Joint Modeling of Survival and Longitudinal Data Measured with Error, with Application to Assessing Immune Correlates of Protection in Vaccine Efficacy Trials

Rong Fu

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

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Peter B. Gilbert, Chair

Scott S. Emerson

Ching-Yun Wang

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Joint Modeling of Survival and Longitudinal Data Measured with Error, with Application to Assessing Immune Correlates of Protection in Vaccine Efficacy Trials

Rong Fu

Chair of the Supervisory Committee:
Professor Peter B. Gilbert
Department of Biostatistics

Assessing immune correlates of protection, the immune responses that reliably predict the vaccine efficacy on the clinical endpoint, has always been an important objective in vaccine efficacy trials. In this dissertation, we study the continuous and dichotomized trajectory of time-varying immune response as the immune correlate of protection in two-phase sampling design cohort studies. We adopt the joint modeling framework that models the immune response data measured longitudinally and with error and the time-to-event clinical endpoint simultaneously. The inherent evolution of the time-varying immune response is characterized by a random effects model, and its relationship with the instantaneous risk of the clinical event is modeled by the Cox proportional hazards regression. This regression model allows for direct assessment of immune correlates of protection in Prentice's framework. This evaluation is different from the traditional work that is based on measured values of biomarkers. Instead, by studying the underlying trajectory, the application is to generate hypotheses about the biological mechanisms of protection. The main objective of the dissertation is to develop statistical methods to make inference on the regression model accounting for the missing immune response data due to two-phase sampling. For the inference on the continuous immune response trajectory, we extend the existing conditional score method to the two-phase sampling design cohort studies by using the technique of weighting the complete cases by the inverse probabilities of observing the immune response data, and the

augmented inverse probability weighting. For the dichotomized immune response trajectory, we propose estimating equations based on regression calibration method. We also generalize it to two-phase samples by the inverse probability weighting method. We finally apply the proposed methods to the AIDS Clinical Trials Group (ACTG) 175 dataset, a randomized clinical trial comparing monotherapy with combination therapy among HIV-1-infected subjects.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my gratefulness to my advisor, Dr. Peter Gilbert, for his efforts and time spent on mentoring me. Also his brilliant perspectives on big questions and real study related research problems encourage and inspire me to develop my own background in broad fields of science. I am also very grateful to my committee members who provided me generous help and useful suggestion. I would also like to thank my parents and my brother who have always been supportive and caring about me.

DEDICATION

to my family

Chapter 1

INTRODUCTION

Vaccination has been widely used to prevent infectious disease. The basis of how vaccination works is the long-term immunological memory of the adaptive immune system. Our adaptive immune system protects us against the attack from specific pathogens in two aspects of immunity: humoral and cellular immunity. The humoral immunity refers to the production of antibodies (five isotypes in human: IgA, IgD, IgE, IgG and IgM) which bind to specific antigens to block their ability to infect; and the cell mediated immunity involves the function of CD4+ T-cell and CD8+ cytotoxic T lymphocyte which protect the body against the intracellular infection [Letvin, 2005, Pantaleo and Koup, 2004]. After an initial exposure and response to a pathogen, when being attacked by the same pathogen in future, the immune system is able to react more rapidly and more effectively to it. This is how the immunological memory is built up. A good candidate vaccine can provide durable protection against the infection or disease. It is a central goal in vaccine research to investigate the correlates of immunity: which and what type of immune responses are functional and predictive in conferring protection. In Phase IIb or Phase III vaccine efficacy trials, besides the assessment of vaccine efficacy (VE) to prevent or control the infection and disease, it is another very important objective to evaluate and determine the immune correlates of risk (CoRs) and immune correlates of protection (CoPs) [Haynes et al., 1996]. This is particular of interest when evidence of VE of a vaccine has been found.

[Qin et al., 2007] defined the CoR as an immunological biomarker that predicts the clinical endpoint used to assess the VE in some population. Ascertainment of a CoR can be done by statistical association analysis between the immune biomarker and the clinical endpoint. Such assessment requires the observation of variability in the immune biomarkers. However, for example, in HIV-1 vaccine efficacy trial on healthy volunteers, the immune response level could be zero for placebo recipients because of no prior exposure to the virus.

In that case, the CoR analysis is done among vaccinees only. In other vaccine efficacy trials, like the chimeric-yellow fever-dengue (CYD) vaccine, substantial variability in antibody titers can be observed in both the vaccine and placebo group. The CoR analysis can be done in each treatment group or pooled together adjusting for the vaccination status [Gilbert et al., 2008, Plotkin and Gilbert, 2012]. The assessment of a CoR is more straightforward than for a CoP, because the parameters of interest are statistical parameters that do not require extra causal modeling assumptions. It is the usually the first step to establish a CoR before moving forward to ascertainment of it as a CoP.

The concept of CoP has always been confusing. Informally speaking, identifying a CoP is to validate an immunological surrogate biomarker that can be used to reliably predict the protection effect of the vaccine for subgroups. It is of great use in the sense that it could substitute for the clinical endpoint which takes long time or even is unethical to study. It also provides the guidance of the vaccine development once we understand which immune biomarkers explain the vaccine effect on the clinical endpoint. For a CoR to be a CoP, the ambiguity and challenge arise to distinguish between a CoP that predicts the protection effect of the vaccine statistically versus mechanistically. We adopt the recent definition of CoP given by [Plotkin and Gilbert, 2012] that a CoP is an immune biomarker that can be used to reliably statistically predict VE. If a CoP predicts the VE because it is in the causal pathway that vaccine provides protection, it is called a mechanistic correlate of protection (mCoP). Otherwise, it is called a non-mechanistic correlate of protection (nCoP), which is not directly responsible for the protection but is correlated with a mechanistic one to be able to reliably statistically predict VE.

Statistical assessment of a CoP can be done using a number of approaches to evaluate a surrogate endpoint or causal mediator. [Joffe and Greene, 2009] reviewed and compared four major frameworks: Prentice's framework [Prentice, 1989]; direct and indirect causal effects of treatment [Pearl, 2001, Robins and Greenland, 1992]; principal stratification [Frangakis and Rubin, 2002] and meta-analysis. We discuss the first three more in subsequent sections. To further evaluate a CoP as a mCoP is more challenging because we need knowledge of biologic function of immune responses induced by vaccine and the disease process, and we even need to do intervention trial on animals or humans with or without exposure to the

immune responses.

This dissertation is built upon the importance of immune correlates analysis. However, it is not dealing with same problems as many other papers did conventionally to identify an measured immune biomarker as CoR and/or CoP. Observed immune biomarkers are contaminated by measurement errors and our methods assume that their evolution follows underlying trajectories. This dissertation, instead, concentrates on assessing how these unobserved hypothetical immune response processes predict the clinical endpoint and the protection effect of the vaccine. Therefore the dissertation is not aimed to propose a method to validate a biomarker as a CoP to be substituted for the clinical endpoint in real trials. Instead, by studying the underlying trajectories, the application is more to generate hypotheses about the biological mechanisms of protection. We consider it as an addendum to the traditional CoR/CoP analysis of observed immune responses. To address these scientific objectives, we consider statistical methods in the framework of jointly modeling the biomarker trajectories with a random or mixed effects model and the event time data with a Cox model simultaneously [DeGruttola and Tu, 1994, Faucett and Thomas, 1996, Tsiatis et al., 1995, Wang and Taylor, 2001a, Wulfsohn and Tsiatis, 1997]. More discussion on this is in Section 1.1.

In a large Phase IIb/III vaccine efficacy trial, it is costly and even redundant to assess the blood samples of all participants for evaluation of immune biomarkers, especially when there is a low rate of clinical event. Therefore cost-effective sampling designs are applied to reduce the number of participants on whom the biomarkers are collected. For example in HIV-1 vaccine efficacy trials, the serum and plasma samples can be only analyzed on a case-control subsample of participants who acquire HIV-1 infection and on a random subsample of participants who are not infected during the entire study. Also, a random subcohort could be selected at the time initiating the study and biomarkers are collected on the subcohort and all participants who have the clinical endpoint, which is referred to as the prospective case-cohort sampling [Prentice, 1986]. In this dissertation, we consider a general technique, two-phase sampling design, which includes case-control and retrospective case-cohort design as special cases. Since the biomarker measurements are missing for a subset of sample, statistical methods should account for this to achieve consistent estimates. However, the

joint modeling approaches have been rarely applied to the situation where the longitudinal measurements are only available on a subset of the study cohort. One research objective of this dissertation is therefore to apply the joint modeling approach to evaluate the time-dependent CoR or CoP under the two-phase sampling design. In Section 1.2, we introduce the definition of the two-phase sampling considered in this dissertation, and then review and discuss popular statistical strategies employed in two-phase sampling study.

Most joint modeling approaches are built up for continuous time-dependent covariates. Our experience with the immune response data in HIV-1 and dengue trials shows that in some vaccinated participants the level of immune marker declines almost to zero during a short period after the final immunization, while for others the level stays positive until the end of follow-up. This inspires us to look into if having a positive reaction to the presence of an immune response could predict the clinical endpoint. It is also interesting to consider what if the immune response begins to protect only when its value is above some threshold. Therefore the second research objective of this dissertation is motivated to develop statistical model to characterize the association between the time-to-failure data and the dichotomized current status (negative vs. positive or low vs. high) based on the underlying true immune responses.

In the following sections of this chapter, we first explain the motivation and statistical methods developed to address our scientific questions, and then review the two-phase sampling design and related statistical methods. Finally we introduce HIV-1 vaccine efficacy trials and dengue vaccine efficacy trials that motivate this dissertation.

1.1 Time-dependent CoR/CoP

1.1.1 Time-dependent CoR

In randomized vaccine efficacy trials, participants are randomly assigned to either placebo or vaccine group. Then they are followed up until the occurrence of a significant clinical endpoint, drop out, or the termination of study. We focus on the clinical endpoint which is the time to a clinically significant infection or disease. At the same time, their blood samples are collected frequently at multiple visits for measurements of immune responses.

The immune biomarker measured at a certain time point such as at the peak value after full immunization is of great interest in immune correlates analysis. However, additional information can be learned by studying the entire time-dependent biomarker process. Intuitively, it is closer to the mechanism of protection by evaluating how the current value of immune biomarker if exposed at this time to predict the occurrence of an endpoint in next short time period, than using the immune marker at a single time point to predict the endpoint in subsequent months or years of follow-up [Pawitan and Self, 1993]. Even though the CoR analysis is based on post-randomization immune biomarkers and can only identify association which has no causal interpretation, it is still useful in providing the insight and generating hypothesis of biological mechanism to be validated in future intervention experiment. This dissertation is therefore motivated with the first scientific goal of evaluating a immune biomarker as a time-dependent CoR.

One naive approach to model the association between longitudinal covariates and a time-to-event endpoint is the standard Cox proportional hazards model with time-dependent covariate [Kalbfleisch and Prentice, 2002]. It is reasonable and useful in the sense that it approximates the average instantaneous transmission probability of infection or disease per exposure to the current value of biomarker [Halloran et al., 1998, Rhodes et al., 1996]. However, one needs to be very cautious in interpreting this model when the time-dependent covariates are internal or not predictable [Kalbfleisch and Prentice, 2002](p196-199). This is especially an issue when the endpoint is time-to-death, because the measurement of an internal covariate requires the survival of the individual. Having become infected or diseased can also have dramatic influence on the level of immune response.

Another issue with the standard Cox model with a time-dependent covariate is that the covariate history over the entire follow-up period is required to obtain asymptotically consistent estimation [Andersen and Gill, 1982]. However such observations are not feasible in reality due to periodic collection of longitudinal immune biomarkers. One commonly used solution for this is to assume constant biomarker value between measurement time points. However it does not hold for the immune response levels that often decrease after immunization. Moreover, the laboratory assessment of the immune biomarkers are often subject to measurement errors, and ignoring such measurement errors may lead to biased

inference [Prentice, 1982].

Considering all these pros and cons with the standard Cox model, in this dissertation, we develop our statistical method to evaluate the time-dependent CoR in the context of joint modeling framework, i.e. modeling the longitudinal data with a mixed effects model and the event time data with a Cox model simultaneously [DeGruttola and Tu, 1994, Faucett and Thomas, 1996, Tsiatis et al., 1995, Wang and Taylor, 2001a, Wulfsohn and Tsiatis, 1997]. Unlike the standard Cox model assuming a hazard function conditional on the observed longitudinal immune biomarkers, the joint modeling method aims to quantify the effect of underlying true and unobserved evolution of the immune biomarker process on the time-to-event process. It assumes that the underlying true trajectory of the time-dependent immune biomarker has all of its effects on the time-to-event process through the random effects. In this way, by specifying the same functional form of the random effects and time in the Cox model as that in the longitudinal data model, we are able to estimate how the current value of hypothetical true biomarker if exposed at this time predicts the instantaneous rate of clinical endpoint. This evaluation of CoR is different from the traditional work that is based on measured value of biomarkers. We think it is of interest because studying the underlying trajectory may be better for generating hypotheses about the biological mechanisms of protection.

1.1.2 Time-dependent CoP

Statistical assessment of a CoP or an immunological surrogate biomarker can go back to Prentice's landmark paper in 1989 [Prentice, 1989]. By Prentice's definition, testing the null hypothesis of no vaccine effect on an immunological surrogate provides a valid test of the null hypothesis of no vaccine effect on the clinical endpoint. The criteria to validate a surrogate include (i) the vaccine has an effect on both the clinical endpoint and the surrogate biomarker; (ii) the surrogate is correlated with the clinical endpoint; and (iii) conditional on the surrogate, the distribution of the clinical endpoint is independent of the vaccination status. The joint modeling method itself enables the evaluation of statistical surrogates within this Prentice framework, envisioning the "surrogate" as the true underly-

ing biomarker trajectory. In this formulation, the Cox model includes both the vaccination status and the hypothetical immune biomarker trajectory process, as well as adjustment of confounders (assuming not affected by vaccination) of the biomarker-clinical-endpoint relationship. However checking these conditions, especially (c) is difficult in practice because we need to test the null hypothesis of "non-zero coefficient" and alternative hypothesis of "zero coefficient" of the vaccination status in aforementioned adjusted Cox model, which is infeasible in finite sample. An alternative to this hypothesis is to use the proportion of the vaccine effect on clinical endpoint that is explained by the biomarker [Freedman et al., 1992]. The proportion of treatment effect explained (PTE) under the Cox proportional model can be found in [Lin et al., 1997].

However, even though this framework has been widely used in evaluating a statistical surrogate, it may give misleading conclusion. Some literature criticized the attempt to use Prentice's framework to evaluate a surrogate biomarker for time-to-event endpoint because they are sufficient and necessary conditions for Prentice's definition of surrogate only for binary endpoints [Buyse and Molenberghs, 1998, Weir and Walley, 2006]. Moreover, either the hypothesis testing or the PTE estimates are operationally to estimate the vaccine effect adjusting for the biomarker based on observed data. However, this statistical control on the biomarker variable is generally biased estimation of the real estimand we desire, i.e. a measure of the vaccine effect that is not causally explained by the biomarker. Briefly speaking, this is because the adjustment of post-randomization biomarker introduces selection bias and the clinical endpoint is actually compared between individuals with and without vaccination but belong to two different sub-populations [Frangakis and Rubin, 2002, Joffe et al., 2007]. Again, we would like to emphasize that, by utilizing the joint modeling framework in Prentice's surrogate evaluation framework, we tend to deal with the latent biomarker trajectory instead of the observed one, with the purpose of obtaining the insight in generating the hypotheses about the biological mechanisms of protection.

The principal stratification framework is another way to assess a biomarker that statistically predicts the VE [Frangakis and Rubin, 2002]. This is developed based on potential or counterfactual endpoints and aims at estimating how the VE varies across subgroups defined by the vaccine effect of biomarker, or defined by the biomarker in vaccine recipients. For

the purpose of demonstration, we introduce some notations. Let T be the time to a clinical endpoint and Z is the vaccination status. For simplicity, we first do not distinguish between the observed longitudinal immune biomarkers and the unobserved latent time-dependent biomarker trajectory. Let $S(t)$ denote the value of a general post-randomization biomarker at time t which we would like to assess as a surrogate and let $\tilde{S}(t) = \{S(u), 0 \leq u \leq t\}$ be its history. Write $\tilde{S} = \tilde{S}(T)$. We apply the counterfactual notations: $T^{z\tilde{s}}$ is the time to clinical endpoint if the vaccination status and the entire immune biomarker process are assigned to z and $\tilde{s}$ respectively. Similarly a counterfactual surrogate endpoint can be defined for $\tilde{S}(t)$ and $S(t)$ with superscript $z = \{0, 1\}$.

In the principal stratification framework, we are interested in the causal effect of vaccination in a union of basic principal strata subgroups each defined by a pair of potential biomarker values $\{S^0(t_1), S^1(t_2)\}$. At a given time t_0, for example, the estimand defined on a particular stratum of interest is $\mathbb{P}(T^1 > t_0 | S^0(u) = S^1(u) = 0, 0 \leq u \leq t_0) - \mathbb{P}(T^0 > t_0 | S^0(u) = S^1(u) = 0, 0 \leq u \leq t_0)$, which compares the vaccine effect on the probability of no occurrence of clinical endpoint before and include t_0 for individuals who would have had zero immune biomarkers value at all times whether or not they were vaccinated.

We also would like to introduce another set of concepts defined on potential or counterfactual endpoints: "controlled direct effects" and "natural direct/indirect effects" [Pearl, 2001, Robins and Greenland, 1992]. The controlled direct effect is contrasting $T^{1\tilde{s}}$ with $T^{0\tilde{s}}$ where the treatment and biomarker process are jointly manipulated to z and $\tilde{s}$. A natural direct effect contrasts $T^{1\tilde{S}^{z^*}}$ with $T^{0\tilde{S}^{z^*}}$ where the surrogate biomarker level is left the value that would be if vaccination status had been z^*, $z^* \in \{0, 1\}$. The controlled direct effect makes sense when the full intervention on the biomarker is available. The natural direct effect allows for the biomarker to be the natural value if the one treatment had been imposed, so it is relevant to the mechanism how the treatment results in an endpoint. Most existing literatures discuss the identifiability of controlled or natural direct effect based on observed data focus on linear models with continuous clinical endpoints. Until recently there have only been a few paper on time-to-event endpoint and time-independent surrogate biomarker [Lange and Hansen, 2011, Martinussen et al., 2011, Tchetgen Tchetgen, 2011, VanderWeele, 2011]. [VanderWeele, 2011] proved that under certain no-unmeasured confounder assump-

tions as well as the rare event assumption, the natural direct effect comparing the ratio of hazard functions can be written as a complex formula of the Cox regression model coefficients adjusting for biomarker and the biomarker-treatment regression coefficients fitted on observed data. This provides insight that it could be very complicated to estimate and interpret the direct effects defined on the hazard scale. Therefore, a more straightforward estimand could be defined in terms of survival probabilities as above for principal stratification. [Zheng and van der Laan, 2012] investigated a more complex problem with event time endpoint and time-dependent surrogate. The challenge of such a problem is that the event time process may have an implication on the time-dependent biomarker process. The identifiability is troublesome if we also block the back door path from the event time process to the biomarker process.

Without certain assumptions, neither the natural/controlled direct effects nor the estimand defined on principal stratification can be identified from observed data, because only the endpoints under the assigned treatment can be observed [Tchetgen Tchetgen, 2014, Zheng and van der Laan, 2012]. In the case like in HIV-1 vaccine trials where the immune response levels for placebo recipients are zero and thus have no variability, the natural direct effect can be evaluated among the placebo recipients only with a simplified form [Lendle et al., 2013]. For the principal stratification method, we could naturally make a monotonicity assumption that $S^1(t) > S^0(t), t \in [0, \tau]$. This facilitates its identification in observed data [Tchetgen Tchetgen, 2014]. These three types of estimands are all of interest, and may be more or less fitting for different settings. In this dissertation, we are always interested in the latent true biomarker trajectories instead of their observed values. Therefore the definitions and interpretations of these estimands do no fit directly to our setting. Actually, since the latent trajectories are determined by some random effects models, intervention or stratification on $S^z(u)$ could be obtained totally through that on the random effects.

This dissertation is working mainly within the Prentice's framework by using joint modeling approach. We also attempt to relate it to the causal effect framework described above by exploring the definition of estimands and identifiability assumptions, as well as how to do estimation and interpretation.

1.2 Sampling design

1.2.1 Two-phase sampling design

Two-phase sampling is a cost-effective design which was first introduced by [Neyman, 1938]. It is particularly useful and efficient when it is expensive to measure some covariates or redundant to collect them on all subjects, especially in the rare event setting. [Haneuse et al., 2011] described a general two-phase sampling scheme. In vaccine efficacy trials, the first phase sample, or the study cohort, usually consist of all study subjects enrolled into the trial that are sampled from a super population of interest. On them, we follow up for the primary clinical endpoint, and measure covariates such as demographic characteristics, baseline health and medical information. Given these variables collected on all study subjects, we can further divide them into exclusive and exhaustive strata. Then within each stratum, we sample the Phase II subjects and the immune biomarkers of interest are only assessed on these Phase II subjects. The second phase sample could also be just a random subsample from the study cohort, but utilizing the stratified sampling scheme may increase efficiency. Such efficiency gain could be due to oversampling of the most informative individuals. Also it could be due to the retrieving of additional information that is associated with the covariate in the analysis model. Commonly, either Bernoulli sampling scheme or the sampling without replacement scheme are used to sample subjects for measurement of immune biomarkers. By Bernoulli sampling, each subject is examined independently for a Bernoulli indicator of whether or not they will be sampled. The feature of the Bernoulli sampling is that the sampled subjects are independent from each other, but the final number of subjects being sampled is random. By sampling without replacement, we are able to control for the total number of subjects being sampled but we loose the independency between them. In this dissertation, we focus on Bernoulli sampling.

