimators of PE, TPF, FPF, AUC, PCF and PNF (prediction error, true positive fraction, false positive fraction, the area under the receiver operating characteristic curve, proportion of cases followed and proportion of the population to follow, respectively) in a longitudinal setting with a time-dependent disease status under case-cohort and nested case-control study designs. We develop resampling-based variance estimators of these estimators. We evaluate their performance using simulation studies and illustrate their performance on a nested case-control study of hepatocellular carcinoma within the HALT-C clinical trial.

2.4 Perturbation: a resampling-based variance estimation method

Perturbation is a resampling-based variance estimation method [Rao and Zhao, 1992, Parzen et al., 1994, Jin et al., 2001] originally developed as an alternative to the bootstrap [Efron, 1979]. It was studied as early as the 1980's under names such as Bayesian bootstrap, random weighting method or randomly weighted bootstrap [Rao and Zhao, 1992].

In general, resampling-based variance estimation approaches are useful when the estimate of a quantity of interest can be obtained, but an analytical expression for the variance is difficult to estimate, for example when the estimating function for a vector of parameters is not smooth [Tian et al., 2004]. Variance estimation using perturbation involves weighting, or perturbing, the estimand directly with weights that are independent and identically distributed random variables from a known distribution. Conditional on the data, the only random quantities in the perturbed estimand are the perturbation weights [Rao and Zhao, 1992, Jin et al., 2001, Jin et al., 2003].

Perturbation has been used to estimate variance in a wide variety of problems: Rao and Zhao approximated the distribution of M-estimates in linear models using perturbation [Rao and Zhao, 1992], and Parzen, Wei and Ying used it for making inference for a parameter estimated using non-smooth pivotal estimating equations [Parzen et al., 1994]. Jin, Ying and Wei addressed the problem of variance estimation for a vector of parameters in a semi-parametric setting where the parameters are estimated by optimizing an objective function with a U-process structure [Jin et al., 2001]. Specifically, they addressed situations where the estimating function is not continuous and solving the estimating function requires numerical methods, which can be challenging if the parameter vector is large [Jin et al., 2001]. Jin *et al.* used a perturbed convex loss function to approximate the distributions of rank-based estimators in an accelerated failure time (AFT) model with censoring [Jin et al., 2003]. Park and Wei, and Cai, Tian and Wei studied variations of that problem and additionally provided methods for estimating pointwise confidence intervals for the predicted survival

function based on the AFT model with censoring [Park and Wei, 2003, Cai et al., 2005].

Uno *et al.* considered the problem of evaluating prediction rules based on baseline biomarkers, censored time-to-event outcomes and general working models of the type $\mathrm{P}(T \leq t \mid Z) = g(\beta' Z)$ where $g(\cdot)$ is a known, strictly increasing, differentiable function, β is a p-dimensional vector of unknown parameters and Z is a set of baseline predictors [Uno et al., 2007]. They used IPW to account for censoring in the estimation of the model parameters. The evaluation measure considered was the overall misclassification rate (OMR), which is defined as $D_n(c) = \mathrm{E}|\mathrm{I}(T_\circ \leq t) - \mathrm{I}(g(\widehat{\beta}' Z_\circ) > c)|$ where the expectation is taken over $\{(X_i, Z_i, \Delta_i)\}$ and $(T_\circ, Z_\circ)$, and $\circ$ represents data for a new individual. Suppose that as $n \to \infty$, $\widehat{\beta}$ converges to a constant vector β_0, then $D_n(c) \to D(c)$, where $D(c) = \mathrm{E}|\mathrm{I}(T_\circ \leq t) - \mathrm{I}(g(\beta_0' Z_\circ) > c)|$. The authors provided an IPW estimator for $D(c)$:

$$\widehat{D}(c) = \frac{1}{n} \sum_{i=1}^{n} \frac{w_i}{\widehat{G}(X_i \wedge t)} |\mathrm{I}(X_i \leq t) - \mathrm{I}(g(\widehat{\beta}' Z_i) > c)|$$

where $w_i = \mathrm{I}(T_i \wedge t \leq C_i) = \mathrm{I}(X_i \leq t)\Delta_i + \mathrm{I}(X_i > t)$ and $\widehat{G}(\cdot)$ is the Kaplan-Meier estimator of $G(\cdot)$. Perturbing that estimator, they provided $\widehat{D}^*(c)$, a perturbed version of $\widehat{D}(c)$:

$$\widehat{D}^*(c) = \frac{1}{n} \sum_{i=1}^{n} V_i \left(\frac{w_i}{\widehat{G}^*(X_i \wedge t)} |\mathrm{I}(X_i \leq t) - \mathrm{I}(g(Z_i' \widehat{\beta}^*) > c)| \right)$$

where $\{V_i, i = 1, \ldots, n\}$ represent n independent positive random variables V_i from a known distribution with mean and variance equal to 1, $\widehat{G}^*(\cdot)$ and $\widehat{\beta}^*$ are the corresponding perturbed versions of $\widehat{G}(\cdot)$ and $\widehat{\beta}$.

To construct $\widehat{G}^*(\cdot)$, [Uno et al., 2007] *et al.* used the martingale representation formula for the Kaplan-Meier estimate [Fleming and Harrington, 1991]. Specifically, for $t > 0$, the unconditional distribution of $\widehat{G}(t) - G(t)$ can be approximated by the conditional distribution (given the data) of $-\widehat{G}(t) \sum_{i=1}^{n} V_i \int_0^t \left(\sum_{j=1}^{n} \mathrm{I}(X_j > s) \right)^{-1} d\widehat{M}_i(s)$ where $\widehat{M}_i(t) = \mathrm{I}(X_i \leq t, \Delta_i = 0) - \int_0^t \mathrm{I}(X_i > s) d\widehat{\Lambda}(s)$ and $\widehat{\Lambda}(\cdot)$ is the standard Nelson-Aalen estimator

of the cumulative hazard function for the censoring variable C. Then, $\widehat{G}^*(t) = \widehat{G}(t) - \widehat{G}(t)\sum_{i=1}^n V_i \int_0^t \big(\sum_{j=1}^n \mathrm{I}(X_j > s)\big)^{-1} d\widehat{M}_i(s)$.

A perturbed $\widehat{\beta}$, $\widehat{\beta}^*$, can be estimated by solving $\widehat{U}^*(\beta) = \frac{1}{n}\sum_{i=1}^n \frac{w_i}{\widehat{G}^*(X_i \wedge t)} Z_i \big(\mathrm{I}(X_i \le t) - g(\beta' Z_i)\big) V_i = 0$. The authors note that because $U(\beta)$ is a differentiable function in β, $(\widehat{\beta}^* - \widehat{\beta})$ could also be obtained by perturbing the first-order expansion of $\sqrt{n}(\widehat{\beta} - \beta_0)$.

Uno *et al.* show that a standardized transformation of $\widehat{D}(\widehat{c})$

$$\sqrt{n}\left(\log(-\log(\widehat{D}(\widehat{c}))) - \log(-\log(D_0))\right) \tag{2.4}$$

is asymptotically equivalent to

$$\big(\widehat{D}(\widehat{c})\log(\widehat{D}(\widehat{c}))\big)^{-1} W$$

where $W = \sqrt{n}(\widehat{D}(\widehat{c}) - D_0)$, $D(c_0) = D_0$, c_0 is a minimizer of $D(c)$ and $\widehat{c}$ is a minimizer of $\widehat{D}(c)$, $0 \le c \le 1$. Using arguments in [Park and Wei, 2003, Cai et al., 2005], Uno *et al.* show that the distribution of 2.4 can be approximated by the conditional distribution of

$$\big(\widehat{D}(\widehat{c})\log(\widehat{D}(\widehat{c}))\big)^{-1} \widehat{W}^*$$

given the data, where $\widehat{W}^* = \sqrt{n}\left(\widehat{D}^*(\widehat{c}) - \widehat{D}(\widehat{c})\right)$. In practice, the distribution of W can be approximated by generating a large number P of random samples $\widehat{W}^*$.

Uno *et al.* used similar arguments to show that perturbation inference procedures provide valid variance estimates for other measures of predictive accuracy, such as sensitivity and specificity. Specifically, for the prediction rule $\mathrm{I}(g(\widehat{\beta}' Z) > c)$, the sensitivity is $\mathrm{SE}(c) = \mathrm{P}(g(\beta_0' Z_\circ) > c \mid T_\circ \le t)$ and the specificity is $\mathrm{SP}(c) = \mathrm{P}(g(\beta_0' Z_\circ) \le c \mid T_\circ > t)$.

These conditional probabilities can be estimated consistently by

$$\widehat{\mathrm{SE}}(c) = \frac{\sum_{i=1}^{n} \Delta_i \mathrm{I}(g(\widehat{\beta}' Z_i) > c, X_i \leq t)/\widehat{G}(X_i)}{\sum_{i=1}^{n} \Delta_i \mathrm{I}(X_i \leq t)/\widehat{G}(X_i)}$$

and

$$\widehat{\mathrm{SP}}(c) = \frac{\sum_{i=1}^{n} \mathrm{I}(g(\widehat{\beta}' Z_i) \leq c, X_i > t)}{\sum_{i=1}^{n} \mathrm{I}(X_i > t)}$$

Processes $\sqrt{n}(\widehat{\mathrm{SE}}(c) - \mathrm{SE}(c))$ and $\sqrt{n}(\widehat{\mathrm{SP}}(c) - \mathrm{SP}(c))$ converge weakly to a mean-0 Gaussian process ([Uno et al., 2007] Appendix D).

Suppose that a threshold c^+ is chosen such that $SE(c^+) = \gamma$. Uno *et al.* show that when $\mathrm{P}(T \leq t \mid \beta_0' Z = y)$ is positive for y in the support of $\beta_0' Z$, c^+ is unique between 0 and 1.

Let $\widehat{c}^+$ be a solution to $\widehat{\mathrm{SE}}(c) = \gamma$, then $\widehat{c}^+$ is consistent for c^+. Furthermore, $\widehat{\mathrm{SE}}(c^+) \to \mathrm{SE}(c^+)$. Uno *et al.* use perturbation to obtain the estimated standard error of $\widehat{\mathrm{SE}}(\widehat{c}^+)$, or its transformation. Specifically, the perturbed $\widehat{\mathrm{SE}}(c)$ is

$$\widehat{\mathrm{SE}}^*(c) = \frac{\sum_{i=1}^{n} \Delta_i \mathrm{I}(g(Z_i' \widehat{\beta}^*) > c, X_i \leq t) V_i / \widehat{G}^*(X_i)}{\sum_{i=1}^{n} \Delta_i \mathrm{I}(X_i \leq t) V_i / \widehat{G}^*(X_i)}$$

The perturbed $\widehat{\mathrm{SP}}(c)$, $\widehat{\mathrm{SP}}^*(c)$, can be obtained similarly.

Now, let $\widehat{c}^*$ be a solution to the equation $\widehat{\mathrm{SE}}^*(\widehat{c}^*) = \gamma$. Following similar arguments as those given for the OMR and the weak convergence of the process $\sqrt{n}(\widehat{\mathrm{SE}}(c) - \mathrm{SE}(c))$, Uno *et al.* show that when n is large, the distribution of $\sqrt{n}(\widehat{\mathrm{SE}}(\widehat{c}^+) - \mathrm{SE}(c^+))$ can be well approximated by the conditional distribution $\sqrt{n}(\widehat{\mathrm{SE}}^*(\widehat{c}^*) - \widehat{\mathrm{SE}}(\widehat{c}^+))$ given the data. Similarly for $\sqrt{n}(\widehat{\mathrm{SP}}(\widehat{c}^+) - \mathrm{SP}(c^+))$.

Finally, Uno *et al.* argue that any reasonable summary of prediction precision constructed from $\mathrm{SE}(c)$ and $\mathrm{SP}(c)$, for example area under the ROC, can be estimated consistently through $\widehat{\mathrm{SE}}(c)$ and $\widehat{\mathrm{SP}}(c)$ and a large-sample approximation can be obtained based on $\widehat{\mathrm{SE}}^*(c)$ and $\widehat{\mathrm{SP}}^*(c)$ [Uno et al., 2007]. These results paved the way for development of in-

ference procedures for semiparametric and nonparametric time-dependent predictive values of prognostic biomarkers with time-to-event outcomes [Zheng et al., 2008]. Also, this is the approach that we base our inference procedures on for making inference under longitudinal cohort studies (Chapter 4).

Recently, Cai and Zheng used perturbation to estimate the variance of accuracy measures estimated nonparametrically under nested case-control (NCC) sampling [Cai and Zheng, 2013]. Finite sampling in the second phase of two-phase sampling induces a correlation between the ξ_i's, the sampling indicators, where $\xi_i = 1$ if individual i was sampled into the second phase of the study. This poses a challenge for resampling-based variance estimation methods such as the bootstrap, which is unable to account for this induced dependence [Cai and Zheng, 2013]. Recently, Saegusa proposed a weighted bootstrap that involves weighting the bootstrap samples by a product of two weights that correspond to randomness from each sampling phase and each stratum [Saegusa, 2014]. The results are applicable to exposure stratified case-control studies and stratified case-cohort (sCCH) studies, but not NCC studies. NCC sampling is performed repeatedly from a risk set at each event time, thus it induces a much more complex dependence structure in the data [Cai and Zheng, 2013] than does stratified case-cohort sampling. Cai and Zheng showed that using perturbation for variance estimation we can recover the correct variance of IPW estimators based on two-phase samples under finite sampling, including NCC studies [Cai and Zheng, 2013]. This provides the theoretical justification for our proposed perturbation methods under longitudinal two-phase studies (Chapter 5).

2.5 Summary

The vast majority of research on evaluation of biomarkers deals with diagnostic markers and tests, where both the marker and the outcome are ascertained at the same, single time point. Interest in evaluation of predictive markers let to research not only on risk prediction but also in evaluation of predictions. Thus, methods to evaluate predictive markers measured

at baseline, where the outcome was ascertained at some later time began to appear in the literature in the early 2000's. Interest in evaluation of predictiveness of markers measured longitudinally to inform about events in the future is even more recent, and that is the setting we are interested in.

Table 2.1 summarizes the literature describing the progress made recently in the area of evaluation of predictive biomarkers measured at baseline, or at a later time s, and highlights the areas or research which we aim to make progress in though the work described in this dissertation, as indicated by * in Table 2.1.

We tackle various aspects of the longitudinal prediction and evaluation, starting with the development of flexible and robust models for longitudinal predictions (Chapter 3), evaluation of longitudinal risk predictions in a cohort setting (Chapter 4) and case-cohort and nested case-control study designs (Chapter 5). Conclusions, remarks and questions generated by the work described in the following chapters to be addressed in the future constitute Chapter 6.

Table 2.1: Summary table of recent literature on evaluation of predictive biomarkers measured at baseline or longitudinally, under cohort, case-cohort and nested case-control study designs. PE = prediction error, TPF = true positive fraction, FPF = false positive fraction, ROC = receiver operating characteristic curve, AUC = area under the ROC curve, PCF = proportion of cases followed, PNF = proportion of the population followed. We denote the areas that we address in this dissertation with *.

	Cohort			
	Marker measured at baseline (Time-dependent setting)		Marker measured up to time s (Longitudinal setting)	
Measure	Estimation	Inference	Estimation	Inference
PE	[Schoop et al., 2008]	-	[Schoop et al., 2008]	*
	[Schoop et al., 2011]	-	[Schoop et al., 2011]	*
TPF/FPF/ROC/AUC	[Heagerty et al., 2000]	-	[Cai et al., 2006][a]	
	[Zheng and Heagerty, 2004]		[Zheng and Heagerty, 2007][a]	
	[Heagerty and Zheng, 2005]	[Uno et al., 2007]	*	*
PCF/PNF	-	-	*	*
	Case-cohort			
	Time-dependent setting		Longitudinal setting	
Measure	Estimation	Inference	Estimation	Inference
PE	-	-	*	*
TPF/FPF/ROC/AUC	[Liu et al., 2012]	-	*	*
PCF/PNF	-	-	*	*
	Nested case-control			
	Time-dependent setting		Longitudinal setting	
Measure	Estimation	Inference	Estimation	Inference
PE	-	-	*	*
TPF/FPF/ROC/AUC	[Cai and Zheng, 2011, Cai and Zheng, 2013]		*	*
PCF/PNF	-	-	*	*

[a]evaluation of markers measured at the prediction time, $Y(s)$, without incorporating any other covariate information.

Chapter 3

PREDICTION BASED ON LONGITUDINAL AND TIME-TO-EVENT DATA: MODELING CHOICES AND EVALUATION

In this chapter we develop flexible and robust estimation for partly conditional survival models for longitudinal data with time-to-event outcomes. We evaluate risk predictions obtained using these models with the current gold standard, the joint model, using simulation studies and illustrate their performance on the ESRDS dataset.

3.1 Introduction

Long term follow-up is common in many medical investigations where the interest lies in predicting patients' risks for a future adverse outcome using biomarkers repeatedly measured over time. Serial measurements provide information on the marker level and its changes over time, thus are often more informative regarding the occurrence of an outcome of interest at a future time, compared with prediction based only on information collected at baseline. A key quantity of risk prediction in longitudinal setting is the probability of developing an adverse outcome in the next τ_0 time interval given survival up to time s and covariate information available up to time s:

$$R\big(\tau_0 \mid s, \mathbf{H}(s)\big) = \mathrm{P}\big(T \leq s + \tau_0 \mid T > s, \mathbf{H}(s)\big),$$

where T is the time to the outcome of interest and $\mathbf{H}(s)$ contains both time-constant and longitudinal covariate history up to time s. Such information can be used to identify subjects at high risk who can be targeted for preventive strategy or treatment plan. Similarly, less

frequent follow-up may be recommended to low risk individuals. The primary inferential objective is to identify predictive algorithms for risk of an adverse outcome at a future time based on relevant longitudinal processes up to the time of prediction, and to assess the utility of such an algorithm for use in aiding medical decisions regarding disease monitoring plan and treatment choices.

The use of repeated biomarkers presents some unique challenges in outcome prediction analysis. The effects of biomarkers often vary with time. Variation due to measurement error in biomarkers often leads to attenuated estimates of the relative risk parameter in a Cox model that is biased towards zero [Prentice, 1982]. Little is known about how such bias affects the predictions of future risk. In survivorship analysis, serial markers are typically incorporated in a time-varying covariate Cox regression model [Crowley and Hu, 1977]. Since the marker is usually measured only at discrete time points, time-specific measurements for the marker may not be available for members in the risk set when a failure occurs. To address such difficulties, joint modeling (JM) for both the marker process and survival data have been developed in recent years [Tsiatis et al., 1995, Faucett and Thomas, 1996, Wulfsohn and Tsiatis, 1997]. Estimation of the parameters can be done through a two-stage approach [Pawitan, 1993, Bycott and Taylor, 1998] or a full likelihood-based approach [De Gruttola and Tu, 1994, Faucett and Thomas, 1996, Henderson et al., 2000, Wang and Taylor, 2001]. Both approaches involve specifying the longitudinal marker process via a random effect submodel or a latent class submodel, and then linking the failure time process via a time-varying covariate hazard submodel of the latent biomarker values without measurement error.

The main focus in the joint modeling literature has been on inferring how biomarkers change over time and the estimation of coefficients of the association between the marker and the disease, such as the hazard ratio, of a clinical outcome based on biomarkers. Adopting the joint modeling approach for prediction of disease progression has recently been considered in the literature [Proust-Lima and Taylor, 2009, Rizopoulos, 2011, Taylor et al., 2013]. However the calculation of updated risks, $R(\tau_0 \mid s, \mathbf{H}(s))$, in this setting is not straightfor-

ward and involves fairly complicated model specification and a computational procedure that requires integration over the marker process, and the computational burden increases drastically when multiple biomarkers are under consideration. Furthermore, valid prediction depends critically on the correct specification of both submodels within the JM framework. Methodology that is computationally simple yet sufficiently flexible to capture updated risks over time is appealing from a practical standpoint. A reliable procedure is also needed for calculating the updated accuracy summaries.

In order to develop a survival model that directly facilitates the estimation of $R(\tau_0 \mid s, \mathbf{H}(s))$ for any specific pair of (s, τ_0), Zheng and Heagerty (2005) proposed a class of semi-parametric models, called 'partly conditional survival models' (PC), where the conditional hazard of failure is specified by conditioning on only a part of the marker trajectory at select 'landmark' times. Specifically, a marker measured at time s, $Y(s)$, generates a data record with time origin for T reset to s, i.e., the time scale is now in terms of 'time since measurement'. The marker measurement at time s can now be regarded as a baseline or 'external' measure. As a result, there can be multiple derived event times for each individual corresponding to her/his repeatedly measured marker values. Thus, the PC approach casts the problem of using longitudinal marker values for predicting survival within the general framework of multivariate survival models, and avoids using a time-varying covariate Cox model for risk prediction. The PC approach does not require a modeling specification regarding the event time as a function of the entire marker process. Implementation of the partly conditional model is relatively simple using slightly modified standard statistical software [Zheng and Heagerty, 2005]. A similar approach based on the idea of a landmark time was considered by [van Houwelingen, 2007]. A thorough investigation of different approaches to prediction using longitudinal markers is needed in order to provide practical guidance to researchers in terms of the choice of estimation methods.

To make a comparison among longitudinal prediction tools, it is important to gauge model performance based on their intended clinical utilities. The evaluation of the performance of a

medical test has been based on receiver operating characteristic (ROC) curves and calculation of time-dependent ROC curves to evaluate a longitudinal marker has been considered in [Zheng and Heagerty, 2004]. More recently other metrics have been proposed for risk model evaluation with a binary outcome (e.g., [Pencina et al., 2008, Gu and Pepe, 2009, Pfeiffer and Gail, 2011]). In this manuscript we will compare the performance of updated prediction rules developed based on either a JM or various PC approaches, adopting metrics for assessing the performance of prediction algorithms to the setting with longitudinal markers and survival outcomes.

The goals of this chapter are to provide flexible tools for calculating updated risk for patients under medical monitoring over time, and to investigate the performance of predictive algorithms derived from either the joint modeling or the partly conditional modeling approaches. In particular, we propose two simple, yet flexible, modeling methods within the partly conditional modeling framework. The methods have the advantage of ease of implementation and are robust to violations in modeling assumptions. To provide practical guidance in the choice among the proposed methods, we conduct extensive numerical studies to compare prediction rules derived from the different approaches using an array of summary measures of predictive performance. We introduce general models in Section 3.2. We describe estimation and inference procedures of proposed methods in Section 3.3 and 3.4. Measures of predictiveness for model comparison are presented in section 3.5.4. The results of numerical studies of finite sample performance of our proposed procedures and predictive performance comparison are presented in Section 3.5. We illustrate the performance of our methods on the End Stage Renal Disease Study (ESRDS) conducted at the Community Health Network in San Francisco in Section 3.6, and provide a summary in the final section. We refer to the notation introduced in section 2.1.1, and introduce additional notation as needed.

3.2 Modeling longitudinal data with time-to-event outcomes

3.2.1 Joint model

Joint modeling framework for longitudinal and time-to-event data [Tsiatis et al., 1995, Faucett and Thomas, 1996, Wulfsohn and Tsiatis, 1997, Henderson et al., 2000] allows the full representation of the joint likelihood given the observed data, $\mathcal{D}_n$ [Tsiatis and Davidian, 2004]. It comprises two linked sub-models, one for the underlying 'true' biomarker process as a function of time, and one linking the failure time T_i to the underlying 'true' biomarker process.

To model the biomarker process, a mixed effects model is often considered: $\mathbf{Y}_i = \mathbf{Y}_i^* + \mathbf{e}_i$, where $\mathbf{Y}_i^* = \mathbf{U}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i$ is a vector of the underlying 'true' biomarker values for subject i, $\boldsymbol{\beta}$ denotes a $(p \times 1)$ vector of fixed effects, $\mathbf{b}_i$ is a $(q \times 1)$ vector of the random effects assumed to be normally distributed with mean 0 and variance $\boldsymbol{\Sigma}$, $\mathbf{U}_i$ is a $(m_i \times p)$ matrix of covariates and $\mathbf{W}_i$ is a $(m_i \times q)$ matrix of covariate data for individual i. The measurement error $\mathbf{e}_i = \{e_i(s_{i1}), \ldots, e_i(s_{im_i})\}$ and $e_i(s_{ij})$ is assumed to be distributed $N(0, \sigma_e^2)$, $\text{cov}(e_i(s_{ij}), e_i(s_{ij'})) = 0$ for $j \neq j'$, and is independent of $\mathbf{b}_i$.

The failure time process is usually modeled using a time-varying covariate proportional hazards model [Cox, 1972] with hazard function

$$\lambda_i(t|Y_i^*(t), \mathbf{Z}_i) = \frac{1}{\Delta t} \lim_{\Delta t \to 0} P(t \leq T_i < t + \Delta t \mid T_i \geq t,\, Y_i^*(t), \mathbf{Z}_i)$$

typically modeled by a proportional hazards regression model $\lambda_i(t) = \lambda_0(t)\exp(Y_i^*(t)\eta + \mathbf{Z}_i\boldsymbol{\gamma})$ [Wulfsohn and Tsiatis, 1997, Tsiatis and Davidian, 2004]. Since both model specifications involve unknown quantities, fitting involves iteration between the linear mixed effects sub-model and the survival submodel.

For a future individual with $\mathbf{H}(s) = \mathbf{H}_\circ(s) = [\mathbf{Z}_\circ, \mathbf{Y}_\circ(s), \mathbf{s}_\circ(s)]$ and event free at time s, one may predict $R(\tau_0|s, \mathbf{H}_\circ(s))$ as $R^{\text{JM}}(\tau_0|s, \mathbf{H}_\circ(s)) = P(s < T_\circ \leq s + \tau_0 \mid T_\circ > s, \mathbf{H}(s) =$

$\mathbf{H}_\circ(s), \mathcal{D}_n, \theta)$, with θ denoting the parameter vector of the joint model. [Rizopoulos, 2011] showed that $R^{\text{JM}}(\tau_0|s, \mathbf{H}_\circ(s))$ can be rewritten as:

$$\begin{aligned} R^{\text{JM}}(\tau_0|s, \mathbf{H}_\circ(s)) &= 1 - P(T > s + \tau_0 \mid T > s, \mathbf{H}(s) = \mathbf{H}_\circ(s), \mathcal{D}_n, \theta) \\ &= 1 - \int \frac{S_\circ(s + \tau_0 \mid \mathbf{Y}^*_\circ(\tau_0 + s), \mathbf{Z}_\circ; \theta)}{S_\circ(s \mid \mathbf{Y}^*_\circ(s), \mathbf{Z}_\circ; \theta)} p(b \mid T > s, \mathbf{H}(s) = \mathbf{H}_\circ(s); \theta) db \end{aligned}$$

where $S_\circ(\cdot) = \exp\left(-\int_0^t \lambda(u|Y^*_\circ(u),\, \mathbf{Z}_\circ;\, \theta)du\right)$. Point and interval estimates can be obtained using Monte Carlo simulations [Proust-Lima and Taylor, 2009, Rizopoulos, 2011], thus prediction for an individual requires complex computation.

3.2.2 PC models

In a partly conditional survival model framework, at time s, we are interested in predicting the residual lifetime $\mathcal{T}_s = T - s \mid \mathcal{T}_s > 0$ using covariate history up to s based on $\mathbf{H}(s)$. To directly model such an updated risks, the analysis time of a PC model is such that the measurement time becomes the baseline time and estimation is based on the derived failure times calculated from each measurement time. Specifically, at s_{ij}, the observed data vector for subject i is $\mathcal{O}_{ij} = \{X_{ij}, \delta_i, \mathbf{H}(s_{ij})\}$, where $X_{ij} = X_i - s_{ij}$. A PC model then specifies the relationship between $\mathcal{T}_{ij} = T_i - s_{ij}$ and $\mathbf{H}(s_{ij})$ without having to specify the full marker process. One advantage of such an approach is its robustness, since we only need to specify a working model for the residual lifetime as a function of s_{ij} and $\mathbf{H}(s_{ij})$. Computationally it is also much easier to implement compared to a JM model. Below we introduce two estimating procedures for PC models that vary in efficiency and flexibility.

3.3 Two estimating approaches for PC models

3.3.1 Partly conditional Cox-type model (PC_{Cox})

To make a prediction about an event within $\mathcal{T}_s \leq \tau_0$ for a given τ_0, one approach considered in [Zheng and Heagerty, 2005] is to assume

$$\lambda(\tau|\mathbf{H}_i(s_{ij})) = \lambda_0(\tau)\exp\big(\boldsymbol{\alpha}'\mathbf{B}(s_{ij}) + \boldsymbol{\gamma}'\mathbf{Z}_i + \eta(\tau)'h(\mathbf{Y}_i(s_{ij}))\big) \tag{3.1}$$

where $\lambda_0(\cdot)$ is an unknown baseline hazard, $\mathbf{B}(s)$ is a spline basis function of s which captures the potentially non-linear effect of measurement time s_{ij} on the risk. $\eta(\tau)$ allows the marker effect to vary over τ, the time since measurement. $h(\mathbf{Y}_i(s_{ij}))$ is a functional of $\mathbf{Y}_i(s_{ij})$, for example it can be simply $Y_i(s_{ij})$, or a predicted response profile approximating $Y_i^*(s_{ij})$ at s_{ij} based on LMEM estimates of fixed and random effects, their variance components and observed data up to time s_{ij} for individual i. A kernel-based estimating procedure [Cai and Sun, 2003] can be used to obtain estimators in model (3.1). To further reduce the computational burden in fitting the model, we consider instead a similar model:

$$\begin{aligned}\lambda(\tau|\mathbf{H}_i(s_{ij})) &= \lambda_0(\tau)\exp\big(\boldsymbol{\alpha}'\mathbf{B}(s_{ij}) + \boldsymbol{\gamma}'\mathbf{Z}_i + \eta'\mathbf{B}(\tau)h(\mathbf{Y}_i(s_{ij}))\big) \\ &= \lambda_0(\tau)\exp(\boldsymbol{\theta}'_{\text{Cox}}\mathbf{H}_i^B(s_{ij},\tau))\end{aligned} \tag{3.2}$$

where $\boldsymbol{\theta}_{\text{Cox}} = [\boldsymbol{\alpha}', \boldsymbol{\gamma}', \eta']'_{P\times 1}$ and $\mathbf{H}_i^B(s_{ij},\tau) = [\mathbf{B}(s_{ij})', \mathbf{Z}_i', \mathbf{B}(\tau)h(\mathbf{Y}_i(s_{ij}))']'$.

Estimators of the model parameters, $\boldsymbol{\theta}_{\text{Cox}}$, can be obtained by maximizing the following marginal log partial likelihood:

$$\mathcal{L}(\boldsymbol{\theta}_{\text{Cox}}) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\int\left(\boldsymbol{\theta}'_{\text{Cox}}\mathbf{H}_i^B(s_{ij},X_{ij}) - log\left[\sum_{k=1}^{n}\sum_{l=1}^{m_k}\mathrm{I}(X_{kl}\geq X_{ij})\exp\left(\boldsymbol{\theta}'_{\text{Cox}}\mathbf{H}_i^B(s_{kl},X_{ij})\right)\right]\right)dN_{ij}(u) \tag{3.3}$$

where $N_{ij}(u) = \mathrm{I}(\delta_i = 1, X_{ij}\leq u)$.

For a new subject with covariate history $\mathbf{H}_\circ(s)$ up to time s, our key predictive probability within time τ_0 from s can be estimated as

$$\widehat{R}^{\mathrm{PC}}_{\mathrm{Cox}}(\tau_0 \mid s, \mathbf{H}_\circ(s), \widehat{\boldsymbol{\theta}}_{\mathrm{Cox}}) = 1 - \exp(-\widehat{\Lambda}(\tau_0 \mid s, \mathbf{H}_\circ(s), \widehat{\boldsymbol{\theta}}_{\mathrm{Cox}})),$$

with

$$\widehat{\Lambda}(\tau_0|\mathbf{H}_\circ(s), \widehat{\boldsymbol{\theta}}_{\mathrm{Cox}}) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\int_0^{\tau_0} \frac{\exp(\widehat{\boldsymbol{\theta}}'_{\mathrm{Cox}}\mathbf{H}^B_\circ(s_{ij}, u))}{\sum_{k=1}^{n}\sum_{l=1}^{m_k}\mathrm{I}(X_{kl} \geq X_{ij})\exp(\widehat{\boldsymbol{\theta}}'_{\mathrm{Cox}}\mathbf{H}^B_k(s_{kl}, u))} dN_{ij}(u),$$

3.3.2 Partly conditional generalized linear model (PC_{GLM})

The above model based on a Cox-type partly conditional model can be further relaxed to allow effect of measurement time to vary with prediction time τ. To this end, we propose to approximate $R(\tau_0 \mid s, \mathbf{H}_i(s))$ through a working *marginal* partly conditional generalized linear model with a binary outcome ($\mathrm{PC}_{\mathrm{GLM}}$),

$$\begin{aligned} P\big(\mathcal{T}_{ij} \leq \tau_0 | s_{ij}, \mathbf{H}_i(s_{ij}), \mathcal{T}_{ij} > 0\big) &= g\big(\boldsymbol{\alpha}'\mathbf{B}(s_{ij}) + \boldsymbol{\gamma}'\mathbf{Z}_i + \eta' h(\mathbf{Y}_i(s_{ij}))\big) \\ &= g\big(\boldsymbol{\theta}'_{\mathrm{GLM}}\mathbf{H}^B_i(s_{ij})\big), \end{aligned} \tag{3.4}$$

where $\mathcal{T}_{ij} = T_i - s_{ij}$, and $\mathbf{H}^B_i(s_{ij}) = [\mathbf{B}(s_{ij})', \mathbf{Z}'_i, h(\mathbf{Y}_i(s_{ij}))']'$, $g(\cdot)$ is a link function such as anti-logit, and $\boldsymbol{\theta}_{\mathrm{GLM}} = [\boldsymbol{\alpha}', \boldsymbol{\gamma}', \eta']'_{P\times 1}$. We may also include possible interactions between $\mathbf{B}(s_{ij})$ and the other predictors to allow for time-varying effects of the predictors. In the presence of censoring, we propose to estimate $\boldsymbol{\theta}_{\mathrm{GLM}}$ by solving a weighted estimating equation:

$$\mathcal{U}(\boldsymbol{\theta}_{\mathrm{GLM}}) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\widehat{w}^{(C)}_{ij}\frac{\dot{g}(\boldsymbol{\theta}'_{\mathrm{GLM}}\mathbf{H}^B_i(s_{ij}))}{g(\boldsymbol{\theta}'_{\mathrm{GLM}}\mathbf{H}^B_i(s_{ij}))\left(1 - g(\boldsymbol{\theta}'_{\mathrm{GLM}}\mathbf{H}^B_i(s_{ij}))\right)}\left(\mathrm{I}(X_{ij} \leq \tau_0) - g(\boldsymbol{\theta}'_{\mathrm{GLM}}\mathbf{H}^B_i(s_{ij}))\right), \tag{3.5}$$

where

$$\dot{g}(\boldsymbol{\theta}'_{\text{GLM}}\mathbf{H}_i^B(s_{ij})) = \frac{\partial g(\boldsymbol{\theta}'_{\text{GLM}}\mathbf{H}_i^B(s_{ij}))}{\partial \boldsymbol{\theta}_{\text{GLM}}}$$

and

$$\widehat{w}_{ij}^{(C)} = \delta_i \mathrm{I}(T_i \leq s_{ij} + \tau_0)\frac{1}{\widehat{G}(T_i)} + \mathrm{I}(X_i > s_{ij} + \tau_0)\frac{1}{\widehat{G}(s_{ij} + \tau_0)}$$

accounts for censoring and $\widehat{G}(\cdot)$ is the Kaplan-Meier estimate of $G(\cdot)$, the survival function of C. Such time-specific logistic model has been previously studied, for example, by [Zheng et al., 2006] and [Uno et al., 2007] with baseline covariates.

In both PC models, If the dimension of $\boldsymbol{\theta}_{\text{Cox}}$ or $\boldsymbol{\theta}_{\text{GLM}}$ is not small relative to the number of events, we will maximize a penalized log-likelihood $\mathcal{L}(\boldsymbol{\theta}) - \lambda \sum_{j=2}^{P} \mathcal{P}(|\theta_j|)$, where $\mathcal{P}(\cdot)$ is a pre-specified penalty function such as the L_1 and L_2 penalty. The tuning parameter $\lambda \geq 0$ controls the amount of penalty and can be selected via standard procedures such as the AIC, BIC or the cross-validation [Hastie et al., 2003].

Let $\widehat{\boldsymbol{\theta}}_{\text{GLM}}$ denote the final estimator of $\boldsymbol{\theta}$. Then for a future subject with covariate history up time s, $\mathbf{H}_\circ(s)$, the probability of an event within τ_0 time from s, $P(\mathcal{T}_s \leq \tau_0 | \mathcal{T}_s > 0, \mathbf{H}_\circ(s))$ can be estimated as

$$\widehat{R}^{\text{PC}}_{\text{GLM}}(\tau_0 \mid s, \mathbf{H}_\circ(s), \widehat{\boldsymbol{\theta}}_{\text{GLM}}) = g\left(\widehat{\boldsymbol{\theta}}'_{\text{GLM}}\mathbf{H}_\circ^B(s)\right).$$

3.3.3 Novel two-stage PC models

Biomarkers are often, if not always, measured with error. When the error term is nontrivial, using observed values at the time of prediction, $Y_i(s_{ij})$, often leads to attenuated effects and naturally it might impact the risk prediction as well. In the longitudinal setting, replacing $Y_i(s_{ij})$ with an estimated quantity that is less prone to error may result in better approximation of the true underlying risk. We therefore propose a two-stage procedure, where in the first stage we model the longitudinal process and calculate a fitted $\widehat{Y}_i(s_{ij})$ based on the

best linear unbiased predictor (BLUP) estimator. Specifically, under the following model for $\mathbf{Y}_i$:

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{U}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, n \\ \boldsymbol{\epsilon}_i &\sim N(\mathbf{0}, \sigma^2\mathbf{I}), \quad \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}(\boldsymbol{\phi})), \quad \mathbf{Y}_i \sim N\big(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i\big), \end{aligned}$$

where $\mathbf{U}_{i(m_i\times p)}$ and $\mathbf{W}_{i(m_i\times q)}$ are matrices of covariates, which are subsets of $(\mathbf{Z}_i, \mathbf{s}_i)$ for individual i, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_q)'$ denotes the parameter vector of the variance components of the random effects and $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{I} + \mathbf{W}_i\mathbf{D}(\boldsymbol{\phi})\mathbf{W}_i'$. Therefore the BLUP estimator is

$$\widehat{\mathbf{Y}}_i = \mathbf{U}_i\widehat{\boldsymbol{\beta}} + \mathbf{W}_i\mathbf{D}(\widehat{\boldsymbol{\phi}})\mathbf{W}_i'(\boldsymbol{\Sigma}_i)^{-1}(\mathbf{Y}_i - \mathbf{U}_i\widehat{\boldsymbol{\beta}}). \tag{3.6}$$

In the second stage, we obtain BLUP-based estimators of risk model parameters $\widehat{\boldsymbol{\theta}}_{\text{Cox}}^{\text{BLUP}}$ and $\widehat{\boldsymbol{\theta}}_{\text{GLM}}^{\text{BLUP}}$ by replacing $h(\mathbf{Y}_i(s_{ij}))$ with $h(\widehat{\mathbf{Y}}_i(s_{ij}))$ in Equation 3.3 or Equation 3.5, respectively.

For a new subject with $\mathbf{H}_\circ(s)$ observed up to time $s_{\circ j}$, the predicted random effect is

$$\widehat{\mathbf{b}}_\circ | s_{\circ j} = \mathbf{D}(\widehat{\boldsymbol{\phi}})\mathbf{W}_\circ'(s_{\circ j})(\widehat{\sigma}^2\mathbf{I} + \mathbf{W}_\circ(s_{\circ j})\mathbf{D}(\widehat{\boldsymbol{\phi}})\mathbf{W}_\circ'(s_{\circ j}))^{-1}(\mathbf{Y}_\circ(s_{\circ j}) - \mathbf{U}_\circ(s_{\circ j})\widehat{\boldsymbol{\beta}}),$$

where $\mathbf{Y}_\circ(s_{0j})$, $\mathbf{U}_\circ(s_{\circ j})$ and $\mathbf{W}_\circ(s_{\circ j})$ denotes the covariate data available up to time $s_{\circ j}$ for the new subject '$\circ$'. Then the fitted marker value is

$$\widehat{Y}_\circ(s_{\circ j}) = \mathbf{U}_\circ(s_{\circ j})\widehat{\boldsymbol{\beta}} + \mathbf{W}_\circ(s_{\circ j})(\widehat{\mathbf{b}}_\circ \mid s_{\circ j}). \tag{3.7}$$

Replacing $\mathbf{H}_\circ(s)$ with $\widehat{\mathbf{H}}_\circ(s) = [\mathbf{Z}_\circ, \widehat{\mathbf{Y}}_\circ(s), \mathbf{s}_\circ(s)]$ in $\widehat{R}_{\text{Cox}}^{\text{PC}}(\tau_0 \mid \mathbf{H}_\circ(s))$ or $\widehat{R}_{\text{GLM}}^{\text{PC}}(\tau_0 \mid \mathbf{H}_\circ(s))$ we can obtain BLUP-based predicted risk estimators $\widehat{R}_{\text{Cox}}^{\text{BLUP}}(\tau_0 \mid \mathbf{H}_\circ(s))$ or $\widehat{R}_{\text{GLM}}^{\text{BLUP}}(\tau_0 \mid \mathbf{H}_\circ(s))$ for the new subject. Note that such a two-stage approach is different from the two-stage estimation procedure in the classic joint modeling literature, in that a PC model rather than a time-varying covariate survival model is considered here for prediction. Furthermore, we

provide inference procedure in the following sections to account for the variations generated across all stages.

3.4 Inference for PC models

It is important to note that PC models are 'working' models that are used to approximate risk as a function of a given covariate. Likely, they are not correctly specified. On the other hand, under mild regularity conditions, the standard maximum likelihood (or partial likelihood) estimators of model parameters converge to a constant vector, as $n \to \infty$ (Hjort, 1992). This stability feature is essential for developing the large sample properties of estimators for $R_\circ(\tau_0 \mid s, \boldsymbol{\theta}_0, \mathbf{H}_\circ(s))$, where $\boldsymbol{\theta}_0$ is the true parameter value.

3.4.1 Inference for ordinary PC models

To make inference on $\widehat{R}^{\mathrm{PC}}_{\mathrm{Cox}}(\tau_0 \mid s, \mathbf{H}_\circ(s), \widehat{\boldsymbol{\theta}}_{\mathrm{Cox}})$, we note that the process for some fixed τ_0 is

$$\begin{aligned}
\mathcal{W}(\tau_0) &= n^{\frac{1}{2}} \left(\widehat{R}^{\mathrm{PC}}_{\mathrm{Cox}}(\tau_0 \mid s, \widehat{\boldsymbol{\theta}}_{\mathrm{Cox}}, \mathbf{H}_\circ(s)) - R^{\mathrm{PC}}_{\mathrm{Cox}}(\tau_0 \mid s, \boldsymbol{\theta}_0, \mathbf{H}_\circ(s)) \right) \\
&= n^{-\frac{1}{2}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \zeta_{ij}(\tau_0, s, \boldsymbol{\theta}_0, \mathbf{H}_\circ(s)) + o_p(1)
\end{aligned}$$

where

$$\begin{aligned}
\zeta_{ij}(\tau_0, s, \boldsymbol{\theta}_0, \mathbf{H}_\circ(s)) &= \left(1 - R^{\mathrm{PC}}_{\mathrm{Cox}}(\tau_0 \mid \boldsymbol{\theta}_0, \mathbf{H}_\circ(s))\right) \Bigg[\int_0^{\tau_0} \frac{\exp(\boldsymbol{\theta}_0' \mathbf{H}^B_\circ(s,u)) dM_{ij}(u)}{\mathbb{S}^{(0)}(\boldsymbol{\theta}_0, u)} \\
&+ V(\boldsymbol{\theta}_0, \tau_0)' \mathbf{I}^{-1}(\boldsymbol{\theta}_0) \int_0^{\infty} \left(\mathbf{H}^B_i(s_{ij}, u) - \mathcal{E}(\boldsymbol{\theta}_0, u) \right) dM_{ij}(u) \Bigg], \\
V(\boldsymbol{\theta}_0, \tau_0) &= -\int_0^{\tau_0} \left(\frac{\mathbf{H}^B_\circ(s,u) - \mathcal{E}(\boldsymbol{\theta}_0, u)}{\mathbb{S}^{(0)}(\boldsymbol{\theta}, u)} \right) dE(N_{ij}(u)), \\
\mathbf{I}(\boldsymbol{\theta}_0) &= \int_0^{\infty} \left(\frac{\mathbb{S}^{(2)}(\boldsymbol{\theta}_0, u)}{\mathbb{S}^{(0)}(\boldsymbol{\theta}_0, u)} - \mathcal{E}(\boldsymbol{\theta}_0, u)^{\otimes 2} \right) dE(N_{ij}(u))
\end{aligned}$$

where

$$\mathbb{S}^{(b)}(\boldsymbol{\theta}_0, u) = E(\mathrm{I}(u \leq X_{ij})\mathbf{H}_i^B(s_{ij}, u)^{\otimes b} \exp\left(\boldsymbol{\theta}_0' \mathbf{H}_i^B(s_{ij}, u)\right))$$

and $\mathcal{E}(\boldsymbol{\theta}_0, u) = \frac{\mathbb{S}^{(1)}(\boldsymbol{\theta}_0, u)}{\mathbb{S}^{(0)}(\boldsymbol{\theta}_0, u)}$ for any vector $\boldsymbol{a}$, $\boldsymbol{a}^{\otimes 0} = 1$, $\boldsymbol{a}^{\otimes 1} = \boldsymbol{a}$ and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}'$ and

$$M_{ij}(t) = N_{ij}(t) - \int_0^t \mathrm{I}(X_{ij} \geq u) d\Lambda(\boldsymbol{\theta}_0, u).$$

It can be shown that $\mathcal{W}(\tau_0)$ converges weakly to a Gaussian process.

To approximate the distribution of $\mathcal{W}(\tau_0)$, we propose a resampling-based method via stochastic perturbation [Parzen et al., 1994, Jin et al., 2001, Uno et al., 2007]. Let $V^{(b)} = \{V_1^{(b)}, \ldots, V_n^{(b)}\}$ be independent random variables from the exponential distribution with mean and variance equal to 1, then for $b = 1, \ldots B_p$, one may obtain the perturbed estimates

$$\mathcal{W}_{(b)}^*(\tau_0) = \sum_{i=1}^{n} \left(1 - V_i^{(b)}\right) \sum_{j=1}^{m_i} \zeta_{ij}(\tau_0, s, \boldsymbol{\theta}_0, \mathbf{H}_\circ(s))$$

With a large number of realizations of $\mathcal{W}_{(b)}^*(\tau_0)$, one may obtain the variance of $\widehat{R}_{\mathrm{Cox}}^{\mathrm{PC}}(\tau_0 | s, \widehat{\boldsymbol{\theta}}, \mathbf{H}_\circ(s))$ as $\widehat{\sigma}^2 = \frac{1}{B_p - 1} \sum_{b=1}^{B_p} \widehat{\mathcal{W}}_{(b)}^{*2}(\tau_0)$, where $\widehat{\mathcal{W}}_{(b)}^*(\tau_0)$ is obtained by replacing all the unknown parameters in $\mathcal{W}_{(b)}^*(\tau_0)$ by their estimated counterparts.

To make inference on $\widehat{R}_{\mathrm{GLM}}^{\mathrm{PC}}(\tau_0 | s, \mathbf{H}_\circ(s), \widehat{\boldsymbol{\theta}}_{\mathrm{GLM}})$, one may first obtain the perturbed estimates $\widehat{\boldsymbol{\theta}}_{\mathrm{GLM}}^{(b)}$ as the solution to

$$\begin{aligned} \mathcal{U}^{(b)}(\boldsymbol{\theta}_{\mathrm{GLM}}) &= \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{V_i^{(b)}\, \widehat{w}_{ij}^{(C)(b)}\, \dot{g}(\boldsymbol{\theta}_{\mathrm{GLM}}' \mathbf{H}_i^B(s_{ij}))}{g(\boldsymbol{\theta}_{\mathrm{GLM}}' \mathbf{H}_i^B(s_{ij}))\left(1 - g(\boldsymbol{\theta}_{\mathrm{GLM}}' \mathbf{H}_i^B(s_{ij}))\right)} \qquad (3.8) \\ &\quad \times \left(\mathrm{I}(X_{ij} \leq \tau_0) - g(\boldsymbol{\theta}_{\mathrm{GLM}}' \mathbf{H}_i^B(s_{ij}))\right) \qquad (3.9) \end{aligned}$$

where

$$\widehat{w}_{ij}^{(C)(b)} = \delta_i\, \mathrm{I}(T_i < s_{ij} + \tau_0) \frac{1}{\widehat{G}^{(b)}(T_i)} + \mathrm{I}(X_i > s_{ij} + \tau_0) \frac{1}{\widehat{G}^{(b)}(s_{ij} + \tau_0)}$$

and $\widehat{G}^{(b)}(\cdot)$ corresponds to $\widehat{G}(\cdot)$ estimated with weights $\mathbf{V}^{(b)}$. The variance for $\widehat{R}^{\text{PC}}_{\text{GLM}}(\tau_0 \mid s, \mathbf{H}_\circ(s), \widehat{\boldsymbol{\theta}}_{\text{GLM}})$ can be calculated as the variance from the B_p realizations of $\widehat{R}^{\text{PC}}_{\text{GLM}}\big(\tau_0 \mid s, \mathbf{H}_\circ(s), \widehat{\boldsymbol{\theta}}^{(b)}_{\text{GLM}}\big)$.

3.4.2 Inference for BLUP-based PC Models

For two-stage based approaches, additional steps are needed to account for the variation in $\widehat{\mathbf{Y}}_i$, $\widehat{\mathbf{Y}}_\circ(s)$ due to variation in $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Phi}}_{((1+q)\times 1)} = (\widehat{\sigma}^2_{1\times 1}, \widehat{\boldsymbol{\phi}}'_{q\times 1})'$. We can again consider the perturbation approach. Specifically, we first obtain the b^{th} perturbed estimators as: $\widehat{\boldsymbol{\beta}}^{(b)}_{p\times 1} = \widehat{\boldsymbol{\beta}} + (\mathbf{U}'\boldsymbol{\Sigma}^{-1}\mathbf{U})^{-}\mathbf{U}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}^{(b)}$ with $\mathbf{U}_{N\times 1} = (\mathbf{U}'_1, \ldots, \mathbf{U}'_n)'$,

$$\boldsymbol{\Sigma}_{N\times N} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \cdots & \vdots \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_n \end{bmatrix}, \qquad \boldsymbol{\Psi}^{(b)}_{N\times 1} = \begin{bmatrix} V_1^{(b)}(\mathbf{Y}_1 - \mathbf{U}_1\widehat{\boldsymbol{\beta}}) \\ \vdots \\ V_n^{(b)}(\mathbf{Y}_n - \mathbf{U}_n\widehat{\boldsymbol{\beta}}) \end{bmatrix}$$

$\boldsymbol{\Sigma}_{i_{(m_i\times m_i)}} = \widehat{\sigma}^2\mathbf{I} + \mathbf{W}_i\mathbf{D}(\widehat{\boldsymbol{\phi}})\mathbf{W}_i$, n denotes the number of subjects and $N = \sum_{i=1}^{n} m_i$ is the number of observations. $\boldsymbol{\Phi}^{(b)}_{(1+q)\times 1} = \widehat{\boldsymbol{\Phi}} + \mathbb{I}^{-1}_{\boldsymbol{\Phi}\boldsymbol{\Phi}}\Omega^{(b)}_{\boldsymbol{\Phi}}$, where $\mathbb{I}_{\boldsymbol{\Phi}\boldsymbol{\Phi}}$ is a $(1+q)\times(1+q)$ information matrix for the variance components from a restricted or a regular maximum likelihood, $L(\boldsymbol{\beta}, \boldsymbol{\Phi})$, and $\Omega^{(b)}_{\boldsymbol{\Phi}}$ is the corresponding perturbed score equation. With a restricted maximum likelihood,

$$\boldsymbol{\Omega}^{(b)}_{\boldsymbol{\Phi}} = \begin{bmatrix} \Omega^{(b)}_{\Phi_1} = \Omega^{(b)}_{\sigma^2} \\ \Omega^{(b)}_{\Phi_2} = \Omega^{(b)}_{\phi_1} \\ \vdots \\ \Omega^{(b)}_{\Phi_{1+q}} = \Omega^{(b)}_{\phi_q} \end{bmatrix}$$

$$\mathbb{I}_{\mathbf{\Phi\Phi}} = \begin{bmatrix} \mathbb{I}_{\Phi_1\Phi_1} = \mathbb{I}_{\sigma^2\sigma^2} & \mathbb{I}_{\Phi_1\Phi_2} = \mathbb{I}_{\sigma^2\phi_1} & \cdots & \mathbb{I}_{\Phi_1\Phi_{1+q}} = \mathbb{I}_{\sigma^2\phi_q} \\ \mathbb{I}_{\Phi_2\Phi_1} = \mathbb{I}_{\phi_1\sigma^2} & \mathbb{I}_{\Phi_2\Phi_2} = \mathbb{I}_{\phi_1\phi_1} & \cdots & \mathbb{I}_{\Phi_2\Phi_{1+q}} = \mathbb{I}_{\phi_1\phi_q} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{I}_{\Phi_{1+q}\Phi_1} = \mathbb{I}_{\phi_q\sigma^2} & \mathbb{I}_{\Phi_{1+q}\Phi_2} = \mathbb{I}_{\phi_q\phi_1} & \cdots & \mathbb{I}_{\Phi_{1+q}\Phi_{1+q}} = \mathbb{I}_{\phi_q\phi_q} \end{bmatrix}$$

$\Omega^{(b)}_{\Phi_k} = -\frac{1}{2}tr[\mathbf{P}(\partial\mathbf{\Sigma}/\partial\Phi_k)] + \frac{1}{2}\sum_{i=1}^{n} V_i^{(b)}(\mathbf{Y}_i - \mathbf{U}_i\widehat{\boldsymbol{\beta}})'\mathbf{\Sigma}_i^{-1}(\partial\mathbf{\Sigma}_i/\partial\Phi_k)\mathbf{\Sigma}_i^{-1}(\mathbf{Y}_i - \mathbf{U}_i\widehat{\boldsymbol{\beta}})$, $k = 1,\ldots,(1+q)$. $\mathbb{I}_{\Phi_i,\Phi_j} = -\frac{1}{2}tr[\mathbf{P}(\partial\mathbf{\Sigma}/\partial\Phi_i)\mathbf{P}(\partial\mathbf{\Sigma}/\partial\Phi_j)]$ with $i = (1,\ldots,(1+q))$ and $j = (1,\ldots,(1+q))$, $\mathbf{P} = \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{U}(\mathbf{U}'\mathbf{\Sigma}^{-1}\mathbf{U})^{-}\mathbf{U}'\mathbf{\Sigma}^{-1}$. Plugging in $\boldsymbol{\beta}^{(b)}$ and $\mathbf{\Phi}^{(b)}$ in Equation 3.6 and 3.7, respectively, yields perturbed $\widehat{\mathbf{Y}}_i^{(b)}$ and $\widehat{\mathbf{Y}}_\circ^{(b)}(s)$.

To obtain confidence intervals for $\widehat{R}^{\text{BLUP}}_{\text{Cox}}(\tau_0 \mid \mathbf{H}_\circ(s))$, we note that

$$\begin{aligned}
\mathcal{W}(\tau_0)^{\text{BLUP}}_{\text{Cox}} &= n^{\frac{1}{2}}\left(\widehat{R}^{\text{BLUP}}_{\text{Cox}}(\tau_0 \mid s, \mathbf{H}_\circ(s)) - R^{\text{BLUP}}_{\text{Cox}}(\tau_0 \mid s, \mathbf{H}_\circ(s))\right) \\
&= n^{\frac{1}{2}}\left(\widehat{R}^{\text{BLUP}}_{\text{Cox}}(\tau_0 \mid s, \mathbf{H}_\circ(s), \widehat{\boldsymbol{\theta}}^{\text{BLUP}}_{\text{Cox}}, \boldsymbol{\beta}_0, \mathbf{\Phi}_0) - R^{\text{BLUP}}_{\text{Cox}}(\tau_0 \mid s, \mathbf{H}_\circ(s), \boldsymbol{\theta}_0, \boldsymbol{\beta}_0, \mathbf{\Phi}_0)\right) \\
&+ n^{\frac{1}{2}}\left(\widehat{R}^{\text{BLUP}}_{\text{Cox}}(\tau_0 \mid s, \mathbf{H}_\circ(s), \widehat{\boldsymbol{\theta}}^{\text{BLUP}}_{\text{Cox}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{\Phi}}) - \widehat{R}^{\text{BLUP}}_{\text{Cox}}(\tau_0 \mid s, \mathbf{H}_\circ(s), \widehat{\boldsymbol{\theta}}^{\text{BLUP}}_{\text{Cox}}, \boldsymbol{\beta}_0, \mathbf{\Phi}_0,)\right) + o_p(1) \\
&= n^{-\frac{1}{2}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\zeta_{ij}(\tau_0, s, \mathbf{H}_\circ(s), \boldsymbol{\theta}_0, \boldsymbol{\beta}_0, \mathbf{\Phi}_0) \\
&+ n^{\frac{1}{2}}\left(R^{\text{BLUP}}_{\text{Cox}}(\tau_0 \mid s, \mathbf{H}_\circ(s), \boldsymbol{\theta}_0, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{\Phi}}) - R^{\text{BLUP}}_{\text{Cox}}(\tau_0 \mid s, \mathbf{H}_\circ(s), \boldsymbol{\theta}_0, \boldsymbol{\beta}_0, \mathbf{\Phi}_0)\right) + o_p(1)
\end{aligned}$$

The distribution of $\mathcal{W}(\tau_0)^{\text{BLUP}}_{\text{Cox}}$ can be approximated with B_p random variates with its bth component as:

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{m_i}\zeta_{ij}(\tau_0, s, \mathbf{H}_\circ(s), \boldsymbol{\theta}_0, \boldsymbol{\beta}_0, \mathbf{\Phi}_0)(1 - V_i^{(b)}) \\
+ &\left(\widehat{R}^{\text{BLUP}}_{\text{Cox}}(\tau_0 \mid s, \mathbf{H}^{(b)}_\circ(s), \widehat{\boldsymbol{\theta}}^{\text{BLUP(b)}}_{\text{Cox}}, \widehat{\boldsymbol{\beta}}^{(b)}, \widehat{\mathbf{\Phi}}^{(b)}) - \widehat{R}^{\text{BLUP}}_{\text{Cox}}(\tau_0 \mid s, \mathbf{H}_\circ(s), \widehat{\boldsymbol{\theta}}^{\text{BLUP}}_{\text{Cox}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{\Phi}})\right),
\end{aligned}$$

where $\widehat{\boldsymbol{\theta}}^{\text{BLUP(b)}}_{\text{Cox}}$ is obtained by replacing $\mathbf{Y}_i(s_{ij})$ in Equation 3.3 with $\widehat{\mathbf{Y}}_i^{(b)}(s_{ij})$.

To make inference on $\widehat{R}^{\text{BLUP}}_{\text{GLM}}(\tau_0 \mid \mathbf{H}_\circ(s), \widehat{\boldsymbol{\theta}}^{\text{BLUP}}_{\text{GLM}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Phi}})$, we first estimate $\widehat{\boldsymbol{\theta}}^{\text{BLUP(b)}}_{\text{GLM}}$ by replacing $\mathbf{Y}_i(s_{ij})$ in Equation 3.8 with $\widehat{\mathbf{Y}}^{(b)}_i(s_{ij})$. Then, the variance for

$$\widehat{R}^{\text{BLUP}}_{\text{GLM}}(\tau_0 \mid s, \mathbf{H}_\circ(s), \widehat{\boldsymbol{\theta}}^{\text{BLUP}}_{\text{GLM}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Phi}})$$

can be calculated as the variance of the B_p realizations of

$$\widehat{R}^{\text{BLUP}}_{\text{GLM}}\big(\tau_0 \mid s, \mathbf{H}^{(b)}_\circ(s), \widehat{\boldsymbol{\theta}}^{\text{BLUP(b)}}_{\text{GLM}}, \widehat{\boldsymbol{\beta}}^{(b)}, \widehat{\boldsymbol{\Phi}}^{(b)}\big).$$

3.5 Simulations

Our simulations addressed two goals: to evaluate the finite sample performance of PC model-based estimation and inference procedures proposed in Sections 3.3 and 3.4; and to compare the prediction performance of the proposed PC models with that of the JM in terms of accuracy and discrimination.

3.5.1 Data generation

The longitudinal marker data was generated following the general setup used in the JM literature (e.g., [Tsiatis et al., 1995]). We assume a linear mixed effects model with measurement error: $Y_i(u) = W_i(u) + e_i(u) = \alpha_{0i} + \alpha_{1i} f(u) + e_i(u)$ where $f(u) = log(u/\nu_1)$, with a Weibull baseline hazard: $\lambda_0(u) = v/\nu_2(u/\nu_2)^{v-1}$ and $\nu_1 = 30$, scale $\nu_2 = 15$, shape $v = 1.4$. The random components $\alpha_i = (\alpha_{0i}, \alpha_{1i})$ were generated as a bivariate normal with mean $(\mu_{\alpha_0}, \mu_{\alpha_1})^T = (0.6, -0.1)^T$ and the covariance matrix $\boldsymbol{\Sigma}_\alpha = \begin{bmatrix} 0.83^2 & -0.005 \\ -0.005 & 0.13^2 \end{bmatrix}$. The measurement error $e_i(u) \sim_{iid} N(0, \sigma_e^2)$, $i = 1, ..., n$, $j = 1, ..., m_i$ and $\sigma_e = 0.1$ and 1.0, representing small and large measurement errors. Figure 3.1 shows the biomarker trajectories for subjects under the two simulated measurement error settings. Failure time was assumed to depend on the covariate (without error) through a proportional hazards

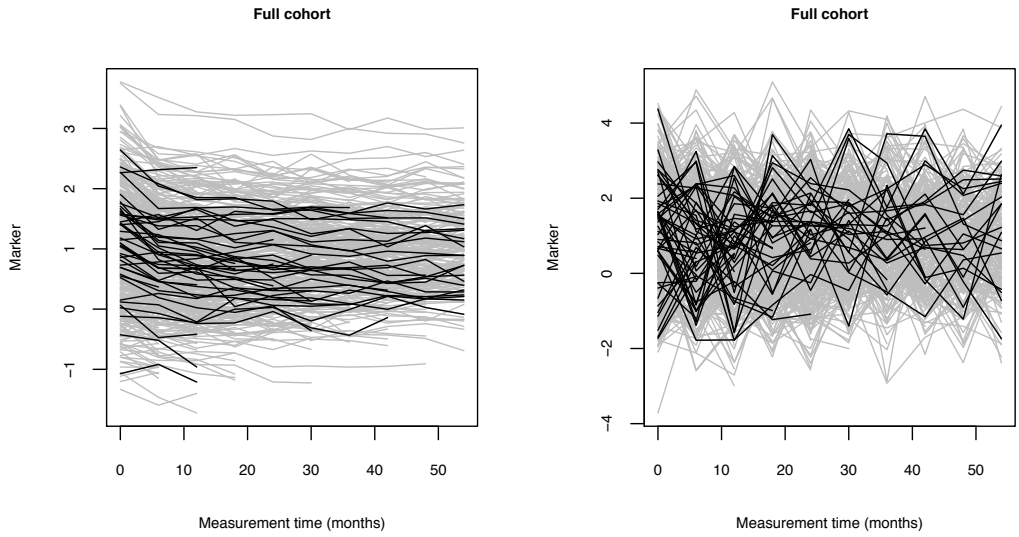

Figure 3.1: Example simulated dataset with $\mu_{\alpha_1} = -0.1$ and standard deviation of the measurement error $\sigma_e = 0.1$ (left panel) and $\sigma_e = 1.0$ (right panel).

relationship: $\lambda_i(u) = \lambda_0(u)\exp(\beta W_i(u))$, where $\beta = -1.5$. Censoring time was generated from a uniform distribution $c_i \sim U(0, 300)$, resulting in about 30% censoring. There were up to 10 measurements per subject taken at 6 month intervals and n = 500. We were interested in predicting the risk of an event at time $s + \tau_0$ given biomarker data up to time s: $R_i(\tau_0 \mid s) = P\big(T_i \leq s + \tau_0 \mid T_i > s, \mathbf{Y}_i(s)\big)$. All simulations were run in `R` (www.r-project.org). The joint model was fit using the `JM` package [Rizopoulos, 2010].

3.5.2 Models

In all simulations we used five models to estimate the conditional survival probability: a partly conditional Cox-type model (PC_{Cox}) (section 3.3.1), partly conditional generalized linear model with a logit link function (PC_{GLM}) (section 3.3.2), two-stage PC_{Cox} model (PC_{Cox} BLUP) and a two-stage PC_{GLM} model (PC_{GLM} BLUP) (section 3.3.3) and a joint

model with a linear mixed effects submodel with random intercepts and slopes (JM) (section 3.2.1). For the joint model fitting we used a function provided in the `R` package `JM` [Rizopoulos, 2010] with the *weibull-PH-GH* option, which fits a relative risk model with a Weibull baseline risk function and uses the Gauss-Hermite integration rule to approximate the integrals. Note that in all simulations considered, JMs were fit with specifications that corresponded to the true models.

3.5.3 Longitudinal prediction and inference

Longitudinal prediction and inference for $\mathbf{PC_{GLM}}$ and $\mathbf{PC_{Cox}}$: simulation setup. In the first set of simulations we investigated the point and variance estimates of predicted partly conditional risk (PCR) as derived by PC_{GLM} and PC_{Cox}. For PC_{GLM} and PC_{Cox} we compare the $R(\tau_0|s, \mathbf{H}_\circ(s))$ for an individual with $Y_\circ = 0$ or $Y_\circ = 1$ (Figures 3.2 and 3.3) at time s, for four sets of s and τ_0, the time of prediction and length of the prediction window, respectively, for small and large magnitudes of the measurement error ($\sigma_e = \{0.1, 1.0\}$). For PC_{GLM} and PC_{Cox} the true values of $R(\tau_0|s, Y_\circ = 0)$ and $R(\tau_0|s, Y_\circ = 1)$ were obtained empirically, by simulating a large dataset (n = 5,000,000) without censoring and empirically estimating the proportion of survivors at $s + \tau_0$ given survival up to time s, for subjects with marker values close to 0, (0 ± 0.05) and 1, (1 ± 0.05).

Longitudinal prediction and inference for $\mathbf{PC_{GLM}}$ and $\mathbf{PC_{Cox}}$: results. As seen in Tables 3.1 and 3.2, both PC_{GLM} and PC_{Cox}, even though they are 'working' models and only approximating the true model, produce PCR that are quite close to the truth, and the perturbation based asymptotic standard error (ASE) estimators approximate the empirical standard errors (ESD) closely. PC_{GLM} tends to have smaller bias, however PC_{Cox} tends to have a smaller mean squared error (MSE). The performance of the two models is similar for $\sigma_e = 0.1$ (Table 3.1) and $\sigma_e = 1.0$ (Table 3.2).

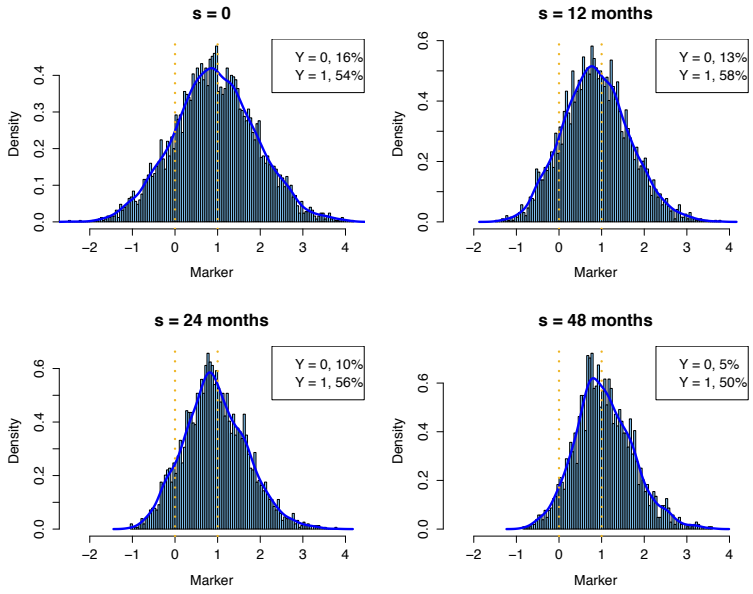

Figure 3.2: Density of marker values from a simulated dataset (n = 5000) with $\sigma_e = 0.1$ and $\mu_{\alpha_1} = -0.1$. Proportions of individuals with marker values below 0 and 1 for different conditioning times s are show in the top right of each panel, with the four panels corresponding to $s = \{0, 12, 24, 48\}$ months.

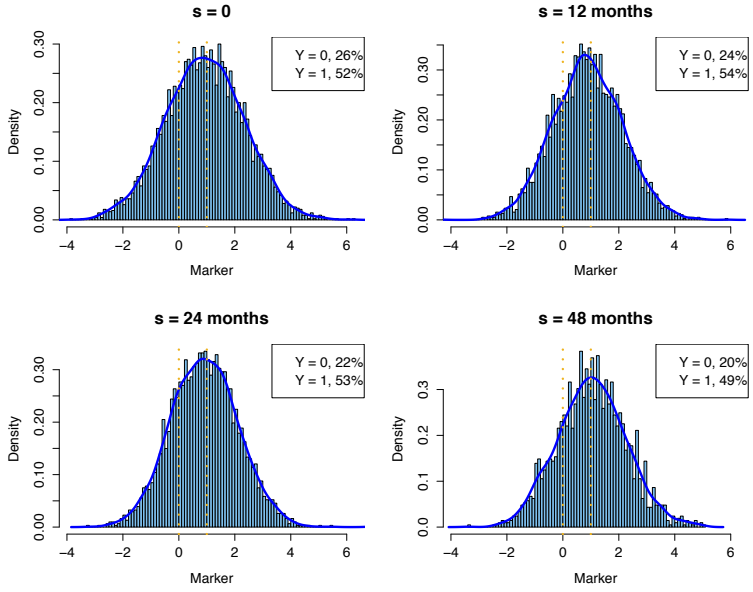

Figure 3.3: Density of marker values from a simulated dataset (n = 5000) with $\sigma_e = 1.0$ and $\mu_{\alpha_1} = -0.1$. Proportions of individuals with marker values below 0 and 1 for different conditioning times s are show in the top right of each panel, with the four panels corresponding to $s = \{0, 12, 24, 48\}$ months.

Longitudinal prediction and inference for two-stage $\mathbf{PC_{GLM}}$ (BLUP) model: simulation setup. We then evaluated the point and interval estimates of PCR based on a two-stage $\mathrm{PC_{GLM}}$ (BLUP) model. For $\mathrm{PC_{GLM}}$ BLUP we estimated $R_{\text{o}}(\tau_0|s)$ for two randomly selected subjects with marker trajectories up to s = 12 and s = 24 months, and $6 \leq \tau_0 \leq 36$ months. Based on the individual's observed values up to $s = 12$ and 24, we obtained the BLUP marker values and made risk predictions for each individual based on those. To approximate true values, we estimated PCRs for the two subjects based on their trajectories up to s using a joint model specified according to the true underlying model and used a sample size of n=10,000.

Longitudinal prediction and inference for two-stage $\mathbf{PC_{GLM}}$ (BLUP) model: results. The PCRs are quite close to the 'truth' and the perturbation-based point-wise CIs trace the empirical CIs estimated over 500 simulated datasets quite well. This adds confidence that the proposed variance estimation procedure can be used in practice for making inference regarding patient's risk at the personal level (Figure 3.4).

3.5.4 Comparison and evaluation of predictive performance

In this section we compare the predictive performance of five longitudinal models for estimating the conditional survival probability ($\mathrm{PC_{Cox}}$, $\mathrm{PC_{GLM}}$, $\mathrm{PC_{Cox}}$ BLUP, $\mathrm{PC_{GLM}}$ BLUP and a joint model), using measures of predictive accuracy and discrimination. We first introduce the evaluation measures used in our simulations, the simulation setup and results follow.

The development of the accuracy and discrimination measures will be described in detail in Chapter 4. We now briefly introduce the measures that will be used in the simulations described in this section and refer the reader to Chapter 4 for more details on their development.