Case-control study is one of the applications of two-phase sampling that has been widely used for rare binary outcomes in epidemiology. It is retrospectively sampling based on outcome status, where all cases who develop the disease, and a number of controls who do not have disease at the same time point are sampled at random or by stratum.

Another application is the case-cohort design in analyzing event time data, especially

in a large cohort study for rare event disease. It was first proposed in [Prentice, 1986] in a manner of prospective sampling where a simple random sample from the study cohort is taken at baseline (called subcohort) and the Phase II covariates are collected on this subcohort and all cases. [Borgan et al., 2000] extended it to the exposure stratified case-cohort design where the subcohort is taken independently from each stratum defined by Phase I covariates. The simple case-cohort or exposure stratified case-cohort design is considered as the stratified two-phase sampling design by considering all cases as a separate stratum and being sampled with probability one [Nan, 2004]. Note that, in this dissertation, we consider a general two-phase sampling design that does not necessarily include all cases.

1.2.2 Statistical methods

Regardless of whether the each individual is independently sampled or not, as long as the probability of being selected depends on the variables in the analysis model, the subsample selected is unrepresentative of the study population [Seaman and White, 2011]. The naive complete case (CC) analysis based on the complete observations is generally biased. One popular way to deal with missing data is multiple imputation (MI), where the missing observations are estimated by assumed distribution of observed and missing variables. Another widely used technique is inverse probability weighting (IPW) complete-case method. The concept of IPW was first proposed by [Horvitz and Thompson, 1952], where the complete observation is weighted by the inverse of the probability it being sampled. The intuition behind the IPW is to try to reconstruct the entire study population. MI does not need a model for the missing probability but does need a model for how the missing variables can be predicted from the observed data, while IPW requires the missing probability model only. The IPW estimator using pre-specified sampling probabilities is generally less efficient than the MI estimator, because it only makes use of the complete observations, and discards the subjects with missing data. However, the IPW estimator is still a favorable approach because it is easy to implement and easy to interpret. The IPW estimator provides unbiased inference if the score function and the sampling probability model are correctly specify. The unbiasedness of an MI estimator however requires the correct specification of the joint

distribution of observed and missing covariates to make correct imputation. Comparatively, the sampling probability model is easier to specified. Also in the two-phase sampling study where the longitudinal measurements are missing entirely for subjects outside the Phase II sample, the MI approach might not be very helpful since it is very hard to predict the entire course of the time-dependent covariate process.

[Breslow and Wellner, 2007] discussed the simple semi-parametric IPW estimators for both Bernoulli sampling and sampling without replacement in a general likelihood setting. They took the Cox proportional hazards model with time-independent covariates as a special case. As pointed out above, weighting the complete observations by the known sampling probabilities generally leads to inefficient estimation. So there has forthcoming a rich number of literatures to improve the efficiency. One popular way is to use estimated weights or calibrated weights leveraging Phase I covariates [Breslow et al., 2009a,b]. [Saegusa and Wellner, 2013] provided theoretical work on the asymptotic properties for them in a general semi-parametric model. They also compared the Bernoulli sampling to the sampling without replacement in terms of asymptotic variance of estimates. The estimated weights or calibrated weights help in a sense to account for the variability of the actual sampling fractions by utilizing the information from variables observed on all subjects. How much efficiency gain in finite sample setting may depend on the sample size and the correlation of Phase I covariates with the influence function. [Kulich and Lin, 2004] proposed a doubly weighted estimator specifically for Cox model with time-dependent variables. In their method, two levels of time-dependent weights were used. They offered a way in determining the second-level weight matrix using phase I covariates that led to an approximately optimal estimator.

Another direction is based on the augmented inverse probability weighting (AIPW) method proposed by [Robins et al., 1994], where an additional augmentation term as a function of the Phase I data (which are also called as auxiliary variables in such models) is added. This class of estimators has been almost exclusively focused on Bernoulli sampling. They demonstrated that an augmentation term as the conditional expectation of the influence function given the auxiliary variables achieved the optimal efficiency within the class of estimators with arbitrary forms of augmentation term. In this method, the IPW

estimator can be considered as a special case with augmentation term always equal to zero. AIPW estimators have the property of double robustness, which means the estimators are consistent as long as either the sampling probability model or the augmentation term as the conditional expectation of estimating score are correctly specified. The IPW estimators are however biased if the sampling probability model is misspecified. The application of AIPW specifically for the Cox proportional hazards model can be found in [Luo et al., 2009, Qi et al., 2005, Wang and Chen, 2001]. [Qi et al., 2005] proposed to estimate the sampling probability and the augmentation term via non-parametric kernel estimators, whose estimator have been proved to achieve the optimal efficiency. However, they all focused on time-independent covariates.

Several other semi-parametric methods were developed specifically for making inference on Cox regression model with time-independent covariates in an efficient manner under Bernoulli sampling[Chatterjee and Chen, 2007, Chatterjee et al., 2003, Nan, 2004, Nan et al., 2004, Scheike and Martinussen, 2004].

1.3 Motivating studies

1.3.1 VAX004 HIV-1 trial

VAX004, completed in 2003, was the world's first phase III placebo-controlled HIV-1 vaccine efficacy trial. The study was conducted to test the efficacy of AIDSVAX B/B, a recombinant HIV-1 envelope glycoprotein subunit (rgp120) vaccine[Flynn et al., 2005]. A total of 5,403 HIV-1-uninfected volunteers in North America and The Netherlands were included in the study. Participants were randomly assigned in 2:1 allocation to receive injections of vaccine or placebo at months 0, 1, 6, 12, 18, 24 and 30 and were followed up until Month 36. During the follow-up, 368 participants became HIV-1-infected. The VE, defined as (1 - hazard ratio of infection) $\times$ 100%, was estimated as 6% (95% CI: -17 to 24, p-value 0.59). Immune response biomarkers were measured at and two weeks after each immunization visit.

A follow-up study was published in 2005 evaluating the correlation of risk of eight vaccine-induced binding or neutralizing antibody responses to the hazard of HIV-1 infection

[Gilbert et al., 2005]. The antibody levels for immune correlates analysis were evaluated at the last peak time point (two weeks after vaccination) before HIV-1 infection for all HIV-1 infected vaccine recipients, and at all peak time points for a prospectively defined case-cohort sample that included 5% of vaccine recipients. It used the classic case-cohort sampling [Prentice, 1986] where the subcohort was determined by Bernoulli sampling at baseline and the immune response were measured in all cases of the last HIV-1 negative time-point only. The antibody levels for vaccinees were modeled both quantitatively and in discretized quartiles. For analyses with the quartile antibody levels, the relative risks, estimated by hazard ratios, comparing the higher response quartiles to the first quartile, and comparing each response quartile of vaccinees to the placebo were estimated independently for each antibody response variable. The Self-Prentice [Self and Prentice, 1988] method accommodating the case-cohort sampling was used to estimate the hazard ratios. Generally, a pattern of inverse correlation between antibody responses and the risk of HIV-1 infection was identified. Because the vaccine did not protect against HIV-1 infection, these correlates are interpreted as markers of susceptibility to HIV-1 infection.

[Forthal et al., 2007] performed the immune correlates analysis for the antibody-dependent, cell mediated virus inhibition (ADCVI) antibody generated by the vaccine. The antibody level measured at week 12.5 was used for uninfected vaccinees, and the antibody level measured at two weeks after the last vaccination before infection was used for infected vaccinees. The case-control sample for immune correlate analysis here consisted of all infected vaccinees and 5% uninfected vaccinees as well as an enriched sample of high-risk uninfected vaccinees. The hazard ratio of infection associated with ADCVI activity was estimated using Borgan II Estimator [Borgan et al., 2000], which respected the stratified and outcome-depending sampling. High level of vaccine-induced ADCVI activity was found to be correlated low HIV-1 infection rate.

[Li et al., 2008] investigated the same antibody as in [Forthal et al., 2007], but considered its longitudinal peak measurements after Month 6, i.e. at months 6.5, 12.5, 18.5, 24.5 and 30.5. The scientific objective of this immune correlate analysis was to evaluate the association of current ADCVI levels with HIV-1 infection over the next 6 months. The data were collected for a two-phase sample consisting of all infected vaccinees and a stratified

sample of uninfected vaccinees, with the strata defined by sex, race and high risk status. For infected vaccinees only the measurements taken at the visit before the diagnosis of infection were used. Unlike the former two immune correlates analyses, they treated the time-to-event data as grouped discrete failure time data. This had advantages in vaccine trial studies where the immune responses were tested at some pre-specified visits. The actual time-to-infection was only able to be identified within a time interval between two visits and the immune responses are assumed to be constant within each time interval and varying between them (a simplifying assumption known to be false). So this method was able to capture the time-dependence of the immune response somehow. The estimator of hazard ratio was obtained by maximizing the IPW likelihood of second phase subjects given such grouped event data structure. Multiple imputation was used for the missing biomarker data in the subcohort. They identified that the antibody was inversely associated with the risk of HIV-1 infection as well.

1.3.2 RV144 HIV-1 trial

RV144 was a randomized, placebo-controlled efficacy trial on a prime-boost HIV-1 vaccine (a combination of vaccines ALVAC-HIV and AIDSVAS B/E) in Thailand [Rerks-Ngarm et al., 2009]. In this study, 16,402 HIV-uninfected volunteers were randomized to the vaccine or placebo group in 1:1 allocation. The vaccine or placebo were administrated at weeks 0, 4, 12, and 24, with ALVAC-HIV administrated at all four visits and a boosting with AIDSVAX B/E at weeks 12 and 24 for vaccinees. Volunteers were then followed up for 42 months after entry. The testing for HIV-1 infection was made at weeks 0, 26, and every 6-month follow-up visit until the termination of study. A total of 125 HIV-1 infections were diagnosed in the modified intention-to-treat analysis set (excluding 7 participants who were HIV-1 infected at baseline). The corresponding VE from Cox model was estimated as 31.2% (95% CI=1.1 to 52.1; log-rank test p-value=0.04), suggesting a modest protective effect of the vaccine against HIV-1 infection. It was the first supporting evidence of a partially efficacious HIV-1 vaccine.

The subsequent immune correlates analysis on six pre-selected immune responses mea-

sured at Week 26 (two weeks after the final immunization) identified two of them were significantly correlated with the risk of HIV-1 infection in vaccinees [Haynes et al., 2012]. So the focus of the immune correlates analysis is to evaluate the CoR in vaccine recipients. The immune responses were taken on a two-phase sample, including all vaccinees who were diagnosed with HIV-1 infection after week 26 and a stratified subsample of vaccinees who were not infected throughout the study. The stratification variables included sex, the number of vaccinations received, and per-protocol status. The hazard ratio of infection was estimated using Borgan II estimator [Borgan et al., 2000]. Two immune variables, IgA binding antibody and the binding of IgG antibody to V1V2 of the gp120 Env, were identified as significantly correlated with HIV-1 infection. This analysis was unable to evaluate the immune correlates of protection within the Prentice's frame work [Prentice, 1989] since the immune responses were not variable in the placebo arm. The positive findings of RV144 in vaccine efficacy and correlates of risk have been interpreted as being very important for the HIV-1 vaccine and general vaccine fields, providing guidance for the design of future vaccines and vaccine trials.

There are forthcoming data which include measurements of immune responses at all 6-monthly visits (from Month 0 to 36) prior to HIV-1 infection diagnosis for all vaccine recipients who acquired HIV-1 infection during the trial, and measuring immune responses at all 6-monthly visits for a selected random sample of vaccine recipients who reached the Month 42 terminal study visit HIV-1 negative. The longitudinal data are anticipated from 41 infected and 205 uninfected vaccine recipients. This dataset allows for the time-dependent immune correlates analyses.

1.3.3 CYD14 and CYD15 dengue vaccine trials

In 2014, two phase III vaccine efficacy trials on live attenuated tetravalent dengue vaccine (CYD-TDV) demonstrated substantial vaccine efficacy in preventing the dengue primary disease, virologically confirmed symptomatic dengue of any serotype. The CYD14 trial consisted of 10,275 children in five Asian countries with an estimated VE of 56.5% (95% CI = 43.8 to 66.4) [Capeding et al., 2014]. The CYD15 trial was conducted in Latin America

with a total of 20,869 participants and an estimated VE of 60.8% (95% CI = 52.0 to 68.0) in a recent press release [Villar et al., 2014]. In both trials, study participants were randomized in 2:1 allocation to receive vaccine or a placebo at months 0, 6 and 12, and then were followed for 13 months as active phase for dengue disease. The immune responses, four anti-dengue serotype-specific neutralizing antibody titers, were measured at months 0, 7, 13 and 25 on a random immunogenicity subset including both placebo and vaccine recipients. In the dengue trials, because many of the trial participants were previously exposed to dengue viruses and hence the antibody responses substantially vary in the placebo arm. Therefore, we are able to evaluate the time-dependent immune correlates of protection for these dengue trials using Prentice's framework.

1.4 The outline of this dissertation

This dissertation is motivated to evaluate the time-dependent CoRs and CoPs in vaccine efficacy trials where the immune response variables are measured under the two-phase sampling design. The subjects to be sampled for measurement of immune response data are by Bernoulli sampling. Statistically, we could evaluate the time-dependent CoPs in the frameworks of Prentice's criteria based on Cox the proportional hazards model with the (continuous or dichotomized) time-varying process of the immune biomarker as a covariate. Considering the measurement error of the immune responses, we adopt the "joint modeling" framework to make inference on such Cox models and account for the missing biomarker data by design.

The structure of the dissertation is as follows: Chapter 2 develops the IPW and AIPW conditional score estimator for the joint model of continuous longitudinal biomarker and event time data under two-phase sampling design. Results on asymptotic properties are provided. Chapter 3 presents simulation studies to evaluate the performance of the IPW and AIPW conditional score estimators in terms of consistency and efficiency. Chapter 4 develops the risk set recalibration method and related theories for the model with dichotomized biomarker process and Chapter 5 includes the corresponding simulation studies. In Chapter 6, we applied the proposed method to AIDS Clinical Trials Group (ACTG) 175 study [Hammer et al., 1996]. In Chapter 7 there are discussions on the proposed methods as well

as open questions for future research.

Chapter 2

JOINT MODELING FOR CONTINUOUS TIME-DEPENDENT BIOMARKERS IN TWO-PHASE SAMPLING DESIGN COHORT STUDIES

2.1 Background

This chapter focuses on evaluating the continuous underlying trajectory of immune biomarkers as time-dependent CoRs and CoPs in vaccine efficacy trials. As mentioned in Section 1.1, there are several complications in analyzing the classic Cox regression with time-dependent covariate. First, a valid inference of the model requires the functional form of the time-varying trajectory of the covariate. One most commonly used solution to this is assuming constant covariate values between two subsequent measurement time-points. However, this approach fails to capture any variation of the covariate values between two time points, especially for long intervals. Also the measured immune responses from assays are subject to measurement errors and the classic Cox regression ignoring such errors could lead to biased inference. This inspires us to adopt the "joint modeling" strategy that models the underlying true trajectory of the time-dependent covariate and the event time endpoint simultaneous.

The fundamental setup of a joint model consists of two sub-models: one for the inherent trajectory of the time-dependent covariate and one for the time-to-event process. The covariate sub-model characterizes the hypothetical underlying trajectory of the time-dependent covariate. Commonly used models include the linear mixed effects model [Guo and Carlin, 2004, Henderson et al., 2000, Rizopoulos et al., 2009] or linear random effects model [Dafni and Tsiatis, 1998, Tsiatis and Davidian, 2001, Wulfsohn and Tsiatis, 1997]. Great flexibilities can be achieved to model the evolution of the covariate process over time by using a polynomial or a spline function of time. Several other papers dealt with the generalized mixed effects model [Xu and Zeger, 2001] or nonlinear mixed effects model [Wu et al., 2010]. The model of [Wu et al., 2010] also accounted for the biological understanding

of the biomarker process in response to treatment. With the help of such trajectory models, the covariate value can be obtained at any time point continuously. Most literatures use a Cox proportional hazards model for the sub-model of time-to-event data process. Accelerated failure time (AFT) models are also studied [Hanson et al., 2011, Tseng et al., 2005]. As discussed in Section 1.1, we focus on the Cox regression model in this dissertation. This model can be a good fit to the infectious disease setting where there is major interest in understanding how the current level of biomarkers affect the instantaneous risk, a setting where instantaneous hazard rates are interpretable and the proportional hazards assumption may be reasonable. Various ways have also been proposed to link the two sub-models together, by incorporating different functional forms of the random effects (and other time-dependent predictors) in the random effects model (or mixed effects model) to the hazard function. In other words, these two sub-models are linked together via the random effects. For example the majority of the literatures take the hazard function depending on the current underlying covariate value [Tseng et al., 2005, Wang and Taylor, 2001b] and some others use only several components of the random effects and/or their interaction with time [Guo and Carlin, 2004, Henderson et al., 2000, Song et al., 2002, Wang, 2006].

In the early stage, the two-stage method is used for making inference on the joint model. The time-dependent covariate models are fitted first and the covariate values are imputed at desirable time points to fit the Cox regression separately [Pawitan and Self, 1993]. Such a naive imputation method suffers from non-eliminated bias since it ignores the relationship between the measured longitudinal covariates and the event time data. For example, more measurements may indicate longer time to event. Another class of methods are recalibration methods based on [Prentice, 1982], which are aimed to reduce the bias by estimating the hazard function given the observed covariate values [Dafni and Tsiatis, 1998, Tsiatis et al., 1995, Wang et al., 1997, 2000, 2001]. However since it is complex to derive the analytical form of the observed-covariate-hazard, such methods are generally based on strong model assumptions and approximations, thus still failing to reduce the bias entirely. Likelihood approaches have also been developed to making inference for joint models [DeGruttola and Tu, 1994, Rizopoulos et al., 2009, Wulfsohn and Tsiatis, 1997]. The likelihood approach is most often considered due to its efficiency. However, despite the requirement of specifying

the joint distribution of the random effects and the event time data, the likelihood function usually involves no-closed form of the integral over unknown random effects. Therefore considerable computational burden is anticipated due to the numerical integration. R packages are available for the likelihood inference: `JM` [Rizopoulos, 2010] and `joineR` [Philipson et al., 2012]. Some other researchers developed a set of Bayesian procedures for inference using Markov Chain Monte Carlo (MCMC), which relies further on the specification for distribution of parameters and requires the examination for convergence [Faucett and Thomas, 1996, Xu and Zeger, 2001]. The above mentioned methods are easy to interpret but have their own drawbacks such as the complication to implement and the needs for distributional assumptions. [Tsiatis and Davidian, 2001] developed a conditional score method with a notable innovation that it does not rely on the distributional assumption of the random effects at all. The rationale of the method is to derive the intensity density of the event time process conditional on a complete and sufficient statistic of the unknown random effects. Therefore, the induced conditional hazard function given it does not depend on the unknown random effects at all. Then estimating equations for coefficients of the Cox regression model are constructed in a spirit similar to that for partial likelihood score. This method is much less computationally intensive and is easy to generalize to handle multiple time-dependent covariates [Song et al., 2002]. [Wang, 2006] developed a corrected score method by constructing estimating equations whose conditional expectation given the random effects are asymptotically equivalent to the partial likelihood score equations in terms of the true underlying time-dependent covariates. Both conditional score and corrected score methods are consistent and asymptotically normal under regularity conditions. For reviews of more joint modeling research please see [McCrink et al., 2013, Tsiatis and Davidian, 2004, Wu et al., 2012].

This dissertation favors the advantages of conditional score estimator. In particular the motivating studies include multiple immune biomarkers indicating an interest of analyzing the association of one single biomarker to the clinical endpoint adjusting for others, a setting where the likelihood approach may computationally fail. However, to the best of our knowledge, no joint modeling approaches focus on the situation where the longitudinal biomarkers are measured on a random or a biased subsample of the full study cohort, which is usually

the case in vaccine efficacy trials and generally in prevention efficacy trials. As a result, in this chapter, we adopt the conditional score method, and develop the corresponding IPW and AIPW estimator to accommodating the missingness due to the two-phase sampling. We only consider the Bernoulli sampling for the second phase sample. For demonstration simplicity, we concentrate on the model with a single longitudinal biomarker. The theories and deviations apply to the model with multiple longitudinal biomarkers immediately.