Measures of prediction calibration *Prediction error* (PE) is a measure of calibration of a risk prediction model, that is it quantifies the distance between risk predictions and the true risk. In a longitudinal setting, it is defined as the expected quadratic loss between the predicted probability of an event in a τ_0 time interval past s and the observed event in that timeframe conditional on being at risk at time s [Schoop et al., 2008, Schoop et al., 2011]. The estimator of PE is defined as:

$$\begin{aligned}\widehat{\mathrm{PE}}(\tau_0 \mid s) &= \frac{1}{\sum_{i=1}^{n} \widehat{w}_i^c(\tau_0 \mid s)} \sum_{i=1}^{n} \Big(\mathrm{I}(s < X_i \leq s + \tau_0) \left(1 - \widehat{R}_i(\tau_0 \mid s)\right)^2 \widehat{w}_i^c(\tau_0 \mid s) \\ &\qquad + \mathrm{I}(X_i > s + \tau_0) \left(\widehat{R}_i(\tau_0 \mid s)\right)^2 \widehat{w}_i^c(\tau_0 \mid s) \Big)\end{aligned}$$

where

$$\widehat{w}_i^c(\tau_0 \mid s) = \delta_i \, \mathrm{I}(s < X_i \leq s + \tau_0) \, \frac{\widehat{G}(s)}{\widehat{G}(X_i)} + \mathrm{I}(X_i > s + \tau_0) \, \frac{\widehat{G}(s)}{\widehat{G}(s + \tau_0)}$$

is the inverse probability weight (IPW) to account for censoring and $\widehat{G}(\cdot)$ is the Kaplan-Meier estimate of $G(\cdot)$, the censoring distribution.

Measures of prediction discrimination In clinical practice, a decision at time s is often made based on a subject's risk of experiencing an event in the next time interval τ_0 based on $\mathbf{H}_\circ(s)$, $R(\tau_0, s, \mathbf{H}_\circ(s))$, rather than on specific values of the multivariate $\mathbf{H}_\circ(s)$. Key summary indices for characterizing the performance of a risk prediction model are therefore dependent on specifying a risk threshold ψ for a positive test, and are defined as

$$\begin{aligned}\mathrm{TPF}(\tau_0 \mid s, \psi) &= P\big(R_i(\tau_0 \mid s) \geq \psi \mid s < T_i \leq s + \tau_0\big) \\ \mathrm{FPF}(\tau_0 \mid s, \psi) &= P\big(R_i(\tau_0 \mid s) \geq \psi \mid T_i > s + \tau_0\big).\end{aligned}$$

When estimated over the full range of risk thresholds, these can be displayed using an ROC curve: $\text{ROC}(\tau_0 \mid s, \cdot) = \{\text{TPF}\big(\tau_0 \mid s, \text{FPF}^{-1}(\tau_0 \mid s, \psi)\big),\ \psi \in (0,1)\}$. When no specific risk thresholds are of key interest, a measure such as the area under the ROC curve (AUC) provides a summary of the ability of the risk prediction model to discriminate between cases and non-cases across the full range of risk thresholds: $\text{AUC}(\tau_0 \mid s) = \int_0^1 \text{ROC}(\tau_0 \mid s, \psi) d\psi$.

For a robust evaluation of various longitudinal risk prediction rules, we consider estimators

$$\begin{aligned}
\widehat{\text{TPF}}(\tau_0 \mid s, \psi) &= \frac{\sum_{i=1}^n \delta_i \, \text{I}(\widehat{R}_i(\tau_0 \mid s) > \psi \mid s < T_i \leq s + \tau_0) \, \widehat{w}_i^c(\tau_0 \mid s)}{\sum_{i=1}^n \delta_i \, \text{I}(s < T_i \leq s + \tau_0) \, \widehat{w}_i^c(\tau_0 \mid s)}, \\
\widehat{\text{FPF}}(\tau_0 \mid s, \psi) &= \frac{\sum_{i=1}^n \text{I}(\widehat{R}_i(\tau_0 \mid s) > \psi \mid X_i > s + \tau_0)}{\sum_{i=1}^n \text{I}(X_i > s + \tau_0)},
\end{aligned}$$

and $\widehat{AUC}(\tau_0 \mid s) = \int_0^1 \widehat{ROC}(\tau_0 \mid s, \psi) d\psi$.

Comparison and evaluation of predictive performance: simulation setup. In the second set of simulations we compare the predictive performance of five longitudinal models for estimating the conditional survival probability (PC_{Cox}, PC_{GLM}, PC_{Cox} BLUP, PC_{GLM} BLUP and a joint model), using measures of predictive accuracy and discrimination: $\text{PE}(\tau_0 \mid s)$, $\text{TPF}(\tau_0 \mid s, \psi)$ and $\text{FPF}(\tau_0 \mid s, \psi)$ for $\psi = 0.25$, $\text{AUC}(\tau_0 \mid s)$. For the joint model fitting we used a function provided in the `R` package `JM` [Rizopoulos, 2010]. Note that in all simulations considered, JMs were fit with specifications that corresponded to the true model.

We varied the magnitude of the measurement error ($\sigma_e = \{0.1, 1.0\}$), and the predictions were made for four sets of s and τ_0 with $s = \{24, 48\}$ and $\tau_0 = \{12, 24\}$. All simulations used $n = 500$ and the estimates were obtained from 500 replications.

For each simulation run, all models were fit to a training dataset. A validation dataset was generated with the same parameters as the corresponding training dataset. PCRs were predicted for individuals in the validation set who were are risk at time s. We used the last

available marker value (marker at time s) for prediction in the PC models, and used all the marker information up to time s in prediction using the two stage PC_{Cox} BLUP model and the JM. Accuracy performance metrics were calculated nonparametrically conditioning on time s.

Comparison and evaluation of predictive performance: results. In the scenario where measurement error is small, $\sigma = 0.1$, the predictive performance of all models is comparable (Table 3.3). In the case of a large measurement error, $\sigma = 1.0$, JM using individual trajectories up to time s performed better in terms of both calibration and discrimination compared to the PC_{GLM}, PC_{Cox}, PC_{GLM} BLUP and PC_{Cox} BLUP models. However, we observed a substantial improvement in predictive performance of the PC_{GLM} BLUP and PC_{Cox} BLUP over PC_{GLM} and PC_{Cox}. The performance of the PC_{GLM} BLUP and PC_{Cox} BLUP models was often comparable to that of the JM, and was achieved with much simpler computational steps at model fitting and risk prediction stages (Table 3.4).

3.5.5 Summary of simulation results

In our simulation studies we evaluated the accuracy of risk predictions with the true risks when possible (PC_{GLM} or PC_{Cox}), as well as an example comparing risk predictions obtained using the PC_{GLM} BLUP and the current gold standard, the joint model, which is also the data generating model in our simulations (JM). The predicted risks are close to the true risks (or the gold standard) and the estimated standard errors are very close to the empirical standard errors.

Further evaluation of predicted risks in terms of their calibration and discrimination suggests that when measurement error is small, the performance of all models is comparable (Table 3.3). When measurement error is large (Table 3.4), an improvement in model calibration (reduction in prediction error (PE)) and improvement in performance of the discrimination rule based on predicted risk (larger area under the ROC (AUC)) can be achieved

by using the two-stage PC models when a fitted marker value at time s incorporating prior information is used for prediction. The fitted value is obtained using marker values up to time s for a new subject and estimates from a linear mixed effects model fitted to a full cohort, not including the new subject. The performance of the two-stage models as evaluated by PE and AUC is comparable to JM. We note that the improvement in performance of PC_{GLM} BLUP and PC_{Cox} BLUP hinges on the assumption that the longitudinal marker data is modeled correctly. This same assumption is required by the JM.

The values of the true and false positive fractions (TPF and FPF) for fixed risk thresholds differ slightly between the models, suggesting that the distributions of conditional risk differ between those with and those without events in the τ_0 time frame from s. However, the integrated difference between TPF and FPF over all possible risk thresholds is quite stable even if TPF and FPF for a single threshold differ (for example, see PC_{GLM} BLUP, PC_{Cox} BLUP and JM for $s = 40$ and $\tau_0 = 12$, Table 3.3). This suggests that the risk distributions obtained from the different models for those with (D) and without events ($\bar{\text{D}}$) are slightly shifted, but the area of overlap of the distributions of D and $\bar{\text{D}}$ is largely unaffected, since the AUC is stable between the models. Thus, though we see some small differences in the distributions of predicted risks from the different models, the models are comparable in their ability to discriminate between D and $\bar{\text{D}}$.

Based on these results, we conclude that PC models are adequate for use in practice. We recommend using PC_{GLM} BLUP or PC_{Cox} BLUP models when the biomarkers under consideration have large measurement errors and when the marker trajectory can be modeled well. When there is no strong variation observed over time, we expect the simpler approach of PC_{GLM} or PC_{Cox} to perform well.

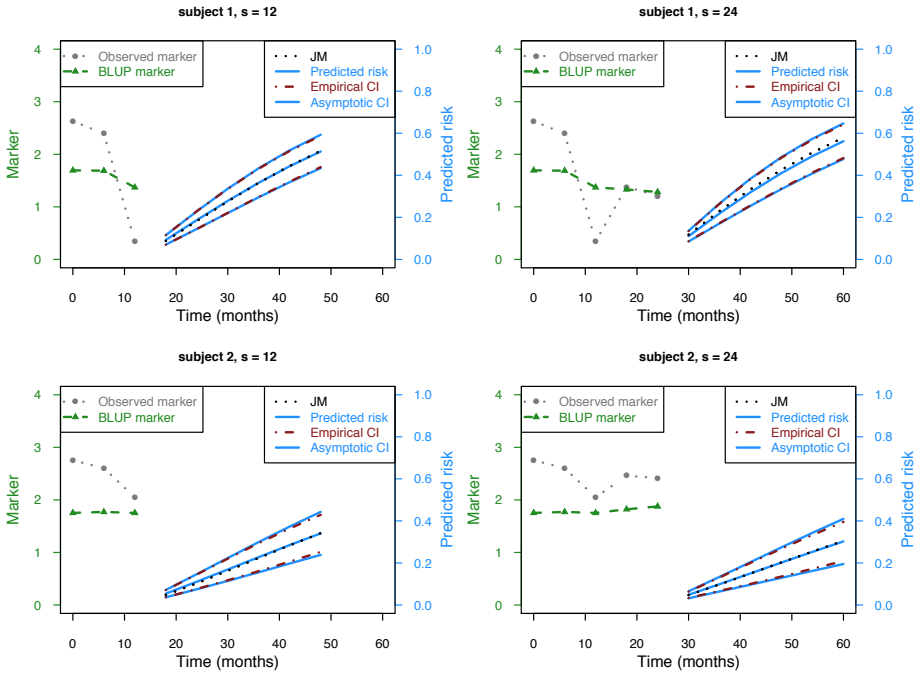

Figure 3.4: Simulation results for individual partly conditional survival probability predictions using two-stage BLUP estimator (PC_{GLM} BLUP) for two randomly selected example subjects from a simulated dataset.

Table 3.1: ($\sigma_e = 0.1$) The conditional predicted risk, $R(\tau_0 \mid s)$, for four sets of s and τ_0 for subjects with marker $= \{0, 1\}$ estimated using the partly conditional generalized linear model with the logit link function (PC_{GLM}) and the partly conditional Cox-type model (PC_{Cox}). The true risks were obtained empirically. Empirical standard errors (ESD), asymptotic standard errors (ASE) and mean-squared errors (x1000) (MSE) are also provided. Data simulated with standard deviation of the measurement error $\sigma_e = 0.1$, replications = 500, n = 500 at baseline.

	$\tau_0 = 12$				$\tau_0 = 24$			
	s = 24		s = 48		s = 24		s = 48	
	y = 0	y = 1	y = 0	y = 1	y = 0	y = 1	y = 0	y = 1
True	0.332	0.089	0.319	0.088	0.565	0.174	0.549	0.171
PC_{GLM}								
EST	0.334	0.089	0.330	0.087	0.552	0.178	0.543	0.173
ESD	0.029	0.011	0.038	0.014	0.037	0.021	0.045	0.024
ASE	0.029	0.011	0.039	0.014	0.036	0.021	0.046	0.025
MSE(x1000)	0.859	0.119	1.636	0.187	1.497	0.449	2.189	0.605
PC_{Cox}								
EST	0.319	0.090	0.307	0.086	0.545	0.179	0.528	0.171
ESD	0.019	0.008	0.018	0.007	0.030	0.015	0.029	0.013
ASE	0.018	0.008	0.017	0.007	0.029	0.015	0.028	0.013
MSE(x1000)	0.489	0.066	0.447	0.055	1.185	0.257	1.230	0.177

3.6 Real data example: ESRDS dataset

3.6.1 Study design and participants

The ESRDS is a retrospective cohort of 698 individuals aged 18-60 years with severe non-dialysis requiring chronic kidney disease (CKD) who received ambulatory care in the Community Health Network (CHN) in San Francisco from January 1, 1996 to February 29, 2008. CKD was defined based on at least two outpatient estimated glomerular filtration rate (eGFR) measurements $\leq 60\, mL/min/1.73^2$ that were separated by at least three months,

Table 3.2: ($\sigma_e = 1.0$) The conditional predicted risk, $R(\tau_0 \mid s)$, for four sets of s and τ_0 for subjects with marker $= \{0, 1\}$ estimated using the partly conditional generalized linear model with the logit link function (PC_{GLM}) and the partly conditional Cox-type model (PC_{Cox}). The true conditional risks (True) were obtained empirically. Empirical standard errors (ESD), asymptotic standard errors (ASE) and mean-squared errors ($\times 1000$) (MSE) are also provided. Data simulated with standard deviation of the measurement error $\sigma_e = 1.0$, 500 replications, $n = 500$ at baseline.

	$\tau_0 = 12$				$\tau_0 = 24$			
	s = 24		s = 48		s = 24		s = 48	
	y = 0	y = 1	y = 0	y = 1	y = 0	y = 1	y = 0	y = 1
True	0.189	0.131	0.157	0.109	0.333	0.237	0.280	0.207
PC_{GLM}								
EST	0.190	0.123	0.166	0.107	0.334	0.230	0.298	0.201
ESD	0.017	0.013	0.021	0.016	0.026	0.021	0.031	0.025
ASE	0.017	0.012	0.021	0.015	0.026	0.021	0.032	0.025
MSE(x1000)	0.289	0.210	0.538	0.23	0.671	0.497	1.323	0.658
PC_{Cox}								
EST	0.191	0.125	0.171	0.111	0.332	0.229	0.301	0.206
ESD	0.011	0.009	0.009	0.007	0.019	0.015	0.016	0.013
ASE	0.011	0.008	0.009	0.007	0.018	0.015	0.015	0.013
MSE(x1000)	0.116	0.113	0.287	0.055	0.317	0.292	0.662	0.163

of these subjects we identified those with severe CKD defined based on the baseline eGFR measurement of $\leq 30\, mL/min/1.73^2$. The outcome of interest was time to ESRD (initiation of dialysis or kidney transplantation) or death. The longitudinal marker of interest was eGFR and was recorded at least once in every 6-month time interval from study entry until an event or censoring. Other covariates available were demographic (age, sex, ethnicity) and comorbidities at baseline (diabetes, hypertension, chronic viral disease, substance abuse).

3.6.2 Analysis methods

We were interested in predicting the conditional probability of progression to ESRD or death within a given timeframe, τ_0, given event-free survival and covariate information up to time s, $R_i(\tau_0|s) = P\big(T_i \leq s+\tau_0|T_i > s, \mathbf{H}_i(s)\big)$. We used clinically relevant prediction timeframes, τ_0, of 1 and 3 years basing our predictions on data available for the last 1 or 2 years. A subject found to be at high risk of ESRD in a 1-year timeframe would likely be recommended to meet with their healthcare team in order to start preparing for dialysis. A three-year timeframe would provide a subject at high risk with preventive treatment options, starting with controlling comorbid conditions such as hypertension and diabetes, lifestyle changes, among others.

In order to make risk predictions for every subject in our dataset, but avoid overfitting by training and validating our models on the same data, we used cross-validation. We split the dataset into five subsets using stratified sampling stratifying on baseline eGFR level ([0, 15], (15, 30]), with the five subsets being mutually exclusive. Each of the five subsets was treated as a validation set, with the remaining 4/5 of the data used as a training set.

We fit the five models introduced in sections 3.2 and 3.3. The first model was a partly conditional generalized linear model with the logit link function with eGFR as the main predictor and measurement time modeled with a natural cubic spline with df = 3 (PC_{GLM}). Our second model was a partly conditional Cox-type model with measurement time modeled with a natural cubic spline with df = 3 and eGFR as the main predictor (PC_{Cox}). The third model was the two-stage PC_{Cox} BLUP model: for each individual in the validation set, fitted marker values were obtained via the empirical BLUP (best linear unbiased predictor) from the given individual's data up to time s and REML (restricted maximum likelihood) estimates of covariance parameters from the LMEM fit to the training set. The observed marker trajectories were non-linear with a tendency to increase shortly after entry into the cohort and stabilizing after about 0.5 years (Figure 3.5). To model that we fit a linear spline

with a knot at 0.5 years for the fixed and random effects in the LME model. This approach provided smoothing to individual marker data while 'shrinking' each individual's marker trajectory towards the population-averaged mean marker trajectory. For each prediction time s, the empirical BLUP of the random effect was updated to take into account any additional data that has become available since the previous prediction time (PC_{Cox} BLUP). The fourth model was a two-stage PC_{GLM} model (PC_{GLM} BLUP). The fifth model was a joint model, where the LMEM sub-model was as that used for PC_{Cox} BLUP, in two-stage PC models for modeling eGFR as a linear spline with a knot at 0.5 years for both the fixed and random effects (JM).

For each training set we fit each of the four models and obtained risk predictions based on the corresponding validation set. Repeating this process for each of the five training-validation pairs, we obtained risk predictions for every individual in our dataset for a given s and τ_0. From these, using the estimators of prediction evaluation measures in section 4, we estimated the prediction error (PE), the true positive fraction (TPF) at a risk threshold of 0.25, the false positive fraction (FPF) at a risk threshold of 0.25 and the area under the ROC curve (AUC). The standard errors were estimated empirically by perturbing the risk vector with weights $\sim exp(1)$ 200 times. To estimate the time-dependent ROC, we estimated the time-dependent TPF and FPF using the predicted risks for all individuals for a given s and τ_0 using each individual's risk as a threshold.

3.6.3 Results

The characteristics of the ESRDS dataset are summarized in Table 3.7. The mean (sd) eGFR among the subjects in the study cohort was 19.6 (7.6) $mL/min/1.73^2$, mean age was 45.7 (9.4) years, with a quarter of subjects under the age of 40 years at study entry, 64% male, 38% were black, and more than 26% of the subjects presented with at least one of the risk factors for ESRD (hypertension substance abuse, chronic viral disease). During the

Figure 3.5: Observed individual trajectories of eGFR over time, stratified by eGFR groups at baseline ([0,15], (15,20], (20,25], (25,30] mL/min/1.73m^2) and age groups at baseline ((16, 40], (40, 50], (50, 60] years). Trajectories for subjects with 3 or more observations are plotted.

study, there were 321 (46%) composite events observed (220 (31.5%) ESRD events and 101 (14.5%) deaths without prior ESRD diagnosis). The average follow-up time was 6.2 years (max = 14.3 years) with the average time to event of 2.8 years. Composite event counts and ESRD event counts stratified by baseline eGFR and age are summarized in Table 3.5.

The composite events of ESRD or death were dominated by ESRD events in all eGFR and age strata. Composite event counts in the relevant prediction timeframes are summarized in Table 3.6.

The number of eGFR measurements available for a given individual ranged from 1-24, and were taken at approximately 6-month time intervals, with over 50% of the individuals providing four or more measurements. A description of the characteristics of the dataset are summarized in Table 3.7.

The estimates of the prediction capacity measures were summarized in Table 3.8. The prediction errors for the three year timeframe were twice those for predictions for one year, with PC_{Cox} and PC_{Cox} BLUP achieving the lowest prediction errors of the five methods (Table 3.8), but the confidence intervals of PE for all methods for a given s and τ_0 overlapped. TPF and FPF at the risk threshold of 0.25 were considerably different between the methods, suggesting that the risk distributions of those with events and those without events differed between the methods. Despite these differences in risk distributions, the ability of the models to discriminate between the two groups of subjects was relatively comparable.

The AUC estimates ranged from 0.65 (PC_{GLM} BLUP, $s = 1$, $\tau_0 = 3$) to 0.86 (PC_{Cox}, $s = 2$, $\tau_0 = 1$). Overall, the best discrimination between those with and without events betwen s and $s + \tau_0$, given being at risk at time s, was achieved for $s = 2$ and $\tau_0 = 1$ with the AUC of 0.71 for the JM and between 0.78 and 0.86 for the PC models. The PC models fit to the observed data (PC_{Cox} and PC_{GLM}) performed better in terms of calibration and discrimination compared to the two-stage models (PC_{Cox} BLUP and PC_{GLM} BLUP) and the JM.

Figure 3.6 shows that the time-dependent ROC curves estimated using PC Cox-type models are very similar across s and τ_0 pairs, which is consistent with the AUC estimates in Table 3.8. The best prediction performance for eGFR was achieved with PC_{Cox} for $s = 2$ and $\tau_0 = 1$ with AUC of 0.86 (95% CI: (0.78, 0.94)). Both the PC_{GLM} BLUP and the JM performed poorly compared to other models, especially for $\tau_0 = 1$. PC_{GLM} BLUP was

outperformed by all others for $s = 1$ and $\tau_0 = 3$. The shape of the ROC curve for JM at $s = 2$ and $\tau_0 = 1$ suggests that though the model does appear to assign higher risk to those with events than those without events overall, about 5% of the subjects with highest predicted risks did not, in fact, experience events in that timeframe. We note a similar trend with PC_{GLM} BLUP for $s = 1$ and $\tau_0 = 3$ (Figure 3.6, Table 3.6).

The plots of predicted risk for one example individual (Figure 3.7) with predictions up to 9 years based on marker data for 2, 4 and 6 years showed that, in general, the PC models behaved similarly to one another. With two years of data, the BLUP predicted response profile resulted in a value very close to the observed value at $s = 2$. The resulting predictions from 2.5 to 9 years increase slowly over time, with prediction curves from different models are almost parallel. At $s = 4$, the BLUP marker value and the observed value differ substantially, and this results in a larger spread of predictions between different models. Risk curves again increase over time, except for those corresponding to PC_{GLM} and PC_{GLM} BLUP which level off at around 6.5 years and start to slowly decrease after that. For $s = 6$, prediction paths for all the PC models are very similar and increasing, the trend for the JM is similar, but about 0.2 lower on the risk scale throughout. This patient experienced an event at 11.4 years after study entry, thus the higher predicted risks are likely to be closer to the true risk for this particular individual (Figure 3.7). The trends seen in this plot were consistent with results summarized in Table 3.8.

3.6.4 Remarks on the ESRD analysis and results

The estimated glomerular filtration rate (eGFR) trajectories proved to be quite challenging to model. We chose the linear mixed effects (LME) model with a linear spline with a knot at 0.5 years after exploring several models with linear and natural cubic splines. Our goal was to find a parsimonious model that accounted for the sharp change in marker trajectory observed for most subjects in the dataset. Parsimony was important because achieving

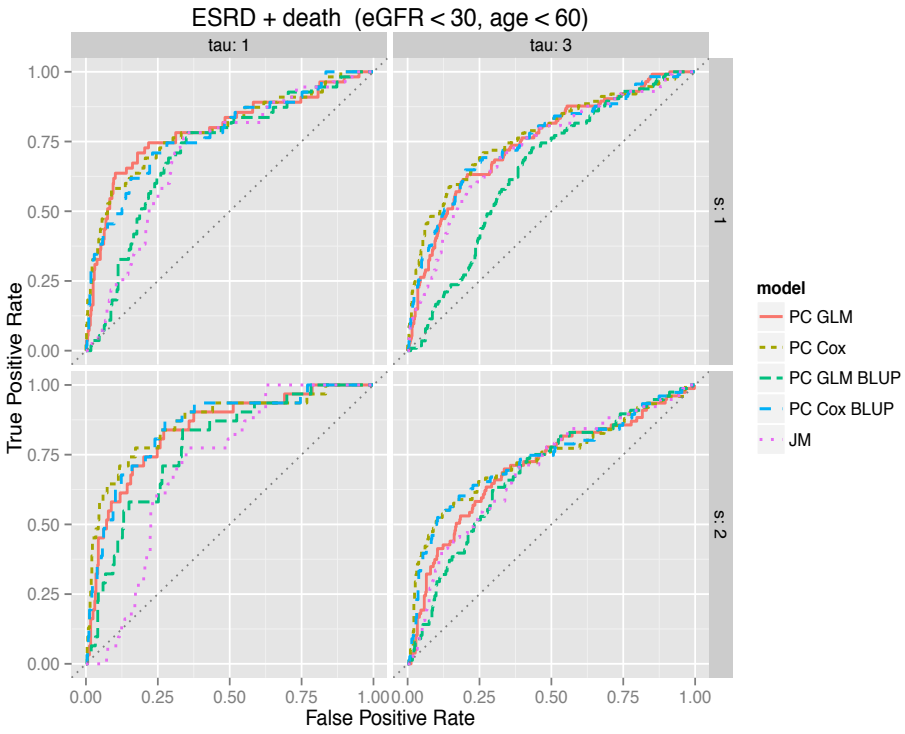

Figure 3.6: (ESRD Study) Time-dependent ROC curves for predicted risk of a composite outcome of end stage renal disease (ESRD) and death in the ESRD Study from the Community Health Network in San Francisco. The risk was predicted using five models: the partly conditional generalized linear model with the logit link function (PC GLM), partly conditional Cox-type model (PC Cox), two-stage PC models (PC Cox BLUP and PC GLM BLUP) and the joint model (JM). The rows represent $s = 1$ and 2 years, the columns refer to $\tau_0 = 1$ and 3 years.

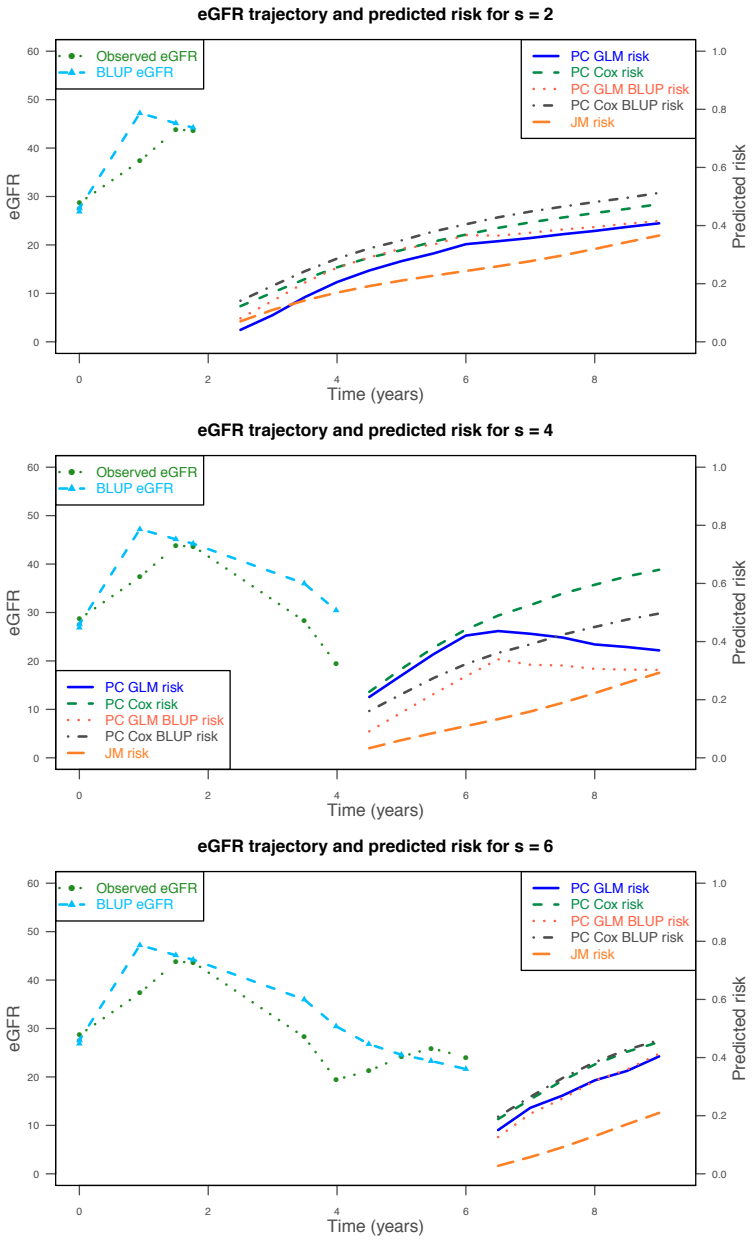

Figure 3.7: (ESRD study) Individual risk predictions for subject 531 from the ESRD Study. We fit five models to the ESRDS dataset excluding subject 531, and obtained risk predictions at 0.5 year increments starting at time $s = 2$, 4 and 6 years (rows) with maximum prediction time of 9 years. The curves on the left show the observed data the predicted response profiles of eGFR the risk prediction curves shown on the right were based on. The five models fit were the partly conditional logistic model(PC GLM), partly conditional Cox-type model (PC Cox), two-stage PC models (PC Cox BLUP and PC GLM BLUP) and the joint model (JM).

convergence when fitting the JM with a complex LME model is challenging. We opted to use the same LME model for the two-stage PC models and the JM in order to compare them in as similar a situation as possible. However, using a more complex model for the longitudinal trajectories would not pose a challenge for the two-stage PC models. We expect that the improvement, had we used a better fitting LME model in the two-stage PC models, would result in some improvement in the results, though it is difficult to speculate on the magnitude of the expected improvement.

Overall, the eGFR is a good predictor of risk of ESRD or death. It performed especially well in our analysis for the 1 year prediction timeframe ($\tau_0 = 1$) for subjects still at risk at 2 years ($s = 2$) with the PC_{Cox} model, which does not require any assumptions about the marker process.

3.7 Summary

Updated predictions of risk using accumulated information among patients under active surveillance are important for disease management. Development of dynamic prediction rules is often challenging, requiring the consideration of both longitudinal process of the predictors and their relationship with disease outcome that can change over time. Statistical methods focusing on longitudinal prediction have appeared in the literature in recent years, however, much work is still needed.

In this chapter we developed flexible approaches that can accommodate complex practical settings, and evaluated modeling choices in order to provide practical guidance for using the available methods in outcome prediction research. Specifically, we proposed estimation and inference procedures for estimating partly conditional risk, $R(\tau_0 \mid s, \mathbf{H}(s))$, based on the covariate history $\mathbf{H}(s)$ up to time s. When large within-individual variability is observed for longitudinal covariates, it is assumed to be non-systematic and the marker trajectories can be modeled well, the BLUP approach to obtain predicted response profiles is a nice tool for smoothing the marker data in an attempt to reduce noise in the marker trajectories.

In that situation, the predictions based on the two-stage PC_{GLM} BLUP and PC_{Cox} BLUP models can achieve better prediction performance compared to those obtained from PC_{GLM} and PC_{Cox}.

The partly conditional models provide a versatile, flexible and robust framework for longitudinal risk prediction and are a practical alternative to joint models. One of the strengths of the PC models is that they are computationally relatively simple, and can be easily extended to more complex situations as dictated by the applied problem at hand. For example, multiple longitudinal biomarkers can be easily incorporated without having to resort to multi-level integration and strong modeling assumptions, as can markers with complex longitudinal trajectories that cannot be modeled adequately with simple linear mixed effects models.

In the next chapter (Chapter 4) we focus our attention on estimation and inference for measures of calibration and discrimination in cohort studies. Some of these measures have been briefly introduced in order to carry out some of the evaluations presented in this chapter. The full treatment of why these measures were chosen, their definitions, estimation, inference and evaluation follows in Chapter 4.

Table 3.3: ($\sigma_e = 0.1$) Estimates (EST) and empirical standard errors (ESD) of measures of predictive capacity summarizing predictions based on the partly conditional generalized linear model with the logit link function (PC_{GLM}), partly conditional Cox-type model using observed data (PC_{Cox}) and smoothed marker values (PC_{Cox} (BLUP)), and the joint model (JM). The estimates were obtained for four sets of s and τ_0 for small measurement error ($\sigma_e = 0.1$), $n = 500$ at baseline, 500 replications.

	$\tau_0 = 12$		$\tau_0 = 24$	
	s = 24	s = 48	s = 24	s = 48
	EST (ESD)	EST (ESD)	EST (ESD)	EST (ESD)
PC_{GLM}				
PE	0.178 (0.014)	0.171 (0.022)	0.198 (0.014)	0.212 (0.022)
TPF(0.25)	0.791 (0.059)	0.706 (0.117)	0.936 (0.030)	0.903 (0.056)
FPF(0.25)	0.422 (0.063)	0.366 (0.090)	0.663 (0.067)	0.632 (0.100)
AUC	0.760 (0.034)	0.730 (0.059)	0.789 (0.029)	0.749 (0.048)
PC_{Cox}				
PE	0.179 (0.014)	0.172 (0.024)	0.197 (0.013)	0.212 (0.021)
TPF(0.25)	0.764 (0.057)	0.615 (0.114)	0.948 (0.026)	0.911 (0.050)
FPF(0.25)	0.385 (0.056)	0.274 (0.067)	0.701 (0.061)	0.645 (0.087)
AUC	0.760 (0.034)	0.730 (0.059)	0.789 (0.029)	0.749 (0.048)
PC_{GLM} (BLUP)				
PE	0.177 (0.014)	0.170 (0.022)	0.197 (0.014)	0.211 (0.022)
TPF(0.25)	0.793 (0.058)	0.710 (0.117)	0.936 (0.030)	0.906 (0.056)
FPF(0.25)	0.420 (0.063)	0.364 (0.089)	0.661 (0.067)	0.632 (0.098)
AUC	0.763 (0.034)	0.733 (0.059)	0.792 (0.029)	0.752 (0.048)
PC_{Cox} (BLUP)				
PE	0.178 (0.014)	0.171 (0.024)	0.196 (0.013)	0.210 (0.020)
TPF(0.25)	0.768 (0.057)	0.623 (0.114)	0.948 (0.025)	0.914 (0.050)
FPF(0.25)	0.386 (0.056)	0.276 (0.065)	0.701 (0.062)	0.647 (0.087)
AUC	0.763 (0.034)	0.733 (0.059)	0.792 (0.029)	0.752 (0.048)
JM				
PE	0.176 (0.014)	0.170 (0.021)	0.196 (0.014)	0.209 (0.020)
TPF(0.25)	0.789 (0.056)	0.708 (0.090)	0.944 (0.024)	0.918 (0.046)
FPF(0.25)	0.409 (0.047)	0.350 (0.070)	0.682 (0.057)	0.648 (0.081)
AUC	0.765 (0.034)	0.737 (0.052)	0.791 (0.031)	0.755 (0.047)

PE = prediction error, TPF(0.25) = true positive fraction at $R(\tau_0 \mid s) = 0.25$, FPF(0.25) = false positive fraction at $R(\tau_0 \mid s) = 0.25$, AUC = area under the ROC curve

Table 3.4: ($\sigma_e = 1.0$) Estimates (EST) and empirical standard errors (ESD) of measures of predictive capacity summarizing predictions based on the partly conditional generalized linear model (PC_{GLM}), partly conditional Cox-type model (PC_{Cox}) using observed data and smoothed marker values (PC_{GLM} BLUP, PC_{Cox} BLUP), and the joint model (JM). The estimates were obtained for four sets of s and τ_0 for large measurement error ($\sigma_e = 1.0$), $n = 500$ at baseline, 500 replications.

	$\tau_0 = 12$		$\tau_0 = 24$	
	s = 24	s = 48	s = 24	s = 48
	EST (ESD)	EST (ESD)	EST (ESD)	EST (ESD)
PC_{GLM}				
PE	0.209 (0.014)	0.196 (0.023)	0.248 (0.012)	0.254 (0.018)
TPF(0.25)	0.759 (0.085)	0.608 (0.150)	0.975 (0.023)	0.945 (0.046)
FPF(0.25)	0.572 (0.089)	0.443 (0.129)	0.918 (0.045)	0.859 (0.083)
AUC	0.639 (0.042)	0.614 (0.063)	0.651 (0.039)	0.623 (0.056)
PC_{Cox}				
PE	0.209 (0.014)	0.193 (0.023)	0.248 (0.011)	0.253 (0.017)
TPF(0.25)	0.717 (0.074)	0.543 (0.117)	0.985 (0.015)	0.960 (0.036)
FPF(0.25)	0.527 (0.073)	0.384 (0.086)	0.941 (0.037)	0.892 (0.065)
AUC	0.636 (0.041)	0.609 (0.064)	0.650 (0.037)	0.619 (0.057)
PC_{GLM} BLUP				
PE	0.194 (0.014)	0.183 (0.022)	0.224 (0.013)	0.228 (0.020)
TPF(0.25)	0.782 (0.071)	0.679 (0.124)	0.956 (0.028)	0.927 (0.054)
FPF(0.25)	0.491 (0.074)	0.394 (0.103)	0.802 (0.069)	0.734 (0.099)
AUC	0.709 (0.038)	0.693 (0.056)	0.727 (0.034)	0.709 (0.053)
PC_{Cox} BLUP				
PE	0.195 (0.013)	0.180 (0.022)	0.223 (0.012)	0.229 (0.018)
TPF(0.25)	0.762 (0.065)	0.629 (0.122)	0.972 (0.020)	0.947 (0.042)
FPF(0.25)	0.466 (0.065)	0.344 (0.078)	0.852 (0.061)	0.788 (0.082)
AUC	0.708 (0.038)	0.692 (0.060)	0.729 (0.033)	0.705 (0.053)
JM				
PE	0.192 (0.014)	0.183 (0.020)	0.223 (0.014)	0.227 (0.019)
TPF(0.25)	0.794 (0.057)	0.715 (0.094)	0.961 (0.020)	0.938 (0.039)
FPF(0.25)	0.506 (0.053)	0.430 (0.073)	0.813 (0.054)	0.759 (0.073)
AUC	0.710 (0.039)	0.694 (0.059)	0.728 (0.037)	0.710 (0.053)

PE = prediction error, TPF(0.25) = true positive fraction at $R(\tau_0 \mid s) = 0.25$, FPF(0.25) = false positive fraction at $R(\tau_0 \mid s) = 0.25$, AUC = area under the ROC curve

Table 3.5: Distribution of composite event counts and ESRD events stratified by eGFR and age categories. Numbers of subjects who did not experience an event during the study are summarized as non-events.