2.2 Notation and modeling

2.2.1 Longitudinal data model and survival data model

Let T and C be the event time and censoring time. The observed right-censored data is $V = \min(T, C)$ and $\Delta = I(T \leq C)$. Let $\tilde{Z} = (Z, L^T)^T$ where Z is the treatment group (1 for vaccination and 0 for placebo) and L is a $p-1$ dimensional vector of baseline time-independent confounding variables. We denote the time-dependent biomarker process which is not observed directly by $\widetilde{X}(\tau) = \{X(u), 0 \leq u \leq \tau\}$, with τ being the time when the follow-up ends and $X(u)$ being the value of biomarker at time u. We assume the following random effects model representing the inherent trajectory of $X(u)$

$$X(u) = \alpha^T f(u) \tag{2.1}$$

where $f(u)$ is a q dimensional vector of known functions of time u and α are the subject-specific random effects. Flexible models (e.g., polynomial or spline model) are obtainable via different specifications of $f(u)$. For example $f(u) = (1, u)^T$ specifies a simple linear model. The observed longitudinal biomarker values are from an additive measurement error model

$$W(u) = \alpha^T f(u) + e(u) \tag{2.2}$$

The measurement errors $e(u)$ are Normal distributed with mean zero and variance $\mathbb{C}ov(e(u), e(s)) = I(u = s)\sigma^2$. Also we assume $e(u)$ is independent of α. Suppose the set of measurement time points are $T^m = (T_1^m, \cdots, T_J^m)^T$ with $0 \leq T_1^m < T_2^m < \cdots < T_J^m \leq V$, and J being the total number of time points. We allow T^m to be varying by subjects. Then let

$W = (W_1, \cdots, W_J)^T$, where $W_j = W(T_j^m)$ and let $e = (e_1, \cdots, e_J)$ where $e_j = e(T_j^m)$.

For any fixed time u, let $J(u)$ be the maximum number of measurement time points up to and including time u (i.e. $0 \leq T_1^m < \cdots < T_{J(u)}^m \leq u < T_{J(u)+1}^m$) and $T^m(u) = (T_1^m, \cdots, T_{J(u)}^m)^T$ be the corresponding vector of ordered measuring time points. We consider the following proportional hazards model of the event time

$$\begin{aligned} \lambda(u) &= \lim_{du \to 0} \frac{1}{du} \mathbb{P}(u \leq T < u + du | T \geq u, \alpha, \tilde{Z}, T^m(u), C) \\ &= \lim_{du \to 0} \frac{1}{du} \mathbb{P}(u \leq T < u + du | T \geq u, \alpha, \tilde{Z}) \\ &= \lambda_0(u) \exp\{X(u)\beta + \tilde{Z}^T \eta + X(u) Z \gamma\} \end{aligned} \tag{2.3}$$

In this model, we assume non-informative censoring and non-informative measuring time to the event time given the information already provided by α and $\tilde{Z}$. [Song et al., 2002] considered a more generalized hazard function than (2.3), where they assumed $\lambda(u) = \lambda_0(u) \exp\{\beta^T G(u) \alpha + \eta^T \tilde{Z}\}$, with $G(u)$ being a matrix of functions of u. Such specification links the event time to the time-dependent biomarker through the $G(u)\alpha$. For example, when $G(u) = f^T(u)$, it reduces to the hazard function (2.3) we are considering; when $G(u)$ is a $q \times q$ identity matrix, it becomes the Cox proportional hazards model taking the random effects as covariates, which is the model considered in [Wang, 2006]. Modeling the random effects as covariates is particularly of interest when the trend of $X(u)$ is believed to dominate the association between time-dependent biomarker and the event time. This dissertation considers the hazard function (2.3) and intends to evaluate how the current hypothetical true value of biomarker predicts the instantaneous hazard of interest, which intuitively is closer to the mechanism of protection. Our derivations based on (2.3) are ready to be extended to Song's proportional hazards model.

2.2.2 Two-phase sampling model

We consider the vaccine efficacy trials where immune biomarker data are not collected on all participants by design. For example, in the RV144 HIV-1 Thai trial, the immune responses were assessed on a case-control sample. In CYD dengue vaccine trials, the longitudinal pro-

file of the antibody titers for each serotype were only measured on a random immunogenicity set ($\sim 18\%$) from the study cohort. To obtain valid inference, we must apply appropriate statistical technique to account for such sampling design, which is called two-phase sampling [Breslow and Wellner, 2007, Haneuse et al., 2011, Neyman, 1938]. Now we introduce the sampling design we consider in this dissertation and the associated notations.

Generally, in the first phase, a large sample with size N is taken from the study population. Let $(V_i, \Delta_i, Z_i, L_i^T, A_i^T)^T, i = 1, \cdots, N$ be the data collected on N independent subjects. The vector A_i are auxiliary variables which might be predictive of the immune biomarker. In the second phase, a random Bernoulli sample is taken from the N subjects, with sampling probabilities given by $\pi(O_i)$, where O_i are (a subset of) the variables collected at the first phase. Let ξ_i be the binary indicator of being sampled ($\xi_i = 1$). Then the longitudinal immune biomarkers $\{W_i, T_i^m, J_i\}$ are assessed only on subjects with $\xi_i = 1$, i.e. the observed data for $i = 1, \cdots, N$ are $\{V_i, \Delta_i, Z_i, L_i, A_i, \xi_i, \xi_i W_i, \xi_i T_i^m, \xi_i J_i\}$.

To emphasize, the sampling probability model is characterized by a parametric model in terms of finite-dimensional parameter ρ

$$\mathbb{P}(\xi = 1|O, \alpha, W, T^m, J) = \mathbb{P}(\xi = 1|O) = \pi(O; \rho) \tag{2.4}$$

This is the missing at random (MAR) assumption. We also assume positive sampling probabilities with $0 < \delta < \pi_i(O_i; \rho) \leq 1$ for some constant value $\delta > 0$ and for all $i = 1, \cdots, N$.

Now considering the special situation where the second phase sample is taken by a stratified Bernoulli sampling. That is to say, suppose the N subjects can be divided exclusively and exhaustively into S strata based on O: $\{\mathcal{O}_1, \cdots, \mathcal{O}_S\}$. We use $I(O \in \mathcal{O}_s)$ to indicate whether a subject belongs to stratum $\mathcal{O}_s$. In a case-control sampling, the strata are defined by Δ. Let $N_1, \cdots, N_S$ be the size of each stratum such that $N_1 + \cdots + N_S = N$. In the second phase, if a subject belongs to stratum $\mathcal{O}_s$, then with a probability ρ_s the subject will be sampled, i.e.

$$\pi(O; \rho) = \sum_{s=1}^{S} I(O \in \mathcal{O}_s)\rho_s \tag{2.5}$$

We also let the probability of belonging to a stratum as $\nu_s = \mathbb{P}(O \in \mathcal{O}_s) > 0$, $s = 1, \cdots, S$.

2.3 Methods to evaluate time-dependent CoP

The central application of this dissertation is to evaluate how the current level/status of immune response is associate with the instantaneous risk of the clinical endpoint and to assess the validity of current level/status of immune response as a CoP for measuring VE. The assessment of association is straightforward based on model (2.3). For the assessment of a CoP, this dissertation mainly focuses on the Prentice's framework [Prentice, 1989] in terms of statistical parameters. As a complement, we also introduce the framework based on causal effects, which leads to causal interpretation. This section therefore describes these two frameworks which the joint model (2.3) can be applied to assess the time-dependent CoP.

2.3.1 Prentice's criteria

The implementation of joint modeling approach in two-phase sampling provides the way to analyze the underlying hypothetical trajectory of time-dependent immune biomarkers as correlates of risk directly in vaccine efficacy trials where the biomarker data are only measured in a subset of subjects. This approach also enables the evaluation of time-dependent immune correlates of protection in the framework of Prentice's approach [Prentice, 1989]. By Prentice's definition, for an immune biomarker to be an immunological surrogate, is one on which the test of the null hypothesis of no vaccine effect is also a valid test of the null hypothesis of no vaccine effect on the clinical endpoint. To apply the Prentice's framework to assess the time-dependent CoP based on model (2.3), we need to check

(i) Z has an effect on T, and Z has an effect on the immune biomarker $\tilde{X}(\tau)$.

(ii) $\tilde{X}(\tau)$ is correlated with the clinical endpoint T in both treatment groups.

(iii) Z has no effect on T given the immune biomarker $\tilde{X}(\tau)$. To check this, we need first to rule out that the vaccine effect on T is modified by the biomarker, i.e. there is a significant interaction effect in (2.3). Otherwise, the biomarker fails to meet Prentice's criteria. If we do declare $\gamma = 0$, then we fit $\lambda(u) = \lambda_0(u) \exp\{X(u)\beta + Z\eta_{Z|X} + L^T\eta_L\}$

including all dual predictors L for $X(u)$ and T , and to see if $\eta_{Z|X} = 0$ is plausibly close to zero.

However checking condition (iii) is difficult statistically because conceptually we need to test the null hypothesis of $\eta_{Z|X} \neq 0$ versus the alternative $\eta_{Z|X} = 0$, which requires infinity sample size. An insignificant p-value for the regression coefficient $\eta_{Z|X}$ could be lack of evidence to reject $\eta_{Z|X} = 0$, instead of evidence to accept $\eta_{Z|X} = 0$. We could use the confidence interval of $\eta_{Z|X}$ to judge the precision with which it is near zero. An alternative way is to use a value to measure the proportion of treatment effect explained (PTE) by the biomarker, defined as $1 - \eta_{Z|X}/\eta_Z$, where η_Z is the regression coefficient of Z without adjustment of the biomarker [Freedman et al., 1992, Lin et al., 1997]. However, PTE is not guaranteed to be in [0,1] and could be quite variable, suggesting it is only very useful when the η_Z is very large [Flandre and Saidi, 1999]. A recent published paper defined a new measurement, the proportion of treatment effect captured by the potential surrogate (PCS), still with 1 indicating a perfect surrogate and 0 indicating a useless surrogate [Kobayashi and Kuroki, 2014]. PCS is guaranteed to be in [0,1] and is less variable.

2.3.2 Causal effects framework

Another framework to evaluate the CoP which confers causal effects interpretation is based on the concepts of natural direct and indirect effects.

In Section 1.1.2, we define the conterfactual underlying biomarker history up to and including time t as $\tilde{X}^z(t) = \{X^z(u), 0 \leq u \leq t\}$ if the vaccination status had taken $Z = z$. To evaluate the underlying biomarker trajectory as a surrogate, we would like to look at the potential time to clinical endpoint $T^{z\tilde{x}}$ when the vaccination status had taken z and the underlying biomarker trajectory history had taken $\tilde{x}$. Note that, under the random effects model assumption (2.1), the hypothetical biomarker trajectory is entirely determined by the random effects. Besides, the time to clinical endpoint depends on this trajectory all through the random effects. Therefore, in this section, instead of using the notation of a biomarker process $\tilde{X}^z(t)$, we use notation X^z_α to emphasize that the whole biomarker trajectory is determined all by α.

Under this potential-outcome framework, the concepts of direct and indirect causal effects of treatment have been defined. The controlled direct effect (CDE) contrasts $T^{1\tilde{x}_\alpha}$ with $T^{0\tilde{x}_\alpha}$, i.e. comparing the time to clinical endpoint when the treatment had taken 1 to that when the treatment had taken 0, manipulating the biomarker process to $\tilde{x}_\alpha$. It is often of interest in policy making. The natural direct effect (NDE_z) contrasts $T^{1X^z_\alpha}$ with $T^{0X^z_\alpha}$, where the biomarker level is allowed to be the value it would be if treatment had been z, $z \in \{0, 1\}$. The total effect (TE) can be decomposed into the sum of NDE_z and the natural indirect effect (NIE_z). This provokes the measurement NIE_z/TE, or the PCS_z defined in [Kobayashi and Kuroki, 2014] as

$$PCS_z = \frac{NIE_z^2}{NDE_z^2 + NIE_z^2}$$

to quantify the proportion of causal effect captured by X_α. A proportion close to one could indicate a good surrogate. The problem with these definitions based on counterfactual endpoints is that only the ones under the assigned treatment can be observed. Thus in order to make inference on the causal effects based on observed dataset, we need to make some identification assumptions given in Assumption I. Recall that L is a vector of potential confounders measured at baseline.

Assumption I

I1. Consistency. $T = T^{zx_\alpha}$ if $Z = z$ and $X_\alpha = x_\alpha$. $X_\alpha = X^z_\alpha$ if $Z = z$.

I2. $(Z, X_\alpha) \perp T^{zx_\alpha}|L$

I3. $Z \perp X^z_\alpha|L$

I4. $T^{zx_\alpha} \perp X^{z^*}_\alpha|Z, L$, for any z, $z^* \in \{0, 1\}$

I5. Zero biomarker level among the untreated $\mathbb{P}(X_\alpha = 0|Z = 0) = 1, a.s$

For a fixed time point $t_0 \in [0, \tau]$, we define the natural direct effect for treatment $z = \{0, 1\}$ as

$$NDE_z(t_0) = \mathbb{E}[I(T^{1X^z_\alpha} \geq t_0) - I(T^{0X^z_\alpha} \geq t_0)] \tag{2.6}$$

and the natural indirect effect for treatment as

$$NIE_z(t_0) = \mathbb{E}[I(T^{1-z,X^1_\alpha} \geq t_0) - I(T^{1-z,X^0_\alpha} \geq t_0)] \tag{2.7}$$

So that $NDE_z(t_0) + NIE_z(t_0) = TE(t_0) \equiv \mathbb{E}[I(T^1 \geq t_0) - I(T^0 \geq t_0)]$. From the following theorems, we are able to estimate $NDE_z(t_0)$ and $NIE_z(t_0)$ based on observed data and then further evaluate the $PCS_z(t_0)$.

Theorem 2.3.1. *Under Assumptions I1 - I4, $NDE_z(t_0)$ and $NIE_z(t_0)$ are identifiable from the observed data.*

Proof. Note that by Assumption I1 - I3 we have

$$\begin{aligned}
\mathbb{E}[I(T^{z^*x_\alpha} \geq t_0)|L] &= \mathbb{E}[I(T^{z^*x_\alpha} \geq t_0)|Z = z^*, X_\alpha = x_\alpha, L] \\
&= \mathbb{E}[I(T \geq t_0)|Z = z^*, X_\alpha = x_\alpha, L] \\
&= S_T(t_0|Z = z^*, X_\alpha = x_\alpha, L) \\
d\mathbb{P}_{X^z_\alpha}(x_\alpha|L) &= d\mathbb{P}(X^z_\alpha \leq x_\alpha|Z = z, L) \\
&= d\mathbb{P}(X_\alpha \leq x_\alpha|Z = z, L)
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}[I(T^{z^*X^z_\alpha} \geq t_0)] &= \mathbb{E}\{\mathbb{E}\{\mathbb{E}[I(T^{z^*X^z_\alpha} \geq t_0)|Z, X^z_\alpha, L]Z, L\}\} \\
&= \mathbb{E}\{\int_{x_\alpha} \mathbb{E}[I(T^{z^*X^z_\alpha} \geq t_0)|Z, X^z_\alpha = x_\alpha, L]d\mathbb{P}_{X^z_\alpha}(x_\alpha|Z, L)\} \\
&= \mathbb{E}\{\int_{x_\alpha} \mathbb{E}[I(T^{z^*x_\alpha} \geq t_0)|Z, X_\alpha = x_\alpha, L]d\mathbb{P}_{X^z_\alpha}(x_\alpha|L)\} \quad \text{(Assumption I1,I3)} \\
&= \mathbb{E}\{\int_{x_\alpha} \mathbb{E}[I(T^{z^*x_\alpha} \geq t_0)|Z, L]d\mathbb{P}_{X^z_\alpha}(x_\alpha|L)\} \quad \text{(Assumption I1,I4)} \\
&= \mathbb{E}\{\int_{x_\alpha} \mathbb{E}[I(T^{z^*x_\alpha} \geq t_0)|L]d\mathbb{P}_{X^z_\alpha}(x_\alpha|L)\} \quad \text{(Assumption I1,I2)} \\
&= \mathbb{E}\{\int_{x_\alpha} S_T(t_0|Z = z^*, X_\alpha = x_\alpha, L)p_{X_\alpha}(x_\alpha|Z = z, L)dx_\alpha\} \\
&= \mathbb{E}\left[g_{z^*,z}(L; t_0, \lambda_0, \theta, \theta_\alpha)\right]
\end{aligned}$$

where $\lambda_0 = \lambda_0(u)$, $\theta = (\beta, \eta^T, \gamma)^T$ are parameters involved in the hazard function of observed

data given in (2.3), and θ_α are the parameters involved in the density $p_{X_\alpha}(x_\alpha|Z=z,L)$. The first term $S_T(t_0|\cdot)$ inside the integration over x_α can be fitted from observed data using (2.3). The second term is the conditional density function of X_α, which is indeed determined by α, given $Z=z,L$. The conditional score estimator, however, dose not require any distributional assumption on the random effects. Therefore, in order to further identify $NDE_z(t_0)$ and $NIE_z(t_0)$, we have to make the distributional assumption for α, and estimate θ_α through likelihood approach based on data W,Z,L. Finally $NDE_z(t_0)$ and $NIE_z(t_0)$ can be estimated by

$$\widehat{NDE}_z(t_0) = N^{-1}\sum_{i=1}^{N}\Big\{g_{1,z}(L_i;t_0,\hat{\lambda}_0,\hat{\theta},\hat{\theta}_{\alpha_i}) - g_{0,z}(L_i;t_0,\hat{\lambda}_0,\hat{\theta},\hat{\theta}_{\alpha_i})\Big\}$$
$$\widehat{NIE}_z(t_0) = N^{-1}\sum_{i=1}^{N}\Big\{g_{1-z,1}(L_i;t_0,\hat{\lambda}_0,\hat{\theta},\hat{\theta}_{\alpha_i}) - g_{1-z,0}(L_i;t_0,\hat{\lambda}_0,\hat{\theta},\hat{\theta}_{\alpha_i})\Big\}$$

where the calculation of $g_{z^*,z}$ needs the help of numerical integration. □

In HIV vaccine trials, it is reasonable to assume that the underlying biomarker level is zero if the participants in the placebo group are healthy and have no prior exposure to the virus. That is to say $\mathbb{P}(X_\alpha=0|Z=0)=1,a.s$, or $\mathbb{P}(X_\alpha^0=0)=1,a.s$. In this case with constant biomarker in the placebo group, we could consider the parameter of natural direct effect among the untreated proposed by [Lendle et al., 2013]. We consider the natural direct effect among the untreated

$$NDU(t_0) \quad = \quad \mathbb{E}\Big\{\Big[I(T^{1X_\alpha^0}\geq t_0) - I(T^{0X_\alpha^0}\geq t_0)\Big]\,|Z=0\Big\} \tag{2.8}$$

One good property of such parameter known from [Lendle et al., 2013] is that under complete randomization assumption, $Z \perp (X_\alpha^z, T^{zx_\alpha}, L)$, and conditions I1,I4, we have $NDE_0(t_0)=NDU(t_0)$, i.e. the natural direct effect among the placebo group equals the total natural direct effect of placebo.

Theorem 2.3.2. *Under Assumptions I1 - I3, $NDU(t_0)$ is identifiable from observed data.*

Proof. Similar as in the proof of Theorem 2.3.1, we have

$$\begin{aligned}
& \mathbb{E}[I(T^{z^* X_\alpha^0} \geq t_0)|Z=0] \\
= \; & \mathbb{E}\{\mathbb{E}\{\mathbb{E}[I(T^{z^* X_\alpha^0} \geq t_0)|Z=0, X_\alpha^0, L]Z=0, L\}|Z=0\} \\
= \; & \mathbb{E}\{\int_{x_\alpha} \mathbb{E}[I(T^{z^* X_\alpha^0} \geq t_0)|Z=0, X_\alpha^0 = x_\alpha, L] d\mathbb{P}_{X_\alpha^0}(x_\alpha|Z=0, L)|Z=0\} \\
= \; & \mathbb{E}\{\int_{x_\alpha} \mathbb{E}[I(T^{z^* x_\alpha} \geq t_0)|Z=0, X_\alpha = x_\alpha, L] d\mathbb{P}_{X_\alpha^0}(x_\alpha|L)|Z=0\} \quad \text{(Assumption I1,I3)} \\
= \; & \mathbb{E}\{\int_{x_\alpha} \mathbb{E}[I(T^{z^* x_\alpha} \geq t_0)|L] d\mathbb{P}_{X_\alpha^0}(x_\alpha|L)|Z=0\} \quad \text{(Assumption I1,I2)} \\
= \; & \mathbb{E}\{\int_{x_\alpha} S_T(t_0|Z=z^*, X_\alpha = x_\alpha, L) p_{X_\alpha}(x_\alpha|Z=0, L) dx_\alpha|Z=0\} \\
= \; & \mathbb{E}\left[g_{z^*,0}(L; t_0, \lambda_0, \theta, \theta_\alpha)|Z=0\right]
\end{aligned}$$

□

Similarly $NDU(t_0)$ can be estimated by

$$\widehat{NDU}(t_0) = \sum_{i=1}^{N} I(Z_i=0) \left\{ g_{1,0}(L_i; t_0, \hat{\lambda}_0, \hat{\theta}, \hat{\theta}_{\alpha_i}) - g_{0,0}(L_i; t_0, \hat{\lambda}_0, \hat{\theta}, \hat{\theta}_{\alpha_i}) \right\} / \sum_{i=1}^{N} I(Z_i=0)$$

Under Assumption I5 additionally, $g_{z^*,0}$ reduces to

$$g_{z^*,0}(L; t_0, \lambda_0, \eta) = \exp\{-\Lambda_0(t_0) \exp\{z^* \eta_Z + L^T \eta_L\}\}$$

which can be directly estimated by fitting model (2.3).

To evaluate the confidence interval of these quantities, since it is hard to obtain the analytical form of the standard errors, we suggest using the bootstrap method. Note for the time-dependent immune response process, we are only interested in the immune response level measured before the event. We know that if a subject becomes infected or develops the disease, the pattern of his/her immune response level could alter dramatically. Therefore even though in our setting the trajectory is fully determined by the time-independent random effects, it is crucial to clearly define the natural direct and indirect effects in terms of the "random effects" that quantify the trajectory before the occurrence of an event. It is

always a complication in dealing with the time-dependent variable and event time process simultaneously. Here we propose the initial work with the definitions of estimands and procedure for estimation. More detailed work is needed along this direction.

2.4 Methods of IPW and AIPW

In this section, we discuss the general models and properties for IPW and AIPW estimators. In next section, we will apply them to make inference for the joint model (2.3).

Suppose the parameter of interest θ can be estimated by solving the estimating equation $M_F(\theta) = \sum_{i=1}^{N} M_i(\theta) = 0$ when there is no missing data. In following sections, we will derive specific forms of estimating equations for the conditional score methods under two-phase sampling. Here the notation $M_i(\theta)$ is used to denote for a general estimating equations satisfying regularity conditions. We use it here to review and describe some general properties for IPW and AIPW estimators.. Let θ_0 be the true parameter with $\mathbb{E}_0[M(\theta_0)] = 0$, where $\mathbb{E}_0[\cdot]$ denotes the expectation evaluated under the truth. We also use $\dot{f}_\theta = \partial f/\partial\theta$ to denote the derivative of function f with respect to the parameter θ. We sometimes omit θ in the subscript and use $\dot{f}$ when f is fully parameterized by θ.

Under the situation of two-phase sampling, we first consider a class of estimating equations defined by

$$M_\pi = \left\{ h : M_h(\theta, \pi, h) = \sum_{i=1}^{N} \frac{\xi_i}{\pi(O_i)} M_i(\theta) + \sum_{i=1}^{N} \left(1 - \frac{\xi_i}{\pi(O_i)}\right) h(\tilde{O}_i; \theta) = 0 \right\} \tag{2.9}$$

where $h(\tilde{O}; \theta)$ is a function of $\tilde{O}$, and $\tilde{O}$ is a union of the sampling variables O and possibly other predictor variables from $\{V, \Delta, Z, L, A\}$. We assume the sampling probabilities involved in this class of estimating equations are fully and correctly specified. This is a reasonable assumption because the sampling are usually conducted by design. Note when $h \equiv 0$, $M_h(\theta, \pi, h) = 0$ leads to the IPW estimator, and $h = \mathbb{E}[M(\theta)|\tilde{O}]$ leads to the AIPW estimator proposed by [Robins et al., 1994].

Similarly to the Proposition 2.2 in [Robins et al., 1994], under some regularity conditions

we can show that $\hat{\theta}_h(\pi) \xrightarrow{p} \theta_0$ as $N \to \infty$ where $\hat{\theta}_h(\pi)$ solves $M_h(\theta, \pi, h) = 0$ and

$$N^{1/2}\left(\hat{\theta}_h(\pi) - \theta_0\right) = -\left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} M_{h,i}(\theta_0, \pi, h) + o_p(1)$$

with

$$M_{h,i}(\theta_0, \pi, h) = \frac{\xi_i}{\pi(O_i)} M_i(\theta_0) + \left(1 - \frac{\xi_i}{\pi(O_i)}\right) h(\tilde{O}_i; \theta_0)$$

By the Law of Total Variance $\mathbb{V}ar(X) = \mathbb{E}[\mathbb{V}ar(X|Y)] + \mathbb{V}ar[\mathbb{E}(X|Y)]$, the covariance matrix for $M_{h,i}(\theta_0, \pi, h)$ is

$$\begin{aligned}
&\mathbb{E}_0\left[\frac{1}{\pi} M M^T\right] - \mathbb{E}_0\left[\frac{1-\pi}{\pi} h_0 h_0^T\right] \\
&+\mathbb{E}_0\left\{\frac{1-\pi}{\pi}\left[h - h_0\right]\left[h - h_0\right]\right]^T\right\} + \mathbb{E}_0\left\{\frac{1-\pi}{\pi}\left[h_0 h^T - h h_0^T\right]\right\}
\end{aligned}$$

where for notational simplicity, we let $h_0(\tilde{O}; \theta_0) = \mathbb{E}_0[M(\theta_0)|\tilde{O}]$. Apparently the variance is minimized with $h = \mathbb{E}[M(\theta)|\tilde{O}]$. The result also implies that, for the IPW estimator $\hat{\theta}_{IPW}(\pi)$ which solves $M_h(\theta, \pi, 0) = 0$ with $h \equiv 0$, the asymptotic variance of $N^{1/2}\left(\hat{\theta}_{IPW}(\pi) - \theta_0\right)$ is

$$\Sigma_{IPW}(\pi) = \left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1} \mathbb{E}_0\left[\frac{1}{\pi} M M^T\right] \left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1}$$

And for the AIPW estimator $\hat{\theta}_{AUG}(\pi)$ which solves $M_h(\theta, \pi, \mathbb{E}) = 0$ with $h = \mathbb{E}[M(\theta)|\tilde{O}]$, the asymptotic variance of $N^{1/2}\left(\hat{\theta}_{AUG}(\pi) - \theta_0\right)$ is

$$\begin{aligned}
\Sigma_{AUG}(\pi) &= \left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1} \left\{\mathbb{E}_0\left[\frac{1}{\pi} M M^T\right] - \mathbb{E}_0\left[\frac{1-\pi}{\pi} h_0 h_0^T\right]\right\} \left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1} \\
&= \Sigma_{IPW}(\pi) - \left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1} \mathbb{E}_0\left[\frac{1-\pi}{\pi} h_0 h_0^T\right] \left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1}
\end{aligned}$$

It shows that that the IPW estimator is inefficient in this class of estimates. It can be improved by using $h = \mathbb{E}[M(\theta)|\tilde{O}]$. Or, there is an alternative way to use the estimated sampling probabilities [Breslow and Wellner, 2007]. Naturally we also consider another set

of estimating equations defined by

$$M_{\hat{\pi}} = \left\{ h : M_h(\theta, \hat{\pi}, h) = \sum_{i=1}^{N} \frac{\xi_i}{\hat{\pi}(O_i)} M_i(\theta) + \sum_{i=1}^{N} \left(1 - \frac{\xi_i}{\hat{\pi}(O_i)}\right) h(\tilde{O}_i; \theta) = 0 \right\} \tag{2.10}$$

where $\hat{\pi}(O_i) = \pi(O_i; \hat{\rho})$ where $\hat{\rho}$ maximizes the likelihood based on the correctly specified probability model (2.4), i.e.

$$\hat{\rho} = \arg\max_{\rho} \prod_{i=1}^{N} \pi(O_i; \rho)^{\xi_i} \left(1 - \pi(O_i; \rho)\right)^{1-\xi_i} \tag{2.11}$$

or solves the score equations

$$\begin{aligned} S_{\pi,F}(\rho) = \sum_{i=1}^{N} S_{\pi,i}(\rho) &= \sum_{i=1}^{N} \frac{\partial}{\partial \rho} \log \left\{ \pi(O_i; \rho)^{\xi_i} \left(1 - \pi(O_i; \rho)\right)^{1-\xi_i} \right\} \\ &= \sum_{i=1}^{N} \frac{\xi_i - \pi(O_i; \rho)}{\pi(O_i; \rho)(1 - \pi(O_i; \rho))} \frac{\partial \pi(O_i; \rho)}{\partial \rho} = 0 \end{aligned} \tag{2.12}$$

Suppose ρ_0 are the true parameters for the sampling probability model. Let $\hat{\theta}_h(\hat{\pi})$ denote the solution to $M_h(\theta_0, \hat{\pi}, h) = 0$. Still, under regularity conditions $\hat{\theta}_h(\hat{\pi}) \xrightarrow{p} \theta_0$ as $N \to \infty$ and

$$N^{1/2} \left(\hat{\theta}_h(\hat{\pi}) - \theta_0 \right) = - \left\{ \mathbb{E}_0 \left[\dot{M} \right] \right\}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \phi_{h,i}(\theta_0, \pi, h) + o_p(1)$$

with

$$\phi_{h,i}(\theta_0, \pi, h) = M_{h,i}(\theta_0, \pi, h) + \mathbb{E}_0 \left[(M - h) \frac{\dot{\pi}}{\pi} \right] \left\{ \mathbb{E}_0 \left[\dot{S}_\pi \right] \right\}^{-1} S_{\pi,i}(\rho_0)$$

The covariance matrix of $\phi_{h,i}(\theta_0, \pi, h)$ is

$$\begin{aligned} & \mathbb{E}_0 \left[M_h M_h^T \right] + \mathbb{E}_0 \left[(M - h) \frac{\dot{\pi}}{\pi} \right] \left\{ \mathbb{E}_0 \left[\dot{S}_\pi \right] \right\}^{-1} \mathbb{E}_0 \left[(M - h) \frac{\dot{\pi}}{\pi} \right]^T \\ = \; & \mathbb{E}_0 \left[\frac{1}{\pi} M M^T \right] - \mathbb{E}_0 \left[\frac{1 - \pi}{\pi} h_0 h_0^T \right] \end{aligned}$$

$$+\mathbb{E}_0\left[\frac{1-\pi}{\pi}\left[h-h_0\right]\left[h-h_0\right]^T\right]+\mathbb{E}_0\left[\frac{1-\pi}{\pi}\left[h_0h^T-hh_0^T\right]\right]$$
$$-\mathbb{E}_0\left[\left[\mathbb{E}_0[M|\tilde{O}]-h\right]\frac{\dot{\pi}}{\pi}\right]\left\{\mathbb{E}_0\left[\frac{\dot{\pi}\dot{\pi}^T}{\pi(1-\pi)}\right]\right\}^{-1}\mathbb{E}_0\left[\left[h_0-h\right]\frac{\dot{\pi}}{\pi}\right]^T$$

It is not straightforward to seek the function h at which the minimal variance is achieved, because it is hard to compare the third to the fifth term without further information about the sampling probability model. However, we can still have the asymptotic variance for the IPW estimator $\hat{\theta}_{IPW}(\hat{\pi})$ which solves $M_h(\theta,\hat{\pi},0)=0$ with $h\equiv 0$, and for the AIPW estimator $\hat{\theta}_{AUG}(\hat{\pi})$ which solves $M_h(\theta,\hat{\pi},\mathbb{E})=0$ with $h=\mathbb{E}[M(\theta)|\tilde{O}]$. The asymptotic variance of $N^{1/2}\left(\hat{\theta}_{IPW}(\hat{\pi})-\theta_0\right)$ is

$$\begin{aligned}\Sigma_{IPW}(\hat{\pi}) &= \Sigma_{IPW}(\pi)\\ &\quad -\left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1}\left\{\mathbb{E}_0\left[M\frac{\dot{\pi}}{\pi}\right]\left\{\mathbb{E}_0\left[\frac{\dot{\pi}\dot{\pi}^T}{\pi(1-\pi)}\right]\right\}^{-1}\mathbb{E}_0\left[M\frac{\dot{\pi}}{\pi}\right]^T\right\}\left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1}\end{aligned}$$

And for $N^{1/2}\left(\hat{\theta}_{AUG}(\hat{\pi})-\theta_0\right)$ is

$$\Sigma_{AUG}(\hat{\pi}) = \Sigma_{IPW}(\pi)-\left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1}\mathbb{E}_0\left[\frac{1-\pi}{\pi}h_0h_0^T\right]\left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1}$$

So far, we can tell that $\hat{\theta}_{IPW}(\pi)$ is the least efficient among the four estimators considered above, and $\hat{\theta}_{AUG}(\hat{\pi})$ and $\hat{\theta}_{AUG}(\pi)$ are asymptotically equal.

Now we consider the special case with stratified Bernoulli sampling where the probabilities are given by (2.5), the scores for subject i are simplified as

$$S_{\pi,i}(\rho)=(S_{\pi_1,i}(\rho),\cdots,S_{\pi_S,i}(\rho))^T \tag{2.13}$$
$$S_{\pi_s,i}(\rho)=I(O_i\in\mathcal{O}_s)\frac{\xi_i-\rho_s}{\rho_s(1-\rho_s)} \tag{2.14}$$

Then it can be verified that

$$\mathbb{E}_0[\dot{S}_\pi] \equiv \left(\mathbb{E}_0\left[\frac{\partial S_{\pi_k}}{\partial\rho_l}\right]\right)_{k,l}=\text{Diag}\left\{-\frac{\nu_1}{\rho_{01}(1-\rho_{01})},\cdots,-\frac{\nu_S}{\rho_{0S}(1-\rho_{0S})}\right\} \tag{2.15}$$

$$\mathbb{E}_0[M\frac{\dot{\pi}}{\pi}] = \left(\frac{\nu_1}{\rho_{01}}\mathbb{E}_{0|1}[M],\cdots,\frac{\nu_S}{\rho_{0S}}\mathbb{E}_{0|S}[M]\right) \tag{2.16}$$

where $\mathbb{E}_{0|s}[\cdot]=\mathbb{E}_0[\cdot|O\in\mathcal{O}_s]$ is the expectation evaluated given the membership in stratum $\mathcal{O}_s$. Now we have

$$\begin{aligned}
\Sigma_{IPW}(\pi) &= \left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1}\mathbb{E}_0\left[\frac{1}{\pi}MM^T\right]\left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1}\\
\Sigma_{AUG}(\pi) &= \Sigma_{IPW}(\pi)-\left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1}\left\{\sum_{s=1}^{S}\frac{1-\rho_{0s}}{\rho_{0s}}\nu_s\mathbb{E}_{0|s}\left[h_0h_0^T\right]\right\}\left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1}\\
\Sigma_{IPW}(\hat{\pi}) &= \Sigma_{IPW}(\pi)-\left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1}\left\{\sum_{s=1}^{S}\frac{1-\rho_{0s}}{\rho_{0s}}\nu_s\mathbb{E}_{0|s}[M]\mathbb{E}_{0|s}[M]^T\right\}\left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1}\\
&= \Sigma_{IPW}(\pi)-\left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1}\left\{\sum_{s=1}^{S}\frac{1-\rho_{0s}}{\rho_{0s}}\nu_s\mathbb{E}_{0|s}\left[h_0\right]\mathbb{E}_{0|s}\left[h_0\right]^T\right\}\left\{\mathbb{E}_0\left[\dot{M}\right]\right\}^{-1}\\
\Sigma_{AUG}(\hat{\pi}) &= \Sigma_{AUG}(\pi)
\end{aligned}$$

On the other hand, for the class of estimates given by $M_{\hat{\pi}}$, the covariance of their influence function $\phi_{h,i}(\theta_0,\pi,h)$ is

$$\begin{aligned}
&\mathbb{E}_0\left[\frac{1}{\pi}MM^T\right]-\mathbb{E}_0\left[\frac{1-\pi}{\pi}h_0h_0^T\right]+\mathbb{E}_0\left\{\frac{1-\pi}{\pi}\left[h_0h^T-hh_0^T\right]\right\}\\
&+\sum_{s=1}^{S}\frac{1-\rho_{0s}}{\rho_{0s}}\nu_s\mathbb{E}_{0|s}\left[[h-h_0][h-h_0]^T\right]\\
&-\sum_{s=1}^{S}\frac{1-\rho_{0s}}{\rho_{0s}}\nu_s\mathbb{E}_{0|s}\left[h-h_0\right]\mathbb{E}_{0|s}\left[h-h_0\right]^T
\end{aligned}$$

The last two terms are actually $\sum_{s=1}^{S}\frac{1-\rho_{0s}}{\rho_{0s}}\nu_s\mathbb{V}ar_{0|s}\left[h-h_0\right]$.

Since $\nu_s>0$ and $0<\rho_{0s}<1$, in order to minimize the variance above, we need to find a function h that satisfies $\mathbb{V}ar_{0|s}\left[h-h_0\right]=0$ for all $s=1,\cdots,S$. This implies that $h=\mathbb{E}[M|\tilde{O}]$, a.s.. Therefore, under stratified Bernoulli sampling, $\Sigma_{AUG}(\pi)$ and $\Sigma_{AUG}(\hat{\pi})$ achieve the minimal variance within the class of estimates yielded by M_π and $M_{\hat{\pi}}$ respectively and they are asymptotically equivalent. Therefore, if the correct model $\mathbb{E}[M|\tilde{O}]$ is available, or a set of estimating equations equivalent to $M_h(\theta,\pi,\mathbb{E})$ can be found, the resulting estimates are efficient. However, it is unusual to specify a correct model for $\mathbb{E}[M|\tilde{O}]$, thus resulting in

even less efficient estimators than IPW. IPW estimators are relatively easy to implement, but could give unstable estimate if some sampling probabilities are outliers [Kang and Schafer, 2007]. [Cao et al., 2009] discussed this issue based on a simple mean model to find the optimal γ_{opt} for a known function $h(\tilde{O};\theta,\gamma)$, and proposed a relatively stable method even if the sampling probabilities are close to zero. [Han, 2012] also provided a way to search for a better variance given one estimated function $h(\tilde{O};\theta,\hat{\gamma})$. [Qi et al., 2005] applied the non-parametric local constant regression to estimate $\mathbb{E}[M|\tilde{O}]$ on which the optimal efficiency is obtained. Their method was developed for the Cox regression model with time-independent covariate X, so it only needs to estimate $\mathbb{E}[f(X)|\tilde{O}]$ once regardless of the time. However in this dissertation, it is the model with time-dependent covariate and unknown random effects, so the augmentation term in the form of $\mathbb{E}[\cdot|\tilde{O}]$ needs to be taken care of over time. In following sections of this chapter, we will develop the IPW and AIPW estimators for the joint model (2.3). We first review and generalize the conditional score method with interaction term in the Cox regression model.

2.5 Conditional score estimator

The conditional score method was developed by [Tsiatis and Davidian, 2001] and then was generalized for multiple time-dependent biomarkers by [Song et al., 2002]. This estimator does not require specific distributional assumption for the random effects α other than Normal measurement errors. The derivations are parallel to that in [Tsiatis and Davidian, 2001], so we do not put too much details here and only outline the key steps in constructing the estimating equations.

Let $\theta = (\beta, \eta^T, \gamma)^T$ be the regression coefficients in (2.3). Define the event process as $N(u) = I(V \leq u, \Delta = 1, J(u) \geq q)$ and the at risk process as $Y(u) = I(V \geq u, J(u) \geq q)$, where $J(u) \geq q$ indicates that at least q measurements have been observed up to and including time u. We also define the design matrix, the vector of observed longitudinal measurements and the vector of measurement errors for each subject up to and including time u as

$$\widetilde{F}(u) = \begin{pmatrix} f^T(T_1^m) \\ \vdots \\ f^T(T_{J(u)}^m) \end{pmatrix}, \widetilde{W}(u) = \begin{pmatrix} W_1 \\ \vdots \\ W_{J(u)} \end{pmatrix}, \widetilde{e}(u) = \begin{pmatrix} e_1 \\ \vdots \\ e_{J(u)} \end{pmatrix} \tag{2.17}$$

Then (2.2) implies that $\widetilde{W}(u) = \widetilde{F}(u)\alpha + \widetilde{e}(u)$. Let $\hat{\alpha}(u) = [\widetilde{F}^T(u)\widetilde{F}(u)]^{-1}\widetilde{F}^T(u)\widetilde{W}(u)$ be the least squares estimate for α using data up to and including time u. Conditional on $\{\alpha, \tilde{Z}, T^m(u), J(u), Y(u) = 1\}$, the least squares estimate of $X(u)$, $\hat{X}(u) = \hat{\alpha}^T(u)f(u)$ is Normal distributed as $N(X(u), d(u, \sigma^2))$, where $d(u, \sigma^2) = \sigma^2 f^T(u)\ [\widetilde{F}^T(u)\widetilde{F}(u)]^{-1}f(u)$. Similarly as in [Tsiatis and Davidian, 2001], we have the following conditional intensity for $N(u)$.