End Stage Renal Disease Study						
eGFR	Non-events		ESRD or death		ESRD	
category	n	%	n	%	n	%
[0,5]	9	2.4	27	8.4	26	11.8
(5,15]	75	19.9	84	26.2	60	27.3
(15,20]	58	15.4	65	20.2	47	21.4
(20,25]	91	24.1	75	23.4	47	21.4
(25,30]	144	38.2	70	21.8	40	18.2
Age	Non-events		ESRD or death		ESRD	
category	n	%	n	%	n	%
(16, 40]	104	27.6	73	22.7	56	25.5
(40, 50]	140	37.1	113	35.2	75	34.1
(50, 60]	133	35.3	135	42.1	89	40.5

Table 3.6: Number of composite (ESRD or death) events observed within followup-time intervals of interest (in years).

End Stage Renal Disease Study		
	Prediction timeframe	
Conditioning time	$\tau_0 = 1$	$\tau_0 = 3$
s = 1	55	114
s = 2	31	77

Table 3.7: Baseline characteristics of the severe chronic kidney disease (CKD) Community Health Network (CHN) subcohort defined based on age and eGFR at baseline (age ≤60 and eGFR ≤30mL/min/1.73m2) (n = 698).

End Stage Renal Disease Study, CHN severe CKD subcohort (n = 698)		
Age (years) (n, %)		
— (16, 40]	177	25.4
— (40, 50]	253	36.2
— (50, 60]	268	38.4
Male (n, %)	449	64.3
Ethnicity (n, %)		
— White	193	27.7
— Black	269	38.5
— Hispanic	127	18.2
— Asian	96	13.8
— Other ethnicity	13	1.9
eGFR mL/min/1.73m^2 (mean, sd)	19.6	7.6
eGFR mL/min/1.73m^2 (n, %)		
— [0,5]	36	5.2
— (5,15]	159	22.8
— (15,20]	123	17.6
— (20,25]	166	23.8
— (25,30]	214	30.7
Homeless status (n, %)	84	12.0
Diabetes (n, %)	98	14.0
Hypertension (n, %)	188	26.9
Cardiovascular disease (n, %)	54	7.7
Substance abuse (n, %)	229	32.8
Chronic viral disease (n, %)	244	35.0

Table 3.8: Estimates (EST) and standard errors (ESD) of measures of predictive capacity summarizing predictions of a composite outcome (death or ESRD) based on eGFR and a partly conditional logistic model with the logit link function (PC_{GLM}), partly conditional Cox-type model using observed data (PC_{Cox}), smoothed marker values obtained using predicted random effects based on estimates from a linear mixed effects model and individual's data up to time s (PC_{GLM} BLUP, PC_{Cox} BLUP) and the joint model (JM). The estimates were obtained for four sets of s and τ_0. The number of events between s and $s+\tau_0$, and the number of subjects at risk at time s are denoted by n_e and n, respectively.

End Stage Renal Disease Study				
	$\tau_0 = 1$ year		$\tau_0 = 3$ years	
	s = 1 year ($n_e/n = 55/574$) EST (ESD)	s = 2 years ($n_e/n = 31/519$) EST (ESD)	s = 1 year ($n_e/n = 114/574$) EST (ESD)	s = 2 years ($n_e/n = 77/519$) EST (ESD)
PC_{GLM}				
PE	0.082 (0.015)	0.062 (0.016)	0.152 (0.033)	0.150 (0.037)
TPF(0.25)	0.745 (0.062)	0.742 (0.080)	0.736 (0.041)	0.695 (0.049)
FPF(0.25)	0.249 (0.013)	0.205 (0.014)	0.366 (0.016)	0.350 (0.017)
AUC	0.791 (0.024)	0.838 (0.029)	0.749 (0.022)	0.703 (0.027)
PC_{Cox}				
PE	0.075 (0.009)	0.053 (0.011)	0.132 (0.024)	0.128 (0.031)
TPF(0.25)	0.582 (0.060)	0.710 (0.085)	0.719 (0.043)	0.694 (0.050)
FPF(0.25)	0.135 (0.012)	0.135 (0.013)	0.333 (0.018)	0.355 (0.020)
AUC	0.791 (0.033)	0.861 (0.041)	0.771 (0.024)	0.731 (0.029)
PC_{GLM} BLUP				
PE	0.100 (0.016)	0.076 (0.018)	0.179 (0.033)	0.165 (0.038)
TPF(0.25)	0.691 (0.061)	0.645 (0.088)	0.816 (0.046)	0.777 (0.049)
FPF(0.25)	0.312 (0.016)	0.264 (0.016)	0.600 (0.021)	0.502 (0.022)
AUC	0.714 (0.034)	0.782 (0.039)	0.650 (0.027)	0.682 (0.030)
PC_{Cox} BLUP				
PE	0.078 (0.006)	0.057 (0.011)	0.139 (0.017)	0.130 (0.028)
TPF(0.25)	0.436 (0.058)	0.645 (0.090)	0.719 (0.046)	0.707 (0.052)
FPF(0.25)	0.064 (0.013)	0.121 (0.015)	0.366 (0.021)	0.374 (0.023)
AUC	0.777 (0.039)	0.851 (0.047)	0.756 (0.028)	0.731 (0.032)
JM				
PE	0.086 (0.007)	0.069 (0.011)	0.147 (0.020)	0.139 (0.023)
TPF(0.25)	0.218 (0.060)	0.323 (0.084)	0.728 (0.044)	0.635 (0.047)
FPF(0.25)	0.096 (0.015)	0.201 (0.016)	0.386 (0.021)	0.350 (0.020)
AUC	0.706 (0.038)	0.708 (0.048)	0.727 (0.027)	0.691 (0.031)

PE = prediction error, TPF(0.25) = true positive fraction at $R(\tau_0 \mid s) = 0.25$, FPF(0.25) = false positive fraction at $R(\tau_0 \mid s) = 0.25$, AUC = area under the ROC curve

Chapter 4

ASSESSING PREDICTION PERFORMANCE UNDER LONGITUDINAL COHORT STUDIES

4.1 Introduction

In the previous chapter we tackled the problem of predicting risk of an event in the next τ_0 time interval from s for a new subject given their covariate information up to time s. The quality of such predictions depends on the predictiveness of the longitudinal marker, on how well the model relating the longitudinal marker to the event of interest approximates their true relationship, and in the case of the joint model and the two-stage models discussed in the previous chapter, longitudinal biomarker data.

Each of these aspects needs to be carefully evaluated before deciding on the marker and the model. In most situations, as in both our illustrative datasets, the marker to be evaluated has already been selected, but modeling options remain. Practical guidance as to how to evaluate these modeling options was discussed in Chapter 3. We now turn our attention to methods for evaluating the quality of risk predictions obtained from an arbitrary longitudinal marker and an arbitrary predictive model.

Our choice of evaluation measures was driven by their clinical utility and clinically meaningful interpretation. In this chapter we focus on evaluation of predictions in a cohort setting, while keeping in mind their potential utility in evaluating risk predictions under two-phase study designs. This was one of the motivations for our choice of inference procedures. In other words, the inference procedures developed in this chapter can be extended and applied to variance estimation of risk prediction evaluation measures in two-phase studies.

In this chapter we develop estimators of measures of calibration and discrimination to

evaluate risk predictions in a longitudinal setting with time-to-event outcomes. We develop resampling-based inference procedures for variance estimation of the estimators of the evaluation measures. We evaluate the performance of our estimators using simulations and illustrate them on the ESRDS dataset.

In the next section we discuss the reasons behind our choice of measures to evaluate risk predictions in a longitudinal setting, their definitions follow in section 4.3, estimation procedures are presented in section 4.4, inference in section 4.5. Simulation studies (section 4.6), results (section 4.6.5) and an illustration of our methods on the ESRD study dataset is presented in section 4.7. The summary section (4.8) concludes the chapter.

4.2 Choice of measures to evaluate predictive capacity in a longitudinal setting

True and false positive fractions (TPF and FPF) are standard measures for evaluating tests, diagnostic or predictive. They also commonly appear in other areas of study, under possibly different names [Pepe, 2003]. They are widely used in practice, where TPF quantifies the sensitivity of a test, and 1-FPF the specificity of a test. These measures serve as building blocks for various other measures in wide clinical use, such as the receiver operating characteristic (ROC) curves and the area under the ROC (AUC) (section 2.2). TPF and FPF have also been used as a basis for developing measures such as *net benefit*, which additionally incorporates the disease prevalence into the calculation and is used to quantify the utility of risk predictions to guide treatment on a population level [Vickers and Elkin, 2006, Pepe et al., 2013]. Therefore, TPF and FPF, and thus ROC and AUC, are fundamental for work on measures for evaluation of any tests, including predictive tests in a longitudinal setting.

Proportion of cases followed and proportion of the population to follow (PCF and PNF) are risk prediction summaries recently proposed by Pfeiffer and Gail [Pfeiffer and

Gail, 2011]. They are also based on TPF and FPF, respectively, and additionally require estimation of a risk threshold for a given quantile of risk in a population. The PCF(q) represents the estimated proportion of cases that would be captured if we followed proportion q of the population at highest risk, and PNF(p) represents the estimated proportion of the population at highest risk that would need to be followed in order to capture proportion p of the cases (section 2.2). We chose them because of their clinically relevant interpretation and potential for use in evaluating the feasibility of preventive interventions on a health care system level.

Prediction error (PE) is a measure of model calibration rooted in the 1950's [Brier, 1950]. It has a structure that is intuitive, comparing the predicted risks with the observed outcomes. We chose the definition which uses the squared loss function, which provides a proper scoring rule and ensures that PE attains its minimum at the true underlying risk, which may not be the case with other loss functions [Schoop et al., 2011].

4.3 Definitions of measures of predictive capacity in a longitudinal setting

The clinical utility of biomarkers has traditionally been quantified with a receiver operating characteristic (ROC) curve and the area under the ROC (AUC) estimated using markers at baseline or, in a longitudinal setting, using marker value at time s [Heagerty et al., 2000, Zheng and Heagerty, 2004]. In our approach, our goal is to incorporate all information available up to time s in making the prediction at s. Such information may include baseline covariates, longitudinal covariates, as well as multiple longitudinal markers. More formally, we want to have the ability to make a decision at time s based on a subject's risk of experiencing an event in the next time interval τ_0, $R_i(\tau_0 \mid s) = R(\tau_0 \mid s, \mathbf{H}_i(s)) = P\big(s < T_i \leq s + \tau_0 \mid T_i \geq s, \mathbf{H}_i(s)\big)$, given their covariate information up to time s, rather than on specific values of the longitudinal covariate of subject i at time s, $Y_i(s)$. Therefore, we define a test based on risk at time s, which can be thought of as a summary of various covariate

information available up to time s.

Definition of a test A test is said to be *positive* if $R(\tau_0 \mid s, \mathbf{H}_i(s)) > \psi$, $\psi \in (0,1)$, and *n*egative otherwise.

Definition of an event A subject i is said to have experienced an *event* if $s < T_i \leq s + \tau_0$ and not experienced an event if $T_i > s + \tau_0$. We make a distinction between *cases* and subjects who have experienced an *event*. A subject who has experienced an event at some point during the entire study will be referred to as a *case*. A subject who has experienced an event between s and $s + \tau_0$ will be referred to as *a subject who has experienced an event* in a given timeframe. This distinction will become especially important in the next chapter where we deal with two-phase sampling designs.

Definitions of the true and false positive fractions (TFP and FPF) The true positive and false positive fractions are just that, fractions of positive tests among subjects with and without events, respectively. Depending on the setting, diagnostic or predictive, time-dependent or longitudinal, the definition of a test and event in a given setting determines the definition of these fractions. In the longitudinal setting, [Zheng and Heagerty, 2007] defined the $\text{TPF}(\tau_0 \mid s)$ and $\text{FPF}(\tau_0 \mid s)$ as

$$\begin{aligned} \text{TPF}(\tau_0 \mid s, \psi) &= \text{P}(Y(s) > c \mid s < T \leq s + \tau_0) \quad \text{and} \\ \text{FPF}(\tau_0 \mid s, \psi) &= \text{P}(Y(s) > c \mid T > s + \tau_0). \end{aligned}$$

Thus, they considered the marker value at time s in defining the test. This definition allows for the estimate of TPF and FPF to be reevaluated when new information becomes available, but it does not use biomarker information prior to s, nor does it allow for adjustment for baseline covariates. Lastly, it allows for evaluation of a single biomarker at a time. Building on that definition with the goal of addressing the above shortcomings, we define our test in

terms of risk at time s, where any number of longitudinal markers and baseline covariates can be used to estimate that risk.

We define the true and false positive fractions in a longitudinal setting as:

$$\begin{aligned} \mathrm{TPF}(\tau_0 \mid s, \psi) &= \mathrm{P}(R(\tau_0 \mid s) > \psi \mid s < T \leq s + \tau_0) \quad \text{and} \\ \mathrm{FPF}(\tau_0 \mid s, \psi) &= \mathrm{P}(R(\tau_0 \mid s) > \psi \mid T > s + \tau_0) \end{aligned}$$

Definitions of the ROC and the area under to ROC (AUC) The longitudinal definitions of the receiver operating characteristic curve, $\mathrm{ROC}(\tau_0 \mid s, \cdot)$, and the area under the receiver operating characteristic curve, $\mathrm{AUC}(\tau_0 \mid s)$, follow:

$$\begin{aligned} \mathrm{ROC}(\tau_0 \mid s, \cdot) &= \{\mathrm{TPF}\big(\tau_0 \mid s, \mathrm{FPF}^{-1}(\tau_0 \mid s, \psi)\big),\ \psi \in (0,1)\} \quad \text{and} \\ \mathrm{AUC}(\tau_0 \mid s) &= \int \mathrm{ROC}(\tau_0 \mid s, \psi) d\psi \end{aligned}$$

Definitions of proportion of cases followed (PCF) and proportion of population needed to be followed (PNF) In the longitudinal setting, we define the proportion of cases followed, $(\mathrm{PCF}(\tau_0 \mid s, q))$, as the proportion of subjects who experience an event in time τ_0 from s, conditional on being at risk at time s, if we follow a proportion q of individuals at highest conditional risk, $R(\tau_0 \mid s)$, in the population (Figure 4.1). Let ψ_q denote a risk threshold such that $\mathrm{P}(R(\tau_0 \mid s) > \psi_q) = q$. Then

$$\mathrm{PCF}(\tau_0 \mid s, q) = \mathrm{P}(R(\tau_0 \mid s) > \psi_q \mid s < T \leq s + \tau_0)$$

A related measure, the proportion of the population needed to be followed, $\mathrm{PNF}(\tau_0 \mid s, p)$, denotes the proportion of the population at highest conditional risk that needs to be followed in order to capture a proportion p of cases (Figure 4.1). Let ψ_p denote a risk threshold such

that

$$\mathrm{P}(R(\tau_0 \mid s) > \psi_p \mid s < T \leq s + \tau_0) = p$$

Then

$$\mathrm{PNF}(\tau_0 \mid s, p) = \mathrm{P}(R(\tau_0 \mid s) > \psi_p \mid T > s)$$

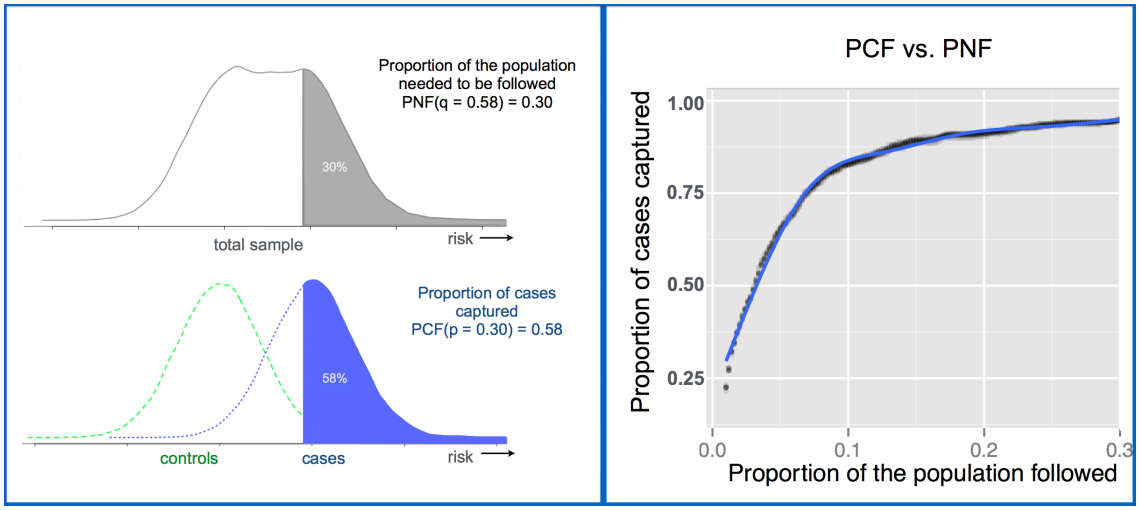

Figure 4.1: The relationship between the proportion of cases followed (PCF) and the proportion of population to be followed (PNF).

Definition of prediction error Definition of prediction error (PE) based on risk in a longitudinal setting and its estimation in a cohort setting was developed by Schoop *et al.* [Schoop et al., 2008, Schoop et al., 2011]. We emphasize that it is their definition and estimator of PE that we present in this chapter, and we chose to introduce it here for clarity of presentation. Prediction error is a measure of calibration that we would like to make available for evaluating model calibration in two-phase studies. We introduce all measures that we provide estimators for under two-phase study designs in this chapter, as we only present the estimation of additional weights that are needed to account for sampling bias

in the next chapter. Thus, the presentation of the estimators in this chapter is general in the sense that the form of the estimators is the same under two-phase sampling; only the estimation of the weights is different in two-phase studies. In the cohort, the weights only account for censoring, whereas in two-phase studies the weights account for censoring and sampling bias. The latter is the focus of Chapter 5.

$\mathrm{PE}(\tau_0 \mid s)$, as defined by [Schoop et al., 2008, Schoop et al., 2011], is

$$\mathrm{PE}(\tau_0 \mid s) = E\big((\mathrm{I}(s < T \leq s + \tau_0) - R(\tau_0 \mid s))^2 \mid T > s\big)$$

4.4 Estimation of prediction performance measures under longitudinal cohort studies

In this section we describe nonparametric estimation procedures of our performance measures. Previous work done on TPF and FPF in a longitudinal setting used semiparametric estimation methods [Zheng and Heagerty, 2004, Zheng and Heagerty, 2007]. We opted for nonparametric estimation as a way to increase robustness and require fewer assumptions.

We assume that censoring time is independent of covariates and event time. That assumption is required since we use Kaplan-Meier to estimate the censoring distribution, which we then use to estimate the inverse probability weights to account for censoring. The assumption of independence between the censoring time and covariates can be relaxed if the censoring distribution were to be estimated with a method allowing for covariate adjustment, such as Cox regression.

Estimation of the true and false positive fractions

$$\widehat{\mathrm{TPF}}(\tau_0 \mid s, \psi) = \frac{\sum_{i=1}^{n} \mathrm{I}(\widehat{R}_i(\tau_0 \mid s) > \psi)\, \mathrm{I}(s < X_i \leq s + \tau_0)\, \widehat{w}_i(\tau_0 \mid s)}{\sum_{i=1}^{n} \mathrm{I}(s < X_i \leq s + \tau_0)\, \widehat{w}_i(\tau_0 \mid s)} \tag{4.1}$$

$$\widehat{\mathrm{FPF}}(\tau_0 \mid s, \psi) = \frac{\sum_{i=1}^{n} \mathrm{I}(\widehat{R}_i(\tau_0 \mid s) > \psi)\, \mathrm{I}(X_i > s + \tau_0)\, \widehat{w}_i(\tau_0 \mid s)}{\sum_{i=1}^{n} \mathrm{I}(X_i > s + \tau_0)\, \widehat{w}_i(\tau_0 \mid s)} \tag{4.2}$$

where

$$\begin{aligned}\widehat{w}_i(\tau_0 \mid s) &= \widehat{w}_i^{cens}(\tau_0 \mid s) \\ &= \delta_i \, \mathrm{I}(s < X_i \leq s + \tau_0) \, \frac{\widehat{G}(s)}{\widehat{G}(X_i)} + \mathrm{I}(X_i > s + \tau_0) \, \frac{\widehat{G}(s)}{\widehat{G}(s + \tau_0)} \end{aligned} \tag{4.3}$$

is the inverse probability weight (IPW) to account for censoring and $\widehat{G}(\cdot)$ is the Kaplan-Meier estimate of $G(\cdot)$, the censoring distribution.

Estimation of the area under the ROC curve (AUC) The $AUC(\tau_0 \mid s)$ is estimated by

$$\widehat{AUC}(\tau_0|s) = \frac{1}{2} \sum_{i=1}^{n} \left[\left(\widehat{\mathrm{TPF}}(\tau_0|s, \psi_{(i)}) + \widehat{\mathrm{TPF}}(\tau_0|s, \psi_{(i+1)}) \right) \cdot \left(\widehat{FPF}(\tau_0|s, \psi_{(i)}) - \widehat{\mathrm{FPF}}(\tau_0|s, \psi_{(i-1)}) \right) \right] \tag{4.4}$$

where $\psi_{(i)}$ is the i^{th} ordered risk $\widehat{R}_{(i)}(\tau_0 \mid s)$, $\psi_{(0)} = 0$, $\psi_{(n+1)} = 1$, $\widehat{\mathrm{FPF}}(\tau_0 \mid s, \psi_{(0)}) = 0$ and $\widehat{\mathrm{TPF}}(\tau_0 \mid s, \psi_{(n+1)}) = 1$.

Estimation of the proportion of cases followed (PCF) proportion of population needed to be followed (PNF) To estimate the $\mathrm{PCF}(\tau_0 \mid s, \cdot)$ we first sort the data according to decreasing risk, $\widehat{R}_{\cdot}(\tau_0 \mid s)$. We find the largest k satisfying the following inequality:

$$\frac{\sum_{i=1}^{k} \mathrm{I}(X_i > s) \, \widehat{w}_i(\tau_0 \mid s)}{\sum_{i=1}^{n} \mathrm{I}(X_i > s) \, \widehat{w}_i(\tau_0 \mid s)} \leq q \tag{4.5}$$

Then the estimator of $\mathrm{PCF}(\tau_0 \mid s, \cdot)$ is defined as

$$\widehat{\mathrm{PCF}}(\tau_0 \mid s, q) = \frac{\sum_{i=1}^{k} \mathrm{I}(s < X_i \leq s + \tau_0) \, \widehat{w}_i(\tau_0 \mid s)}{\sum_{i=1}^{n} \mathrm{I}(s < X_i \leq s + \tau_0) \, \widehat{w}_i(\tau_0 \mid s)} \tag{4.6}$$

To estimate the $\mathrm{PNF}(\tau_0 \mid s, \cdot)$ we sort the data by decreasing risk, $\widehat{R}_{\cdot}(\tau_0 \mid s)$, and we find

the largest k satisfying the following inequality:

$$\frac{\sum_{i=1}^{k} \mathrm{I}(s < X_i \leq s + \tau_0)\, \widehat{w}_i(\tau_0 \mid s)}{\sum_{i=1}^{n} \mathrm{I}(s < X_i \leq s + \tau_0)\, \widehat{w}_i(\tau_0 \mid s)} \leq p \tag{4.7}$$

then the $\mathrm{PNF}(\tau_0 \mid s, \cdot)$ is estimated by

$$\widehat{\mathrm{PNF}}(\tau_0 \mid s, p) = \frac{\sum_{i=1}^{k} \mathrm{I}(X_i > s)\, \widehat{w}_i(\tau_0 \mid s)}{\sum_{i=1}^{n} \mathrm{I}(X_i > s)\, \widehat{w}_i(\tau_0 \mid s)} \tag{4.8}$$

Estimation of prediction error The estimator of the prediction error, $\mathrm{PE}(\tau_0 \mid s)$, is

$$\begin{aligned} \widehat{\mathrm{PE}}(\tau_0 \mid s) \;=\; & \frac{1}{\sum_{i=1}^{n} \widehat{w}_i(\tau_0 \mid s)} \sum_{i=1}^{n} \Big(\mathrm{I}(s < X_i \leq s + \tau_0) \left(1 - \widehat{R}_i(\tau_0 \mid s)\right)^2 \widehat{w}_i(\tau_0 \mid s) \\ & + \mathrm{I}(X_i > s + \tau_0) \left(\widehat{R}_i(\tau_0 \mid s)\right)^2 \widehat{w}_i(\tau_0 \mid s) \Big) \end{aligned} \tag{4.9}$$

4.4.1 Estimation of prediction performance measures under longitudinal cohort studies when measurement times vary

When measurement times vary substantially from those specified in the study protocol, but can still be assumed to be independent of censoring and event times, estimation procedures incorporating last-value-carried-forward (LVCF) can be improved upon with the Best Linear Unbiased Predictor (BLUP) approach. Let s denote the time of interest. In the LVCF approach, the last available measurement before time s is carried forward and imputed for any missing observations at time s. This approach can cause bias in the estimation of model parameters [Tsiatis and Davidian, 2004]. This bias is likely to be substantial in settings where the slope of the marker trajectory is steep, or negligible if the marker trajectories are relatively stable.

The BLUP approach developed in section 3.3.3 uses the marker trajectory information in the full cohort to inform about the marker trajectory of a new subject at some future

time point $s+\tau_0$ given information up to s (equation 3.7). Thus, in settings where the slope of the marker trajectory is nonzero, and when the marker trajectory can be modeled well, using BLUP to impute the observations at time s can lead to reduced bias in the estimation of model parameters.

4.5 Inference for estimators of prediction performance measures under longitudinal cohort studies

The *bootstrap* was introduced in 1979 as a resampling-based method for estimating the standard error of arbitrarily mathematically complicated estimators [Efron, 1979, Efron and Tibshirani, 1993]. It has been widely used in practice to estimate standard errors of estimators that themselves can be estimated using standard methods, but for which the derivation of an explicit expression for the variance is difficult or not feasible.

In the bootstrap, the sampling is done with replacement. The weights have a multinomial distribution with parameters n and $(1/n, ..., 1/n)$, where n is the sample size. Thus, for a given bootstrap sample, some individuals may not be sampled at all, others may be sampled up to n times. The estimation of the quantity of interest, Q, proceeds in the usual way using the data in the bootstrapped sample. Repeating this procedure B times results in B estimates $\widehat{Q}^b$, $b = 1, \ldots, B$, and the empirical variance of those is the variance of $\widehat{Q}$.

A different approach, which is referred to in the literature as *perturbation* or simply a *resampling method* for variance estimation, was originally developed as an alternative to the bootstrap [Rao and Zhao, 1992, Parzen et al., 1994, Jin et al., 2001]. Similarly to the bootstrap, perturbation is a resampling-based variance estimation method. Intuitively, perturbation adds error of known magnitude to the estimation of a quantity of interest, Q, and estimates the magnitude of the variability in $\widehat{Q}$ that the perturbation resulted in. In perturbation, the weights are continuous, positive random variables with expectation and variance equal to 1. Thus, for a given iteration of the perturbation, each individual's data has a continuos weight associated with it and though a given individual's contribution to

the estimation may be up- or down-weighted, it is always there.

Either the bootstrap or the perturbation could be used for variance estimation of estimators developed in this chapter, where the estimation is based on a cohort. Thus, in the early phase of the project we considered using the bootstrap for this part of the project. As the project matured, we decided to use perturbation for two reasons. One, our estimators are nonparametric, thus not likely to be smooth. There appears to be some emphasis in the literature on the fact that perturbation works well even for non-smooth estimators. We note, however, that the bootstrap tends to work quite well in practice in such situations as well. The second reason, and this was the driving one, was that the bootstrap does not provide valid inference in two-phase studies sampled under finite sampling, whereas perturbation does [Cai and Zheng, 2013]. Thus, since we wanted to develop inference procedures that could be used in a wide variety of settings, we opted for perturbation as the variance estimation method for all parts of this work.

Next, we describe one general approach to the perturbation. This was the basis for the perturbation procedures used for variance estimation in our work, and was based on the work of [Uno et al., 2007]. There are other approaches to using perturbation to estimate the variance of a quantity of interest, see section 2.4 for additional details.

In general, the placement of the weights can be thought of as recursive. Consider the following simple example (adapted from [Uno et al., 2007]) where we explain how to estimate a perturbed overall misclassification rate (OMR). For a working model $\mathrm{P}(T \leq t \,|\, Z) = g(\beta Z)$, where $g(\cdot)$ is a known, strictly increasing, differentiable function, T denotes the event time *without censoring* (only for the purposes of this example), Z denotes a baseline covariate and β is an unknown parameter. β can be estimated by solving the following equation:

$$U(\beta) = \frac{1}{n} \sum_{i=1}^{n} Z_i \left(\mathrm{I}(T_i \leq t) - g(\beta Z_i) \right) = 0$$

To obtain a perturbed estimate of β, we would generate V_{ip}, $i = 1, \ldots, n$, from a known

distribution with expectation and variance equal to 1, for example $V_i \sim$ exponential(1). Then, we can estimate the p^{th} perturbed β, $\widehat{\beta}_p^*$, $p = 1, \ldots, P$, by solving

$$U^*(\beta) = \frac{1}{n}\sum_{i=1}^{n} V_{ip}\big(Z_i \left(\mathrm{I}(T_i \leq t) - g(\beta Z_i)\right)\big) = 0$$

In recursion terms, this estimating equation would be our baseline, since it only uses the observed data and the perturbation weights in the estimation. Next, we will use $\widehat{\beta}_p^*$ in estimation of the perturbed OMR. Uno *et al.* define OMR as

$$D(c) = \mathrm{E}|\mathrm{I}(T_\circ \leq t) - \mathrm{I}(g(\beta Z_\circ) > c)|$$

where $\cdot_\circ$ denotes a future individual. For some threshold c, $D(c)$ can be estimated as

$$\widehat{D}(c) = \frac{1}{n}\sum_{i=1}^{n} |\mathrm{I}(T_i \leq t) - \mathrm{I}(g(\widehat{\beta} Z_i) > c)|$$

To estimate the p^{th} perturbed $\widehat{D}(c)$, $\widehat{D}_p^*(c)$, we will use the perturbed estimate of $\widehat{\beta}_p^*$ from the earlier step and perturb each individual's contribution to the estimation of $\widehat{D}_p^*(c)$. Thus

$$\widehat{D}_p^*(c) = \frac{1}{n}\sum_{i=1}^{n} V_{ip}\, |\mathrm{I}(T_i \leq t) - \mathrm{I}(g(\widehat{\beta}_p^* Z_i) > c)|$$

This concludes our example. Estimation for more complex scenarios proceeds in a similar fashion, starting with estimation based on the data and perturbation weights, and possibly using those estimates as part of the estimator of more complex quantities.

Uno *et al.* [Uno et al., 2007] provided theoretical justification for perturbation and used it to estimate standard errors of nonparametric estimators of true and false positive fractions in a time-dependent setting for a marker measured at baseline. The authors dealt with a situation where the marker is measured at baseline, event status is time-dependent and they

used inverse probability weighting (IPW) to account for censoring in estimation [Uno et al., 2007].

A general inference procedure for performance measures estimated under longitudinal cohort studies The variance of each of our performance measures can be estimated as follows:

1. Generate $n \times P$ independent and identically distributed random variables V_{ip} from a known distribution with $\mathrm{E}(V_{ip}) = 1$ and $\mathrm{Var}(V_{ip}) = 1$, and $\mathbf{V}_{n\times P} = \{V_{ip},\ i = 1, \ldots, n,\ p = 1, \ldots, P\}$.

2. Use $\mathbf{V}_{n\times P}$ to obtain P perturbed estimates of

   (a) the censoring distribution $\widehat{G}^*_p(\cdot)$

   (b) the IPW conditional on s to account for censoring, $\widehat{w}^{*cens}_{ip}(\tau_0 \mid s)$. We can then estimate

   (c) the perturbed conditional risk, $\widehat{R}^*_{ip}(\tau_0 \mid s)$ and

   (d) the evaluation measures $\widehat{\mathrm{M}}^*_p(\tau_0 \mid s)$, where M denotes any of {TPF, FPF, AUC, PCF, PNF, PE} and $*$ denotes a perturbed estimate.

3. The empirical variance of the P estimates $\widehat{\mathrm{M}}^*_p(\tau_0 \mid s)$, $p = 1, \ldots, P$, is the estimated variance of $\widehat{\mathrm{M}}(\tau_0 \mid s)$, M = {TPF, FPF, AUC, PCF, PNF, PE}.

The details of estimation for each iteration $p = 1, \ldots, P$ in step 2 are described next.

Inference for the true positive fraction (TPF) and false positive fraction (FPF) The p^{th} perturbed estimator of $\text{TPF}(\tau_0 \mid s, \psi)$ and $\text{FPF}(\tau_0 \mid s, \psi)$ is

$$\widehat{\text{TPF}}_p^*(\tau_0 \mid s, \psi) = \frac{\sum_{i=1}^n \text{I}(\widehat{R}_{ip}^*(\tau_0 \mid s) > \psi)\, \text{I}(s < X_i \leq s + \tau_0)\, \widehat{w}_{ip}^*(\tau_0 \mid s)}{\sum_{i=1}^n \text{I}(s < X_i \leq s + \tau_0)\, \widehat{w}_{ip}^*(\tau_0 \mid s)} \tag{4.10}$$

$$\widehat{\text{FPF}}_p^*(\tau_0 \mid s, \psi) = \frac{\sum_{i=1}^n \text{I}(\widehat{R}_{ip}^*(\tau_0 \mid s) > \psi)\, \text{I}(X_i > s + \tau_0)\, \widehat{w}_{ip}^*(\tau_0 \mid s)}{\sum_{i=1}^n \text{I}(X_i > s + \tau_0)\, \widehat{w}_{ip}^*(\tau_0 \mid s)} \tag{4.11}$$

where $\widehat{R}_{ip}^*(\tau_0 \mid s)$ denotes the p^{th} perturbed risk estimate for individual i conditional on data available up to time s and

$$\begin{aligned} \widehat{w}_{ip}^*(\tau_0 \mid s) &= \widehat{w}_{ip}^{*cens}(\tau_0 \mid s) \\ &= V_{ip}\left(\delta_i\, \text{I}(s < X_i \leq s + \tau_0)\frac{\widehat{G}_p^*(s)}{\widehat{G}_p^*(X_i)} + \text{I}(X_i > s + \tau_0)\frac{\widehat{G}_p^*(s)}{\widehat{G}_p^*(s + \tau_0)}\right) \end{aligned} \tag{4.12}$$

is the p^{th} perturbed estimated inverse probability weight (IPW) to account for censoring and $\widehat{G}_p^*(\cdot)$ is the p^{th} perturbed Kaplan-Meier estimate of $G(\cdot)$, the censoring distribution. For details on estimation of $\widehat{R}_{ip}^*(\tau_0 \mid s)$, see Chapter 3.

Inference for the area under the ROC curve (AUC) The p^{th} perturbed estimator of $AUC(\tau_0 \mid s)$ is

$$\begin{aligned} \widehat{\text{AUC}}_p^*(\tau_0 \mid s) = \frac{1}{2}\sum_{i=1}^n \Big[&\big(\widehat{\text{TPF}}_p^*(\tau_0 \mid s, \psi_{(i)}) + \widehat{\text{TPF}}_p^*(\tau_0 \mid s, \psi_{(i+1)})\big) \\ &\times \big(\widehat{\text{FPF}}_p^*(\tau_0 \mid s, \psi_{(i)}) - \widehat{\text{FPF}}_p^*(\tau_0 \mid s, \psi_{(i-1)})\big)\Big] \end{aligned} \tag{4.13}$$

where $\psi_{(i)}$ is the i^{th} ordered p^{th} perturbed risk $\widehat{R}_{(i)p}^*(\tau_0 \mid s)$, $\psi_{(0)} = 0$, $\psi_{(n+1)} = 1$, $\widehat{\text{FPF}}_p^*(\tau_0 \mid s, \psi_{(0)}) = 0$ and $\widehat{\text{TPF}}_p^*(\tau_0 \mid s, \psi_{(n+1)}) = 1$.