Lemma 2.5.1. *Define* $Q(u, \theta, \sigma^2) = \hat{\alpha}^T(u)f(u) + dN(u)d(u, \sigma^2)(\beta + \gamma Z)$ *if* $Y(u) = 1$. *Assuming the conditional independency* $T \perp (C, T^m, J)$ *given* α *and* $\tilde{Z}$. *Then conditioning on* $\{Q(u, \theta, \sigma^2), \tilde{Z}, T^m(u), J(u), Y(u) = 1\}$ *the intensity process for* $dN(u)$ *is*

$$\begin{aligned} & \lim_{du \to 0} \frac{1}{du} \mathbb{P}(dN(u) = 1 | Q(u, \theta, \sigma^2), \tilde{Z}, T^m(u), J(u), Y(u) = 1) \\ = \ & \lambda_0(u) \exp\left\{ \beta Q(u, \theta, \sigma^2) + \eta^T \tilde{Z} + \gamma Z Q(u, \theta, \sigma^2) - \frac{1}{2}(\beta + \gamma Z)^2 d(u, \sigma^2) \right\} \end{aligned} \tag{2.18}$$

which does not depend on the unknown random effects α.

Proof. Let $\mathcal{C} = \{\tilde{Z}, T^m(u), J(u), Y(u) = 1\}$. At any time u, like in [Tsiatis and Davidian, 2001], under $T \perp (C, T^m, J)|(\alpha, \tilde{Z})$ we have

$$\begin{aligned} & \mathbb{P}(dN(u) = r, \hat{X}(u) = x | \alpha, \mathcal{C}) \\ = \ & \mathbb{P}(dN(u) = r | \hat{X}(u) = x, \alpha, \mathcal{C}) \mathbb{P}(\hat{X}(u) = x | \alpha, \mathcal{C}) \\ = \ & \Big[\lambda_0(u) du \exp\{\beta X(u) + \eta^T \tilde{Z} + \gamma X(u) Z\} \Big]^r \\ & \Big[1 - \lambda_0(u) du \exp\{\beta X(u) + \eta^T \tilde{Z} + \gamma X(u) Z\} \Big]^{1-r} \frac{1}{\sqrt{2\pi d(u, \sigma^2)}} \exp\{ - \frac{(x - X(u))^2}{2d(u, \sigma^2)} \} \\ \propto \ & \Big[1 - \lambda_0(u) du(u) \exp\{\beta X(u) + \eta^T \tilde{Z} + \gamma X(u) Z\} \Big]^{1-r} \exp\{ \frac{X(u)}{d(u, \sigma^2)} \\ & \big[x + d(u, \sigma^2)(\beta + \gamma Z) r \big] \} \end{aligned}$$

The conditional intensity process is therefore derived as follows.

$$
\begin{aligned}
& \mathbb{P}(dN(u)=1, Q(u,\theta,\sigma^2)=q|\mathcal{C}) \\
= & \int \mathbb{P}(dN(u)=1, Q(u,\theta,\sigma^2)=q|\alpha,\mathcal{C})p(\alpha|\mathcal{C})d\alpha \\
= & \int \mathbb{P}(dN(u)=1, \hat{X}(u)=q-d(u,\sigma^2)(\beta+\gamma Z)|\alpha,\mathcal{C})p(\alpha|\mathcal{C})d\alpha \\
= & \frac{1}{\sqrt{2\pi d(u,\sigma^2)}}\lambda_0(u)du\exp\{\eta^T\tilde{Z}-\frac{q^2}{2d(u,\sigma^2)}+(\beta+\gamma Z)q-\frac{1}{2}d(u,\sigma^2)(\beta+\gamma Z)^2\} \\
& \int \exp\{\frac{2qX(u)-X^2(u)}{2d(u,\sigma^2)}\}p(\alpha|\mathcal{C})d\alpha \\
= & o_p(1)
\end{aligned}
$$

as $du \to 0$.

$$
\begin{aligned}
& \mathbb{P}(dN(u)=0, Q(u,\theta,\sigma^2)=q|\mathcal{C}) \\
= & \int \mathbb{P}(dN(u)=0, Q(u,\theta,\sigma^2)=q|\alpha,\mathcal{C})p(\alpha|\mathcal{C})d\alpha \\
= & \int \mathbb{P}(dN(u)=0, \hat{X}(u)=q|\alpha,\mathcal{C})p(\alpha|\mathcal{C})d\alpha \\
= & \frac{1}{\sqrt{2\pi d(u,\sigma^2)}}\exp\{-\frac{q^2}{2d(u,\sigma^2)}\} \\
& \int \Big\{1-\lambda_0(u)du\exp\{\beta X(u)+\eta^T\tilde{Z}+\gamma X(u)Z\}\Big\}\exp\{\frac{2qX(u)-X^2(u)}{2d(u,\sigma^2)}\}p(\alpha|\mathcal{C})d\alpha \\
= & \frac{1}{\sqrt{2\pi d(u,\sigma^2)}}\exp\{-\frac{q^2}{2d(u,\sigma^2)}\}\int \exp\{\frac{2qX(u)-X^2(u)}{2d(u,\sigma^2)}\}p(\alpha|\mathcal{C})d\alpha+o_p(1)
\end{aligned}
$$

as $du \to 0$. And

$$
\begin{aligned}
& \frac{1}{du}\mathbb{P}(dN(u)=1|Q(u,\theta,\sigma^2)=q,\mathcal{C}) \\
= & \frac{1}{du}\frac{\mathbb{P}(dN(u)=1, Q(u,\theta,\sigma^2)=q|\mathcal{C})}{\mathbb{P}(dN(u)=1, Q(u,\theta,\sigma^2)=q|\mathcal{C})+\mathbb{P}(dN(u)=0, Q(u,\theta,\sigma^2)=q|\mathcal{C})} \\
= & \frac{\lambda_0(u)\exp\{\eta^T\tilde{Z}-\frac{q^2}{2d(u,\sigma^2)}+(\beta+\gamma Z)q-\frac{1}{2}d(u,\sigma^2)(\beta+\gamma Z)^2\}\int\exp\{\frac{2qX(u)-X^2(u)}{2d(u,\sigma^2)}\}p(\alpha|\mathcal{C})d\alpha}{\exp\{-\frac{q^2}{2d(u,\sigma^2)}\}\int\exp\{\frac{2qX(u)-X^2(u)}{2d(u,\sigma^2)}\}p(\alpha|\mathcal{C})d\alpha} \\
& +o_p(1) \\
= & \lambda_0(u)\exp\{\beta q+\eta^T\tilde{Z}+\gamma qZ)-\frac{1}{2}d(u,\sigma^2)(\beta+\gamma Z)^2\}+o_p(1)
\end{aligned}
$$

as $du \to 0$. □

For notational simplicity, throughout this chapter, we let

$$H_i(u,\theta,\sigma^2) = \left[Q_i(u,\theta,\sigma^2), \tilde{Z}_i^T, Q_i(u,\theta,\sigma^2)Z_i)\right]^T$$

If the variance of measurement errors σ^2 is known, the unbiased estimating equations for θ can be therefore derived as

$$U_F(\theta,\sigma^2) = \sum_{i=1}^{N} \int_0^\tau \left\{ H_i(u,\theta,\sigma^2) - \frac{E_F^{(1)}(u,\theta,\sigma^2)}{E_F^{(0)}(u,\theta,\sigma^2)} \right\} dN_i(u) = 0 \tag{2.19}$$

where for $r = 0, 1$

$$\begin{aligned}
E_i^{(0)}(u,\theta,\sigma^2) &= \exp\{H_i(u,\theta,\sigma^2)^T\theta - \frac{1}{2}(\beta+\gamma Z_i)^2 d_i(u,\sigma^2)\} \\
E_i^{(1)}(u,\theta,\sigma^2) &= H_i(u,\theta,\sigma^2)E_i^{(0)}(u,\theta,\sigma^2) \\
E_F^{(r)}(u,\theta,\sigma^2) &= N^{-1}\sum_{i=1}^{N} Y_i(u)E_i^{(r)}(u,\theta,\sigma^2)
\end{aligned}$$

However σ^2 is usually unknown. We can estimate it as $\hat{\sigma}^2$ by solving $S_{e,F}(\sigma^2) \equiv \sum_{i=1}^N S_{e,i}(\sigma^2) = 0$ where

$$S_{e,i}(\sigma^2) \;=\; J_i I(J_i \geq q) \left\{ \left[\widetilde{W}_i(V_i) - \widetilde{F}_i(V_i)\hat{\alpha}_i(V_i)\right]^T \left[\widetilde{W}_i(V_i) - \widetilde{F}_i(V_i)\hat{\alpha}_i(V_i)\right] - \sigma^2(J_i - q) \right\} \tag{2.20}$$

And we can estimate θ by solving $U_F(\theta,\hat{\sigma}^2) = 0$. The baseline hazards are estimated by

$$d\widehat{\Lambda}_0^F(u) = \frac{\sum_i dN_i(u)/N}{E_F^{(0)}(u,\hat{\theta},\hat{\sigma}^2)} \tag{2.21}$$

2.6 IPW conditional score estimator

2.6.1 Prespecified sampling probabilities

We start with the case with correctly and fully specified sampling probabilities. We define the IPW conditional score estimator $\hat{\theta}_{IPW}(\pi)$ for θ as the solution to $U_{IPW}(\theta, \hat{\sigma}^2_{IPW}(\pi), \pi) = 0$ where

$$U_{IPW}(\theta, \hat{\sigma}^2_{IPW}(\pi), \pi) \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} \int_0^\tau \left\{ H_i(u, \theta, \sigma^2) - \frac{E^{(1)}_{IPW}(u, \theta, \hat{\sigma}^2_{IPW}(\pi), \pi)}{E^{(0)}_{IPW}(u, \theta, \hat{\sigma}^2_{IPW}(\pi), \pi)} \right\} dN_i(u) \tag{2.22}$$

$\hat{\sigma}^2_{IPW}(\pi)$ estimates σ^2 by solving $S_{e,IPW}(\sigma^2, \pi) \equiv \sum_{i=1}^{N} (\xi_i/\pi_i) S_{e,i}(\sigma^2) = 0$, and

$$E^{(r)}_{IPW}(u, \theta, \sigma^2, \pi) \;=\; \frac{1}{N} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} Y_i(u) E^{(r)}_i(u, \theta, \sigma^2), \quad r = 0, 1 \tag{2.23}$$

The baseline hazards are estimated via

$$d\widehat{\Lambda}^{IPW}_0(u) = \frac{\sum_i dN_i(u)/N}{E^{(0)}_{IPW}(u, \hat{\theta}_{IPW}(\pi), \hat{\sigma}^2(\pi), \pi)} \tag{2.24}$$

Define

$$M_i(\theta, \sigma^2) \;=\; \int_0^\tau \left\{ H_i(u, \theta, \sigma^2) - \frac{e^{(1)}(u, \theta, \sigma^2)}{e^{(0)}(u, \theta, \sigma^2)} \right\} dD_i(u) \tag{2.25}$$

$$dD_i(u) \;=\; dN_i(u) - \lambda_0(u) Y_i(u) E^{(0)}_i(u, \theta, \sigma^2) du \tag{2.26}$$

$$e^{(r)}(u, \theta, \sigma^2) \;=\; \mathbb{E}[Y(u) E^{(r)}(u, \theta, \sigma^2)], r = 0, 1 \tag{2.27}$$

By Lemma 2.5.1 and the same arguments as for (8a) in [Tsiatis and Davidian, 2001], we know if θ_0 is the true parameter of (2.3) and σ^2_0 is the true variance of measurement errors, then $\mathbb{E}[M(\theta_0, \sigma^2_0)] = 0$. In the following regularity conditions, we also assume that they are also the unique solutions. We shall show next that the estimating equations $N^{-1} U_{IPW}(\theta, \hat{\sigma}^2_{IPW}(\pi), \pi)$ are asymptotically equivalent to $N^{-1} \sum_{i=1}^{N} (\xi_i/\pi_i) M_i(\theta, \sigma^2)$. The latter is a sum of i.i.d. random variates on which the empirical theories are readily applied

under the following regularity conditions. Let $(\theta_0^T, \sigma_0^2, \rho_0^T)^T$ be the true parameters. For any parameter, for example θ, We use $\mathcal{N}(\theta_0)$ to denote the compact neighborhood of θ_0 and $\mathcal{N}(\tau, \theta_0)$ for $[0, \tau] \times \mathcal{N}(\theta_0)$. Let $\mathbb{E}_0[\cdot]$ and $\mathbb{V}ar_0[\cdot]$ denote the expectation and variance evaluated under the truth.

Assumption A:

A1. The event time T is independent of the censoring time and the measuring schedule information (C, T^m, J), given α and $\tilde{Z}$.

A2. $\Lambda_0(\tau) < \infty$, $\mathbb{P}(Y(\tau) = 1) > 0$.

A3. The parameter space for $(\theta^T, \sigma^2, \rho^T)^T$ is compact and the true values $(\theta_0^T, \sigma_0^2, \rho_0^T)^T$ lie in the interior.

A4. $\mathbb{P}(\xi = 1|O, \alpha, W, T^m, J) = \mathbb{P}(\xi = 1|O) = \pi(O; \rho) > \delta > 0$, for all ρ and some constant $\delta > 0$.

A5. $\sup_{u\in[0,\tau]} \left|H(u, \theta, \sigma^2)\right|$, $\sup_{u\in[0,\tau]} \left|\dot{H}_{\theta,\sigma^2}(u, \theta, \sigma^2)\right|$, and $\sup_{u\in[0,\tau]} \left|d(u, \sigma^2)\right|$ are bounded and $\mathbb{V}ar \int_0^\tau |dE_i^{(r)}(u, \theta, \sigma^2)| < \infty$.

A6. $\mathbb{E}[M(\theta, \sigma^2)] \neq 0$ if $(\theta^T, \sigma^2)^T \neq (\theta_0^T, \sigma_0^2)^T$.

A7. $\mathbb{V}ar_0[M]$ is finite and positive definite. $\mathbb{E}_0[\dot{M}]$ exists and is invertible.

A8. $\mathbb{E}\left[\sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)} \left|M(\theta, \sigma^2)\right|\right] < \infty$, $\mathbb{E}\left[\sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)} \left|M(\theta, \sigma^2)M(\theta, \sigma^2)^T\right|\right] < \infty$, and $\mathbb{E}\left[\sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)} \left|\dot{M}(\theta, \sigma^2)\right|\right] < \infty$.

Lemma 2.6.1. *Under conditions A1-A5, as $N \to \infty$,*

$$\sup_{(u,\theta,\sigma^2)\in\mathcal{N}(\tau,\theta_0,\sigma_0^2)} \left| \frac{E_{IPW}^{(1)}(u, \theta, \sigma^2, \pi)}{E_{IPW}^{(0)}(u, \theta, \sigma^2, \pi)} - \frac{e^{(1)}(u, \theta, \sigma^2)}{e^{(0)}(u, \theta, \sigma^2)} \right| \xrightarrow{p} 0, \quad r = 0, 1$$

Proof. By the Double Expectation Theorem $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$, we have

$$\mathbb{E}\left[E_{IPW}^{(r)}(u, \theta, \sigma^2, \pi)\right] = \mathbb{E}\left[\frac{\xi_i}{\pi_i} E_i^{(r)}(u, \theta, \sigma^2)\right] = e^{(r)}(u, \theta, \sigma^2), \quad r = 0, 1$$

Apparently $E^{(r)}(u, \theta, \sigma^2)$ is a continuous function of $H(u, \theta, \sigma^2)$ and $d(u, \sigma^2)$, and they are all continuous in $(\theta^T, \sigma^2)^T$. Therefore by condition A4-A5 we have the uniform convergence

$$\sup_{(u,\theta,\sigma^2)\in\mathcal{N}(\tau,\theta_0,\sigma_0^2)} \left| E_{IPW}^{(r)}(u, \theta, \sigma^2, \pi) - e^{(r)}(u, \theta, \sigma^2) \right| \xrightarrow{p} 0, \quad r = 0, 1$$

Also we can prove that $e^{(0)}(u,\theta,\sigma^2)$ is bounded away from zero on $\mathcal{N}(\tau,\theta_0,\sigma_0^2)$, following similar arguments as in [Fleming and Harrington, 1991] (page 305-306). Therefore the uniform convergence stated in the Lemma holds. □

Lemma 2.6.2. *Under conditions A1-A5, $N^{-1/2}U_{IPW}(\theta_0,\sigma_0^2,\pi)$ is asymptotically equivalent to a sum of i.i.d. mean zero random variates,*

$$N^{-1/2}U_{IPW}(\theta_0,\sigma_0^2,\pi) \quad = \quad N^{-1/2}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}M_i(\theta_0,\sigma_0^2)+o_p(1)$$

Proof. We rewrite $N^{-1/2}U_{IPW}(\theta,\sigma^2,\pi)$ as

$$\begin{aligned}
& N^{-1/2}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\int_0^\tau\left\{H_i^T(u,\theta,\sigma^2)-\frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)}\right\}dD_i(u)du \\
& -N^{-1/2}\int_0^\tau\left\{\frac{E_{IPW}^{(1)}(u,\theta,\sigma^2,\pi)}{E_{IPW}^{(0)}(u,\theta,\sigma^2,\pi)}-\frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)}\right\}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}dD_i(u) \\
\equiv \quad & N^{-1/2}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}M_i(\theta,\sigma^2)+N^{-1/2}U_{N2}(\theta,\sigma^2,\pi) \qquad (2.28)
\end{aligned}$$

It suffices to show that $N^{-1/2}U_{N2}(\theta_0,\sigma_0^2,\pi)=o_p(1)$. Let $dD_{0,i}(u)=\{dN_i(u)-\lambda_0(u)Y_i(u)$ $E_i^{(0)}(u,\theta_0,\sigma_0^2)du\}$ and $d\bar{D}_{0,N}(u) = N^{-1}\sum_{i=1}^{N}(\xi_i/\pi_i)dD_{0,i}(u)$. By the Double Expectation Theorem and Lemma 2.5.1 we have $\mathbb{E}\left[(\xi_i/\pi_i)dD_{0,i}(u)\right]=0$. The Proposition A.1 in [Kulich and Lin, 2004] implies that $N^{1/2}d\bar{D}_{0,N}(u)$ converges weakly in $l^\infty[0,\tau]$ to a mean-zero Gaussian process uniformly in u. Then the convergence in probability to zero of $N^{-1/2}U_{N2}(\theta_0,\sigma_0^2,\pi)$ follows from Lemma 2.6.1 and Lemma 4.2 in [Kosorok, 2008]. □

Theorem 2.6.3. *Under conditions A1-A8, as $N\to\infty$, (i) $\hat{\theta}_{IPW}(\pi)\xrightarrow{p}\theta_0$; and (ii) $\sqrt{N}\left(\hat{\theta}_{IPW}(\pi)-\theta_0\right)$ converges weakly to a Normal random variate with mean zero and covariance $A^{-1}B(A^{-1})^T$, where*

$$A=\mathbb{E}_0\left[\dot{M}_\theta\right] \quad B=\mathbb{E}_0\left[\frac{1}{\pi}RR^T\right]$$
$$R=M(\theta_0,\sigma_0^2)-\mathbb{E}_0\left[\dot{M}_{\sigma^2}\right]\left\{\mathbb{E}_0\left[\dot{S}_e\right]\right\}^{-1}S_e(\sigma_0^2)$$

Proof. (i) Consistency. The proof is similar to that in [Tsiatis and Davidian, 2001]. We first demonstrate that $N^{-1}U_{IPW}(\theta,\sigma^2,\pi) = N^{-1}\sum_{i=1}^{N}(\xi_i/\pi_i)M_i(\theta,\sigma^2) + o_p(1)$ uniformly in $(\theta^T,\sigma^2)^T \in \mathcal{N}(\theta_0,\sigma_0^2)$. Actually,

$$\begin{aligned}
& \sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)} \left| N^{-1}U_{IPW}(\theta,\sigma^2,\pi) - N^{-1}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}M_i(\theta,\sigma^2)\right| \\
= & \sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)} \left| \int_0^\tau \left\{ \frac{E_{IPW}^{(1)}(u,\theta,\sigma^2,\pi)}{E_{IPW}^{(0)}(u,\theta,\sigma^2,\pi)} - \frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)} \right\} \left\{ N^{-1}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}dD_i(u)du \right\} \right| \\
\leq & \left\{ \sup_{(u,\theta,\sigma^2)\in\mathcal{N}(\tau,\theta_0,\sigma_0^2)} \left| \frac{E_{IPW}^{(1)}(u,\theta,\sigma^2,\pi)}{E_{IPW}^{(0)}(u,\theta,\sigma^2,\pi)} - \frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)} \right| \right\} \\
& \times \left\{ \frac{1}{\delta} + \frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i} \sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)} \left\{ \int_0^\tau \lambda_0(u)Y_i(u)E_i^{(0)}(u,\theta,\sigma^2)du \right\} \right\}
\end{aligned}$$

Lemma 2.6.1 yields the convergence to zero of the first term. The second term, by the Double Expectation Theorem and Law of Large Numbers, converges to

$$\frac{1}{\delta} + \mathbb{E}\left[\sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)} \int_0^\tau \lambda_0(u)Y_i(u)E_i^{(0)}(u,\theta,\sigma^2)du \right] < \infty$$

Therefore

$$N^{-1}U_{IPW}(\theta,\sigma^2,\pi) = N^{-1}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}M_i(\theta,\sigma^2) + o_p(1)$$

uniformly in $(\theta^T,\sigma^2)^T$. On the other hand, since

$$N^{-1}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}M_i(\theta,\sigma^2) = \mathbb{E}\left[M(\theta,\sigma^2)\right] + o_p(1)$$

uniformly in $(\theta^T,\sigma^2)^T$, then for any consistent estimator $\hat{\sigma}^2 \xrightarrow{p} \sigma_0^2$,

$$\begin{aligned}
N^{-1}U_{IPW}(\theta,\hat{\sigma}^2,\pi) &= N^{-1}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}M_i(\theta,\hat{\sigma}^2) + o_p(1) \\
&= \mathbb{E}\left[M(\theta,\sigma^2)\right]_{\sigma^2=\hat{\sigma}^2} + o_p(1) \\
&= \mathbb{E}\left[M(\theta,\sigma_0^2)\right] + o_p(1)
\end{aligned}$$

uniformly in θ. Condition A6 implies the uniqueness of θ_0 as the root of $\mathbb{E}\left[M(\theta, \sigma_0^2)\right] = 0$. It follows from Theorem 5.9 in [van der Vaart, 1998] that $\hat{\theta}_{IPW}(\pi) \xrightarrow{p} \theta_0$ as $N \to \infty$.

(ii) Asymptotic normality. Consider

$$N^{-1}\widetilde{U_{IPW}}(\theta, \sigma^2, \pi) \equiv N^{-1}\begin{pmatrix} U_{IPW}(\theta, \sigma^2, \pi) \\ S_{e,IPW}(\sigma^2, \pi) \end{pmatrix} = N^{-1}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\begin{pmatrix} M_i(\theta, \sigma^2) \\ S_{e,i}(\sigma^2) \end{pmatrix} + o_p(1)$$

By Taylor expansion we have

$$N^{-1/2}\widetilde{U_{IPW}}(\theta_0, \sigma_0^2, \pi) = -N^{-1}\widetilde{\dot{U}_{IPW}}(\theta^*, \sigma^{*2}, \pi)N^{1/2}\begin{pmatrix} \hat{\theta}_{IPW}(\pi) - \theta_0 \\ \hat{\sigma}^2(\pi) - \sigma_0^2 \end{pmatrix}$$

where $(\theta^{*T}, \sigma^{*2})^T$ lies on the segment between $\left(\hat{\theta}_{IPW}(\pi)^T, \hat{\sigma}^2_{IPW}(\pi)\right)^T$ and $\left(\theta_0^T, \sigma_0^2\right)^T$. Since we can prove the uniform convergence of $N^{-1}\widetilde{\dot{U}_{IPW}}(\theta, \sigma^2, \pi)$ in the same way as that for $N^{-1}\widetilde{U_{IPW}}(\theta, \sigma^2, \pi)$, together with the consistency of the estimates, we have

$$\begin{aligned} N^{-1}\widetilde{\dot{U}_{IPW}}(\theta^*, \sigma^{*^2}, \pi) &= N^{-1}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\begin{pmatrix} \dot{M}_i(\theta^*, \sigma^{*^2}) \\ \dot{S}_{e,i}(\sigma^{*^2}) \end{pmatrix} + o_p(1) \\ &= \begin{pmatrix} \mathbb{E}_0[\dot{M}_\theta] & \mathbb{E}_0[\dot{M}_\sigma^2] \\ 0 & \mathbb{E}_0[\dot{S}_e] \end{pmatrix} + o_p(1) \end{aligned}$$

It leads to

$$N^{1/2}\begin{pmatrix} \hat{\theta}_{IPW}(\pi) - \theta_0 \\ \hat{\sigma}^2_{IPW}(\pi) - \sigma_0^2 \end{pmatrix} = -\begin{pmatrix} \mathbb{E}_0[\dot{M}_\theta] & \mathbb{E}_0[\dot{M}_{\sigma^2}] \\ 0 & \mathbb{E}_0[\dot{S}_e] \end{pmatrix}^{-1}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\begin{pmatrix} M_i(\theta_0, \sigma_0^2) \\ S_{e,i}(\sigma_0^2) \end{pmatrix}\right\} + o_p(1)$$

and further

$$\begin{aligned} & N^{1/2}\left(\hat{\theta}_{IPW}(\pi) - \theta_0\right) \\ = & -\left\{\mathbb{E}_0[\dot{M}_\theta]\right\}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\left\{M_i(\theta_0, \sigma_0^2) - \mathbb{E}_0[\dot{M}_{\sigma^2}]\left\{\mathbb{E}_0[\dot{S}_e]\right\}^{-1}S_{e,i}(\sigma_0^2)\right\} + o_p(1) \end{aligned}$$

$$= \quad -A^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}R_i + o_p(1)$$

The asymptotic variance for $\hat{\theta}_{IPW}(\pi)$ is $A^{-1}B(A^{-1})^T$. □

2.6.2 Estimated sampling probabilities

We have shown in Section 2.4 that the IPW estimator with prespecified sampling probabilities π are inefficient. Therefore it is often suggested to use the estimated $\hat{\pi} = \hat{\pi}(\hat{\rho})$ to improve efficiency. We still apply here the sampling probability model (2.4), and the resulting likelihood and score functions (2.13) discussed in Section 2.4. The IPW estimators $\hat{\sigma}^2_{IPW}(\hat{\pi})$ and $\hat{\theta}_{IPW}(\hat{\pi})$ are obtained by solving $S_{e,IPW}(\sigma^2, \hat{\pi}) = 0$ and $U_{IPW}(\theta, \hat{\sigma}^2_{IPW}(\hat{\pi}), \hat{\pi}) = 0$, respectively.

Lemma 2.6.4. *Under conditions A1-A5, as $N \to \infty$,*

$$\sup_{(u,\theta,\sigma^2)\in\mathcal{N}(\tau,\theta_0,\sigma_0^2)}\left|\frac{E^{(1)}_{IPW}(u,\theta,\sigma^2,\hat{\pi})}{E^{(0)}_{IPW}(u,\theta,\sigma^2,\hat{\pi})} - \frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)}\right| \xrightarrow{p} 0, \quad r = 0, 1$$

Proof. Actually, consider $(\xi/\pi(\rho))\, E^{(r)}(u,\theta,\sigma^2)$ as a function of $(\theta^T, \sigma^2, \rho^T)^T$. Then we can also prove its empirical mean uniformly converges to $e^{(r)}(u,\theta,\sigma^2)$ by Glivenko-Cantelli Theorem. Therefore it follows naturally that

$$\begin{aligned}
&\sup_{(u,\theta,\sigma^2)\in\mathcal{N}(\tau,\theta_0,\sigma_0^2)}\left|E^{(r)}_{IPW}(u,\theta,\sigma^2,\hat{\pi}(\hat{\rho})) - e^{(r)}(u,\theta,\sigma^2)\right| \\
\leq \quad &\sup_{\rho}\left\{\sup_{(u,\theta,\sigma^2)\in\mathcal{N}(\tau,\theta_0,\sigma_0^2)}\left|E^{(r)}_{IPW}(u,\theta,\sigma^2,\pi(\rho)) - e^{(r)}(u,\theta,\sigma^2)\right|\right\} \\
= \quad &0
\end{aligned}$$

□

Theorem 2.6.5. *(i) $\hat{\theta}_{IPW}(\hat{\pi}) \xrightarrow{p} \theta_0$; and (ii) $\sqrt{N}\left(\hat{\theta}_{IPW}(\hat{\pi}) - \theta_0\right)$ converges weakly to a Normal random variate with mean zero and covariance $A^{-1}B^*(A^{-1})^T$, where*

$$A = \mathbb{E}_0\left[\dot{M}_\theta\right] \quad B^* = B - \mathbb{E}_0[R\frac{\dot{\pi}}{\pi}]\left\{\mathbb{E}_0[S_\pi S_\pi^T]\right\}^{-1}\mathbb{E}_0[R\frac{\dot{\pi}}{\pi}]^T$$

B and R are defined in Theorem 2.6.3.

Proof. (i) Consistency. Still, consider $N^{-1}U_{IPW}(\theta,\sigma^2,\pi(\rho))$ as a function of $(\theta^T,\sigma^2,\rho^T)^T$, which can be shown as in Theorem 2.6.3 to satisfy

$$\begin{aligned} N^{-1}U_{IPW}(\theta,\sigma^2,\pi(\rho)) &= N^{-1}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i(\rho)}M_i(\theta,\sigma^2)+o_p(1) \\ &= \mathbb{E}[M(\theta,\sigma^2)]+o_p(1) \end{aligned}$$

uniformly in $(\theta^T,\sigma^2,\rho^T)^T$. Therefore it follows naturally that

$$\begin{aligned} & \sup_{(u,\theta,\sigma^2)\in\mathcal{N}(\tau,\theta_0,\sigma_0^2)}\left|N^{-1}U_{IPW}(\theta,\sigma^2,\hat{\pi}(\hat{\rho}))-\mathbb{E}[M(\theta,\sigma^2)]\right| \\ \leq & \sup_{\rho}\left\{\sup_{(u,\theta,\sigma^2)\in\mathcal{N}(\tau,\theta_0,\sigma_0^2)}\left|N^{-1}U_{IPW}(\theta,\sigma^2,\pi(\rho))-\mathbb{E}[M(\theta,\sigma^2)]\right|\right\} \\ = & 0 \end{aligned}$$

Paralell result holds also for σ^2: $N^{-1}S_{e,IPW}(\sigma^2,\hat{\pi})=\mathbb{E}\left[S_e(\sigma^2)\right]+o_p(1)$ uniformly in σ^2. Then the consistency of $\hat{\theta}_{IPW}(\hat{\pi})$ and $\hat{\sigma}^2_{IPW}(\hat{\pi})$ follows from Theorem 5.9 in [van der Vaart, 1998].

(ii) Asymptotic normality. This can be demonstrated in the same way as for $\hat{\theta}_{IPW}(\pi)$ in Theorem 2.6.3. We only outline the key steps. Consider

$$\widetilde{U_{IPW}}(\theta,\sigma^2,\pi(\rho)) \equiv \begin{pmatrix} U_{IPW}(\theta,\sigma^2,\pi(\rho)) \\ S_{e,IPW}(\sigma^2,\pi(\rho)) \\ S_{\pi,F}(\rho) \end{pmatrix}$$

By Taylor expansion, finally we have

$$N^{1/2}\begin{pmatrix} \hat{\theta}_{IPW}(\hat{\pi})-\theta_0 \\ \hat{\sigma}^2_{IPW}(\hat{\pi})-\sigma_0^2 \\ \hat{\rho}-\rho_0 \end{pmatrix} = -\begin{pmatrix} \mathbb{E}_0[\dot{M}_\theta] & \mathbb{E}_0[\dot{M}_{\sigma^2}] & -\mathbb{E}_0[M\frac{\dot{\pi}}{\pi}] \\ 0 & \mathbb{E}_0[\dot{S}_e] & -\mathbb{E}_0[S_e\frac{\dot{\pi}}{\pi}] \\ 0 & 0 & \mathbb{E}_0[\dot{S}_\pi] \end{pmatrix}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\begin{pmatrix} \frac{\xi_i}{\pi_i(\rho_0)}M_i(\theta_0,\sigma_0^2) \\ \frac{\xi_i}{\pi_i(\rho_0)}S_{e,i}(\sigma_0^2) \\ S_{\pi,i}(\rho_0) \end{pmatrix}$$

$$+o_p(1)$$

This further implies that

$$\begin{aligned} & N^{1/2}\left(\hat{\theta}_{IPW}(\hat{\pi})-\theta_0\right) \\ = \; & -A^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{R_i-\mathbb{E}_0[R\frac{\dot{\pi}}{\pi}]\left\{\mathbb{E}_0[S_\pi S_\pi^T]\right\}^{-1}S_{\pi,i}(\rho_0)\right\}+o_p(1) \end{aligned}$$

with asymptotic variance $A^{-1}B^*(A^{-1})^T$. □

2.7 AIPW conditional score estimator

The IPW estimator is easy to implement but could be unstable with sampling probabilities close to zero. Also in above sections we assume the sampling model is correctly specified, but if it is not, the IPW estimator is biased. In Section 2.4 we also show that using estimated sampling probabilities for IPW estimators can improve the efficiency, but still do not achieve the minimal variance bound, which is achieved by AIPW estimator with correct model of full data given $\tilde{O}$.

The AIPW estimator has the property of double robustness. That is as long as either the sampling probability model for $\pi(O;\rho)$ or $\mathbb{E}[M(\theta)|\tilde{O}]$ in (2.4)(2.4) is correct, then the estimating equations are unbiased for θ_0.

$$p(U(\theta)|\xi,O)=\frac{p(U(\theta),\xi|O)}{p(\xi|O)}=\frac{p(\xi|U(\theta),O)p(U(\theta)|O)}{p(\xi|O)}=p(U(\theta)|O)$$

Since in practice it is hard to derive the correct form of $\mathbb{E}[M(\theta)|\tilde{O}]$, we need to estimate it as close as possible the truth. Given the MAR assumption in (2.4), we have $\mathbb{E}[M(\theta)|\xi=1,\tilde{O}]=\mathbb{E}[M(\theta)|\tilde{O}]$. Therefore we could build a model to estimate it based on using complete data. In following sections, we first develop the AIPW conditional score estimator and its asymptotic properties assuming π and $\mathbb{E}[\cdot|\tilde{O}]$ are fully and correctly specified. Then we move on to the situations where either or both of them are estimated.

2.7.1 Prespecified sampling probabilities and $\mathbb{E}[\cdot|\tilde{O}]$

We start with the simplest case where π and $\mathbb{E}[\cdot|\tilde{O}]$ are fully and correctly specified and no unknown parameters are involved. Similar as in [Qi et al., 2005] we define the AIPW conditional score estimating functions for θ in the form of

$$\begin{aligned} & U_{AUG}(\theta, \sigma^2, \pi, \mathbb{E}) \\ = & \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} \int_0^\tau \left\{ H_i(u, \theta, \sigma^2) - \frac{E^{(1)}_{AUG}(u, \theta, \sigma^2, \pi)}{E^{(0)}_{AUG}(u, \theta, \sigma^2, \pi)} \right\} dN_i(u) \\ & + \sum_{i=1}^{N} \left(1 - \frac{\xi_i}{\pi_i}\right) \int_0^\tau \left\{ \mathbb{E}\left\{ H_i(u, \theta, \sigma^2) dN_i(u) | \tilde{O}_i \right\} - \frac{E^{(1)}_{AUG}(u, \theta, \sigma^2, \pi)}{E^{(0)}_{AUG}(u, \theta, \sigma^2, \pi)} \mathbb{E}\left[dN_i(u) | \tilde{O}_i \right] \right\} \end{aligned} \tag{2.29}$$

where for $r = 0, 1$,

$$E^{(r)}_{AUG}(u, \theta, \sigma^2, \pi) = \frac{1}{N} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} Y_i(u) E^{(r)}_i(u, \theta, \sigma^2) + \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{\xi_i}{\pi_i}\right) \mathbb{E}\left[Y_i(u) E^{(r)}_i(u, \theta, \sigma^2) | \tilde{O}_i \right]$$

Note here unlike in the partial likelihood for the classic Cox regression, we define the at risk process as $Y_i(u) = I(V_i \geq u, J_i(u) \geq q)$ which contains the incomplete data of measurement time-points. Therefore even if the event time information is included in $\tilde{O}_i$, we still need to leave it inside the expectation. The estimate $\hat{\theta}_{AUG}(\pi, \mathbb{E})$ for θ solves $U_{AUG}(\theta, \hat{\sigma}^2_{AUG}(\pi, \mathbb{E}), \pi, \mathbb{E}) = 0$, with $\hat{\sigma}^2_{AUG}(\pi, \mathbb{E})$ solves

$$S_{e,AUG}(\sigma^2, \pi, \mathbb{E}) = \sum_{i=1}^{N} \left(\frac{\xi_i}{\pi_i}\right) S_{e,i}(\sigma^2) + \sum_{i=1}^{N} \left(1 - \frac{\xi_i}{\pi_i}\right) \mathbb{E}[S_{e,i}(\sigma^2) | \tilde{O}_i] = 0 \tag{2.30}$$

Define

$$M_{AUG,i}(\theta, \sigma^2, \pi) \quad = \quad \frac{\xi_i}{\pi_i} M_i(\theta, \sigma^2) + \left(1 - \frac{\xi_i}{\pi_i}\right) \mathbb{E}\left[M_i(\theta, \sigma^2) | \tilde{O}_i \right] \tag{2.31}$$

Apparently $\mathbb{E}[M_{AUG}(\theta, \sigma^2, \pi)] = \mathbb{E}[M(\theta, \sigma^2)]$. We will demonstrate that under the following regularity conditions in additional to those in Assumption A, $N^{-1} U_{AUG}(\theta, \sigma^2, \pi, \mathbb{E})$ is asymptotically equivalent to $N^{-1} \sum_{i=1}^{N} (\xi_i / \pi_i) M_{AUG,i}(\theta, \sigma^2, \pi)$. The latter is a sum of i.i.d.

random variates on which the empirical theories are readily applied.

Assumption B

B1. The conditional expectations $\mathbb{E}[\cdot|\tilde{O}]$ involved in (2.29) have bounded variation.

B2. $\mathbb{V}ar_0[M_{AUG}]$ is finite and positive definite. $\mathbb{E}_0[\dot{M}_{AUG}]$ exists and is invertible.

B3. $\mathbb{E}\left[\sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)}\left|M_{AUG}(\theta,\sigma^2)\right|\right] < \infty$,
$\mathbb{E}\left[\sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)}\left|M_{AUG}(\theta,\sigma^2)M_{AUG}(\theta,\sigma^2)^T\right|\right] < \infty$, and
$\mathbb{E}\left[\sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)}\left|\dot{M}_{AUG}(\theta,\sigma^2)\right|\right] < \infty$.

Lemma 2.7.1. *Under conditions A and B, as $N \to \infty$,*

$$\sup_{(u,\theta,\sigma^2)\in\mathcal{N}(\tau,\theta_0,\sigma_0^2)}\left|\frac{E_{AUG}^{(1)}(u,\theta,\sigma^2,\pi)}{E_{AUG}^{(0)}(u,\theta,\sigma^2,\pi)} - \frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)}\right| \xrightarrow{p} 0$$

Proof. By Lemma 2.6.1, it is sufficient to show that for $r = 0, 1$

$$\sup_{(u,\theta,\sigma^2)\in\mathcal{N}(\tau,\theta_0,\sigma_0^2)}\left|E_{AUG}^{(r)}(u,\theta,\sigma^2,\pi) - E_{IPW}^{(r)}(u,\theta,\sigma^2,\pi)\right| \xrightarrow{p} 0$$

Actually

$$\begin{aligned}
E_{AUG}^{(r)}(u,\theta,\sigma^2,\pi) - E_{IPW}^{(r)}(u,\theta,\sigma^2,\pi) &= \frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{\xi_i}{\pi_i}\right)\mathbb{E}\left[Y_i(u)E_i^{(r)}(u,\theta,\sigma^2)|\tilde{O}_i\right] \\
&= \mathbb{E}\left[\left(1-\frac{\xi_i}{\pi_i}\right)\mathbb{E}\left[Y_i(u)E_i^{(r)}(u,\theta,\sigma^2)|\tilde{O}_i\right]\right] + o_p(1) \\
&= o_p(1)
\end{aligned}$$

uniformly in $(u, \theta^T, \sigma^2)^T$. □

Lemma 2.7.2. *Under conditions A and B, $N^{-1/2}U_{AUG}(\theta_0, \sigma_0^2, \pi, \mathbb{E})$ is asymptotically equivalent to a sum of i.i.d. mean zero random variates*

$$N^{-1/2}U_{AUG}(\theta_0,\sigma_0^2,\pi,\mathbb{E}) = N^{-1/2}\sum_{i=1}^{N}M_{AUG,i}(\theta_0,\sigma_0^2,\pi) + o_p(1)$$

Proof. We actually can replace $dN_i(u)$ in $U_{AUG}(\theta, \sigma^2, \pi, \mathbb{E})$ (2.29) with $dD_i(u) = dN_i(u) -$

$\lambda_0(u)E_i^{(0)}(u,\theta,\sigma^2)du$. Moreover, we further rewrite $N^{-1/2}U_{AUG}(\theta,\sigma^2,\pi,\mathbb{E})$ as

$$\begin{aligned}
&N^{-1/2}U_{AUG}(\theta,\sigma^2,\pi,\mathbb{E}) = N^{-1/2}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\int_0^\tau\left\{H_i^T(u,\theta,\sigma^2)-\frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)}\right\}dD_i(u)\\
&+N^{-1/2}\sum_{i=1}^{N}\left(1-\frac{\xi_i}{\pi_i}\right)\int_0^\tau\left\{\mathbb{E}\left\{H_i^T(u,\theta,\sigma^2)dD_i(u)|\tilde{O}_i\right\}-\frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)}\mathbb{E}\left[dD_i(u)|\tilde{O}_i\right]\right\}\\
&-\int_0^\tau\left\{\frac{E_{AUG}^{(1)}(u,\theta,\sigma^2,\pi)}{E_{AUG}^{(0)}(u,\theta,\sigma^2,\pi)}-\frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)}\right\}\left\{N^{-1/2}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}dD_i(u)\right\}\\
&-\int_0^\tau\left\{\frac{E_{AUG}^{(1)}(u,\theta,\sigma^2,\pi)}{E_{AUG}^{(0)}(u,\theta,\sigma^2,\pi)}-\frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)}\right\}\left\{N^{-1/2}\sum_{i=1}^{N}\left(1-\frac{\xi_i}{\pi_i}\right)\mathbb{E}\left[dD_i(u)|\tilde{O}_i\right]\right\}\\
\equiv\;& N^{-1/2}M_{AUG,i}(\theta,\sigma^2,\pi)-N^{-1/2}U_{N2}(\theta,\sigma^2,\pi)-N^{-1/2}U_{N3}(\theta,\sigma^2,\pi) \qquad (2.32)
\end{aligned}$$