Inference for the proportion of cases followed (PCF) and the proportion of population needed to be followed (PNF) To estimate the p^{th} perturbed $\mathrm{PCF}(\tau_0 \mid s, \cdot)$, $\widehat{\mathrm{PCF}}_p^*(\tau_0 \mid s, \cdot)$, we sort the data according to decreasing perturbed risk, $\widehat{R}_{\cdot p}^*(\tau_0 \mid s)$. We find the largest k satisfying the following inequality:

$$\frac{\sum_{i=1}^{k} \mathrm{I}(X_i > s)\, \widehat{w}_{ip}^*(\tau_0 \mid s)}{\sum_{i=1}^{n} \mathrm{I}(X_i > s)\, \widehat{w}_{ip}^*(\tau_0 \mid s)} \leq q \tag{4.14}$$

Then the p^{th} perturbed estimator of $\mathrm{PCF}(\tau_0 \mid s, \cdot)$ is

$$\widehat{\mathrm{PCF}}_p^*(\tau_0 \mid s, q) = \frac{\sum_{i=1}^{k} \mathrm{I}(s < X_i \leq s + \tau_0)\, \widehat{w}_{ip}^*(\tau_0 \mid s)}{\sum_{i=1}^{n} \mathrm{I}(s < X_i \leq s + \tau_0)\, \widehat{w}_{ip}^*(\tau_0 \mid s)} \tag{4.15}$$

To estimate the p^{th} perturbed $\mathrm{PNF}(\tau_0 \mid s, \cdot)$, $\widehat{\mathrm{PNF}}_p^*(\tau_0 \mid s, \cdot)$, we sort the data according to decreasing perturbed risk, $\widehat{R}_{\cdot p}^*(\tau_0 | s)$, and find the largest k satisfying the following inequality:

$$\frac{\sum_{i=1}^{k} \mathrm{I}(s < X_i \leq s + \tau_0)\, \widehat{w}_{ip}^*(\tau_0 \mid s)}{\sum_{i=1}^{n} \mathrm{I}(s < X_i \leq s + \tau_0)\, \widehat{w}_{ip}^*(\tau_0 \mid s)} \leq p \tag{4.16}$$

then the p^{th} perturbed estimator of $\mathrm{PNF}(\tau_0 \mid s, \cdot)$ is

$$\widehat{\mathrm{PNF}}_p^*(\tau_0 \mid s, p) = \frac{\sum_{i=1}^{k} \mathrm{I}(X_i > s)\, \widehat{w}_{ip}^*(\tau_0 \mid s)}{\sum_{i=1}^{n} \mathrm{I}(X_i > s)\, \widehat{w}_{ip}^*(\tau_0 \mid s)} \tag{4.17}$$

Prediction error The p^{th} perturbed estimator of $\mathrm{PE}(\tau_0, |s)$ is

$$\begin{aligned} \widehat{\mathrm{PE}}_p^*(\tau_0 \mid s) \;=\; & \frac{1}{\sum_{i=1}^{n} \widehat{w}_{ip}^*(\tau_0 \mid s)} \sum_{i=1}^{n} \Big(\mathrm{I}(s < X_i \leq s + \tau_0) \left(1 - \widehat{R}_{ip}^*(\tau_0 \mid s)\right)^2 \widehat{w}_{ip}^*(\tau_0 \mid s) \\ & + \mathrm{I}(X_i > s + \tau_0) \left(\widehat{R}_{ip}^*(\tau_0 \mid s)\right)^2 \widehat{w}_{ip}^*(\tau_0 \mid s) \Big) \end{aligned} \tag{4.18}$$

4.6 Simulation studies

4.6.1 Data Generation

The longitudinal marker data was generated following the general setup used in the JM literature ([Tsiatis et al., 1995, Wulfsohn and Tsiatis, 1997]). We assumed a linear mixed effects model with measurement error: $Y_i(u) = W_i(u) + e_i(u) = \alpha_{0i} + \alpha_{1i} f(u) + e_i(u)$ where $f(u) = log(u/\nu_1)$, $\nu_1 = 30$, with a Weibull baseline hazard: $\lambda_0(u) = v/\nu_2(u/\nu_2)^{v-1}$, scale $\nu_2 = 20$ and shape $v = 1.4$. The random components $\alpha_i = (\alpha_{0i}, \alpha_{1i})$ were generated as a bivariate normal with mean $(\mu_{\alpha_0}, \mu_{\alpha_1})^T = (0.6, -0.1)^T$, or $(\mu_{\alpha_0}, \mu_{\alpha_1})^T = (0.6, -1.0)^T$, and a covariance matrix $\mathbf{\Sigma}_\alpha = \begin{bmatrix} 0.83^2 & -0.005 \\ -0.005 & 0.13^2 \end{bmatrix}$. The measurement error $e_i(u) \sim_{iid} N(0, \sigma_e^2)$, $i = 1, ..., n$, $j = 1, ..., m_i$ and $\sigma_e = 0.1$ and 1.0, representing two scenarios: small measurement errors and large measurement errors. Figure 4.2 shows the biomarker trajectories for subjects under the simulation settings varying measurement error σ_e and the slope of the marker trajectory μ_{α_1}. Failure time was assumed to depend on the covariate (without error) through a proportional hazards relationship: $\lambda_i(u) = \lambda_0(u) \exp\{\beta W_i(u)\}$, where $\beta = -1.5$. Censoring time was generated from an exponential distribution (rate = 0.01), with administrative censoring at 180 months. There were up to 10 measurements per subject taken at 6-month intervals (scenario 1) or at 6-month intervals $\pm\, u_{ij}$ (scenario 2) (Figure 4.3), where $U_{ij} \sim$ Unif$(0, 3)$, $i = 1, \ldots, n$, $j = 2, \ldots, m_i$ ensuring that $min(T_i, C_i) \geq (s_{ij} + u_{ij})$. We were interested in predicting the risk of an event at time $s + \tau_0$ given biomarker data up to time s: $R_i(\tau_0 \mid s) = P\big(T_i \leq s + \tau_0 \mid T_i > s, \mathbf{Y}_i(s)\big)$. All simulations were run in `R` (www.r-project.org).

4.6.2 True values

The true values were estimated as follows: we generated two datasets of size n = 500,000, without censoring. One served as a training set and the other as a validation set. We fit

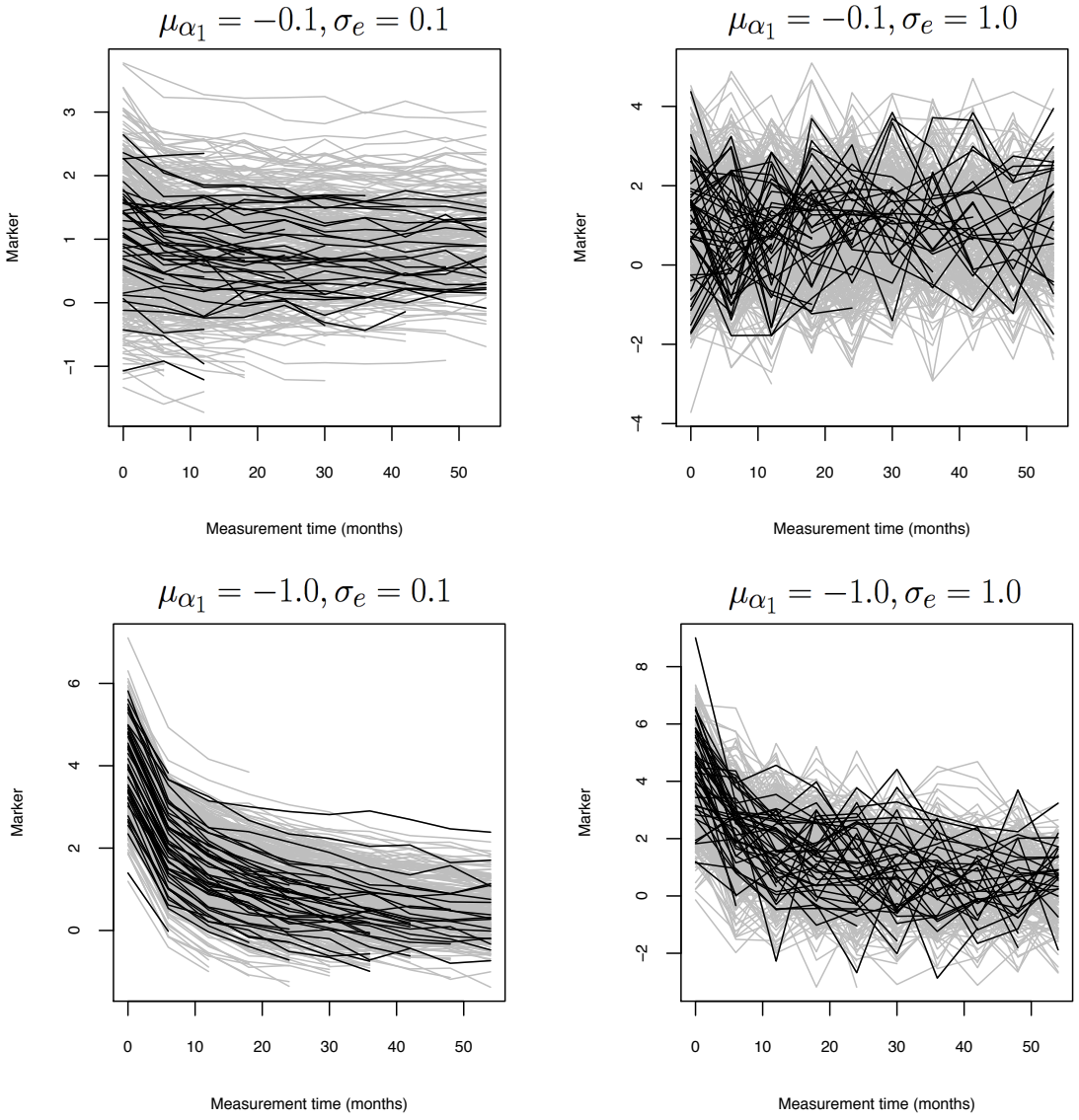

Figure 4.2: Example simulated datasets with $\mu_{\alpha_1} = -0.1$ and standard deviation of the measurement error $\sigma_e = 0.1$ (top left panel), $\mu_{\alpha_1} = -0.1$ and $\sigma_e = 1.0$ (top right), $\mu_{\alpha_1} = -1.0$ and $\sigma_e = 0.1$ (bottom left), $\mu_{\alpha_1} = -1.0$ and $\sigma_e = 1.0$ (bottom right).

the partly conditional logistic (4.19) model to the training set. Using the estimates from the model, observed marker values and the true event status from the validation set we estimated the true values for each of the measures of prediction quality for four sets of $(s, \tau_0) = \{(24, 12), (48, 12), (24, 24), (48, 24)\}$ for every combination of $\sigma_e = \{0.1, 1.0\}$ and $\mu_{\alpha_1} = \{-0.1, -1.0\}$ (Figure 4.2).

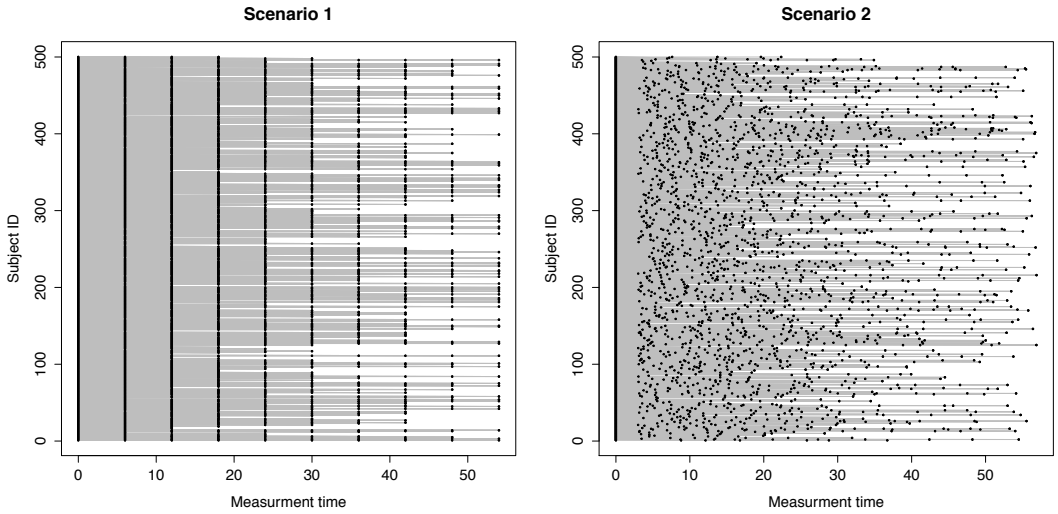

Figure 4.3: Example datasets showing simulated visit times in simulation scenarios 1 and 2. In scenario 1, all subjects have visits and marker measurements at 6-month time intervals as specified in the study protocol (left panel). In scenario 2, all subjects have a baseline visit and a measurement at time zero, but the following visits take place at random times and may occur up to 3 months before or after the scheduled visit (right panel). The visit times are assumed to be independent of the event time, censoring time and covariates, measured or unmeasured. The black dots indicate visits, the grey lines follow each subject until their event time or censoring time.

4.6.3 Evaluation of bias and variance

In this set of simulations we evaluated the performance of the estimation procedures to estimate measures of prediction quality, as well as the variance estimation procedures. We

report the bias, empirical variance, perturbation variance and coverage probability for the true positive fraction at a risk threshold of 0.4 (TPF(0.4)), the false positive fraction at 0.3 risk threshold (FPF(0.3)), the area under the receiver operating characteristic curve (AUC), the proportion of cases followed in the top 20% of subjects at highest risk (PCF(0.20)) and the proportion of the population needed to be followed in order to capture 80% of the cases (PCF(0.80)).

The simulations were run as follows. The data was generated according to scenario 1, where all measurements are taken at equally spaced time intervals. For each iteration, k, of the simulation, we generated two datasets, a training and a validation set with the same simulation parameters and both with censored data. We fit the following model

$$P\big(\mathcal{T}_{ij} \leq \tau_0 \mid s_{ij}, \mathbf{H}_i(s_{ij}), \mathcal{T}_{ij} > 0\big) = g\big(\boldsymbol{\alpha}'\mathbf{B}(s_{ij}) + \beta Y_i(s_{ij})\big) \tag{4.19}$$

where $\mathbf{B}(s)$ is a natural spline of measurement time with df = 3, to the full training set, and predicted the risk of an event for each individual in the validation set using their data at time s and the model parameter estimates. We then estimated the measures of prediction quality using the predicted risk and event status of individuals in the validation set at time s (censored) using estimators described in section 4.4.

To estimate the variance of our estimators, we generated P perturbation weights, V_i, $i = 1, \cdots, P$, for each individual in the validation set, where $V_i \sim exp(1)$. We then estimated P measures of prediction quality using the weighted estimators as described in section 4.5. The variance of a given estimator is then the sample variance of the P perturbed estimators. We ran the following sets of simulations: n = {500, 1000, 2000, 4000}, K = 1000, P = 500, $(s, \tau_0) = \{(24, 12), (48, 12), (24, 24), (48, 24)\}$, $\sigma_e = \{0.1, 1.0\}$ and $\mu_{\alpha_1} = \{-0.1, -1.0\}$. In this set of simulations we report the true value of each measure (True), the estimate (Est), percent bias (Bias %) calculated as ((True-Est)/True$\times$100), empirical standard errors (SE_{emp}), average of the estimated perturbed standard errors (SE_{pert}) and the empirical cov-

erage probability of the *true* values by the 95% confidence intervals ($\times$ 100) constructed as Est $\pm 1.96 \times \text{SE}_{\text{pert}}$ (CP %).

4.6.4 Evaluation of bias when measurement times vary

In this set of simulations we generated data according to scenario 2, where the marker is measured at baseline (time = 0) and then at 6-month intervals $\pm u_{ij}$ months, where $U_{ij} \sim \text{Uniform}(0, 3)$, $i = 1, \ldots, n$, $j = 2, \ldots, m_i$, ensuring that $min(T_i, C_i) \geq (s_{ij} + u_{ij})$. Thus, the measurement time does not depend on any measured or unmeasured covariates, survival time nor censoring time. One approach to deal with irregular measurement times is last value carried forward (LVCF), where the last value observed before the conditioning time, time s, or the time at which the prediction is made, is "carried forward" to time s. We proposed another approach that uses the best linear unbiased predictor (BLUP), where the marker trajectory is modeled and the marker value is predicted from the fitted model and the new individual's data up to time s. The predicted marker value is used for risk prediction, rather than the observed marker value carried forward to time s. We expected there to be most bias when measurement error is small and the marker slope is large, and based on preliminary simulation results that, in fact, is the case. Hence, we focused our simulations on such a scenario and considered $\sigma_e = 0.1$ and the following slopes $\mu_{\alpha_1} = \{-0.1, -1.0, -2.0, -3.0\}$ (see Figure 4.4). In this set of simulations we report the true value of each measure (True), the estimate (Est), percent bias (Bias %) calculated as ((True-Est)/True$\times$100), empirical standard errors (SE_{emp}).

Variable measurement times, LVCF

For each iteration, k, of the simulation, we generated two datasets, a training and a validation set with the same simulation parameters and both with censored data according to measurement spacing described in scenario 2. We fit model 4.19 to the observed

marker values in the full training set. The predicted risk of an event for each individual in the validation set was based on their last observed marker value before time s and the model parameter estimates. Measures of prediction quality and their variance were estimated as described in section 4.6.3. We ran the following sets of simulations: n = 2000, K = 1000, P = 500, $(s, \tau_0) = \{(24, 12), (48, 12), (24, 24), (48, 24)\}$, $\sigma_e = \{0.1, 1.0\}$ and $\mu_{\alpha_1} = \{-0.1, -1.0, -2.0, -3.0\}$.

Variable measurement times, BLUP

For each iteration, k, of the simulation, we generated two datasets, a training and a validation set with the same simulation parameters and both with censored data according to measurement spacing described in scenario 2. We fit the following linear mixed effects model:

$$Y_{ij} = \beta_0 + \beta_1 log(s_{ij} + 1) + b_{0i} + b_{1i} log(s_{ij} + 1) + e_{ij} \tag{4.20}$$

and obtained predicted marker values for all individuals in the training set at regularly spaced 6-month time intervals (BLUP training set). For individuals in the validation set, we used parameter estimates from model 4.20 fit to the training dataset, and individual data up to time s to obtain predicted marker values at time s (BLUP validation set). We then fit the PC_{GLM} model 4.19 to the BLUP training set. The predicted risk of an event for each individual in the BLUP validation set was based on their BLUP marker value at time s and the PC_{GLM} model parameter estimates. Measures of prediction quality and their variance were estimated as described in section 4.6.3. We ran the following sets of simulations: n = 2000, K = 1000, P = 500, $(s, \tau_0) = \{(24, 12), (48, 12), (24, 24), (48, 24)\}$, $\sigma_e = \{0.1, 1.0\}$ and $\mu_{\alpha_1} = \{-0.1, -1.0, -2.0, -3.0\}$.

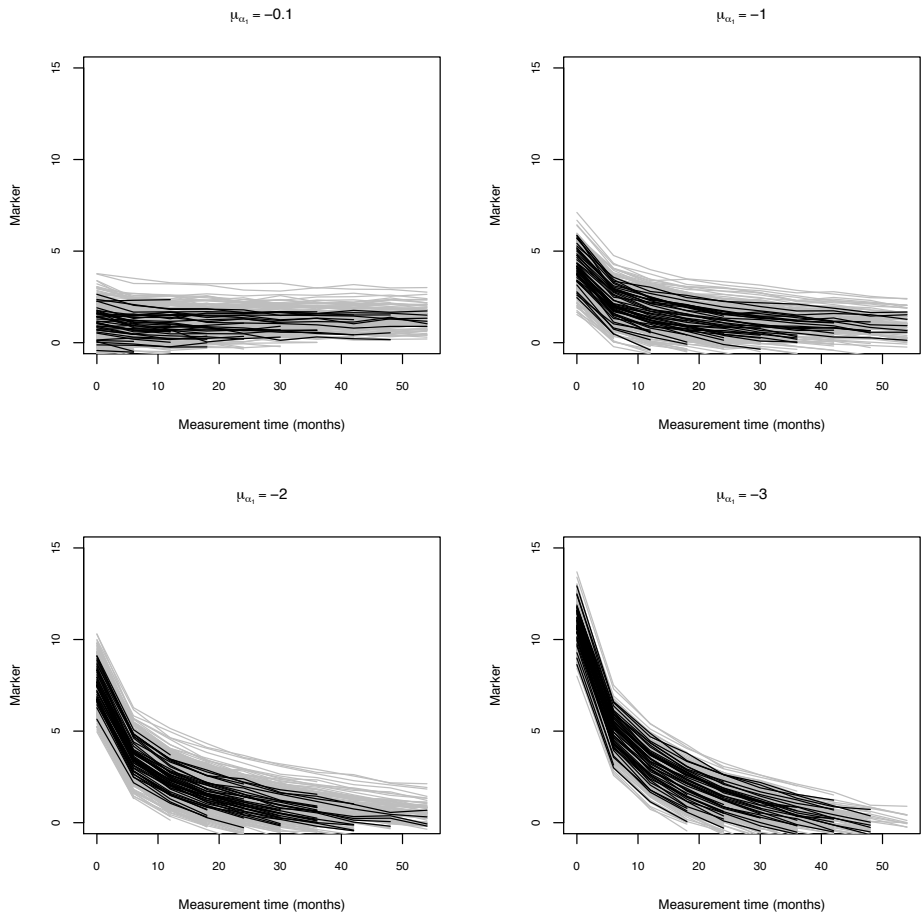

Figure 4.4: Simulated datasets showing the effect of μ_{α_1} on the shape of the trajectory of the marker. The parameters used to generate the data are: $n = 500$, β = -1.5, $\mu_{\alpha_0} = 0.6$, $\sigma_{\alpha_0} = 0.83$, $\sigma_{\alpha_1} = 0.13$, $cov(\alpha_0, \alpha_1)$ = -0.005, $\sigma_e = 0.1$. Maximum number of visits is 10, with visit times every 6 months. For clarity of presentation, all trajectories are plotted in gray, with trajectories of a random sample of 10% of subjects plotted in black.

4.6.5 Simulation results

Simulation results: evaluation of bias and variance For the small error, $\sigma_e = 0.1$, and marker slope of $\mu_{\alpha_1} = -0.1$, the coverage probability (CP) ranged from 93.2-96.4% for all measures and all four sets of (s, τ_0). The bias was $\leq 1.4\%$ for all measures and the perturbation standard error estimates (SE_{pert}) were very close to the empirical standard errors (SE_{emp}) (Table 4.1). In the simulation with $\sigma_e = 1.0$ and $\mu_{\alpha_1} = -0.1$, the bias in all measures was $\leq 1.3\%$, SE_{pert} were very close to SE_{emp} and the CP ranged from 92.4-96.4% (Table 4.2). For the simulation with $\sigma_e = 0.1$ and $\mu_{\alpha_1} = -1.0$ the bias was $\leq 1.4\%$ for all measures and the CP ranged from 93.8-96.5% (Table 4.3). In the setting with a large measurement error, $\sigma_e = 1.0$, and marker slope of $\mu_{\alpha_1} = -1.0$ the estimates were within 1.3% of the truth and the SE_{pert} were very close to SE_{emp}. The CP for FPF(0.3) at $(s, \tau_0) = (48, 24)$ was low at 90.1%, despite the fact that bias was 0.3% (est = 0.965, truth = 0.968), $\text{SE}_{\text{pert}} = 0.024$ and $\text{SE}_{\text{emp}} = 0.023$. This is likely due to the estimate being close to one, and the confidence intervals truncated at one. The CP's for all other measures in this simulation ranged from 92.6-96.1% (Table 4.4).

We also present the results for $(s, \tau_0) = (48, 24)$, $\sigma_e = 0.1$ and $\mu_{\alpha_1} = -1.0$, for four sample sizes at baseline: n = $\{500, 1000, 2000, 4000\}$. This setting, with small variance of the measurement error and mean marker slope of -1.0 seemed to be the most challenging with more variability than other simulation settings. For n = 500 at baseline, the effective sample size at $s = 48$ was relatively small, with 58 events on average in the 24-month timeframe, 28 subjects were still at risk at the end of the 24 months and 14 subjects were censored. For that sample size, we noted up to 3.7% bias and CP's ranged from 92.9 to 98.0%. As the sample size increased, bias decreased, CP's stabilized around 95%. The performance of the estimators was already much better for n = 1000 (at baseline), and for n = 4000 (at baseline) the $\text{SE}_{\text{pert}} = \text{SE}_{\text{emp}}$ for all measures, up to three significant digits, and CP's ranged from 93.5-95.5 (Table 4.5).

Simulation results: evaluation of bias when measurement times vary In this simulation we compared the estimates of the measures and their empirical standard errors for two approaches for imputing missing data at time s, last value carried forward (LVCF) and best linear unbiased predictor (BLUP) when marker measurements deviate from those specified in the protocol by up to 3 months (scenario 2, Figure 4.3). We considered a scenario where the marker slope is large ($\mu_{\alpha_1} = \{2.0, 3.0\}$, Figure 4.4), measurement error is small ($\sigma_e = 0.1$). The LVCF was highly biased, especially in the timeframes when the marker trajectory was steepest, which was early on ($(s, \tau_0) = (24, 12)$). Bias in LVCF during those timeframes ranged from 2.3% to 53.4%, with the direction of the bias consistent with poorer prediction quality: PE increased, AUC decreased, PCF(0.2) decreased, PNF(0.8) increased (Table 4.6). Bias in BLUP for those timeframes ranged from 0-2.5%. For the timeframes corresponding to the leveling off of the marker trajectories, such as $(s, \tau_0) = (48, 12)$, the bias in the LVCF was smaller, but still considerably larger than that for BLUP.

4.6.6 Remarks on the simulation setup and results

We considered a wide range of scenarios in our simulations, with varying marker trajectories, varying magnitudes of measurement error, conditioning times and prediction timeframes. Overall, our estimators performed well, especially when effective sample sizes were not small (Tables 4.1 – 4.4). However, even for small sample sizes (Table 4.5) with n = 500 at baseline and the effective sample size of 100 at $s = 48$, the bias ranged from 0.9-3.7% and coverage probability (CP) ranged from 92.9-98.0%. The performance of the estimation quickly improved, and for n = 1000 at baseline, and effective sample size of 198, the maximum bias observed was 1.8% and CP ranged from 93.2-96.2%. Thus, even for small sample sizes, the performance is quite good. We saw similar trends in all simulation scenarios for n = 500 through n = 4000 at baseline (data not shown). We chose to present the trends in estimation as sample size increases for this particular set of σ_e, μ_{α_1} and (s, τ_0) because our estimators

seemed to perform least well in this setting. That is likely due to the fact that at $s = 48$ the effective sample size is much smaller than at $s = 24$, and for $\tau_0 = 24$ the number of non-events conditional on $s = 48$ is usually small (Table 4.5).

In another set of simulations in which we compared the performance of last value carried forward (LVCF) and best linear unbiased predictor (BLUP) for dealing with missing data when measurement times deviate from those specified in the protocol (Table 4.6). This simulation showed one, possibly extreme, example of the potential bias in the estimation when a naïve approach (LVCF) is used to impute measurements at specified times. Our simulation suggests that it might be advantageous to try to do better than LVCF, and BLUP is one example of an approach that can do better. We note that in our simulation we only considered a scenario where the BLUP response profiles are predicted using the true model. That is an advantage that is not likely to be available in practice. However, even if the true model is not known, in the prediction setting there is no penalty for considering several models and different imputation approaches. The one that results in the best prediction performance as measured by one's performance measures of choice should be selected.

We note that one needs to be careful not to use future information in the modeling, and thus prediction. It is easy to make that mistake by using some of the naïve approaches to imputation. BLUP is an approach that ensures that only past information is used to obtain each point on the predicted marker profile of each individual. Though the BLUP approach is more complex to apply than LVCF, it is likely to offer better prediction performance than LVCF in a variety of settings, while ensuring that the prediction evaluation does not rely on future information. Based on our simulations, those presented in Table 4.6 and others (not shown), we anticipate BLUP outperforming LVCF in situations where nonsystematic noise in the data is high, while the trajectories can be modeled relatively well with a linear mixed effects model.

The number of simulation iterations and perturbations was arrived at by experimentation. The results were very similar with a much larger number of iterations and/or pertur-

bations, suggesting that the simulation results are already stable at 1000 iterations and 500 perturbations.

4.7 Real data example: ESRDS dataset

Study design and analysis End Stage Renal Disease Study (ESRDS) dataset was used to illustrate the methods developed in this chapter. The dataset description was presented in section 3.6.1. To obtain risk predictions we fit a PC_{GLM} model (3.3.2) with a binary outcome of having an event between s and $s + \tau_0$ vs. not, with observed marker values as the main predictor, adjusting for measurement time modeled as a spline with df = 3. Procedures described in section 4.4 were used for estimation and those described in section 4.5 for inference.

Results Some of the results in section 3.6.3, specifically table 3.8, were estimated using the methods presented in this chapter. In table 4.7 we present the estimates for the full set of measures discussed in this chapter, thus these results include PCF(0.2) and PNF(0.8) in addition to those presented earlier.

We observed the best overall prediction performance of estimated glomerular filtration rate (eGFR) in predicting death or end stage renal disease (ESRD) in a $\tau_0 = 1$ year timeframe given up to $s = 2$ years of biomarker information (Table 4.7). For that (s, τ_0) the AUC was 0.84, PE = 0.06, PCF(0.2) = 0.71 and PNF(0.8) = 0.29. All of these indicate better prediction performance than the corresponding estimates obtained for the other three timeframes $(s, \tau_0) = \{(1, 1), (1, 3), (2, 3)\}$. We observed 31 events in $\tau_0 = 1$ year conditional on being at risk at $s = 2$ years. This is the lowest number of events observed in any of the (s, τ_0) timeframes considered (Table 4.7), thus the standard errors of the estimates are highest for $(s, \tau_0) = (2, 1)$ compared to those in other timeframes.

The prediction performance results for eGFR were also very good for $(s, \tau_0) = (1, 1)$ year timeframe, with a small PE (0.08) and a high AUC and PCF(0.2), 0.79 and 0.66,

respectively, showing only a small decrease in prediction performance from the $(s, \tau_0) = (2, 1)$ year timeframe. The PNF(0.8) value was 0.29 at $s = 2$ and deteriorated to 0.51 at $s = 1$, in a $\tau_0 = 1$ timeframe. Otherwise the estimates showed moderate transitions between the different (s, τ_0) timeframes. The prediction performance degraded when we considered the $\tau_0 = 3$ prediction timeframe, with the worst overall predictions observed in $(s, \tau_0) = (3, 2)$ years.

Remarks on the ESRDS analysis and results The effective samples sizes for the different s and τ_0 timeframes were relatively small. Based on our simulations, the estimates can be biased by a few percent for small sample sizes, and the standard errors overestimated. This was not always the case, however, and we found it to depend on the amount of noise in the data. For example, in one of our simulations (results not shown) with a small measurement error, $\sigma_e = 0.1$, and $\mu_{\alpha_1} = -0.1$, $(s, \tau_0) = (4, 1)$ years, with 99 subjects at risk at time s and 21 events observed between s and $s + \tau_0$, we observed 2.8% bias in the estimate of PCF(0.8) and its SE overestimated by 7%. However, other measures were well estimated, as were their standard errors. Thus, we expect the estimates in the ESRDS analysis to be within a few percent of their true values, and their standard errors to be conservative.

4.8 Summary

In this chapter we extended the definitions of several measures of predictive capacity to the longitudinal setting and provided estimation and inference procedures for them. We evaluated their performance using extensive simulation studies and illustrated their use in an analysis of the End Stage Renal Disease Study dataset.

The methods described in this chapter are not only of practical importance, but they are the foundation for the work on estimation and inference for evaluation measures in two-phase studies presented in the next chapter. We attempted to evaluate our methods in simulations that mimic various practical scenarios with varying magnitudes of measurement

error, different marker trajectories and sample sizes. Our estimation and inference methods performed very well in our simulations, providing evidence that they are likely to perform well in practice.

Additionally, we compared the estimation performance of last value carry forward (LVCF) and the best linear unbiased predictor (BLUP) when measurement times deviate from protocol, but the deviations can be assumed to be independent of survival and censoring times, as well as covariates. The BLUP outperformed LVCF substantially in our simulations, and even if the improvement is not as large in practice, we believe that the BLUP is a great approach for using the available information, while ensuring that future information is not used in the estimation, to improve prediction performance.

In the next chapter we build on the work presented in this chapter to develop estimation and inference procedures to under case-cohort, stratified case-cohort and nested case-control study designs.

Table 4.1: Cohort, n = 2000, iterations = 1000, perturbations = 500, σ_e = 0.1 and μ_{α_1} = -0.1. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s + \tau_0$ and n_c is the number of subjects who were censored between s and $s + \tau_0$.

Cohort (n = 2000), iterations = 1000, perturbations = 500, σ_e = 0.1, μ_{α_1} = -0.1						
	s = 24, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 215/591/88					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.155	0.155	0.5	0.007	0.007	95.1
TPF(0.4)	0.464	0.464	0.1	0.045	0.044	94.7
FPF(0.3)	0.248	0.247	0.6	0.027	0.027	94.5
AUC	0.768	0.768	0.0	0.018	0.019	95.7
PCF(0.2)	0.440	0.443	0.7	0.027	0.028	95.5
PNF(0.8)	0.521	0.519	0.4	0.032	0.034	94.5
	s = 48, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 84/274/40					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.150	0.149	0.4	0.011	0.011	95.1
TPF(0.4)	0.334	0.336	0.6	0.067	0.068	94.4
FPF(0.3)	0.206	0.204	0.9	0.037	0.039	94.8
AUC	0.740	0.742	0.3	0.031	0.030	93.2
PCF(0.2)	0.422	0.427	1.3	0.045	0.047	96.4
PNF(0.8)	0.544	0.538	1.0	0.052	0.056	94.1
	s = 24, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 349/398/145					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.183	0.183	0.1	0.007	0.007	94.5
TPF(0.4)	0.768	0.771	0.4	0.030	0.031	94.7
FPF(0.3)	0.494	0.498	0.7	0.039	0.038	94.6
AUC	0.792	0.792	0.1	0.016	0.016	95.0
PCF(0.2)	0.377	0.379	0.5	0.016	0.017	96.0
PNF(0.8)	0.566	0.565	0.1	0.024	0.024	94.3
	s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 138/192/68					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.188	0.189	0.9	0.010	0.010	95.7
TPF(0.4)	0.697	0.695	0.3	0.052	0.054	96.1
FPF(0.3)	0.464	0.471	1.4	0.053	0.055	94.6
AUC	0.767	0.760	0.9	0.026	0.027	95.0
PCF(0.2)	0.379	0.380	0.1	0.029	0.030	94.9
PNF(0.8)	0.572	0.578	1.0	0.038	0.040	94.5

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 4.2: Cohort, n = 2000, iterations = 1000, perturbations = 500, σ_e = 1.0 and μ_{α_1} = -0.1. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s + \tau_0$ and n_c is the number of subjects who were censored between s and $s + \tau_0$.

Cohort (n = 2000), iterations = 1000, perturbations = 500, σ_e = 1.0, μ_{α_1} = -0.1						
	s = 24, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 215/591/88					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.180	0.180	0.1	0.007	0.007	94.9
TPF(0.4)	0.192	0.190	0.8	0.041	0.041	93.1
FPF(0.3)	0.261	0.260	0.7	0.037	0.038	94.5
AUC	0.649	0.648	0.1	0.022	0.022	93.7
PCF(0.2)	0.319	0.319	0.1	0.026	0.027	95.0
PNF(0.8)	0.661	0.661	0.0	0.032	0.032	93.8
	s = 48, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 84/274/40					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.168	0.1	0.011	0.012	95.9
TPF(0.4)	0.102	0.102	0.5	0.043	0.046	93.7
FPF(0.3)	0.179	0.179	0.0	0.046	0.047	92.8
AUC	0.623	0.624	0.1	0.033	0.034	95.6
PCF(0.2)	0.302	0.306	1.3	0.041	0.045	96.4
PNF(0.8)	0.684	0.679	0.8	0.049	0.052	94.6
	s = 24, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 349/398/145					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.227	0.227	0.0	0.005	0.005	94.1
TPF(0.4)	0.698	0.703	0.7	0.047	0.046	93.3
FPF(0.3)	0.738	0.739	0.1	0.046	0.045	92.4
AUC	0.660	0.659	0.1	0.020	0.020	94.6
PCF(0.2)	0.291	0.292	0.5	0.017	0.017	95.3
PNF(0.8)	0.686	0.686	0.0	0.023	0.023	93.6
	s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 138/192/68					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.225	0.226	0.5	0.009	0.009	94.6
TPF(0.4)	0.561	0.564	0.6	0.076	0.077	94.0
FPF(0.3)	0.647	0.652	0.8	0.071	0.070	92.6
AUC	0.635	0.630	0.8	0.031	0.031	95.4
PCF(0.2)	0.284	0.284	0.2	0.029	0.030	95.7
PNF(0.8)	0.699	0.701	0.4	0.035	0.037	94.2

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 4.3: Cohort, n = 2000, iterations = 1000, perturbations = 500, σ_e = 0.1 and μ_{α_1} = -1.0. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s + \tau_0$ and n_c is the number of subjects who were censored between s and $s + \tau_0$.