We shall show that $N^{-1/2}U_{N2}(\theta_0,\sigma_0^2,\pi)=o_p(1)$ and $N^{-1/2}U_{N3}(\theta_0,\sigma_0^2,\pi)=o_p(1)$ as $N\to\infty$. Let $dD_{o,i}(u)=dN_i(u)-\lambda_0(u)Y_i(u)E_i^{(0)}(u,\theta_0,\sigma_0^2)du$. Actually by Lemma 2.5.1 and the Double Expectation Theorem,

$$\begin{aligned}
&\mathbb{E}\left[\frac{\xi_i}{\pi_i}dD_{0,i}(u)\right]=\mathbb{E}\left[dN_i(u)-\lambda_0(u)E_i^{(0)}(u,\theta_0,\sigma_0^2)du\right]=0\\
&\mathbb{E}\left[\left(1-\frac{\xi_i}{\pi_i}\right)\mathbb{E}\left[dD_{0,i}(u)|\tilde{O}_i\right]\right]=\mathbb{E}\left[0\times\mathbb{E}\left[dD_{0,i}(u)|\tilde{O}_i\right]\right]=0
\end{aligned}$$

The Proposition A.1 in [Kulich and Lin, 2004] implies that $N^{-1/2}\sum_{i=1}^{N}(\xi_i/\pi_i)dD_{0,i}(u)$ and $N^{-1/2}\sum_{i=1}^{N}(1-\xi_i/\pi_i)\mathbb{E}\left[dD_{0,i}(u)|\tilde{O}_i\right]$ converge weakly in $l^\infty[0,\tau]$ to a mean-zero Gaussian process uniformly in u. Then we have $N^{-1/2}U_{N2}(\theta_0,\sigma_0^2,\pi)=o_p(1)$ and $N^{-1/2}U_{N3}(\theta_0,\sigma_0^2,\pi)=o_p(1)$ from Lemma 2.7.1 and Lemma 4.2 in [Kosorok, 2008]. Thus

$$N^{-1/2}U_{AUG}(\theta_0,\sigma_0^2,\pi,\mathbb{E})=N^{-1/2}\sum_{i=1}^{N}M_{AUG,i}(\theta_0,\sigma_0^2,\pi)+o_p(1)$$

□

Theorem 2.7.3. *Under conditions A1-A8, B1-B4, as $N\to\infty$, (i) $\hat{\theta}_{AUG}(\pi,\mathbb{E})\xrightarrow{p}\theta_0$; and (ii) $\sqrt{N}\left(\hat{\theta}_{AUG}(\pi,\mathbb{E})-\theta_0\right)$ converges weakly to a Normal random variate with mean zero*

and covariance $A^{-1}C(A^{-1})^T$*, where*

$$A = \mathbb{E}_0[\dot{M}_\theta] \quad C = B - \mathbb{E}_0\left[\frac{1-\pi}{\pi}\mathbb{E}_0[R|\tilde{O}]\mathbb{E}_0[R|\tilde{O}]^T\right]$$

Proof. (i) Consistency. Consider the form of $U_{AUG}(\theta, \sigma^2, \pi, \mathbb{E})$ given in (2.32) in the proof of Lemma 2.7.2:

$$N^{-1}U_{AUG}(\theta, \sigma^2, \pi, \mathbb{E}) = N^{-1}\sum_{i=1}^{N} M_{AUG,i}(\theta, \sigma^2, \pi) - N^{-1}U_{N2}(\theta, \sigma^2, \pi) - N^{-1}U_{N3}(\theta, \sigma^2, \pi)$$

We shall show that $N^{-1}U_{N2}(\theta, \sigma^2, \pi) = o_p(1)$ and $N^{-1}U_{N3}(\theta, \sigma^2, \pi) = o_p(1)$ uniformly in $(\theta^T, \sigma^2)^T$. Actually the former comes from Lemma 2.7.1 and

$$\begin{aligned}
& \sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)} \left| \frac{1}{N}\sum_{i=1}^{N} \frac{\xi_i}{\pi_i} \int_0^\tau dD_i(u) \right| \\
= & \sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)} \left| \frac{1}{N}\sum_{i=1}^{N} \frac{\xi_i}{\pi_i} \int_0^\tau \left\{ dN_i(u) - \lambda_0(u)Y_i(u)E_i^{(0)}(u,\theta,\sigma^2)du \right\} \right| \\
\leq & \frac{1}{\delta} + \frac{1}{N}\sum_{i=1}^{N} \frac{\xi_i}{\pi_i} \sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)} \int_0^\tau \lambda_0(u)Y_i(u)E_i^{(0)}(u,\theta,\sigma^2)du \\
= & \frac{1}{\delta} + \mathbb{E}\left[\sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)} \int_0^\tau \lambda_0(u)Y_i(u)E_i^{(0)}(u,\theta,\sigma^2)du \right] + o_p(1) < \infty
\end{aligned}$$

Similarly $N^{-1}U_{N3}(\theta, \sigma^2, \pi) = o_p(1)$ uniformly in $(\theta^T, \sigma^2)^T$ because

$$\begin{aligned}
& \sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)} \left| \frac{1}{N}\sum_{i=1}^{N} \left(1 - \frac{\xi_i}{\pi_i}\right) \int_0^\tau \mathbb{E}\left[dD_i(u)|\tilde{O}_i \right] \right| \\
\leq & \left(1 + \frac{1}{\delta}\right) \frac{1}{N}\sum_{i=1}^{N} \sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)} \int_0^\tau \left\{ 1 + \lambda_0(u)\mathbb{E}\left[E^{(0)}(u,\theta,\sigma^2)|\tilde{O}_i \right] \right\} du \\
= & \left(1 + \frac{1}{\delta}\right) \mathbb{E}\left[\sup_{(\theta,\sigma^2)\in\mathcal{N}(\theta_0,\sigma_0^2)} \int_0^\tau \left\{ 1 + \lambda_0(u)\mathbb{E}\left[E^{(0)}(u,\theta,\sigma^2)|\tilde{O}_i \right] \right\} du \right] + o_p(1) < \infty
\end{aligned}$$

Thus $N^{-1}U_{AUG}(\theta, \sigma^2, \pi, \mathbb{E}) = N^{-1}\sum_{i=1}^{N} M_{AUG,i}(\theta, \sigma^2, \pi) + o_p(1)$ uniformly in $(\theta^T, \sigma^2)^T$.

On the other hand

$$\begin{aligned} N^{-1}\sum_{i=1}^{N} M_{AUG,i}(\theta,\sigma^2,\pi) &= \mathbb{E}\left[M_{AUG}(\theta,\sigma^2,\pi)\right] + o_p(1) \\ &= \mathbb{E}\left[M(\theta,\sigma^2)\right] + o_p(1) \end{aligned}$$

uniformly in $(\theta^T,\sigma^2)^T$. For any consistent estimator $\hat{\sigma}^2 \xrightarrow{p} \sigma_0^2$, we have $N^{-1}U_{AUG}(\theta,\hat{\sigma}^2,\pi,\mathbb{E}) = \mathbb{E}\left[M(\theta,\sigma_0^2)\right] + o_p(1)$. By assumption A6 and Theorem 5.9 in [van der Vaart, 1998] it yields the consistency that $\hat{\theta}_{AUG}(\pi,\mathbb{E}) \xrightarrow{p} \theta_0$.

(ii) Asymptotic normality. This can be proved in the same way as that in Theorem 2.6.3. We only outline the key steps. Consider

$$\widetilde{U_{AUG}}(\theta,\sigma^2,\pi) = \begin{pmatrix} U_{AUG}(\theta,\sigma^2,\pi,\mathbb{E}) \\ S_{e,AUG}(\sigma^2,\pi,\mathbb{E}) \end{pmatrix}$$

Then $\widetilde{U_{AUG}}(\hat{\theta}_{AUG}(\pi,\mathbb{E}),\hat{\sigma}^2(\pi),\pi) = 0$. By Taylor expansion, finally we have

$$\sqrt{N}\begin{pmatrix} \hat{\theta}_{AUG}(\pi,\mathbb{E}) - \theta_0 \\ \hat{\sigma}^2(\pi) - \sigma_0 \end{pmatrix} = -\begin{pmatrix} \mathbb{E}_0[\dot{M}_\theta] & \mathbb{E}_0[\dot{M}_{\sigma^2}] \\ 0 & \mathbb{E}_0[\dot{S}_e] \end{pmatrix}^{-1} \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\begin{pmatrix} M_{AUG,i}(\theta_0,\sigma_0^2,\pi) \\ S_{e,AUG,i}(\sigma_0^2) \end{pmatrix} + o_p(1)$$

Therefore

$$\begin{aligned} & \sqrt{N}\left(\hat{\theta}_{AUG}(\pi,\mathbb{E}) - \theta_0\right) \\ = & -\frac{1}{\sqrt{N}}\left\{\mathbb{E}_0[\dot{M}_\theta]\right\}^{-1}\sum_{i=1}^{N}\left\{M_{AUG,i}(\theta_0,\sigma_0^2,\pi) - \mathbb{E}_0[\dot{M}_{\sigma^2}]\left\{\mathbb{E}_0[\dot{S}_e]\right\}^{-1} S_{e,AUG,i}(\sigma_0^2)\right\} + o_p(1) \\ = & -\frac{1}{\sqrt{N}}\left\{\mathbb{E}_0[\dot{M}_\theta]\right\}^{-1}\sum_{i=1}^{N}\left\{\frac{\xi_i}{\pi_i}R_i + \left(1-\frac{\xi_i}{\pi_i}\right)\mathbb{E}_0[R_i|\tilde{O}_i]\right\} + o_p(1) \end{aligned}$$

Thus the asymptotic variance of $\hat{\theta}_{AUG}(\pi,\mathbb{E})$ is $A^{-1}C(A^{-1})^T$. □

2.7.2 Estimated sampling probabilities, prespecified $\mathbb{E}[\cdot|\tilde{O}]$

The results in this section are parallel to that in Section 2.6.2.

Lemma 2.7.4. *Under conditions A and B, as $N \to \infty$,*

$$\sup_{(u,\theta,\sigma^2)\in\mathcal{N}(\tau,\theta_0,\sigma_0^2)} \left| \frac{E^{(1)}_{AUG}(u,\theta,\sigma^2,\hat{\pi})}{E^{(0)}_{AUG}(u,\theta,\sigma^2,\hat{\pi})} - \frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)} \right| \xrightarrow{p} 0, \quad r = 0,1$$

Theorem 2.7.5. *Under conditions A and B, as $N \to \infty$, (i) $\hat{\theta}_{AUG}(\hat{\pi},\mathbb{E}) \xrightarrow{p} \theta_0$; and (ii) $\sqrt{N}\left(\hat{\theta}_{AUG}(\hat{\pi},\mathbb{E}) - \theta_0\right)$ converges weakly to a Normal random variate with mean zero and covariance $A^{-1}B(A^{-1})^T$, with A,C defined in Theorem 2.7.3.*

Proof. (i) Consistency. See the proof for Theorem 2.6.5.

(ii) Asymptotic normality. This can be proved in the same way as that in Theorem 2.6.5. We only outline the key steps. Consider

$$\widetilde{U_{AUG}}(\theta,\sigma^2,\pi(\rho),\mathbb{E}) \equiv \begin{pmatrix} U_{AUG}(\theta,\sigma^2,\pi(\rho),\mathbb{E}) \\ S_{e,AUG}(\sigma^2,\pi(\rho),\mathbb{E}) \\ S_\pi(\rho) \end{pmatrix}$$

By Taylor expansion, finally we have

$$\begin{aligned} & N^{1/2} \begin{pmatrix} \hat{\theta}_{AUG}(\hat{\pi},\mathbb{E}) - \theta_0 \\ \hat{\sigma}^2_{AUG}(\hat{\pi},\mathbb{E}) - \sigma_0^2 \\ \hat{\rho} - \rho_0 \end{pmatrix} \\ = & - \begin{pmatrix} \mathbb{E}_0[\dot{M}_\theta] & \mathbb{E}_0[\dot{M}_{\sigma^2}] & 0 \\ 0 & \mathbb{E}_0[\dot{S}_e] & 0 \\ 0 & 0 & \mathbb{E}_0[\dot{S}_\pi] \end{pmatrix}^{-1} \times \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \begin{pmatrix} M_{AUG,i}(\theta_0,\sigma_0^2,\pi(\rho_0)) \\ S_{e,AUG,i}(\sigma_0^2,\pi(\rho_0)) \\ S_{\pi,i}(\rho_0) \end{pmatrix} + o_p(1) \end{aligned}$$

This implies that $\hat{\theta}_{AUG}(\hat{\pi},\mathbb{E})$ is asymptotically equivalent to $\hat{\theta}_{AUG}(\pi,\mathbb{E})$, with asymptotic variance $A^{-1}C(A^{-1})^T$. □

Theorem 2.7.5 indicates that when the augmentation terms in the AIPW formula are fully specified by the correct conditional expectation given $\tilde{O}$, using the estimated sampling probabilies $\hat{\pi}$ to make inference does not further improve the efficiency compared to using the prespeficied probabilities.

2.7.3 Prespeficied sampling probabilities, estimated $\mathbb{E}[\cdot|\tilde{O}]$

In reality it is hard to derive the analytical form of $\mathbb{E}[\cdot|\tilde{O}]$. Commonly used way is to replace it with a function $h(\tilde{O},\gamma)$ in terms of finite parameter γ. However, when $h(\tilde{O};\gamma) \neq \mathbb{E}[\cdot|\tilde{O}]$ for any γ the resulting estimate cannot achieve the optimal asymptotic variance given in Theorem 2.7.3 and Theorem 2.7.5. [Cao et al., 2009] investigated the optimal way to find γ which can finally lead to an estiamte of θ having the minimal variance with h fixed.

[Qi et al., 2005] investigated the AIPW estimator for the Cox regression with time-independent covariates. They proposed to estimate the $\mathbb{E}[\cdot|\tilde{O}]$ by nonparametric Nadaraya-Watson method, and proved that the optimal asymptotic variance can be achieved. In our model, we are however dealing with a more complicated case with time-dependent covariate. So we need to extend their method to our setting.

The expression of estimating scores in (2.29) tells that we need to estimate $\mathbb{E}[G(u,\theta,\sigma^2)|\tilde{O}]$ as a continuous function of u and integrate it over u, where $G(u,\theta,\sigma^2)$ could be $Y(u)E^{(r)}(u,\theta,\sigma^2)$, $dN(u)$ or $H(u,\theta,\sigma^2)dN(u)$. Suppose the predictor variables $\tilde{O}$ are d continuous variables, we estimate it via the non-parametric Nadaraya-Watson estimator, i.e. for any random variate $g(u,\tilde{O};\theta,\sigma^2) = \mathbb{E}[G(u,\theta,\sigma^2)|\tilde{O}]$,

$$\hat{g}(u,\tilde{O};\theta,\sigma^2) = \hat{\mathbb{E}}[G(u,\theta,\sigma^2)|\tilde{O}] = \frac{\sum_{j=1}^{N}\xi_j G_j(u,\theta,\sigma^2)K_H(\tilde{O}-\tilde{O}_j)}{\sum_{j=1}^{N}\xi_j K_H(\tilde{O}-\tilde{O}_j)} \tag{2.33}$$

where $K_H(\cdot)$ is a $s-$th order kernel function and H is the bandwidth which is a $d\times d$ symmetric and positive definite matrix. If any component of $\tilde{O}$ is discrete, we consider the kernel regression with mixed types of predictor variables [Hall et al., 2004]. As the "bandwidth" of the kernel function for the discrete variable goes to zero, the estimator reduces to the kernel estimator with respect to the continuous components within each stratum defined by the discrete variable. The resulting estimating equations for θ is $U_{AUG}(\theta,\sigma^2,\pi,\hat{\mathbb{E}}) = 0$ where replacing $\mathbb{E}$ in (2.29) with $\hat{\mathbb{E}}$ given by (2.33).

However, intuitively such estimation relies on strong assumption on the functional form on u. By examining the two stochastic processes involving $dN(u)$ mentioned above, we find that they are always zero for time u if there is no event occurring at that time. This implies

practically, the estimates of their conditional expectations given $\tilde{O}$ at non-event time points are zero. So we only need to estimate them at event time points. Alternatively, if $\tilde{O}$ include the event time (V, Δ), then conditioning on it we have $U_{AUG}(\theta, \sigma^2, \pi, \hat{\mathbb{E}})$ reduced to

$$\begin{aligned} & U_{AUG}(\theta, \sigma^2, \pi, \hat{\mathbb{E}}) \\ = & \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} \int_0^\tau \left\{ H_i(u, \theta, \sigma^2) - \frac{\hat{E}^{(1)}_{AUG}(u, \theta, \sigma^2, \pi)}{\hat{E}^{(0)}_{AUG}(u, \theta, \sigma^2, \pi)} \right\} dN_i(u) \\ & + \sum_{i=1}^{N} \left(1 - \frac{\xi_i}{\pi_i}\right) \int_0^\tau \left\{ \hat{g}^H(u, \tilde{O}_i; \theta, \sigma^2) - \frac{\hat{E}^{(1)}_{AUG}(u, \theta, \sigma^2, \pi)}{\hat{E}^{(0)}_{AUG}(u, \theta, \sigma^2, \pi)} \hat{g}^J(u, \tilde{O}_i; \theta, \sigma^2) \right\} dN^*_i(u) \end{aligned} \tag{2.34}$$

where $Y^*_i(u) = I(V_i \geq u)$, $dN^*_i(u) = I(V_i = u, \Delta_i = 1)$ and for $r = 0, 1$

$$\begin{aligned} \hat{E}^{(r)}_{AUG}(u, \theta, \sigma^2, \pi) &= \frac{1}{N} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} Y_i(u) E^{(r)}_i(u, \theta, \sigma^2) + \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{\xi_i}{\pi_i}\right) Y^*_i(u) \hat{g}^{(r)}(u, \tilde{O}_i; \theta, \sigma^2) \\ G^J(u, \theta, \sigma^2) &= I(J(u) \geq q), \quad g^J(u, \tilde{O}; \theta, \sigma^2) \equiv \mathbb{E}[G^J(u, \theta, \sigma^2) | \tilde{O}] \\ G^H(u, \theta, \sigma^2) &= I(J(u) \geq q) H(u, \theta, \sigma^2), \quad g^H(u, \tilde{O}; \theta, \sigma^2) \equiv \mathbb{E}[G^Q(u, \theta, \sigma^2) | \tilde{O}] \\ G^{(r)}(u, \theta, \sigma^2) &= I(J(u) \geq q) E^{(r)}(u, \theta, \sigma^2), \quad g^{(r)}(u, \tilde{O}; \theta, \sigma^2) \equiv \mathbb{E}[G^{(r)}(u, \theta, \sigma^2) | \tilde{O}] \end{aligned}$$

Similarly, the variance of measurement error σ^2 can be estimated by solving $S_{e,AUG}(\sigma^2, \pi, \hat{E}) = 0$ defined in the same manner. The corresponding estimators are denoted by $\hat{\theta}_{AUG}(\pi, \hat{\mathbb{E}})$ and $\hat{\sigma}^2_{AUG}(\pi, \hat{\mathbb{E}})$. We need additional regularity conditions to validate the consistency and asymptotic normality of $\hat{\theta}_{AUG}(\pi, \hat{\mathbb{E}})$. We need additional assumptions. The boundary is uniform with respect to u, θ, σ^2. We also use $G(u, \theta, \sigma^2)$ to stand for any of $Y(u)E^{(r)}(u, \theta, \sigma^2)$, $dN(u)$ and $H(u, \theta, \sigma^2)dN(u)$ that need evaluation in the augmentation terms.

Assumption C

C1. The order of the kernel function (the first non-zero moment) is s.

C2. $Nh^{2d} \to \infty$ and $Nh^{2s} \to 0$ as $N \to \infty$.

C3. The marginal probability density function of $\tilde{O}$ and the conditional probability density function of $\tilde{O}$ given ξ are bounded away from zero. They also have s continuous and bounded partial derivatives with respect to the continuous components of $\tilde{O}$.

C4. The conditional expectation $\mathbb{E}[G(u,\theta,\sigma^2)|\tilde{O}]$ in (2.29) have s continuous and uniformly bounded partial derivatives with respect to the continuous components of $\tilde{O}$.

C5. The conditional variance of $\mathbb{V}ar[G(u,\theta,\sigma^2)|\tilde{O}]$ is uniformly bounded.