Cohort (n = 2000), iterations = 1000, perturbations = 500, σ_e = 0.1, μ_{α_1} = -1.0						
	s = 24, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 357/716/113					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.164	0.164	0.3	0.006	0.006	95.2
TPF(0.4)	0.609	0.615	1.0	0.031	0.032	94.5
FPF(0.3)	0.318	0.318	0.0	0.026	0.025	94.2
AUC	0.794	0.794	0.0	0.014	0.014	94.9
PCF(0.2)	0.436	0.439	0.5	0.020	0.020	94.8
PNF(0.8)	0.515	0.514	0.3	0.024	0.025	95.4
	s = 48, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 148/213/36					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.190	0.190	0.2	0.009	0.009	95.2
TPF(0.4)	0.707	0.706	0.2	0.046	0.049	95.8
FPF(0.3)	0.496	0.492	0.7	0.048	0.049	95.0
AUC	0.765	0.763	0.4	0.024	0.025	95.6
PCF(0.2)	0.372	0.376	1.0	0.027	0.028	96.3
PNF(0.8)	0.578	0.578	0.1	0.035	0.038	95.5
	s = 24, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 611/397/178					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.167	0.167	0.2	0.006	0.006	95.8
TPF(0.4)	0.888	0.890	0.3	0.017	0.017	93.8
FPF(0.3)	0.623	0.627	0.6	0.036	0.036	94.6
AUC	0.831	0.828	0.3	0.012	0.013	95.2
PCF(0.2)	0.331	0.332	0.3	0.009	0.009	96.1
PNF(0.8)	0.595	0.595	0.1	0.016	0.016	95.2
	s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 231/112/54					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.170	1.0	0.010	0.010	95.9
TPF(0.4)	0.936	0.932	0.4	0.021	0.022	95.5
FPF(0.3)	0.777	0.770	0.8	0.049	0.052	95.3
AUC	0.813	0.801	1.4	0.024	0.025	94.2
PCF(0.2)	0.297	0.299	0.6	0.013	0.014	96.5
PNF(0.8)	0.640	0.644	0.6	0.024	0.026	95.1

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 4.4: Cohort, n = 2000, iterations = 1000, perturbations = 500, σ_e = 1.0 and μ_{α_1} = -1.0. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s + \tau_0$ and n_c is the number of subjects who were censored between s and $s + \tau_0$.

Cohort (n = 2000), iterations = 1000, perturbations = 500, σ_e = 1.0, μ_{α_1} = -1.0						
	s = 24, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 357/716/113					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.201	0.200	0.4	0.006	0.006	92.8
TPF(0.4)	0.395	0.398	0.9	0.041	0.042	95.1
FPF(0.3)	0.434	0.434	0.0	0.037	0.037	94.5
AUC	0.670	0.670	0.0	0.017	0.017	94.9
PCF(0.2)	0.323	0.324	0.4	0.018	0.019	95.9
PNF(0.8)	0.653	0.652	0.2	0.024	0.025	94.4
	s = 48, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 148/213/36					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.227	0.227	0.1	0.008	0.008	96.1
TPF(0.4)	0.587	0.584	0.5	0.065	0.067	94.4
FPF(0.3)	0.683	0.677	0.9	0.057	0.060	94.8
AUC	0.634	0.632	0.3	0.028	0.029	96.1
PCF(0.2)	0.280	0.284	1.3	0.026	0.028	96.1
PNF(0.8)	0.701	0.700	0.2	0.033	0.035	93.9
	s = 24, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 611/397/178					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.218	0.218	0.2	0.005	0.005	94.2
TPF(0.4)	0.914	0.915	0.2	0.019	0.020	93.0
FPF(0.3)	0.905	0.904	0.0	0.025	0.025	93.3
AUC	0.692	0.689	0.4	0.017	0.017	92.6
PCF(0.2)	0.277	0.278	0.2	0.010	0.010	95.9
PNF(0.8)	0.695	0.697	0.3	0.016	0.015	93.5
	s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 231/112/54					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.214	0.214	0.4	0.009	0.010	95.8
TPF(0.4)	0.965	0.960	0.5	0.019	0.020	94.1
FPF(0.3)	0.968	0.965	0.3	0.023	0.024	90.1
AUC	0.658	0.653	0.8	0.031	0.031	94.4
PCF(0.2)	0.252	0.254	0.8	0.015	0.016	96.4
PNF(0.8)	0.729	0.731	0.3	0.022	0.023	95.4

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 4.5: Cohort, n = 500, 1000, 2000, 4000, iterations = 1000, perturbations = 500, σ_e = 0.1 and μ_{α_1} = -1.0. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s + \tau_0$ and n_c is the number of subjects who were censored between s and $s + \tau_0$.

Cohort, iterations = 1000, perturbations = 500, σ_e = 0.1, μ_{α_1} = -1.0						
	n = 500, s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 58/28/14					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.170	1.3	0.020	0.021	95.3
TPF(0.4)	0.936	0.929	0.7	0.043	0.046	92.9
FPF(0.3)	0.777	0.757	2.5	0.103	0.105	93.6
AUC	0.813	0.782	3.7	0.049	0.055	95.1
PCF(0.2)	0.297	0.305	2.7	0.028	0.033	98.0
PNF(0.8)	0.640	0.646	0.9	0.049	0.054	95.4
	n = 1000, s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 115/56/27					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.170	0.8	0.014	0.015	95.4
TPF(0.4)	0.936	0.930	0.5	0.031	0.032	94.2
FPF(0.3)	0.777	0.763	1.8	0.074	0.074	93.2
AUC	0.813	0.796	2.0	0.034	0.036	95.2
PCF(0.2)	0.297	0.302	1.5	0.019	0.021	96.2
PNF(0.8)	0.640	0.645	0.7	0.035	0.037	94.7
	n = 2000, s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 231/112/54					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.170	1.0	0.010	0.010	95.9
TPF(0.4)	0.936	0.932	0.4	0.021	0.022	95.5
FPF(0.3)	0.777	0.770	0.8	0.049	0.052	95.3
AUC	0.813	0.801	1.4	0.024	0.025	94.2
PCF(0.2)	0.297	0.299	0.6	0.013	0.014	96.5
PNF(0.8)	0.640	0.644	0.6	0.024	0.026	95.1
	n = 4000, s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 461/224/108					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.169	0.4	0.007	0.007	95.5
TPF(0.4)	0.936	0.932	0.4	0.015	0.015	93.5
FPF(0.3)	0.777	0.767	1.3	0.037	0.037	93.8
AUC	0.813	0.807	0.7	0.017	0.017	94.5
PCF(0.2)	0.297	0.298	0.3	0.009	0.009	95.1
PNF(0.8)	0.640	0.642	0.2	0.018	0.018	94.7

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 4.6: Cohort, n = 2000, iterations = 1000, perturbations = 500, $\sigma_e = 0.1$. Comparison of LVCF and BLUP approaches in dealing with measurement times that deviate from protocol. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s + \tau_0$ and n_c is the number of subjects who were censored between s and $s + \tau_0$.

Cohort (n = 2000), iterations = 1000, perturbations = 500, $\sigma_e = 0.1$							
		s = 24, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 463/763/127					
μ_{α_1} = -2		LVCF			BLUP		
	True	Est	Bias %	SE_{emp}	Est	Bias %	SE_{emp}
PE	0.166	0.183	10.0	0.007	0.165	0.3	0.005
TPF(0.4)	0.698	0.501	28.2	0.028	0.694	0.6	0.025
FPF(0.3)	0.366	0.209	42.9	0.020	0.357	2.5	0.023
AUC	0.810	0.792	2.3	0.013	0.811	0.1	0.012
PCF(0.2)	0.429	0.414	3.4	0.016	0.431	0.4	0.015
PNF(0.8)	0.517	0.538	4.1	0.020	0.515	0.3	0.020
		s = 48, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 184/114/24					
μ_{α_1} = -2		LVCF			BLUP		
	True	Est	Bias %	SE_{emp}	Est	Bias %	SE_{emp}
PE	0.180	0.186	3.4	0.010	0.179	0.5	0.010
TPF(0.4)	0.916	0.858	6.4	0.033	0.925	0.9	0.022
FPF(0.3)	0.758	0.662	12.7	0.056	0.771	1.7	0.045
AUC	0.795	0.782	1.7	0.026	0.792	0.4	0.026
PCF(0.2)	0.307	0.308	0.4	0.016	0.311	1.4	0.016
PNF(0.8)	0.636	0.642	1.0	0.028	0.634	0.2	0.028
		s = 24, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 539/769/134					
μ_{α_1} = -3		LVCF			BLUP		
	True	Est	Bias %	SE_{emp}	Est	Bias %	SE_{emp}
PE	0.166	0.202	21.4	0.008	0.166	0.2	0.005
TPF(0.4)	0.744	0.472	36.6	0.026	0.748	0.5	0.022
FPF(0.3)	0.390	0.182	53.4	0.018	0.392	0.4	0.024
AUC	0.821	0.784	4.5	0.013	0.821	0.0	0.012
PCF(0.2)	0.417	0.394	5.6	0.014	0.419	0.5	0.013
PNF(0.8)	0.521	0.563	7.9	0.020	0.520	0.2	0.019
		s = 48, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 171/44/14					
μ_{α_1} = -3		LVCF			BLUP		
	True	Est	Bias %	SE_{emp}	Est	Bias %	SE_{emp}
PE	0.128	0.137	6.8	0.012	0.127	1.1	0.014
TPF(0.4)	0.984	0.948	3.6	0.022	0.983	0.1	0.011
FPF(0.3)	0.911	0.803	11.8	0.071	0.903	0.8	0.050
AUC	0.835	0.803	3.9	0.034	0.821	1.8	0.032
PCF(0.2)	0.254	0.255	0.4	0.010	0.257	1.1	0.010
PNF(0.8)	0.692	0.701	1.4	0.023	0.693	0.2	0.023

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 4.7: Estimates (EST) and standard errors (ESD) of measures of predictive capacity summarizing predictions of a composite outcome (death or ESRD) based on eGFR and a partly conditional logistic model with the logit link function (PC_{GLM}). The estimates were obtained for four sets of s and τ_0. The number of events between s and $s + \tau_0$, and the number of subjects at risk at time s are denoted by n_e and n, respectively.

End Stage Renal Disease Study				
	$\tau_0 = 1$ year		$\tau_0 = 3$ years	
	s = 1 years	s = 2 years	s = 1 years	s = 2 years
	$(n_e/n = 55/574)$ EST (ESD)	$(n_e/n = 31/519)$ EST (ESD)	$(n_e/n = 114/574)$ EST (ESD)	$(n_e/n = 77/519)$ EST (ESD)
PE	0.082 (0.015)	0.062 (0.016)	0.152 (0.033)	0.150 (0.037)
TPF(0.25)	0.745 (0.062)	0.742 (0.080)	0.736 (0.041)	0.695 (0.049)
FPF(0.25)	0.249 (0.013)	0.205 (0.014)	0.366 (0.016)	0.350 (0.017)
AUC	0.791 (0.024)	0.838 (0.029)	0.749 (0.022)	0.703 (0.027)
PCF(0.20)	0.655 (0.057)	0.710 (0.075)	0.482 (0.032)	0.435 (0.042)
PNF(0.80)	0.509 (0.058)	0.291 (0.067)	0.536 (0.044)	0.567 (0.058)

PE = prediction error, TPF(0.25) = true positive fraction at $R(\tau_0 \mid s) = 0.25$, FPF(0.25) = false positive fraction at $R(\tau_0 \mid s) = 0.25$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Chapter 5

ASSESSING PREDICTION PERFORMANCE UNDER LONGITUDINAL TWO-PHASE STUDIES

5.1 Introduction

Before a novel biomarker is introduced into clinical practice, its predictive capacity of future disease status needs to be evaluated. Studies to evaluate longitudinal biomarkers are expensive and, when possible, researches will opt out for more efficient study designs such as case-cohort or nested case-control study designs. If the disease is rare, measuring the marker on everyone in the cohort is not only cost-inefficient, but often infeasible. Case-cohort and nested case-control studies allow investigators to evaluate biomarkers rigorously and at reduced cost, with only a small loss in precision.

In this chapter we develop double inverse probability weighted (DIPW) estimators of measures to evaluate risk predictions under case-cohort (CCH), stratified case-cohort (sCCH) and nested case-control (NCC) study designs (section 5.2), and we develop resampling based variance estimators of those that account for sampling bias and censoring (section 5.3). We evaluate the performance of our estimators using simulations studies (section 5.4) and illustrate them on a nested case-control study within the HALT-C clinical trial (section 5.5).

5.2 Estimation of prediction performance measures under longitudinal two-phase studies

In the estimation of the predictive performance measures we account for the bias introduced by sampling in the second phase of two-phase studies by inverse probability weighting. In the next sections we describe the development of the weights for case-cohort, stratified case-

cohort and nested case-control study designs.

Unless otherwise noted, the notation used in this chapter was introduced in section 2.1.1.

5.2.1 Case-cohort (CCH)

Let ξ_i be a binary indicator of whether ($\xi_i = 1$) or not ($\xi_i = 0$) the i^{th} subject is sampled at phase two and let $\pi_i = \mathrm{P}(\xi_i = 1)$ be the probability of such sampling. The estimation of prediction performance measures in the case-cohort study design setting proceeds as in the cohort setting (equations 4.1 – 4.9), replacing $\widehat{w}_i(\tau_0 \mid s)$ with $\widehat{w}_i^{CCH}(\tau_0 \mid s)$, which includes an additional IPW to account for case-cohort sampling.

The double inverse probability weight (DIPW) in the CCH setting can be estimated as

$$\widehat{w}_i^{CCH}(\tau_0 \mid s) = \widehat{w}_i^{cens^{CCH}}(\tau_0 \mid s) \times w_i^{CCH}$$

where

$$w_i^{CCH} = \frac{\xi_i}{\pi_i^{CCH}}$$

$$\begin{aligned} \pi_i^{CCH} &= \mathrm{P}(\xi_i = 1) \\ &= \delta_i \frac{\sum_{k=1}^n \xi_k \, \delta_k}{\sum_{k=1}^n \delta_k} + (1 - \delta_i) \frac{\sum_{k=1}^n \xi_k \, (1 - \delta_k)}{\sum_{k=1}^n (1 - \delta_k)} \\ &= \frac{\sum_{k=1}^n \xi_k \, \mathrm{I}(\delta_k = \delta_i)}{\sum_{k=1}^n \mathrm{I}(\delta_k = \delta_i)} \end{aligned}$$

and

$$\widehat{w}_i^{cens^{CCH}}(\tau_0 \mid s) = \delta_i \, \mathrm{I}(s < X_i \leq s + \tau_0) \, \frac{\widehat{G}^{CCH}(s)}{\widehat{G}^{CCH}(X_i)} + \mathrm{I}(X_i > s + \tau_0) \, \frac{\widehat{G}^{CCH}(s)}{\widehat{G}^{CCH}(s + \tau_0)}$$

is the estimated IPW accounting for censoring and $\widehat{G}^{CCH}(\cdot)$ is the Kaplan-Meier estimate of $G(\cdot)$, the censoring distribution, estimated using data in the CCH sample with IPW weights,

w_i^{CCH}, accounting for sampling bias.

5.2.2 Stratified case-cohort (sCCH)

If covariate information available for all individuals in the cohort is used to stratify the second phase sampling, the true sampling weights need to be calculated separately for each strata.

Let $g = 1 \dots G$ denote strata of the stratification covariate $\mathbf{U}_{n\times 1}$. Estimation of prediction evaluation measures under a stratified case-cohort study design setting proceeds as in the cohort setting (equations 4.1 - 4.9), replacing $\widehat{w}_i(\tau_0 \mid s)$ with $\widehat{w}_i^{sCCH}(\tau_0 \mid s)$, which includes an additional IPW to account for stratified case-cohort sampling.

This DIPW accounting for censoring and biased sampling in a stratified case-cohort study can be estimated as

$$\widehat{w}_i^{sCCH}(\tau_0 \mid s) = \widehat{w}_i^{cens^{sCCH}}(\tau_0 \mid s) \times w_i^{sCCH}$$

where

$$w_i^{sCCH} = \frac{\xi_i}{\pi_i^{sCCH}}$$

$$\pi_i^{sCCH} = \frac{\sum_{k=1}^{n} \xi_k \, \mathrm{I}(\delta_k = \delta_i)\mathrm{I}(U_k = U_i)}{\sum_{k=1}^{n} \mathrm{I}(\delta_k = \delta_i)\mathrm{I}(U_k = U_i)}$$

and

$$\widehat{w}_i^{cens^{sCCH}}(\tau_0 \mid s) = \delta_i \, \mathrm{I}(s < X_i \leq s + \tau_0) \frac{\widehat{G}^{sCCH}(s)}{\widehat{G}^{sCCH}(X_i)} + \mathrm{I}(X_i > s + \tau_0) \frac{\widehat{G}^{sCCH}(s)}{\widehat{G}^{sCCH}(s + \tau_0)}$$

is the estimated IPW accounting for censoring and $\widehat{G}^{sCCH}(\cdot)$ is the Kaplan-Meier estimate of $G(\cdot)$, the censoring distribution, estimated using data in the sCCH sample with IPW weights, w_i^{sCCH}, accounting for sampling bias.

5.2.3 Nested case-control (NCC)

Samuelsen [Samuelsen, 1997] derives the conditional probability that an individual will ever be chosen as a control in a nested case-control (NCC) study. We incorporate this sampling probability estimator with the estimator of being censored to obtain double inverse probability weights (DIPW) that account for censoring and sampling in estimation of measures of prediction quality. In sampling of the controls we follow a typical nested case-control study design where the cohort is followed prospectively, with respect to membership in the cohort and with respect to occurrence of disease. Thus, the risk sets at the time (X_i) of each event (δ_i) are known. We note that, unlike matched case-control studies in which the matched sets are disjoint, in the NCC study controls may appear in more than one risk set [Thomas, 1977, Goldstein and Langholz, 1992].

For each X_i with δ_i, a set $Q_{i0} = \{q_{i1}, \ldots, q_{im}\}$ of m controls are sampled without replacement from the risk set at time X_i, excluding the case. The sets Q_{i0} are assumed to be independent [Samuelsen, 1997]. We note that a subject who becomes a case at X_i may be sampled as a control for any case j with $X_j < X_i$.

Conditionally on $\mathcal{D}_n$, the probability that individual i is ever sampled as a control in the nested case-control study is given by

$$\pi_{0i}^{NCC} = 1 - \prod_{j:X_j \leq X_i} \left(1 - \frac{m\delta_j}{n_R(X_j)}\right)$$

since the probability of being selected at time X_j is $\frac{m\delta_j}{n_R(X_j)}$ if $X_j < X_i$ and the sampled control sets Q_{j0} are independent [Samuelsen, 1997]. The cases are included with probability 1. Thus, the inverse probability (of sampling) weight for subject i is

$$w_i^{NCC} = \frac{\xi_i}{\pi_i^{NCC}} = \delta_i + (1 - \delta_i)\frac{\xi_{0i}}{\pi_{0i}^{NCC}}$$

where $\pi_i^{NCC} = \delta_i + (1 - \delta_i)\pi_{0i}^{NCC}$ is the probability of subject i being sampled into the NCC.

The probability of sampling could also be estimated. It has been shown that in two-phase stratified studies sampled using Bernoulli sampling, better efficiency is achieved when estimated sampling weights rather than true weights are used in variance estimation [Breslow and Wellner, 2007, Nan et al., 2009, Cai and Zheng, 2011].

One approach which is a good candidate for modeling the sampling probabilities is a generalized additive model (GAM) [Hastie and Tibshirani, 1986]. GAMs are flexible non-parametric models that can be used to identify and characterize nonlinear regression effects [Hastie et al., 2003], such as the relationship between time and selection into the nested case-control study. GAMs take the form:

$$\mathrm{E}(Y \mid C_1, \ldots, C_c) = \alpha + f_1(C_1) + \ldots + f_c(C_c)$$

where Y is the outcome, $C_1, \ldots, C_c$ are the covariates and $f_i(\cdot)$'s are unspecified smooth nonparametric functions, such as locally weighted polynomials, smoothing splines, kernel smoothers, among others [Chambers and Hastie, 1992].

The sampling probability can be modeled using a GAM by fitting the following model to subjects without observed events ($\Delta = 0$):

$$\pi(X, \mathbf{C} \mid \Delta = 0) = g\big(\alpha + f_0(X) + f_1(C_1) + \ldots + f_c(C_c)\big) \tag{5.1}$$

where $g(\cdot)$ is an anti-logit link function, X is the event time, $\mathbf{C} = \{C_1, \ldots, C_c\}$ are the stratifying covariates used in sampling to phase two, and $\pi(X, \mathbf{C} \mid \Delta = 0) = P(\xi = 1 \mid X, \mathbf{C}, \Delta = 0)$ is the probability of being sampled as a control into phase two and $\pi(X, \mathbf{C} | \Delta = 1) = 1$.

Then, the DIPW accounting for censoring and sampling in a nested case-control study

can be estimated as

$$\widehat{w}_i^{NCC}(\tau_0 \mid s) = \widehat{w}_i^{cens^{NCC}}(\tau_0 \mid s) \times \widehat{w}_i^{NCC}$$

where

$$\widehat{w}_i^{NCC} = \frac{\xi_i}{\widehat{\pi}_i^{NCC}}$$

with π_i^{NCC} estimated using a generalized additive model as in equation 5.1 and

$$\widehat{w}_i^{cens^{NCC}}(\tau_0 \mid s) = \delta_i \, \mathrm{I}(s < X_i \leq s + \tau_0) \, \frac{\widehat{G}^{NCC}(s)}{\widehat{G}^{NCC}(X_i)} + \mathrm{I}(X_i > s + \tau_0) \, \frac{\widehat{G}^{NCC}(s)}{\widehat{G}^{NCC}(s + \tau_0)}$$

is the estimated IPW accounting for censoring and $\widehat{G}^{NCC}(\cdot)$ is the Kaplan-Meier estimate of $G(\cdot)$, the censoring distribution, estimated using data in the NCC sample with IPW weights, $\widehat{w}_i^{NCC}$, accounting for sampling bias. We note that the censoring weights could also be estimated from the phase one sample, since all the information used in the estimation of the censoring weights in the NCC study is collected on the full cohort. Then $\widehat{w}_i^{NCC}(\tau_0 \mid s) = \widehat{w}_i(\tau_0 \mid s)$, where $\widehat{w}_i(\tau_0 \mid s)$ is estimated as described in equation 4.3.

5.3 Inference for estimators of prediction performance measures under longitudinal two-phase studies

The bootstrap [Efron, 1979], a widely used resampling-based variance estimation method, is not expected to perform well in two-phase studies due to bias in the sample, as well as induced correlation between sampled individuals in the case of finite sampling [Cai and Zheng, 2011, Cai and Zheng, 2013, Saegusa, 2014]. Recently, Saegusa [Saegusa, 2014] published a document online on bootstrapping in two-phase samples. The author proposed a weighted bootstrap that involves weighting the bootstrap samples by a product of two weights that correspond to randomness from each sampling phase and each stratum. Though it does appear like progress is being made in adapting bootstrap for use in two-phase samples, those approaches are very complex both conceptually and computationally.

Thus, we opted for *perturbation* [Rao and Zhao, 1992, Parzen et al., 1994, Jin et al., 2001], which also involves resampling, but does not require independence between sampled individuals. Hence, it can be used for variance estimation when observations are weakly correlated, such as in two-phase samples under finite sampling. This approach has been shown to produce valid inference in nested case-control studies under finite sampling [Cai and Zheng, 2011, Cai and Zheng, 2013]. Cai *et al.* [Cai and Zheng, 2011] used perturbation to estimate standard errors of estimates of marker evaluation measures based on markers measured at baseline in a nested case-control study. The authors used double inverse probability weighting (DIPW) to account for censoring and sampling.

Our situation differs from this previous work in two ways. First, we evaluate prediction rules based on predicted risk of an event at τ_0 time interval from time s, conditional on being at risk at time s; whereas previous work in two-phase studies evaluated the predictiveness of markers measured at baseline. Second, we are interested in approximating the distribution of $\sqrt{n}\big(\widehat{\mathrm{M}}(\tau_0 \mid s) - \mathrm{M}(\tau_0 \mid s)\big)$, rather than $\sqrt{n}\big(\widehat{\mathrm{M}}(\tau_0) - \mathrm{M}(\tau_0)\big)$ as was the done in [Cai and Zheng, 2011], where M = {TPF, FPF, AUC, PCF, PNF, PE}.

A general inference procedure for performance measures estimated under longitudinal two-phase studies The variance of each of our performance measures under longitudinal CCH, sCCH or NCC can be estimated as follows:

1. Generate $n \times P$ independent and identically distributed random variables V_{ip} from a known distribution with $\mathrm{E}(V_{ip}) = 1$ and $\mathrm{Var}(V_{ip}) = 1$, and $\mathbf{V}_{n\times P} = \{V_{ip},\ i = 1, \ldots, n,\ p = 1, \ldots, P\}$.

2. Use $\mathbf{V}_{n\times P}$ to obtain P perturbed estimates of:

   (a) the sampling weights $\widehat{w}_{ip}^{*CCH}$, $\widehat{w}_{ip}^{*sCCH}$ or $\widehat{w}_{ip}^{*NCC}$, depending on the study design

   (b) the censoring distribution $\widehat{G}_p^*(\cdot)$ weighted by the perturbed sampling weights

(c) the censoring weights, accounting for sampling, $\widehat{w}_{ip}^{*cens}(\tau_0 \mid s)$

(d) the perturbed DIPW accounting for censoring, conditional on s, and sampling: $\widehat{w}_{ip}^{*}(\tau_0 \mid s) = \widehat{w}_{ip}^{*cens}(\tau_0 \mid s) \times \widehat{w}_{ip}^{*s}$, where $s = \{CCH, sCCH, NCC\}$

(e) the risk of an event between s and $s+\tau_0$ conditional on being at risk at s, $\widehat{R}_{ip}^{*}(\tau_0|s)$ and

(f) the evaluation measures $\widehat{\mathrm{M}}_p^{*}(\tau_0 \mid s)$, where M denotes any of {TPF, FPF, AUC, PCF, PNF, PE}

3. The empirical variance of the P estimates $\widehat{\mathrm{M}}_p^{*}(\tau_0 \mid s)$, $p = 1, \ldots, P$, is the estimated variance of $\widehat{\mathrm{M}}(\tau_0|s)$, M = {TPF, FPF, AUC, PCF, PNF, PE} under a given two-phase sampling design.

The details of estimation for each iteration $p = 1, \ldots, P$ in step 2 are described next.

5.3.1 Case-cohort (CCH)

The p^{th} perturbed estimate of the DIPW accounting for censoring and sampling in a longitudinal case-cohort study is:

$$\widehat{w}_{ip}^{*CCH}(\tau_0 \mid s) = \widehat{w}_{ip}^{*cens^{CCH}}(\tau_0 \mid s) \times \widehat{w}_{ip}^{*CCH}$$

where

$$\widehat{w}_{ip}^{*CCH} = \frac{V_{ip}\,\xi_i}{\widehat{\pi}_{ip}^{*CCH}}$$

$$\widehat{\pi}_{ip}^{*CCH} = \frac{\sum_{k=1}^{n} \xi_k\, \mathrm{I}(\delta_k = \delta_i) V_{kp}}{\sum_{k=1}^{n} \mathrm{I}(\delta_k = \delta_i) V_{kp}}$$

and

$$\widehat{w}_{ip}^{*cens^{CCH}}(\tau_0 \mid s) = \delta_i\, \mathrm{I}(s < X_i \leq s + \tau_0)\, \frac{\widehat{G}_p^{*CCH}(s)}{\widehat{G}_p^{*CCH}(X_i)} + \mathrm{I}(X_i > s + \tau_0)\, \frac{\widehat{G}_p^{*CCH}(s)}{\widehat{G}_p^{*CCH}(s + \tau_0)}$$

is the estimated p^{th} perturbed IPW accounting for censoring and $\widehat{G}_p^{*CCH}(\cdot)$ is the p^{th} perturbed Kaplan-Meier estimate of $G_p^*(\cdot)$, the censoring distribution, estimated using data in the CCH sample with p^{th} set of perturbed IPW weights, $\widehat{w}_{ip}^{*CCH}$, accounting for sampling bias.

The variance of $\widehat{\mathrm{M}}(\tau_0 \mid s)$, where M = {TPF, FPF, AUC, PCF, PNF, PE} under a case-cohort study design can be estimated by replacing $\widehat{w}_{ip}^*(\tau_0 \mid s)$ with $\widehat{w}_{ip}^{*CCH}(\tau_0 \mid s)$ in equations 4.10 - 4.18 and step (3) of the algorithm described in section 5.3.

5.3.2 Stratified case-cohort (sCCH)

The p^{th} perturbed estimate of the DIPW accounting for censoring and sampling in a stratified case-cohort study is

$$\widehat{w}_{ip}^{*sCCH}(\tau_0 \mid s) = \widehat{w}_{ip}^{*cens^{sCCH}}(\tau_0 \mid s) \times \widehat{w}_{ip}^{*sCCH}$$

where

$$\widehat{w}_{ip}^{*sCCH} = \frac{V_{ip}\,\xi_i}{\widehat{\pi}_{ip}^{*sCCH}}$$

$$\widehat{\pi}_{ip}^{*sCCH} = \frac{\sum_{k=1}^{n} \xi_k\,\mathrm{I}(\delta_k = \delta_i)\mathrm{I}(U_k = U_i)V_{kp}}{\sum_{k=1}^{n} \mathrm{I}(\delta_k = \delta_i)\mathrm{I}(U_k = U_i)V_{kp}}$$

where $\mathbf{U}_{n\times 1}$ denotes a stratification covariate and

$$\widehat{w}_{ip}^{*cens^{sCCH}}(\tau_0 \mid s) = \delta_i\,\mathrm{I}(s < X_i \leq s + \tau_0)\,\frac{\widehat{G}_p^{*\,sCCH}(s)}{\widehat{G}_p^{*\,sCCH}(X_i)} + \mathrm{I}(X_i > s + \tau_0)\,\frac{\widehat{G}_p^{*\,sCCH}(s)}{\widehat{G}_p^{*\,sCCH}(s + \tau_0)}$$

is the estimated p^{th} perturbed IPW accounting for censoring and $\widehat{G}_p^{*\,sCCH}(\cdot)$ is the p^{th} perturbed Kaplan-Meier estimate of $G_p^*(\cdot)$, the censoring distribution, estimated using data in the sCCH sample with p^{th} set of perturbed IPW weights, $\widehat{w}_{ip}^{*\,sCCH}$, accounting for sampling bias.

The variance of $\widehat{\mathrm{M}}(\tau_0 \mid s)$, where M = {TPF, FPF, AUC, PCF, PNF, PE} under a stratified case-cohort study design can be estimated by replacing $\widehat{w}^*_{ip}(\tau_0|s)$ with $\widehat{w}^{*^{sCCH}}_{ip}(\tau_0|s)$ in equations 4.10 - 4.18 and step (3) of the algorithm described in section 5.3.

5.3.3 Nested case-control

The p^{th} perturbed estimate of the DIPW accounting for censoring and sampling in a nested case-control study is

$$\widehat{w}^{*^{NCC}}_{ip}(\tau_0 \mid s) = \widehat{w}^{*\,cens^{NCC}}_{ip}(\tau_0 \mid s) \times \widehat{w}^{*^{NCC}}_{ip}$$

where

$$\widehat{w}^{*^{NCC}}_{ip} = \frac{V_{ip}\,\xi_i}{\widehat{\pi}^{*^{NCC}}_{ip}} = \delta_i\,V_{ip} + (1-\delta_i)\frac{\xi_{0i}\,V_{ip}}{\widehat{\pi}^{*^{NCC}}_{0ip}}$$

where $\widehat{\pi}^{*^{NCC}}_{0ip}$ denotes the p^{th} perturbed estimate of the probability of being sampled as a control into the NCC subset, $\widehat{\pi}^{*^{NCC}}_{ip}$ is the p^{th} perturbed estimate of the probability of being sampled into the NCC subset and

$$\widehat{w}^{*\,cens^{NCC}}_{ip}(\tau_0 \mid s) = \delta_i\,\mathrm{I}(s < X_i \leq s+\tau_0)\,\frac{\widehat{G}^{*^{NCC}}_p(s)}{\widehat{G}^{*^{NCC}}_p(X_i)} + \mathrm{I}(X_i > s+\tau_0)\,\frac{\widehat{G}^{*^{NCC}}_p(s)}{\widehat{G}^{*^{NCC}}_p(s+\tau_0)}$$

is the estimated p^{th} perturbed IPW accounting for censoring and $\widehat{G}^{*^{NCC}}_p(\cdot)$ is the p^{th} perturbed Kaplan-Meier estimate of $G^*_p(\cdot)$, the censoring distribution, estimated using data in the NCC sample with p^{th} set of perturbed IPW weights, $\widehat{w}^{*^{NCC}}_{ip}$, accounting for sampling bias. We note that the censoring weights could also be estimated from the phase one sample, since all the information used in the estimation of the censoring weights in the NCC study is collected on the full cohort. Then $\widehat{w}^{*^{NCC}}_i(\tau_0 \mid s) = \widehat{w}_i(\tau_0 \mid s)$, where $\widehat{w}_i(\tau_0 \mid s)$ is estimated as described in equation 4.12.

We can estimate $\pi^{*^{NCC}}_{ip}$ using a weighted generalized additive model as in 5.1, but now weighted with the p^{th} set of the perturbation weights. The perturbation weights are incor-

porated into in the final iteration of the local scoring algorithm in the fitting process [Hastie et al., 2003].

5.4 Simulation studies

5.4.1 Simulation setup

The cohort data used in the following simulations was simulated as described in sections 4.6.1 - 4.6.3 under scenario 1, where measurements are taken at 6-month intervals. The true values were obtained as described in section 4.6.2 and were the same as for the simulations in the full cohort. We designed the simulations to be as comparable as possible to those in the full cohort, thus we considered the same set of measurement error magnitudes, $\sigma_e = \{0.1, 1.0\}$, marker slopes, $\mu_{\alpha_1} = \{-0.1, -1.0\}$, and timeframes, $(s, \tau_0) = \{(24, 12), (48, 12), (24, 24), (48, 24)\}$ months. The additional details and sampling schemes in the second phase of case-cohort and nested case-control studies are described in the following sections.

Simulation setup: case-cohort We simulated a large cohort in phase one, with $n = 10000$ for $n^{CCH} = \{250, 500\}$, $n = 15000$ for $n^{CCH} = 1000$ and $n = 20000$ for $n^{CCH} = 2000$, where n^{CCH} denotes the sample size of cases sampled into the case-cohort. We sampled an equal number of cases and controls without replacement (finite sampling) from the full cohort for $n^{CCH} = \{250, 500, 1000, 2000\}$. We refer to subjects who experienced an event at any point in the study as cases, and those who did not as controls. The sample sizes were chosen to correspond to those in the full cohort simulations as much as possible. Risk predictions were obtained using a weighted PC_{GLM} model with a binary outcome (1 if an event between s and $s+\tau_0$ was observed, 0 if still at risk at time $s+\tau_0$), with marker as the main predictor, adjusting for measurement time modeled as a spline with df = 3. The weights used in model fitting were $\widehat{w}_i^{CCH}(\tau_0 \mid s)$ for estimation, and $\widehat{w}_{ip}^{*CCH}(\tau_0 \mid s)$, $p = 1, \ldots, P$, for inference. The

estimates of the prediction evaluaton measures were obtained using methods described in section 5.2.1, and variance estimates as described in sections 5.3 and 5.3.1.

Simulation setup: nested case-control We simulated cohorts of size $n = \{500, 1000, 2000, 4000\}$ as phase 1 samples with administrative censoring of 100 months in order to reduce the number of cases in the full cohort, since we sampled all subjects who experienced an event throughout the study into the second phase sample. We then sampled all the cases into the second phase sample, and for each case we sampled one control from the risk set at the time of the event of the case, stratifying on a dichotomized value of the marker at baseline. The marker at baseline was dichotomized as *high* or *low* at marker = 0.7, corresponding to about 60% of subjects in the *high* group at baseline. Risk predictions were obtained using a weighted PC_{GLM} model with a binary outcome (1 if an event between s and $s+\tau_0$ was observed, 0 if still at risk at time $s+\tau_0$), with marker as the main predictor, adjusting for measurement time modeled as a spline with df = 3. The weights used in model fitting were $\widehat{w}_i^{NCC}(\tau_0 \mid s)$ for estimation, and $\widehat{w}_{ip}^{*NCC}(\tau_0 \mid s)$, $p = 1, \ldots, P$, for inference. The estimates of the prediction evaluaton measures were obtained using methods described in section 5.2.3, and variance estimates as described in sections 5.3 and 5.3.3.

5.4.2 Simulation results

Simulation results: case-cohort Overall, prediction error estimates (PE) performed well, with small bias and the estimated standard errors (SE_{pert}) close to the empirical standard errors (SE_{emp}) and the coverage probability (CP) ranged from 94.9-96.3%. The estimates true positive fraction evaluated at 0.4 risk threshold (TPF(0.4)) performed well, showing little bias with some overestimation of the standard errors in smaller samples. The CP ranged from 93.1-95.5% with two exceptions where CP was 92.0 and 92.2%. In those cases the estimate was close to one, the upper boundary of the range of possible values for TPF. The false positive fraction evaluated at 0.3 risk threshold (FPF(0.3)) performed

well overall, with CP ranging from 92.4-95.6% in all simulations and in most cases it was $> 94\%$, the largest bias observed was 1.5%. The area under the receiver operating characteristic curve (AUC) performed well, with CP ranging from 93.6-96.2% in all simulations. For the proportion of cases captured if 20% of the subjects at highest risk were to be followed (PCF(0.2)) there was up to 1.4% bias in scenarios with small effective sample sizes, and the CP ranged from 95.5-97.4%. For proportion needed to be followed in order to capture 20% of the cases (PNF(0.2)), the performance of the estimators was good overall, with CP 93.8-96.6% in all simulations, and maximum bias of 0.8% (Tables 5.1– 5.4).

Table 5.5 summarizes the performance of measures in the most challenging simulation scenario ($(s, \tau_0) = (48, 24)$, $\sigma_e = 0.1$, $\mu_{\alpha_1} = -1.0$) as sample size increases from $n^{CCH} = 250$ per group to $n^{CCH} = 2000$ per group. We note that those are sample sizes at baseline, and the effective sample sizes at $s = 48$ are substantially smaller. Up to 4% bias seen in the small sample size simulation ($n^{CCH} = 250$), SE_{pert} tended to be overestimated with CP ranging from 92.2-99.1%. As the sample size increased to $n^{CCH} = 2000$, the SE_{pert} converged to within 0.001 of those estimated empirically (SE_{emp}), and the CP ranged from 94.6-96.3% (Table 5.5).

Simulation results: nested case-control Over all simulation scenarios where the phase one sample size at baseline was $n = 2000$, the performance of estimators and variance estimators of all measures performed very well (Tables 5.6 – 5.9). PE showed $< 1\%$ bias with CP ranging from 93.2 to 96.8%. The TPF(0.4) was 0.9% biased in the worst case (Table 5.9) and achieved 92.4-96.1% CP, with the low 92.4% CP likely due to the estimate being close to the boundary, since the estimate was unbiased and the $\text{SE}_{\text{pert}} = \text{SE}_{\text{emp}}$ in that case (Table 5.8). FPF(0.3) was estimated with up to 1.5% bias (Table 5.6) with coverage ranging from 91.7-95.4%, where the low coverage was in a situation where the estimate was close to one, bias was 0.2%, $\text{SE}_{\text{emp}} = 0.022$ and $\text{SE}_{\text{pert}} = 0.024$ (Table 5.9). The AUC was estimated with up to 1.2% bias, with CP ranging from 92.6-95.4%. PCF(0.2) was estimated with up

to 0.7% bias and CP ranging from 94.2-96.9%. Lastly, the PNF(0.8) was estimated with up to 0.7% bias and CP ranging from 92.6-95.7% (Tables 5.6 – 5.9).

In the simulations comparing the bias and variance estimation in the most challenging simulation scenario ($(s, \tau_0) = (48, 24)$, $\sigma_e = 0.1$, $\mu_{\alpha_1} = -1.0$) as sample size increases from n = 500 to 4000 (in phase one at baseline). In the simulation with the smallest sample size, n = 500, the effective sample size at $s = 48$ is 96, including 58 events and 27 non-events still at risk at $s + \tau_0 = 72$, the bias reaches 4.1% (AUC) and CP ranges from 91.2-98.0% for the different measures. When the sample size in phase one at baseline is $n = 4000$, the effective sample size at $s = 48$ is 766, and the estimators converge to within 0.6% of their true values and the CP ranges from 94.1-96.4% (Table 5.10).

5.5 Real data example: HALT-C nested case-control study

5.5.1 Description of the HALT-C dataset

The Hepatitis C Antiviral Long-Term Treatment against Cirrhosis (HALT-C) Trial consisted of 1002 patients with chronic hepatitis C and bridging fibrosis or cirrhosis, who failed to respond or to achieve a sustained virologic response to 20 weeks of combination therapy. Those patients were randomized at 24 weeks to treatment with peginterferon-α-2a or control (no treatment) and were followed every 3 months for 3.5 years after randomization. Blood samples were collected at each visit for subsequent research testing including assays for hepatocellular carcinoma (HCC) biomarkers. To ascertain HCC, ultrasound examinations were performed 6 months after enrollment and every 12 months thereafter. Patients with an elevated or rising α-fetoprotein (AFP), the currently most commonly used marker for detecting HCC, and those with new lesions on ultrasound were evaluated further by CT or MRI. One of the goals of the HALT-C Trial was to identify and validate markers for the surveillance and early diagnosis of HCC. The marker of interest was des-γ-carboxyprothrombin (DCP).

A nested case-control study was used to evaluate the accuracy of DCP in the detection

of HCC. For this study, 39 HCC cases diagnosed between randomization and 3.8 years after randomization were included in the study. For each case, 2 controls without HCC at the time of diagnosis of the case were selected matching on treatment assignment, presence of cirrhosis on baseline biopsy and length of followup. One control was later excluded because of high DCP values due to caumadin (anticoagulant) use, leaving 77 controls. DCP values played no role in diagnosis of HCC.

5.5.2 Analysis of the HALT-C dataset

To estimate the nested case-control (NCC) sampling weights, we fit a generalized additive model (GAM) with a logit link function to the full trial data at baseline (n = 1002) with a binary inclusion in the NCC indicator as the outcome, a smoothing spline function of the event time (df = 4), adjusting for event status and the stratifying selection covariates: cirrhosis (binary) and treatment group assignment (binary). The inverse probability weights of sampling into the NCC were estimated as the inverse of the fitted values from the GAM ($\widehat{w}_i^{NCC}$). The perturbed NCC sampling weights, $\widehat{w}_{ip}^{*NCC}$, were estimated using the above GAM, but weighted with perturbation weights generated from exp(1) distribution, with $P = 500$ perturbation weights generated for each individual.

The censoring weights, $\widehat{w}_i^{cens^{NCC}}(\tau_0 \mid s)$, were then estimated using a weighted Kaplan-Meier estimator of the censoring distribution, accounting for sampling weights in the estimation of censoring weights. For inference, the perturbed NCC sampling weights were used in estimation of the censoring distribution to estimate perturbed censoring weights, $\widehat{w}_{ip}^{*cens^{NCC}}(\tau_0 \mid s)$.

We then fit a weighted $\mathrm{PC}_{\mathrm{GLM}}$ model (section 3.3.2) with a binary outcome (1 if hepatocellular carcinoma was diagnosed between s and $s + \tau_0$, 0 if still at risk at $s + \tau_0$, with individuals censored between s and $s + \tau_0$ excluded from estimation), with the marker as the predictor and measurement time modeled as a spline (df = 3). The weights used in the

fitting were $\widehat{w}_i^{NCC}(\tau_0 \mid s) = \widehat{w}_i^{cens^{NCC}}(\tau_0 \mid s) \times \widehat{w}_i^{NCC}$. We then used procedures described in sections 4.4 and 5.2.3 to estimate the evaluation measures.

For inference, we refit the PC_{GLM} model described above with $\widehat{w}_{ip}^{*NCC}(\tau_0|s) = \widehat{w}_{ip}^{*cens^{NCC}}(\tau_0|s) \times \widehat{w}_{ip}^{*NCC}$ for $p = 1, \ldots, P$, and $P = 500$. For each p, the p^{th} perturbed set of measures was estimated using procedures described in sections 4.5 and 5.3.3. The standard errors of the estimates of the measures were estimated by the square root of the empirical variance of the corresponding $P = 500$ perturbed estimates.

We obtained and evaluated the predictions for conditioning times $s = \{6, 12, 24, 36\}$ and prediction timeframes $\tau_0 = \{6, 12\}$ months.

5.5.3 Results of analysis of the HALT-C nested-case control study dataset

There were 1002 subjects enrolled in the HALT-C clinical trial, with 39 subjects with events and 77 controls selected into the nested case-control (NCC) study. The baseline characteristics of all subjects in the trial and in the NCC study are summarized in Table 5.11. The mean age of the subjects selected into the NCC study was 51.7 years, 22% were female, 60% were white and 57% of the subjects had cirrhosis of the liver. The data in the NCC study is summarized in Figures 5.1 and 5.2. In Figure 5.1 we show the attended visits (circles), event times (filled circles) and censoring times (filled triangles). The subjects are grouped by their risk set with each color corresponding to a given risk set. The inverse probability weights estimated using the generalized additive model are shown to the right of the event or censoring indicators for each individual. In Figure 5.2 we show the marker trajectories for all individuals in the NCC study, stratified by event status (diagnosed with hepatocellular carcinoma during the study vs. not), cirrhosis of the liver and treatment group assignment.

The prediction evaluation results are summarized in Table 5.12. The prediction performance of des-γ-carboxyprothrombin (DCP) in predicting the timeframe of diagnosis with hepatocellular carcinoma (HCC) was good overall. The prediction estimates were especially

notable for $s = 2$ and $\tau_0 = 1$ year prediction timeframe, during which 11 events were observed. The AUC was estimated (standard error) at 0.86 (0.07), with 82% of the events estimated to be captured if 20% of subjects at highest risk were to be followed. That means that we estimate that 9 subjects who would progress to HCC within 1 year would be captured if we followed 21 subjects with highest estimated risks who are still at risk of HCC at 2 years after randomization. For $s = 3$ years and $\tau_0 = 1$ year, we observed 17 events, with the AUC estimated at 0.75 (0.08), PCF(0.2) = 0.65 (0.12) and PNF(0.8) was 0.52 (0.19). Thus, the results deteriorated somewhat by conditioning on $s = 3$ vs. $s = 2$ years, and the prediction error doubled.

The best prediction performance was observed for $s = 2$ years and $\tau_0 = 6$ months. The AUC was estimated at 0.95 (SE = 0.03), however, these estimates are based on 3 events, thus are not likely to be close to the underlying truth. However, these results can be compared with those for the other s and τ_0 timeframes to evaluate trends in the estimates and standard errors of the evaluation measures.

For $s = 1$ year only 1 event was observed in the following 6 months. The results for that s and τ_0 are included in the table for completeness, but are not likely to be meaningful, thus are greyed out.

5.5.4 Remarks on the analysis and results of the HALT-C NCC study dataset

The fitting and prediction was done on the same dataset. Cross-validation was not a good option here due to a small number of events. However, overfitting is not likely to have been an issue in this analysis because the number of data points was much larger than the number of parameters in the models.

The dates of visits were not available, only the dates of scheduled visits. According to the records, the actual visits took place up to 3 weeks before or after the scheduled visit.

Overall, the des-γ-carboxyprothrombin (DCP) biomarker seems to be performing well in

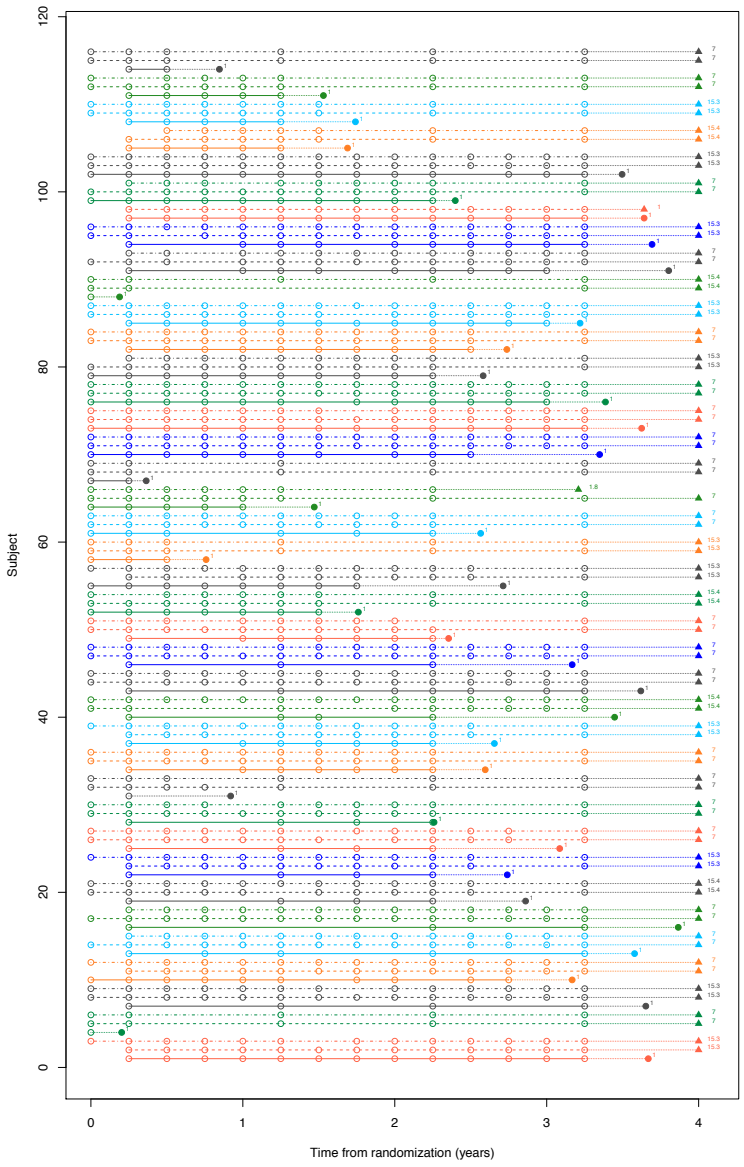

Figure 5.1: An overview of the nested case-control subset of the HALT-C clinical trial. The visits where des-γ-carboxyprothrombin marker was measured are shown as empty circles, the event times are shown as filled circles, and non-events (censoring times) as filled triangles. The subjects are grouped according to their matching in the nested case-control study, with subjects with the same color belonging to the same risk set. The estimated nested case-control inverse probability sampling weights are shown to the right of each event time for the cases, and censoring time for the non-events.

Figure 5.2: The des-γ-carboxyprothrombin (DCP) marker observations in nested case-control subset of the HALT-C clinical trial, stratified by cirrhosis and treatment assignment. The DCP values were truncated at 2000 units and $\log_2$ transformed.

predicting the timeframe of diagnosis with hepatocellular carcinoma (HCC). The numbers of events our estimates were based on were generally small, but these results are promising and warrant further investigation into the evaluation of DCP as a biomarker for predicting hepatocellular carcinoma.

5.6 Summary

In this chapter we presented nonparametric estimation and inference under two-phase study designs for several measures of prediction performance. We evaluated the performance of our estimators using extensive simulation studies and illustrated them on a nested case-control study within the HALT-C clinical trial. Our estimators perform well overall, showing little or no bias and achieving nominal coverage probabilities for reasonable sample sizes. In very small samples we noted some bias and conservative standard errors.

The development of the estimation and inference procedures under two-phase studies built on those developed under the cohort study design. It was our intention to present them in a way that makes that "layering" of estimation procedures clear. Estimators constructed in this way are, we hope, intuitive and elegant. Thus, they lend themselves to further extensions to other, possibly more complex, study designs and applications.

Finally, the methods presented in this chapter can be used in combination with the risk prediction methods presented in chapter 3, thus providing an arsenal of elegant, robust and flexible methods for risk prediction and evaluation of predictions in a wide variety of applications.

Table 5.1: Case-cohort, phase one cohort n = 15000, iterations = 1000, perturbations = 500, CCH (n events/n non-events = 1000/1000), σ_e = 0.1 and μ_{α_1} = -0.1. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s + \tau_0$ and n_c is the number of subjects who were censored between s and $s + \tau_0$.

CCH ($n_e/n_{\bar{e}}$ = 1000/1000), iterations = 1000, perturbations = 500, σ_e = 0.1, μ_{α_1} = -0.1						
	s = 24, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 149/683/128					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.155	0.155	0.4	0.008	0.008	95.2
TPF(0.4)	0.464	0.468	0.8	0.048	0.051	95.5
FPF(0.3)	0.248	0.247	0.5	0.029	0.030	95.3
AUC	0.768	0.770	0.2	0.020	0.021	96.1
PCF(0.2)	0.440	0.445	1.0	0.030	0.032	95.5
PNF(0.8)	0.521	0.519	0.5	0.037	0.040	94.7
	s = 48, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 62/293/63					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.150	0.150	0.5	0.012	0.013	95.1
TPF(0.4)	0.334	0.335	0.1	0.075	0.078	95.0
FPF(0.3)	0.206	0.205	0.7	0.043	0.043	94.3
AUC	0.740	0.742	0.2	0.034	0.034	94.4
PCF(0.2)	0.422	0.426	0.9	0.051	0.054	96.2
PNF(0.8)	0.544	0.543	0.1	0.059	0.065	94.8
	s = 24, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 256/417/232					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.183	0.183	0.2	0.007	0.008	96.0
TPF(0.4)	0.768	0.767	0.0	0.034	0.036	95.2
FPF(0.3)	0.494	0.495	0.2	0.040	0.042	95.2
AUC	0.792	0.791	0.2	0.018	0.018	96.2
PCF(0.2)	0.377	0.378	0.3	0.017	0.019	97.7
PNF(0.8)	0.566	0.567	0.3	0.025	0.027	96.3
	s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 101/209/107					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.188	0.188	0.5	0.011	0.012	96.2
TPF(0.4)	0.697	0.695	0.2	0.060	0.062	94.3
FPF(0.3)	0.464	0.468	0.7	0.056	0.059	95.6
AUC	0.767	0.764	0.4	0.029	0.030	95.8
PCF(0.2)	0.379	0.382	0.8	0.032	0.034	96.6
PNF(0.8)	0.572	0.577	0.8	0.042	0.046	95.0

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 5.2: Case-cohort, phase one cohort n = 15000, iterations = 1000, perturbations = 500, CCH (n events/n non-events ($n_e/n_{\bar{e}}$) = 1000/1000), σ_e = 1.0 and μ_{α_1} = -0.1. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s + \tau_0$ and n_c is the number of subjects who were censored between s and $s + \tau_0$.

CCH ($n_e/n_{\bar{e}}$ = 1000/1000), iterations = 1000, perturbations = 500, σ_e = 1.0, μ_{α_1} = -0.1						
	s = 24, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 158/608/138					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.180	0.180	0.1	0.008	0.008	94.9
TPF(0.4)	0.192	0.191	0.4	0.046	0.047	94.2
FPF(0.3)	0.261	0.259	0.9	0.040	0.042	94.2
AUC	0.649	0.649	0.1	0.023	0.025	96.3
PCF(0.2)	0.319	0.321	0.5	0.028	0.031	95.9
PNF(0.8)	0.661	0.660	0.2	0.035	0.037	94.9
	s = 48, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 62/293/63					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.169	0.9	0.013	0.013	95.0
TPF(0.4)	0.102	0.103	1.0	0.049	0.054	93.1
FPF(0.3)	0.179	0.180	0.6	0.050	0.052	95.0
AUC	0.623	0.623	0.1	0.038	0.039	94.8
PCF(0.2)	0.302	0.304	0.5	0.048	0.051	96.0
PNF(0.8)	0.684	0.682	0.3	0.055	0.060	95.2
	s = 24, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 256/417/232					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.227	0.227	0.3	0.006	0.006	95.1
TPF(0.4)	0.698	0.697	0.2	0.050	0.053	94.8
FPF(0.3)	0.738	0.734	0.4	0.046	0.049	95.6
AUC	0.660	0.659	0.2	0.023	0.022	94.0
PCF(0.2)	0.291	0.291	0.1	0.019	0.020	95.7
PNF(0.8)	0.686	0.686	0.0	0.024	0.025	95.5
	s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 101/209/107					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.225	0.226	0.2	0.009	0.010	95.0
TPF(0.4)	0.561	0.566	1.0	0.085	0.086	94.2
FPF(0.3)	0.647	0.651	0.5	0.073	0.076	94.4
AUC	0.635	0.633	0.3	0.034	0.034	94.7
PCF(0.2)	0.284	0.286	0.9	0.032	0.034	96.1
PNF(0.8)	0.699	0.699	0.0	0.039	0.042	95.7

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 5.3: Case-cohort, phase one cohort n = 15000, iterations = 1000, perturbations = 500, CCH (n events/n non-events ($n_e/n_{\bar{e}}$) = 1000/1000), σ_e = 0.1 and μ_{α_1} = -1.0. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s + \tau_0$ and n_c is the number of subjects who were censored between s and $s + \tau_0$.

CCH ($n_e/n_{\bar{e}}$ = 1000/1000), iterations = 1000, perturbations = 500, σ_e = 0.1, μ_{α_1} = -1.0						
	s = 24, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 265/643/173					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.164	0.164	0.3	0.007	0.007	94.9
TPF(0.4)	0.609	0.615	1.0	0.038	0.037	93.7
FPF(0.3)	0.318	0.319	0.3	0.028	0.028	94.2
AUC	0.794	0.795	0.1	0.017	0.016	93.6
PCF(0.2)	0.436	0.439	0.5	0.022	0.023	95.8
PNF(0.8)	0.515	0.513	0.5	0.027	0.029	93.8
	s = 48, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 111/190/54					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.190	0.190	0.0	0.010	0.011	95.4
TPF(0.4)	0.707	0.705	0.3	0.054	0.056	94.0
FPF(0.3)	0.496	0.492	0.7	0.055	0.055	93.5
AUC	0.765	0.763	0.3	0.028	0.029	94.9
PCF(0.2)	0.372	0.378	1.4	0.031	0.032	96.1
PNF(0.8)	0.578	0.579	0.2	0.041	0.044	94.2
	s = 24, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 453/355/273					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.167	0.168	0.4	0.007	0.007	95.4
TPF(0.4)	0.888	0.889	0.1	0.019	0.020	94.2
FPF(0.3)	0.623	0.626	0.4	0.040	0.040	95.1
AUC	0.831	0.828	0.3	0.015	0.015	95.6
PCF(0.2)	0.331	0.332	0.5	0.010	0.011	95.5
PNF(0.8)	0.595	0.596	0.2	0.018	0.019	94.2
	s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 171/101/83					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.169	0.8	0.011	0.012	96.3
TPF(0.4)	0.936	0.931	0.5	0.024	0.025	94.7
FPF(0.3)	0.777	0.765	1.5	0.055	0.057	95.4
AUC	0.813	0.805	1.0	0.026	0.028	95.7
PCF(0.2)	0.297	0.300	0.9	0.014	0.016	97.4
PNF(0.8)	0.640	0.643	0.4	0.027	0.029	96.6

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 5.4: Case-cohort, phase one cohort n = 15000, iterations = 1000, perturbations = 500, CCH (n events/n non-events ($n_e/n_{\bar{e}}$) = 1000/1000), σ_e = 1 and μ_{α_1} = -1. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s + \tau_0$ and n_c is the number of subjects who were censored between s and $s + \tau_0$.

CCH ($n_e/n_{\bar{e}}$ = 1000/1000), iterations = 1000, perturbations = 500, σ_e = 1.0, μ_{α_1} = -1.0						
	s = 24, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 265/643/173					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.201	0.200	0.5	0.006	0.006	95.1
TPF(0.4)	0.395	0.399	1.1	0.048	0.048	93.8
FPF(0.3)	0.434	0.433	0.2	0.041	0.041	94.6
AUC	0.670	0.671	0.2	0.020	0.020	94.3
PCF(0.2)	0.323	0.326	0.9	0.021	0.022	94.8
PNF(0.8)	0.653	0.650	0.5	0.027	0.028	93.9
	s = 48, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 111/190/54					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.227	0.227	0.0	0.009	0.009	94.9
TPF(0.4)	0.587	0.588	0.2	0.073	0.076	94.9
FPF(0.3)	0.683	0.677	0.8	0.067	0.066	93.1
AUC	0.634	0.634	0.0	0.034	0.034	94.2
PCF(0.2)	0.280	0.284	1.4	0.030	0.032	96.3
PNF(0.8)	0.701	0.701	0.1	0.038	0.041	94.6
	s = 24, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 453/355/273					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.218	0.218	0.3	0.005	0.006	95.2
TPF(0.4)	0.914	0.913	0.0	0.021	0.023	94.8
FPF(0.3)	0.905	0.903	0.2	0.026	0.027	94.2
AUC	0.692	0.688	0.5	0.019	0.019	95.7
PCF(0.2)	0.277	0.278	0.2	0.011	0.012	95.8
PNF(0.8)	0.695	0.698	0.5	0.017	0.017	94.5
	s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 171/101/83					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.214	0.214	0.2	0.011	0.011	95.5
TPF(0.4)	0.965	0.961	0.5	0.021	0.024	93.8
FPF(0.3)	0.968	0.965	0.3	0.022	0.027	92.4
AUC	0.658	0.655	0.5	0.034	0.035	95.8
PCF(0.2)	0.252	0.255	1.3	0.017	0.018	96.6
PNF(0.8)	0.729	0.730	0.1	0.024	0.026	94.7

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 5.5: Case-cohort n events/n non-events ($n_e/n_{\bar{e}}$) are shown in table headings, iterations = 1000, perturbations = 500, σ_e = 0.1 and μ_{α_1} = -1. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s + \tau_0$ and n_c is the number of subjects who were censored between s and $s + \tau_0$.

Case-cohort, iterations = 1000, perturbations = 500, σ_e = 0.1, μ_{α_1} = -1.0						
	n = 250/250, s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 43/25/21					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.171	1.9	0.025	0.025	95.3
TPF(0.4)	0.936	0.929	0.7	0.050	0.054	92.2
FPF(0.3)	0.777	0.764	1.6	0.109	0.114	93.0
AUC	0.813	0.787	3.1	0.056	0.059	94.5
PCF(0.2)	0.297	0.309	4.0	0.030	0.040	99.1
PNF(0.8)	0.640	0.647	1.1	0.055	0.064	96.0
	n = 500/500, s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 85/50/42					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.170	1.1	0.017	0.017	95.8
TPF(0.4)	0.936	0.931	0.5	0.035	0.037	92.0
FPF(0.3)	0.777	0.766	1.4	0.078	0.081	94.6
AUC	0.813	0.799	1.7	0.039	0.040	95.1
PCF(0.2)	0.297	0.303	1.9	0.021	0.025	97.1
PNF(0.8)	0.640	0.645	0.8	0.039	0.042	95.5
	n = 1000/1000, s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 171/101/83					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.169	0.8	0.011	0.012	96.3
TPF(0.4)	0.936	0.931	0.5	0.024	0.025	94.7
FPF(0.3)	0.777	0.765	1.5	0.055	0.057	95.4
AUC	0.813	0.805	1.0	0.026	0.028	95.7
PCF(0.2)	0.297	0.300	0.9	0.014	0.016	97.4
PNF(0.8)	0.640	0.643	0.4	0.027	0.029	96.6
	n = 2000/2000, s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 342/200/166					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.169	0.3	0.008	0.008	96.3
TPF(0.4)	0.936	0.932	0.4	0.018	0.018	94.6
FPF(0.3)	0.777	0.765	1.6	0.039	0.040	95.6
AUC	0.813	0.808	0.5	0.018	0.019	96.1
PCF(0.2)	0.297	0.298	0.3	0.010	0.011	96.3
PNF(0.8)	0.640	0.642	0.2	0.019	0.020	96.2

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 5.6: Nested case-control study simulation results. Total number of subjects at baseline = 2000, simulation iterations = 1000, perturbations = 500, $\sigma_e = 0.1$ and μ_{α_1} = -0.1. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s + \tau_0$ and n_c is the number of subjects who were censored between s and $s + \tau_0$.

NCC sampled from n = 2000, iterations = 1000, perturbations = 500, $\sigma_e = 0.1$, μ_{α_1} = -0.1						
	s = 24, $\tau_0 = 12$, $n_e/n_{\bar{e}}/n_c = 216/509/35$					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.155	0.156	0.3	0.007	0.007	96.4
TPF(0.4)	0.464	0.466	0.6	0.044	0.044	96.1
FPF(0.3)	0.248	0.250	0.6	0.028	0.028	95.4
AUC	0.768	0.767	0.2	0.019	0.019	95.4
PCF(0.2)	0.440	0.440	0.1	0.026	0.028	96.6
PNF(0.8)	0.521	0.520	0.2	0.033	0.034	94.6
	s = 48, $\tau_0 = 12$, $n_e/n_{\bar{e}}/n_c = 85/239/23$					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.150	0.151	0.6	0.011	0.011	93.2
TPF(0.4)	0.334	0.337	0.6	0.066	0.068	95.1
FPF(0.3)	0.206	0.207	0.3	0.039	0.040	94.4
AUC	0.740	0.740	0.1	0.032	0.031	93.8
PCF(0.2)	0.422	0.425	0.8	0.045	0.047	94.8
PNF(0.8)	0.544	0.542	0.4	0.056	0.057	93.4
	s = 24, $\tau_0 = 24$, $n_e/n_{\bar{e}}/n_c = 350/347/64$					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.183	0.183	0.1	0.007	0.007	96.3
TPF(0.4)	0.768	0.771	0.4	0.031	0.031	94.5
FPF(0.3)	0.494	0.499	1.0	0.039	0.040	94.1
AUC	0.792	0.790	0.3	0.016	0.017	95.1
PCF(0.2)	0.377	0.378	0.2	0.016	0.017	95.8
PNF(0.8)	0.566	0.567	0.3	0.024	0.025	95.6
	s = 48, $\tau_0 = 24$, $n_e/n_{\bar{e}}/n_c = 138/166/42$					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.188	0.189	0.7	0.011	0.011	94.3
TPF(0.4)	0.697	0.697	0.0	0.053	0.055	95.2
FPF(0.3)	0.464	0.471	1.5	0.056	0.057	93.7
AUC	0.767	0.760	0.9	0.027	0.027	93.8
PCF(0.2)	0.379	0.379	0.1	0.028	0.030	96.1
PNF(0.8)	0.572	0.576	0.7	0.039	0.041	95.0

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 5.7: Nested case-control study simulation results. Total number of subjects at baseline = 2000, simulation iterations = 1000, perturbations = 500, σ_e = 1.0 and μ_{α_1} = -0.1. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s + \tau_0$ and n_c is the number of subjects who were censored between s and $s + \tau_0$.

NCC sampled from n = 2000, iterations = 1000, perturbations = 500, σ_e = 1.0, μ_{α_1} = -0.1						
	s = 24, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 216/521/40					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.180	0.181	0.5	0.007	0.007	95.6
TPF(0.4)	0.192	0.191	0.4	0.040	0.042	94.4
FPF(0.3)	0.261	0.262	0.3	0.039	0.039	94.3
AUC	0.649	0.647	0.3	0.023	0.022	92.6
PCF(0.2)	0.319	0.318	0.2	0.028	0.027	94.2
PNF(0.8)	0.661	0.661	0.1	0.033	0.033	92.9
	s = 48, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 85/244/26					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.169	1.0	0.011	0.012	94.9
TPF(0.4)	0.102	0.102	0.5	0.043	0.047	94.1
FPF(0.3)	0.179	0.180	0.5	0.047	0.048	93.9
AUC	0.623	0.622	0.1	0.035	0.035	94.5
PCF(0.2)	0.302	0.304	0.6	0.043	0.045	95.4
PNF(0.8)	0.684	0.680	0.7	0.049	0.053	92.6
	s = 24, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 350/355/73					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.227	0.227	0.2	0.005	0.005	95.4
TPF(0.4)	0.698	0.703	0.7	0.047	0.046	93.5
FPF(0.3)	0.738	0.741	0.5	0.046	0.046	93.5
AUC	0.660	0.657	0.4	0.021	0.020	93.6
PCF(0.2)	0.291	0.292	0.4	0.017	0.018	95.5
PNF(0.8)	0.686	0.688	0.2	0.023	0.023	94.4
	s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 138/170/46					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.225	0.226	0.4	0.009	0.009	96.8
TPF(0.4)	0.561	0.565	0.7	0.075	0.077	94.6
FPF(0.3)	0.647	0.652	0.7	0.072	0.071	92.7
AUC	0.635	0.630	0.9	0.031	0.032	95.0
PCF(0.2)	0.284	0.283	0.0	0.029	0.031	95.8
PNF(0.8)	0.699	0.698	0.1	0.036	0.038	95.7

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 5.8: Nested case-control study simulation results. Total number of subjects at baseline = 2000, simulation iterations = 1000, perturbations = 500, $\sigma_e = 0.1$ and μ_{α_1} = -1.0. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s + \tau_0$ and n_c is the number of subjects who were censored between s and $s + \tau_0$.

NCC sampled from n = 2000, iterations = 1000, perturbations = 500, $\sigma_e = 0.1$, μ_{α_1} = -1.0						
	s = 24, $\tau_0 = 12$, $n_e/n_{\bar{e}}/n_c = 357/674/40$					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.164	0.164	0.1	0.006	0.006	95.9
TPF(0.4)	0.609	0.609	0.1	0.033	0.032	94.0
FPF(0.3)	0.318	0.316	0.6	0.026	0.026	94.8
AUC	0.794	0.794	0.1	0.014	0.015	94.3
PCF(0.2)	0.436	0.438	0.4	0.019	0.020	95.8
PNF(0.8)	0.515	0.515	0.0	0.025	0.026	94.4
	s = 48, $\tau_0 = 12$, $n_e/n_{\bar{e}}/n_c = 150/207/26$					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.190	0.190	0.1	0.009	0.009	95.6
TPF(0.4)	0.707	0.708	0.2	0.048	0.049	94.4
FPF(0.3)	0.496	0.496	0.1	0.048	0.049	95.2
AUC	0.765	0.762	0.4	0.025	0.025	94.8
PCF(0.2)	0.372	0.374	0.6	0.027	0.028	95.3
PNF(0.8)	0.578	0.578	0.0	0.035	0.038	95.3
	s = 24, $\tau_0 = 24$, $n_e/n_{\bar{e}}/n_c = 612/383/78$					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.167	0.167	0.2	0.006	0.006	95.3
TPF(0.4)	0.888	0.889	0.1	0.017	0.018	93.7
FPF(0.3)	0.623	0.628	0.7	0.036	0.037	94.4
AUC	0.831	0.828	0.3	0.013	0.013	94.3
PCF(0.2)	0.331	0.332	0.3	0.009	0.009	95.3
PNF(0.8)	0.595	0.596	0.2	0.016	0.016	94.5
	s = 48, $\tau_0 = 24$, $n_e/n_{\bar{e}}/n_c = 231/109/42$					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.169	0.5	0.010	0.010	94.2
TPF(0.4)	0.936	0.935	0.0	0.021	0.021	92.4
FPF(0.3)	0.777	0.777	0.1	0.051	0.052	94.8
AUC	0.813	0.803	1.2	0.025	0.025	94.5
PCF(0.2)	0.297	0.298	0.3	0.013	0.014	95.8
PNF(0.8)	0.640	0.643	0.5	0.024	0.026	94.8

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 5.9: Nested case-control study simulation results. Total number of subjects at baseline = 2000, simulation iterations = 1000, perturbations = 500, σ_e = 1.0 and μ_{α_1} = -1.0. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s + \tau_0$ and n_c is the number of subjects who were censored between s and $s + \tau_0$.

NCC sampled from n = 2000, iterations = 1000, perturbations = 500, σ_e = 1.0, μ_{α_1} = -1.0						
	s = 24, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 358/673/40					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.201	0.201	0.3	0.005	0.006	95.8
TPF(0.4)	0.395	0.398	0.9	0.041	0.041	94.5
FPF(0.3)	0.434	0.434	0.1	0.036	0.037	95.4
AUC	0.670	0.670	0.0	0.018	0.018	94.1
PCF(0.2)	0.323	0.325	0.7	0.019	0.020	95.5
PNF(0.8)	0.653	0.653	0.0	0.024	0.025	94.3
	s = 48, τ_0 = 12, $n_e/n_{\bar{e}}/n_c$ = 150/207/26					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.227	0.228	0.3	0.008	0.008	94.6
TPF(0.4)	0.587	0.587	0.0	0.065	0.067	93.8
FPF(0.3)	0.683	0.681	0.3	0.059	0.060	93.6
AUC	0.634	0.631	0.5	0.031	0.030	93.2
PCF(0.2)	0.280	0.281	0.4	0.027	0.028	96.0
PNF(0.8)	0.701	0.702	0.1	0.034	0.035	94.3
	s = 24, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 611/383/78					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.218	0.218	0.2	0.005	0.005	95.4
TPF(0.4)	0.914	0.914	0.1	0.019	0.020	94.0
FPF(0.3)	0.905	0.905	0.1	0.025	0.025	92.2
AUC	0.692	0.689	0.4	0.017	0.017	94.6
PCF(0.2)	0.277	0.278	0.3	0.010	0.010	95.5
PNF(0.8)	0.695	0.697	0.4	0.015	0.016	93.9
	s = 48, τ_0 = 24, $n_e/n_{\bar{e}}/n_c$ = 232/110/42					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.214	0.214	0.1	0.009	0.010	95.4
TPF(0.4)	0.965	0.961	0.4	0.019	0.021	94.7
FPF(0.3)	0.968	0.966	0.2	0.022	0.024	91.7
AUC	0.658	0.653	0.8	0.031	0.032	95.1
PCF(0.2)	0.252	0.254	0.7	0.014	0.016	96.9
PNF(0.8)	0.729	0.732	0.3	0.021	0.023	95.4

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 5.10: Nested case-control study simulation results showing convergence with increasing sample size, iterations = 1000, perturbations = 500, $\sigma_e = 0.1$ and μ_{α_1} = -1. Effective sample sizes used to estimate the performance measures are summarized in the table as $n_e/n_{\bar{e}}/n_c$, where n_e denotes the number of events observed in a τ_0 timeframe from s, $n_{\bar{e}}$ denotes the number of subjects still at risk at time $s+\tau_0$ and n_c is the number of subjects who were censored between s and $s+\tau_0$.

Nested case-control, iterations = 1000, perturbations = 500, $\sigma_e = 0.1$, μ_{α_1} = -1.0						
	Sampled from n = 500, s = 48, $\tau_0 = 24$, $n_e/n_{\bar{e}}/n_c = 58/27/11$					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.171	1.6	0.021	0.021	95.5
TPF(0.4)	0.936	0.929	0.7	0.044	0.046	91.2
FPF(0.3)	0.777	0.767	1.2	0.107	0.105	91.2
AUC	0.813	0.780	4.1	0.050	0.055	95.2
PCF(0.2)	0.297	0.306	2.8	0.026	0.032	98.0
PNF(0.8)	0.640	0.650	1.6	0.051	0.055	94.1
	Sampled from n = 1000, s = 48, $\tau_0 = 24$, $n_e/n_{\bar{e}}/n_c = 116/55/21$					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.169	0.6	0.015	0.015	94.3
TPF(0.4)	0.936	0.931	0.4	0.031	0.031	92.1
FPF(0.3)	0.777	0.768	1.0	0.075	0.075	93.3
AUC	0.813	0.796	2.0	0.034	0.037	95.5
PCF(0.2)	0.297	0.301	1.3	0.019	0.021	97.0
PNF(0.8)	0.640	0.645	0.7	0.034	0.037	96.6
	Sampled from n = 2000, s = 48, $\tau_0 = 24$, $n_e/n_{\bar{e}}/n_c = 231/109/42$					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.169	0.5	0.010	0.010	94.2
TPF(0.4)	0.936	0.935	0.0	0.021	0.021	92.4
FPF(0.3)	0.777	0.777	0.1	0.051	0.052	94.8
AUC	0.813	0.803	1.2	0.025	0.025	94.5
PCF(0.2)	0.297	0.298	0.3	0.013	0.014	95.8
PNF(0.8)	0.640	0.643	0.5	0.024	0.026	94.8
	Sampled from n = 4000, s = 48, $\tau_0 = 24$, $n_e/n_{\bar{e}}/n_c = 462/219/85$					
	True	Est	Bias %	SE_{emp}	SE_{pert}	CP %
PE	0.168	0.169	0.6	0.007	0.007	96.4
TPF(0.4)	0.936	0.935	0.1	0.015	0.015	94.1
FPF(0.3)	0.777	0.777	0.0	0.037	0.037	94.2
AUC	0.813	0.806	0.8	0.017	0.017	95.7
PCF(0.2)	0.297	0.298	0.2	0.009	0.009	94.8
PNF(0.8)	0.640	0.642	0.3	0.017	0.018	94.6

PE = prediction error, TPF(0.4) = true positive fraction at $R(\tau_0 \mid s) = 0.4$, FPF(0.3) = false positive fraction at $R(\tau_0 \mid s) = 0.3$, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Table 5.11: Baseline characteristics of patients randomized into the HALT-C clinical trial, and those of selected into the nested case-control study (NCC) at 3.8 years after randomization.

Baseline characteristics of patients in the HALT-C clinical trial				
	HALT-C (n = 1002)		HALT-C NCC (n = 116)	
Age at randomization (mean, sd)	50.2	7.2	51.7	7.5
Female (n, %)	289	28.8	26	22.4
Race/ethnicity (n, %)				
— White	717	71.6	70	60.3
— Black	183	18.3	36	31.0
— Hispanic	79	7.9	7	6.0
— Other	23	2.3	3	2.6
Drinks (per week) (mean, sd)	9.6	16.4	9.5	14.5
DCP (log2) (mean, sd)	4.9	0.9	4.9	1.0
Cirrhosis present (n, %)	408	40.7	66	56.9
Randomized to treatment (n, %)	507	50.6	59	50.9
HCC diagnosed (n, %)	39	3.9	39	33.6

Table 5.12: Estimates (EST) and standard errors (ESD) of measures of predictive capacity summarizing predictions of hepatocellular carcinoma based on des-γ-carboxyprothrombin biomarker and a partly conditional logistic model with the logit link function (PC_{GLM}). The estimates were obtained $s = \{6, 12, 24, 36\}$ and $\tau_0 = \{6, 12\}$ months. The number of events between s and $s + \tau_0$, and the number of subjects with no-events before $s + \tau_0$, are denoted by n_e and $n_{\bar{e}}$, respectively. The standard errors were estimated with 500 perturbations.

HALT-C NCC study prediction evaluation results				
	$\tau_0 = 6$ months			
	s = 6 months	s = 1 year	s = 2 years	s = 3 years
	($n_e/n_{\bar{e}} = 3/110$) EST (ESD)	($n_e/n_{\bar{e}} = 1/109$) EST (ESD)	($n_e/n_{\bar{e}} = 3/102$) EST (ESD)	($n_e/n_{\bar{e}} = 8/85$) EST (ESD)
PE(x10)	0.032 (0.018)	0.011 (0.010)	0.028 (0.015)	0.101 (0.032)
AUC	0.757 (0.147)	0.996 (0.002)	0.952 (0.030)	0.806 (0.087)
PCF(0.2)	0.667 (0.237)	1.000 (0.000)	1.000 (0.000)	0.748 (0.146)
PNF(0.8)	0.637 (0.294)	0.004 (0.002)	0.096 (0.048)	0.523 (0.240)
	$\tau_0 = 1$ year			
	s = 6 months	s = 1 year	s = 2 years	s = 3 years
	($n_e/n_{\bar{e}} = 4/109$) EST (ESD)	($n_e/n_{\bar{e}} = 5/105$) EST (ESD)	($n_e/n_{\bar{e}} = 11/94$) EST (ESD)	($n_e/n_{\bar{e}} = 17/76$) EST (ESD)
PE(x10)	0.042 (0.021)	0.055 (0.024)	0.107 (0.035)	0.214 (0.056)
AUC	0.709 (0.125)	0.782 (0.121)	0.858 (0.067)	0.748 (0.077)
PCF(0.2)	0.500 (0.223)	0.600 (0.207)	0.818 (0.112)	0.646 (0.120)
PNF(0.8)	0.637 (0.197)	0.747 (0.285)	0.147 (0.210)	0.523 (0.192)

PE(x10) = prediction error x 10, AUC = area under the ROC curve, PCF(0.2) = the proportion of events in τ_0 timeframe from s captured if 20% of the population at risk at time s were followed, and PNF(0.8) = the proportion of the population at risk at time s that would need to be followed in order to capture 80% of the events in the timeframe τ_0 from s.

Chapter 6

CONCLUSIONS AND FUTURE RESEARCH

Summary and conclusions The work described in this dissertation was motivated by a problem of evaluation of the utility of an anti-idiotypic antibody, suspected to play a role in development of type 1 diabetes, in predicting the onset of type 1 diabetes based on longitudinal measurements of that anti-idiotypic antibody in a case-control study. At the time, methods to address that problem were not available. However, with the tools developed in this dissertation, we are now able to evaluate that marker in terms of its ability to discriminate between those who will and those who will not progress to type 1 diabetes in a given time frame.

We developed estimation techniques for flexible and robust models for risk prediction in a longitudinal setting that are rooted in the partly conditional modeling framework [Zheng and Heagerty, 2005]. The models are robust, requiring no modeling assumptions between event time history and the full marker process, and only model a function of a part of the marker trajectory up to s. They are flexible, because they accommodate multiple covariates in a rich way and are easy to modify and adapt to a given problem.

In the first model, the partly conditional Cox-type model (PC_{Cox}), we modeled the measurement time non-linearly using splines, and the marker effect using a time-dependent coefficient. This model is very flexible, allowing for adjustment for baseline covariates as well as incorporation of additional longitudinal covariates or biomarkers. In the partly conditional generalized linear model (PC_{GLM}), all parameters are effectively time-dependent, since the model is fit to each specific prediction timeframe separately. Again, the measurement time can be modeled in a non-linear fashion using, for example, a spline and covariates

can include baseline covariates as well as longitudinal ones. Thus, the models are robust, making no assumptions as to the marker trajectory and modeling the marker effect with a time-varying coefficient, and flexible in their adaptability to a problem at hand. They can be used to model observed marker data, or some function of the marker trajectory up to a given time point.

In their simplest form, both of these models relate the observed marker information at a given time to the outcome. However, some function of all the available information up to a given time point can be incorporated into the model. This could be as simple as the average of the last two measurements, or could be much more involved. Modeling a function of marker information up to the time s, rather than the observed marker values, is especially desirable when the signal to noise ratio in marker measurements is low. In other words, when marker measurements include non-systematic noise, or measurement error, we might want to mitigate its impact by increasing signal to noise ration and thus, improve the quality of risk predictions.

We suggest one approach to address this issue, the BLUP. In this approach, an LME model is fit to the marker trajectories of the full cohort in order to learn the trends in the relationship between the marker and time. Then, given the estimates from the LME model and the marker data for a new subject, we iterate through their observations conditioning on the time of a given measurement s and obtain a BLUP marker value at that time, or any time point u, $u \leq s$. This method "shrinks" the new individual's marker trajectory towards the mean trajectory and as estimated from the full cohort. The shrinking is desirable, because it reduces the variability in the new individual's trajectory over time, but it does that in a "supervised" fashion. In other words, if the marker data in the full cohort at some time u is highly variable, that variability will be carried through to the BLUP marker values for the new individual, and little information from the cohort will be incorporated into the BLUP marker value. If, on the other hand, there is very little variability in the marker values in the cohort, the BLUP marker values for the new individual will be "shrunk" more towards

the mean estimated from the cohort. We emphasize that the BLUP marker values are only based on the past data of the new individual. That is not always the case in some, more naïve, approaches aimed at reducing the noise, or nonsystematic measurement error, levels in the data. The BLUP marker values can then be used in place of observed marker values in the PC_{Cox} and PC_{GLM} models, a method we refer to as *two-stage modeling*.

In our evaluation of the models via simulations, we compared the risk predicted by PC_{Cox} and PC_{GLM} with the true risk estimated empirically, and we compared the PC_{Cox} BLUP and PC_{GLM} BLUP with predictions and their standard errors with those obtained from the current gold standard approach to risk prediction in a longitudinal setting, the joint model (JM). The reason for that latter was that in the case of PC_{Cox} BLUP or PC_{GLM} BLUP, it is very challenging to obtain an empirical truth, as the predictions depend on the whole marker trajectory up to the prediction time, as well as marker trajectories in the cohort. We also evaluated the models in terms of their ability to discriminate between subjects who will or will not experience an event in a given timeframe using several measures of discrimination and calibration.

The PC models performed well in terms of the accuracy of risk predictions, as compared to the empirically obtained true risks (PC_{Cox}, PC_{GLM}) and compared to the JM (PC_{Cox} BLUP, PC_{GLM} BLUP). The PC_{Cox} BLUP and PC_{GLM} BLUP models, just like the JM, provide an advantage over PC_{Cox} and PC_{GLM} models when the marker measurements are highly variable, with that variability believed to be nonsystematic, and the marker trajectories can be modeled well. If the marker trajectories are complex and difficult to model with an LME model, the two-stage models, as the JM, may perform poorly. In that situation, the PC_{Cox} and PC_{GLM} models should perform better. That is, in fact, what we observed in the analysis of the End Stage Renal Disease Study (ESRDS) dataset.

The PC models have several advantages over the JM. They are simpler to use in practice, easier to modify and fit. Incorporating multiple longitudinal markers into the PC framework (PC_{Cox} and PC_{GLM}) is straightforward, whereas it is a practical impossibility for the JM, as

well as the PC_{Cox} BLUP and PC_{GLM} BLUP in their current form. In our simulations, the performance of the PC_{Cox} BLUP and PC_{GLM} BLUP was comparable to that of the JM in terms of discrimination and calibration. This was a promising finding, as the two-stage PC models are much easier to work with in practice than the JM. Using the `JM` package in `R` [Rizopoulos, 2010], we found the JM difficult to fit, often failing to converge, especially for LME models with more than one random effect.

Thus, the PC models add to the arsenal of tools available to us for risk prediction in a longitudinal setting. They offer significant advantages in situations where few assumptions about the longitudinal marker trajectory can be reasonably made (PC_{Cox}, PC_{GLM}), as well as in situations with considerable non-systematic error in the marker measurements when the marker trajectories can be modeled well, possibly with a complex LME model (PC_{Cox} BLUP, PC_{GLM} BLUP).

Next, we turned our attention to evaluating risk predictions and predictive tests based on risk. Our goal in the second part of the dissertation was to develop nonparametric methods for evaluation of risk predictions in a longitudinal setting under a cohort study design. We selected measures of discrimination that were well established in the diagnostic testing (true and false positive fractions (TPF, FPF)) [Pepe, 2003], as well as measures that are based on TPF and FPF but have a more clinically relevant interpretation, which we view to have potential for use in feasibility studies of preventive interventions (PCF, PNF) [Pfeiffer and Gail, 2011]. Until now, in the prediction setting, these measures were used mainly for evaluation of predictive markers measured at baseline, although some have been extended to the longitudinal setting (TPF, FPF) [Zheng and Heagerty, 2007]. There, the estimators were semiparametric, and used the marker value at the conditioning time to define a test to be evaluated. Our estimators are nonparametric, and we used risk estimated from data up to the conditioning time, s, to define a test. We decided on nonparametric estimation methods to maximize robustness of our estimators by limiting their reliance on modeling assumptions.

We extended the definitions of selected measures to the longitudinal setting. We accounted for censoring using inverse probability weighting (IPW) with the weights estimated nonparametrically. In this setting, all the estimators were conditional on s, meaning the data available up to time s and survival up to time s. We also provided inference procedures for our estimators using a resampling-based approach referred to as perturbation [Parzen et al., 1994, Jin et al., 2001]. It is an elegant approach to variance estimation, and though bootstrap would have been appropriate to use for the estimation in the full cohort as well, we chose perturbation as it can be more easily, and elegantly, applied to variance estimation in two-phase samples.

One of the practical challenges in risk prediction in the longitudinal setting is that longitudinal observations are not available at all times, and often the measurement time deviates substantially from the time specified in the study protocol. One way to address that problem is to carry the last available measurement forward to the time of interest and impute it at that time. This can lead to substantial bias [Tsiatis and Davidian, 2004], especially if the slope of the marker trajectories is steep. If the marker trajectories could be modeled reasonably well with an LME model, then using the BLUP approach to impute the marker measurements at specific times using past information can reduce bias in the estimation of risks and thus, the evaluation measures.

We evaluated our estimators under a cohort study design with censoring by comparing them to the true, estimated empirically from an uncensored large cohort, values of the measures. Our estimators, as well as their variance estimators, performed well achieving nominal coverage probabilities even in settings with effective samples sizes as low as $n = 100$. We also noted a reduction in bias in estimation of evaluation measures when BLUP is used to impute missing data, compared to the LVCF approach. However, the bias in the LVCF is mainly an issue when marker trajectories have steep slopes.

In the third part of the dissertation we extended the estimation procedures discussed in the second part to the two-phase sampling designs. Specifically, we considered case-

cohort, stratified case-cohort and nested case-control study designs, but we note that our estimation procedures could also be used as a guide and be adopted to other, possibly more complex, sampling designs. In this setting we also opted for robust, nonparametric estimation procedures. In addition to accounting for censoring, we account for bias due to sampling by an additional inverse probability weight, which is estimated differently for each study design. Variance estimation using resampling-based approaches poses a challenge in two-phase studies sampled using finite sampling in the second phase, due to correlation induced by sampling between the sampled individuals. Thus, standard approaches such as bootstrap are not appropriate. Perturbation, however, can be used to account for that induced correlation.

We evaluated our estimators and variance estimation procedures with extensive simulation studies. We show that the double inverse probability weighting (DIPW) accounts for censoring and sampling bias in CCH and NCC study designs, with our estimators showing no evidence of meaningful bias and falling within a desirable range of the nominal coverage probability of 95%. We illustrated our methods on ESRDS (cohort) and HALT-C (NCC) studies.

The missing data at specific measurement times is also often an issue in two-phase studies, but it cannot be addressed using the BLUP approach. To the best of our knowledge, methods to estimate parameters on LME models in two-phase samples are not available at this time.

In summary, we developed estimation procedures for flexible and robust partly conditional models that can be used for risk prediction in a longitudinal setting with time-to-event outcomes. We evaluated our models and compared their performance to the current gold standard, the joint model, and offered practical guidance as to the situations in which the different approaches are likely to be most appropriate.

Predictive models and predictive biomarkers have little utility until risk predictions based on them are evaluated and validated. The same evaluation methods can serve to evaluate risk prediction models, as well as the longitudinal predictive markers. In development of our

evaluation methods we considered measures of calibration ($\text{PE}(\tau_0 \mid s)$) and discrimination ($\text{TPF}(\tau_0 \mid s, \cdot)$, $\text{FPF}(\tau_0 \mid s, \cdot)$, $\text{AUC}(\tau_0 \mid s)$, $\text{PCF}(\tau_0 \mid s, \cdot)$, $\text{PNF}(\tau_0 \mid s, \cdot)$). We provided non-parametric estimators of these measures as well as perturbation-based variance estimators, using IPW to account for censoring. We evaluated the performance of our estimators using simulation studies and illustrated them in the analysis of the ESRDS dataset.

It was important to us that our methods be applicable in a wide range of study designs. Thus, we considered case-cohort, stratified case-cohort and nested case-control study designs, which are all commonly used in clinical studies to evaluate biomarkers. We used IPW to account for sampling, in addition to the IPW accounting for censoring, resulting in DIPW estimators. We also provided valid inference procedures for our estimators under the three study designs. We evaluated our methods in simulation studies and illustrated them on a nested case-control study within a HALT-C clinical trial dataset. Our inference procedures are elegant and intuitive and it was our intention to present them in a way that facilitates their extension to other, possibly more complex, study designs.

Future work The work described in this dissertation generated many ideas, some of which we plan to pursue in the near future.

Two-phase sampling can be thought of as a missing data problem. Some study designs, such as case-cohort and nested case-control have been studied extensively and are well understood. In those sampling designs, the units being sampled are individuals, since in most studies each individual would provide one observation. This is not the case in longitudinal studies, as one individual can provide several observations. We are interested in exploring sampling designs where the sampling units are observations, or sets of observations, rather than individuals. Could such sampling offer cost reduction without compromising efficiency?

We plan to explore the adaptability of the perturbation variance estimation approach to these different sampling designs. We expect it to be a nice, relatively easily adaptable tool to use in a wide variety of settings.

In our evaluation we demonstrated that the BLUP is an appropriate, well performing and practical approach to imputing missing data at specified time points in situations where measurement times deviate from the protocol. It uses an LME model to model the relationship between time and marker. Thus, it is currently only available under the cohort study design. Adapting it to two-phase sample designs would be of great interest and of practical consequence. There are two ways in which that could be achieved: one would be to develop a weighted LME, where sampling weights could be incorporated into the fitting of the model. The other would be to explore other modeling options that could be incorporated into the BLUP imputation method instead of the LME, but for which the fitting with sampling weights is already available.

BIBLIOGRAPHY

[Andersen and Gill, 1982] Andersen, P. K. and Gill, R. D. (1982). Cox regression model for counting processes - a large sample study. *Annals of Statistics*, 10(4):1100–1120.

[Anderson et al., 1983] Anderson, J. R., Cain, K. C., and Gelber, R. D. (1983). Analysis of survival by tumor response. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 1(11):710–9.

[Barlow, 1994] Barlow, W. E. (1994). Robust variance estimation for the case-cohort design. *Biometrics*, 50(4):1064–1072.

[Barlow et al., 1999] Barlow, W. E., Ichikawa, L., Rosner, D., and Izumi, S. (1999). Analysis of case-cohort designs. *Journal of Clinical Epidemiology*, 52(12):1165–1172.

[Borgan et al., 2000] Borgan, O., Langholz, B., Samuelsen, S. O., Goldstein, L., and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Analysis*, 6(1):39–58.

[Breslow et al., 2009] Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., and Kulich, M. (2009). Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology*, 169(11):1398–1405.

[Breslow and Wellner, 2007] Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scandinavian Journal of Statistics*, 34(1):86–102.

[Breslow and Wellner, 2008] Breslow, N. E. and Wellner, J. A. (2008). A z-theorem with estimated nuisance parameters and correction note for weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scandinavian Journal of Statistics*, 35(1):186–192.

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

[Bycott and Taylor, 1998] Bycott, P. and Taylor, J. (1998). A comparison of smoothing techniques for cd4 data measured with error in a time-dependent cox proportional hazards model. *Statistics in medicine*, 17(18):2061–2077.

[Cai et al., 2006] Cai, T., Pepe, M. S., Zheng, Y., Lumley, T., and Jenny, N. S. (2006). The sensitivity and specificity of markers for event times. *Biostatistics*, 7(2):182–197.

[Cai et al., 2005] Cai, T., Tian, L., and Wei, L. J. (2005). Semiparametric box-cox power transformation models for censored survival observations. *Biometrika*, 92(3):619–632.

[Cai and Zheng, 2011] Cai, T. and Zheng, Y. (2011). Nonparametric evaluation of biomarker accuracy under nested case-control studies. *Journal of the American Statistical Association*, 106(494):569–580.

[Cai and Zheng, 2013] Cai, T. and Zheng, Y. (2013). Resampling procedures for making inference under nested case-control studies. *Journal of the American Statistical Association*, 108(504):1532–1544.

[Cai and Sun, 2003] Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics*, 30(1):93–111.

[Chambers and Hastie, 1992] Chambers, J. M. and Hastie, T. (1992). *Statistical models in S*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, Calif.

[Chen, 2002] Chen, H. Y. (2002). Double-semiparametric method for missing covariates in cox regression models. *Journal of the American Statistical Association*, 97(458):565–576.

[Cook, 2007] Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115(7):928–35.

[Cox, 1972] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

[Crowley and Hu, 1977] Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, 72(357):27–36.

[Dafni, 1998] Dafni, Urania G., T. A. A. (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, 54(4):1445–1462.

[De Gruttola and Tu, 1994] De Gruttola, V. and Tu, X. M. (1994). Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, 50(4):1003–14.

[Efron, 1979] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

[Efron and Tibshirani, 1993] Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall, New York.

[Ettinger and Lagakos, 1982] Ettinger, D. S. and Lagakos, S. (1982). Phase iii study of ccnu, cyclophosphamide, adriamycin, vincristine, and vp-16 in small-cell carcinoma of the lung. *Cancer*, 49(8):1544–54.

[Etzioni et al., 1999] Etzioni, R., Pepe, M., Longton, G., Hu, C. C., and Goodman, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. *Medical Decision Making*, 19(3):242–251.

[Faucett and Thomas, 1996] Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A gibbs sampling approach. *Statistics in Medicine*, 15(15):1663–1685.

[Fisher and Lin, 1999] Fisher, L. D. and Lin, D. Y. (1999). Time-dependent covariates in the cox proportional-hazards regression model. *Annual Review of Public Health*, 20(1).

[Fleming and Harrington, 1991] Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis*. Wiley, New York.

[Garre et al., 2008] Garre, F. G., Zwinderman, A. H., Geskus, R. B., and Sijpkens, Y. W. J. (2008). A joint latent class changepoint model to improve the prediction of time to graft failure. *Journal of the Royal Statistical Society Series A-Statistics in Society*, 171:299–308.

[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):1029–40.

[Goldstein and Langholz, 1992] Goldstein, L. and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the cox regression model. *Annals of Statistics*, 20(4):1903–1928.

[Graf et al., 1999] Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):15–30.

[Gu and Pepe, 2009] Gu, W. and Pepe, M. (2009). Measures to summarize and compare the predictive capacity of markers. *International Journal of Biostatistics*, 5(1).

[Hastie and Tibshirani, 1986] Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–310.

[Hastie et al., 2003] Hastie, T. J., Tibshirani, R. J., and Friedman, J. (2003). *The elements of statistical learning: data mining, inference and prediction*. Springer, New York.

[Heagerty et al., 2000] Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344.

[Heagerty and Zheng, 2005] Heagerty, P. J. and Zheng, Y. Y. (2005). Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105.

[Henderson et al., 2000] Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

[Jin et al., 2003] Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, 90(2):341–353.

[Jin et al., 2001] Jin, Z., Ying, Z., and Wei, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika*, 88(2):381–390.

[Kalbfleisch and Prentice, 2002] Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data*. Wiley, New York.

[Kerr et al., 2011] Kerr, K. F., McClelland, R. L., Brown, E. R., and Lumley, T. (2011). Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *American Journal of Epidemiology*, 174(3):364–374.

[Korn and Simon, 1990] Korn, E. L. and Simon, R. (1990). Measures of explained variation for survival data. *Statistics in Medicine*, 9:487–504.

[Kulich and Lin, 2004] Kulich, M. and Lin, D. Y. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association*, 99(467):832–844.

[Laird and Ware, 1982] Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–74.

[Law et al., 2002] Law, N. J., Taylor, J. M. G., and Sandler, H. (2002). The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics*, 3(4):547–563.

[Leisenring et al., 1997] Leisenring, W., Pepe, M. S., and Longton, G. (1997). A marginal regression modelling framework for evaluating medical diagnostic tests. *Statistics in medicine*, 16(11):1263–81.

[Lin, 2007] Lin, D. (2007). On the breslow estimator. *Lifetime Data Analysis*, 13(4):471–480.

[Lin, 2000] Lin, D. Y. (2000). On fitting cox's proportional hazards models to survey data. *Biometrika*, 87(1):37–47.

[Lin and Wei, 1989] Lin, D. Y. and Wei, L. J. (1989). The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*, 84(408):1074–1078.

[Lin et al., 2002] Lin, H. Q., Turnbull, B. W., McCulloch, C. E., and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, 97(457):53–65.

[Liu et al., 2012] Liu, D., Cai, T., and Zheng, Y. (2012). Evaluating the predictive value of biomarkers with stratified case-cohort design. *Biometrics*, 68(4):1219–27.

[Lok et al., 2010] Lok, A. S., Sterling, R. K., Everhart, J. E., Wright, E. C., Hoefs, J. C., Di Bisceglie, A. M., Morgan, T. R., Kim, H. Y., Lee, W. M., Bonkovsky, H. L., Dienstag, J. L., and Group, H.-C. T. (2010). Des-gamma-carboxy prothrombin and alpha-fetoprotein as biomarkers for the early detection of hepatocellular carcinoma. *Gastroenterology*, 138(2):493–502.

[Maziarz et al., 2014] Maziarz, M., Black, R. A., Fong, C. T., Himmelfarb, J., Chertow, G. M., and Hall, Y. N. (2014). Evaluating risk of esrd in the urban poor. *Journal of the American Society of Nephrology*.

[Nan, 2004] Nan, B. (2004). Efficient estimation for case-cohort studies. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 32(4):403–419.

[Nan et al., 2009] Nan, B., Kalbfleisch, J. D., and Yu, M. (2009). Asymptotic theory for the semiparametric accelerated failure time model with missing data. *Annals of Statistics*, 37(5 A):2351–2376.

[Park and Wei, 2003] Park, Y. and Wei, L. J. (2003). Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika*, 90(3):717–723.

[Parzen et al., 1994] Parzen, M. I., Wei, L. J., and Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika*, 81(2):341–350.

[Pauler and Finkelstein, 2002] Pauler, D. K. and Finkelstein, D. M. (2002). Predicting time to prostate cancer recurrence based on joint models for non-linear longitudinal biomarkers and event time outcomes. *Statistics in Medicine*, 21(24):3897–3911.

[Pawitan, 1993] Pawitan, Yudi, S. S. (1993). Modeling disease marker processes in aids. *Journal of the American Statistical Association*, 88(423):719–726.

[Pencina et al., 2008] Pencina, M. J., D'Agostino, Ralph B., S., D'Agostino, Ralph B., J., and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the roc curve to reclassification and beyond. *Statistics in Medicine*, 27(2):157–172.

[Pepe, 1998] Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*, 54(1):124–135.

[Pepe, 2000] Pepe, M. S. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association*, 95(449):308–311.

[Pepe, 2003] Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, Oxford.

[Pepe and Janes, 2013] Pepe, M. S. and Janes, H. (2013). Methods for evaluating prediction performance of biomarkers and tests. Technical report, Fred Hutchinson Cancer Research Center.

[Pepe et al., 2013] Pepe, M. S., Kerr, K. F., Longton, G., and Wang, Z. (2013). Testing for improvement in prediction model performance. *Statistics in Medicine*, 32(9):1467–1482.

[Pfeiffer and Gail, 2011] Pfeiffer, R. M. and Gail, M. H. (2011). Two criteria for evaluating risk prediction models. *Biometrics*, 67(3).

[Pollard, 1990] Pollard, D. (1990). *Empirical processes: theory and applications*. Institute of Mathematical Statistics; American Statistical Association, Hayward, Calif.; Alexandria, Va.

[Prentice, 1982] Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69(2):331–342.

[Prentice, 1986] Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11.

[Proust-Lima and Taylor, 2009] Proust-Lima, C. and Taylor, J. M. G. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach. *Biostatistics*, 10(3):535–549.

[Rao and Zhao, 1992] Rao, C. R. and Zhao, L. C. (1992). Approximation to the distribution of m-estimates in linear models by randomly weighted bootstrap. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 54(3):323–331.

[Rizopoulos, 2010] Rizopoulos, D. (2010). Jm: An r package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33.

[Rizopoulos, 2011] Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3).

[Rizopoulos et al., 2013] Rizopoulos, D., Murawska, M., Andrinopoulou, E.-R., Molenberghs, G., Takkenberg, J. J., and Lesaffre, E. (2013). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *arXiv preprint arXiv:1306.6479*.

[Rizopoulos et al., 2009] Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009). Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):637–654.

[Saegusa, 2014] Saegusa, T. (2014). Bootstrapping two-phase sampling. Technical report, Department of Biostatistics, University of Washington.

[Samuelsen, 1997] Samuelsen, S. O. (1997). A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika*, 84(2):379–394.

[Schoop et al., 2008] Schoop, R., Graf, E., and Schumacher, M. (2008). Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics*, 64(2):603–610.

[Schoop et al., 2011] Schoop, R., Schumacher, M., and Graf, E. (2011). Measures of prediction error for survival data with longitudinal covariates. *Biometrical Journal*, 53(2):275–293.

[Self and Prentice, 1988] Self, S. G. and Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*, 16(1):64–81.

[Song and Wang, 2008] Song, X. and Wang, C. Y. (2008). Semiparametric approaches for joint modeling of longitudinal and survival data with time-varying coefficients. *Biometrics*, 64(2):557–566.

[Taylor et al., 2013] Taylor, J. M., Park, Y., Ankerst, D. P., Proust-Lima, C., Williams, S., Kestin, L., Bae, K., Pickles, T., and Sandler, H. (2013). Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*.

[Taylor et al., 2005] Taylor, J. M. J., Yu, M., and Sandler, H. M. (2005). Individualized predictions of disease progression following radiation therapy for prostate cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 23(4):816–25.

[Therneau and Li, 1999] Therneau, T. M. and Li, H. (1999). Computing the cox model for case cohort designs. *Lifetime Data Analysis*, 5(2).

[Thomas, 1977] Thomas, D. C. (1977). Addendum to: Methods of cohort analysis: Appraisal by application to asbestos mining. by f.d.k. liddell, j.c. mcdonald and d.c. thomas. *Journal of the Royal Statistical Society, Series A*, 140:469–491.

[Tian et al., 2004] Tian, L., Liu, J., Zhao, Y., and Wei, L. J. (2004). Statistical inference based on non-smooth estimating functions. *Biometrika*, 91(4):943–954.

[Tsiatis et al., 1995] Tsiatis, A., DeGruttola, V., and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association*, 90(429):27.

[Tsiatis and Davidian, 2001] Tsiatis, A. A. and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88(2):447–458.

[Tsiatis and Davidian, 2004] Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, 14(3):809–834.

[Uno et al., 2007] Uno, H., Cai, T., Tian, L., and Wei, L. J. (2007). Theory and methods - evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478).

[van Houwelingen, 2007] van Houwelingen, H. C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1):70–85.

[Vickers and Elkin, 2006] Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical decision making : an international journal of the Society for Medical Decision Making*, 26(6).

[Volk et al., 2007] Volk, M. L., Hernandez, J. C., Su, G. L., Lok, A. S., and Marrero, J. A. (2007). Risk factors for hepatocellular carcinoma may impair the performance of biomarkers: a comparison of afp, dcp, and afp-l3. *Cancer biomarkers : section A of Disease markers*, 3(2):79–87.

[Wang and Taylor, 2001] Wang, Y. and Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, 96(455):895–905.

[Wu et al., 2012] Wu, L., Liu, W., Yi, G. Y., and Huang, Y. (2012). Analysis of longitudinal and survival data: Joint modeling, inference methods, and issues. *Journal of Probability and Statistics*.

[Wulfsohn and Tsiatis, 1997] Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339.

[Xu and Zeger, 2001] Xu, J. and Zeger, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3):375–387.

[Ye et al., 2008] Ye, W., Lin, X. H., and Taylor, J. M. G. (2008). Semiparametric modeling of longitudinal measurements and time-to-event data-a two-stage regression calibration approach. *Biometrics*, 64(4):1238–1246.

[Yu et al., 2004] Yu, M., Law, N. J., Taylor, J. M. G., and Sandler, H. M. (2004). Joint modeling of longitudinal and survival data - joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*, 14(3):835.

[Yu et al., 2008] Yu, M., Taylor, J. M. G., and Sandler, H. M. (2008). Individual prediction in prostate cancer studies using a joint longitudinal survival cure model. *Journal of the American Statistical Association*, 103(481).

[Zheng et al., 2006] Zheng, Y., Cai, T., and Feng, Z. (2006). Application of the time-dependent roc curves for prognostic accuracy with multiple biomarkers. *Biometrics*, 62(1):279–287.

[Zheng and Heagerty, 2007] Zheng, Y. and Heagerty, P. J. (2007). Prospective accuracy for longitudinal markers. *Biometrics*, 63(2):332–341.

[Zheng et al., 2008] Zheng, Y., Pepe, M. S., Cai, T., and Levy, W. C. (2008). Time-dependent predictive values of prognostic biomarkers with failure time outcome. *Journal of the American Statistical Association*, 103(481):362–368.

[Zheng and Heagerty, 2004] Zheng, Y. Y. and Heagerty, P. J. (2004). Semiparametric estimation of time-dependent roc curves for longitudinal marker data. *Biostatistics*, 5(4):615–632.

[Zheng and Heagerty, 2005] Zheng, Y. Y. and Heagerty, P. J. (2005). Partly conditional survival models for longitudinal data. *Biometrics*, 61(2):379–391.

VITA

Marlena Maziarz majored in Human Biology and Computer Science at the University of Toronto receiving an Honours Bachelor of Science degree with High Distinction. She also holds a Master of Science degree in Computer Science from the University of Toronto. Marlena's interest in public health and her diverse background in biology, computer science and statistics motivated her to pursue a PhD in Biostatistics at the University of Washington.