Lemma 2.7.6. *Under conditions A, B and C, as $N \to \infty$,*

$$\sup_{(u,\theta,\sigma^2)\in\mathcal{N}(\tau,\theta_0,\sigma_0^2)} \left| \frac{\hat{E}^{(1)}_{AUG}(u,\theta,\sigma^2,\pi)}{\hat{E}^{(0)}_{AUG}(u,\theta,\sigma^2,\pi)} - \frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)} \right| \xrightarrow{p} 0$$

Proof. Let $G_i \equiv G_i(u,\theta,\sigma^2) = Y_i(u)E_i^{(0)}(u,\theta,\sigma^2)$, $g_i \equiv g_i(u,\tilde{O}_i;\theta,\sigma^2) = \mathbb{E}[G_i(u,\theta,\sigma^2)|\tilde{O}_i]$, and

$$\hat{g}_i \equiv \hat{g}_i(u,\tilde{O}_i;\theta,\sigma^2) = \frac{\sum_{j=1}^N \xi_j G_j(u,\theta,\sigma^2)K_H(\tilde{O}_i-\tilde{O}_j)}{\sum_{j=1}^N \xi_j K_H(\tilde{O}_i-\tilde{O}_j)}$$

$$\hat{E}^{(0)}_{AUG}(u,\theta,\sigma^2,\pi) = E^{(0)}_{AUG}(u,\theta,\sigma^2,\pi) + \frac{1}{N}\sum_{i=1}^N \left(1-\frac{\xi_i}{\pi_i}\right)\left(\hat{g}_i^{(r)} - g_i^{(r)}\right)$$

Then by Lemma 2.7.1, we only need to prove the second term converges in probability to zero uniformly. The second term is

$$A_N \;=\; \frac{1}{N}\sum_{i=1}^N \left(1-\frac{\xi_i}{\pi_i}\right) \frac{\sum_{j=1}^N \xi_j \{G_j - g_i\} K_H(\tilde{O}_i-\tilde{O}_j)}{\sum_{j=1}^N \xi_j K_H(\tilde{O}_i-\tilde{O}_j)}$$

In Appendix I we show that $\mathbb{E}[A_N] = 0$ and $\mathbb{V}ar[A_N] = o_p(1)$ uniformly in $(u,\theta^T,\sigma^2)^T$. Therefore $A_N = o_p(1)$ uniformly in $(u,\theta^T,\sigma^2)^T$. The proof for $r=1$ is similar. □

Theorem 2.7.7. *Under conditions A, B and C, as $N\to\infty$, (i) $\hat{\theta}_{AUG}(\pi,\hat{\mathbb{E}}) \xrightarrow{p} \theta_0$; and (ii) $\sqrt{N}\left(\hat{\theta}_{AUG}(\pi,\hat{\mathbb{E}}) - \theta_0\right)$ converges weakly to a Normal random variate with mean zero and covariance $A^{-1}C(A^{-1})^T$, where A and C are defined in Theorem 2.7.3.*

Proof. Since (2.33) is a linear operator, $U_{AUG}(\theta,\sigma^2,\pi,\hat{\mathbb{E}})$ in (2.34) can be rewritten with $dN_i(u)$ replaced by $dD_i(u)$ everywhere. We can further rewrite this as

$$U_{AUG}(\theta,\sigma^2,\pi,\hat{\mathbb{E}}) = \sum_{i=1}^N \frac{\xi_i}{\pi_i}\int_0^\tau \left\{ H_i(u,\theta,\sigma^2) - \frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)} \right\} dD_i(u)$$

$$
\begin{aligned}
&+\sum_{i=1}^{N}\left(1-\frac{\xi_i}{\pi_i}\right)\int_0^{\tau}\left\{\mathbb{E}\left[H_i(u,\theta,\sigma^2)dD_i(u)|\tilde{O}_i\right]-\frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)}\mathbb{E}\left[dD_i(u)|\tilde{O}_i\right]\right\}\\
&-\int_0^{\tau}\left\{\frac{\hat{E}^{(1)}_{AUG}(u,\theta,\sigma^2,\pi)}{\hat{E}^{(0)}_{AUG}(u,\theta,\sigma^2,\pi)}-\frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)}\right\}\left\{\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}dD_i(u)\right\}\\
&+\sum_{i=1}^{N}\left(1-\frac{\xi_i}{\pi_i}\right)\int_0^{\tau}\left\{\hat{\mathbb{E}}\left[H_i(u,\theta,\sigma^2)dD_i(u)|\tilde{O}_i\right]-\mathbb{E}\left[H_i(u,\theta,\sigma^2)dD_i(u)|\tilde{O}_i\right]\right\}\\
&-\sum_{i=1}^{N}\left(1-\frac{\xi_i}{\pi_i}\right)\int_0^{\tau}\left\{\frac{\hat{E}^{(1)}_{AUG}(u,\theta,\sigma^2,\pi)}{\hat{E}^{(0)}_{AUG}(u,\theta,\sigma^2,\pi)}\hat{\mathbb{E}}\left[H_i(u,\theta,\sigma^2)dD_i(u)|\tilde{O}_i\right]-\right.\\
&\left.\frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)}\mathbb{E}\left[H_i(u,\theta,\sigma^2)dD_i(u)|\tilde{O}_i\right]\right\}\\
\equiv\;&\sum_{i=1}^{N}M_{AUG,i}(\theta,\sigma^2,\pi)-U_{N2}(\theta,\sigma^2,\pi)+U_{N3}(\theta,\sigma^2,\pi)-U_{N4}(\theta,\sigma^2,\pi)
\end{aligned}
$$

(i) Consistency. We show that $N^{-1}U_{N2}(\theta,\sigma^2,\pi)=o_p(1)$, $N^{-1}U_{N3}(\theta,\sigma^2,\pi)=o_p(1)$ and $N^{-1}U_{N4}(\theta,\sigma^2,\pi)=o_p(1)$ uniformly in $(\theta^T,\sigma^2)^T$. Actually, $N^{-1}U_{N2}(\theta,\sigma^2,\pi)$ converging in probability to zero uniformly follows from Lemma 2.7.6 and the similar arguments in Theorem 2.7.3. For $N^{-1}U_{N3}(\theta,\sigma^2,\pi)=o_p(1)$, we can show it in a spirit similar to the proof of Lemma 2.7.6. And $N^{-1}U_{N4}(\theta,\sigma^2,\pi)$ can be rewritten as

$$
\begin{aligned}
&\int_0^{\tau}\frac{\hat{E}^{(1)}_{AUG}(u,\theta,\sigma^2,\pi)}{\hat{E}^{(0)}_{AUG}(u,\theta,\sigma^2,\pi)}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{\xi_i}{\pi_i}\right)\left(\hat{\mathbb{E}}\left[H_i(u,\theta,\sigma^2)dD_i(u)|\tilde{O}_i\right]-\right.\right.\\
&\qquad\left.\left.\mathbb{E}\left[H_i(u,\theta,\sigma^2)dD_i(u)|\tilde{O}_i\right]\right)\right\}\\
&+\int_0^{\tau}\left\{\frac{\hat{E}^{(1)}_{AUG}(u,\theta,\sigma^2,\pi)}{\hat{E}^{(0)}_{AUG}(u,\theta,\sigma^2,\pi)}-\frac{e^{(1)}(u,\theta,\sigma^2)}{e^{(0)}(u,\theta,\sigma^2)}\right\}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{\xi_i}{\pi_i}\right)\mathbb{E}\left[H_i(u,\theta,\sigma^2)dD_i(u)|\tilde{O}_i\right]\right\}
\end{aligned}
$$

which also converges to zero in probability uniformly in $(\theta^T,\sigma^2)^T$. Thus $N^{-1}U_{AUG}(\theta,\sigma^2,\pi,\hat{\mathbb{E}})$ $=N^{-1}\sum_{i=1}^{N}M_{AUG,i}(\theta,\sigma^2,\pi)+o_p(1)$ uniformly in $(\theta^T,\sigma^2)^T$. This implies that $\hat{\theta}_{AUG}(\pi,\hat{\mathbb{E}})\xrightarrow{p}$ θ_0 as $N\to\infty$.

(ii) Asymptotic normality. Note $O_p(\sqrt{h^{2s}+(Nh^d)^{-2}})=o_p(1)$. Therefore by using the Lemma 1 and Lemma 2 in [Wang and Wang, 2001], we have $N^{-1/2}U_{N3}(\theta_0,\sigma_0^2,\pi)=$ $o_p(1)$. For $N^{-1/2}U_{N2}(\theta_0,\sigma_0^2,\pi)$, the Proposition A.1 in [Kulich and Lin, 2004] implies that $N^{-1/2}\sum_{i=1}^{N}(\xi_i/\pi_i)dD_i(u)$ converges weakly in $l^{\infty}[0,\tau]$ to a mean-zero Gaussian process at

θ_0. Thus together with Lemma 2.7.4 and Lemma 4.2 in [Kosorok, 2008], $N^{-1/2}U_{N2}(\theta_0, \sigma_0^2, \pi) = o_p(1)$. The convergence in probability to zero of $N^{-1/2}U_{N4}(\theta_0, \sigma_0^2, \pi)$ follows similarly. Therefore

$$N^{-1/2}U_{AUG}(\theta_0, \sigma_0^2, \pi, \hat{\mathbb{E}}) = N^{-1/2}\sum_{i=1}^{N} M_{AUG,i}(\theta_0, \sigma_0^2, \pi) + o_p(1)$$

which is also asymptotically equivalent to $N^{-1/2}U_{AUG}(\theta_0, \sigma_0^2, \pi, \mathbb{E})$. Thus the asymptotic variance of $\hat{\theta}_{AUG}(\pi, \hat{\mathbb{E}})$ is still $A^{-1}C(A^{-1})^T$. □

2.7.4 Estimated sampling probabilities, estimated $\mathbb{E}[\cdot|\tilde{O}]$

When the sampling probabilities are estimated via (2.11), parallel results hold as those in Section 2.7.2. We estimate θ and σ^2 by solving $U_{AUG}(\theta, \sigma^2, \hat{\pi}, \hat{\mathbb{E}}) = 0$ and $S_{e,AUG}(\sigma^2, \hat{\pi}, \hat{\mathbb{E}}) = 0$. The obtained estimates are denoted as $\hat{\theta}_{AUG}(\hat{\pi}, \hat{\mathbb{E}})$ and $\hat{\sigma}^2_{AUG}(\hat{\pi}, \hat{\mathbb{E}})$. We have the following Theorem.

Theorem 2.7.8. *Under conditions A, B, and C, as $N \to \infty$, (i) $\hat{\theta}_{AUG}(\hat{\pi}, \hat{\mathbb{E}}) \xrightarrow{p} \theta_0$; and (ii) $\sqrt{N}\left(\hat{\theta}_{AUG}(\hat{\pi}, \hat{\mathbb{E}}) - \theta_0\right)$ converges weakly to a Normal random variate with mean zero and covariance $A^{-1}C(A^{-1})^T$, where A and C are defined in Theorem 2.7.3.*

Proof. The proof is similar to that of Theorem 2.7.5 and Theorem 2.7.7. □

2.7.5 Appendix I

Proof. For simplicity, we only show the proof when predictor variables $\tilde{O}$ are all continuous variables and when the bandwidth matrix H is diagonal with all diagonal elements equal to h. Let

$$\hat{p}_1(\tilde{o}) = \frac{1}{Nh^d}\sum_{j=1}^{N} \xi_j K_H(\tilde{o} - \tilde{O}_j)$$

$$d_{ij} = \left(1 - \frac{\xi_i}{\pi_i}\right)\frac{\xi_j(G_j - g_i)K_H(\tilde{O}_i - \tilde{O}_j)}{p_1(\tilde{O}_i)}$$

Then by the proof in Appendix of [Wang and Wang, 2001] we have $A_N = A_{N1} + O_p(h^{2s} +$

$\frac{1}{Nh^d}) = o_p(1)$, where

$$A_{N1} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(1-\frac{\xi_i}{\pi_i}\right)\frac{\xi_j(G_j-g_i)K_H(\tilde{O}_i-\tilde{O}_j)}{h^d p_1(\tilde{O}_i)} = \frac{1}{N^2h^d}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}$$

Let All_j denote all variables observed for subject j, i.e. $All_j = \{\Delta_j, V_j, \tilde{Z}_j, \xi_j, \xi_j W_j, \xi_j T_j^m\}$. Note that $\mathbb{E}[G_i|\tilde{O}_i] = g_i$. By direct calculation

$$\mathbb{E}[d_{ij}] = \mathbb{E}\left\{\frac{\xi_j(G_j-g_i)K_H(\tilde{O}_i-\tilde{O}_j)}{p_1(\tilde{O}_i)}\mathbb{E}\left[\left(1-\frac{\xi_i}{\pi_i}\right)|All_j,\tilde{O}_i\right]\right\} = 0, \qquad i \neq j$$
$$\mathbb{E}[d_{ii}] = \mathbb{E}\left\{\mathbb{E}\left[\left(1-\frac{1}{\pi_i}\right)\pi_i\frac{(G_i-g_i)K_H(0)}{p_1(\tilde{O}_i)}|\tilde{O}_i\right]\right\} = 0$$

Therefore $\mathbb{E}[A_{1N}] = 0$. Now we look at $\mathbb{E}[A_{1N}A_{1N}^T]$. Let i, j, k, l be four distinct integers. Then

$$\begin{aligned}\mathbb{E}[A_{1N}A_{1N}^T] &= \frac{1}{N^4h^{2d}}\left\{\sum_i d_{ii}d_{ii} + 2\sum_i\sum_j d_{ii}d_{ij} + 2\sum_i\sum_j d_{ij}d_{jj} + 2\sum_i\sum_j\sum_k d_{ii}d_{jk}\right.\\
&+\sum_i\sum_j\sum_k d_{ij}d_{ik} + 2\sum_i\sum_j\sum_k d_{ij}d_{ki} + \sum_i\sum_j\sum_k d_{ij}d_{kj} + \sum_i\sum_j d_{ii}d_{jj}\\
&\left.+\sum_i\sum_j d_{ij}d_{ij} + \sum_i\sum_j d_{ij}d_{ji} + \sum_i\sum_j\sum_k\sum_l d_{ij}d_{kl}\right\} \qquad (2.35)\end{aligned}$$

Note that for any function $f(\tilde{o})$ has s continuous and bounded partial derivative with respect to the continuous components, since the order of the kernel is s, we have

$$\int K_H(z-x)f(z)dz = \int K(u)f(x+hu)du = f(x) + \frac{1}{s!}f^{(s)}(x)h^s\int K(u)u^s du + o(h^s)$$
$$\int K_H^2(z-x)f(z)dz = \frac{1}{h^d}\int K^2(u)f(x+hu)du = \frac{1}{h^d}f(x)\int K^2(u)du + \frac{1}{h^d}O(h)$$

We will use these two expressions repeatedly to show that $\mathbb{E}[A_{1N}A_{1N}^T] = o_p(1)$ uniformly in $(u, \theta^T, \sigma^2)^T$. We examine each of the sums in (2.35).

$$\begin{aligned}\frac{1}{N^4h^{2d}}\sum_i \mathbb{E}[d_{ii}d_{ii}] &= \frac{1}{N^3h^{4d}}\mathbb{E}\left[\left(1-\frac{\xi_i}{\pi_i}\right)^2\xi_i^2\frac{(G_i-g_i)^2K^2(0)}{p_1^2(\tilde{O}_i)}\right]\\ &= \frac{1}{N^3h^{4d}}\mathbb{E}\left[\frac{(1-\pi_1)^2}{\pi_1}\frac{(G_1-g_1)^2K^2(0)}{p_1^2(\tilde{O}_1)}\right]\\ &= O_p(\frac{1}{N^3h^{4d}}) = o_p(1)\end{aligned}$$

$$\begin{aligned}&\frac{1}{N^4h^{2d}}\sum_i\sum_j\sum_k \mathbb{E}[d_{ij}d_{ik}]\\ =\ &\frac{1}{N^4h^{2d}}\sum_i\sum_j\sum_k \mathbb{E}\left[\left(1-\frac{\xi_i}{\pi_i}\right)^2\xi_j\xi_k\frac{(G_j-g_i)(G_k-g_i)K_H(\tilde{O}_i-\tilde{O}_j)K_H(\tilde{O}_i-\tilde{O}_k)}{p_1^2(\tilde{O}_i)}\right]\\ =\ &\frac{1}{N^4h^{2d}}\sum_i\sum_j\sum_k \mathbb{E}\left[\frac{(1-\pi_i)\pi_j\pi_k}{\pi_i}\frac{(G_j-g_i)(G_k-g_i)K_H(\tilde{O}_i-\tilde{O}_j)K_H(\tilde{O}_i-\tilde{O}_k)}{p_1^2(\tilde{O}_i)}\right]\\ =\ &\frac{1}{N^4h^{2d}}\sum_i\sum_j\sum_k \mathbb{E}\left[\frac{(1-\pi_i)\pi_j\pi_k}{\pi_i p_1^2(\tilde{O}_i)}\left\{\mathbb{E}\left[(G_j-g_i)K_H(\tilde{O}_i-\tilde{O}_j)|\tilde{O}_i\right]\right\}^2\right]\\ =\ &\frac{(N-1)(N-2)}{N^3h^{2d}}\mathbb{E}[\pi_1]^2\mathbb{E}\left[\frac{1-\pi_1}{\pi_1 p_1^2(\tilde{O}_1)}\left\{\mathbb{E}\left[(G_2-g_1)K_H(\tilde{O}_1-\tilde{O}_2)|\tilde{O}_1\right]\right\}^2\right]\\ =\ &\frac{(N-1)(N-2)}{N^3h^{2d}}\mathbb{E}[\pi_1]^2\mathbb{E}\left[\frac{1-\pi_1}{\pi_1 p_1^2(\tilde{O}_1)}\left\{\mathbb{E}\left[(g_2-g_1)K_H(\tilde{O}_1-\tilde{O}_2)|\tilde{O}_1\right]\right\}^2\right]\\ =\ &\frac{1}{Nh^{2d}}O_p(h^{2s}) = o_p(1)\end{aligned}$$

since $\mathbb{E}\left[(g_2-g_1)K_H(\tilde{O}_1-\tilde{O}_2)|\tilde{O}_1\right] = O_p(h^s)$.

$$\begin{aligned}\frac{1}{N^4h^{2d}}\sum_i\sum_j \mathbb{E}[d_{ij}d_{ij}] &= \frac{1}{N^4h^{2d}}\sum_i\sum_j \mathbb{E}\left[\left(1-\frac{\xi_i}{\pi_i}\right)^2\xi_j^2\frac{(G_j-g_i)^2K_H^2(\tilde{O}_i-\tilde{O}_j)}{p_1^2(\tilde{O}_i)}\right]\\ &= \frac{1}{N^4h^{2d}}\sum_i\sum_j \mathbb{E}\left[\frac{(1-\pi_i)\pi_j}{\pi_i}\frac{(G_j-g_i)^2K_H^2(\tilde{O}_i-\tilde{O}_j)}{p_1^2(\tilde{O}_i)}\right]\\ &= \frac{1}{N^4h^{2d}}\sum_i\sum_j \mathbb{E}\left\{\mathbb{E}\left[\frac{(1-\pi_i)\pi_j}{\pi_i}\frac{(G_j-g_i)^2K_H^2(\tilde{O}_i-\tilde{O}_j)}{p_1^2(\tilde{O}_i)}|All_j\right]\right\}\end{aligned}$$

$$= \frac{1}{N^2h^{3d}}(O_p(1)+O_p(h)) = o_p(1)$$

$$\begin{aligned}
& \frac{1}{N^4h^{2d}}\sum_i\sum_j \mathbb{E}[d_{ij}d_{ji}] \\
= & \frac{1}{N^4h^{2d}}\sum_i\sum_j \mathbb{E}\left[\left(1-\frac{\xi_i}{\pi_i}\right)\left(1-\frac{\xi_j}{\pi_j}\right)\xi_i\xi_j\frac{(G_j-g_i)(G_i-g_j)K_H^2(\tilde{O}_i-\tilde{O}_j)}{p_1(\tilde{O}_i)p_1(\tilde{O}_j)}\right] \\
= & -\frac{1}{N^4h^{2d}}\sum_i\sum_j \mathbb{E}\left[(1-\pi_i)(1-\pi_j)\frac{(g_j-g_i)^2K_H^2(\tilde{O}_i-\tilde{O}_j)}{p_1(\tilde{O}_i)p_1(\tilde{O}_j)}\right] \\
= & \frac{1}{N^4h^{2d}}\sum_i\sum_j \mathbb{E}\left\{\mathbb{E}\left[(1-\pi_i)(1-\pi_j)\frac{(g_j-g_i)^2K_H^2(\tilde{O}_i-\tilde{O}_j)}{p_1(\tilde{O}_i)p_1(\tilde{O}_j)}|\tilde{O}_j\right]\right\} \\
= & \frac{1}{N^2h^{3d}}(O_p(1)+O_p(h)) = o_p(1)
\end{aligned}$$

The rest terms in (2.35) are zero by noting that $\mathbb{E}[G_i-g_i|\tilde{O}_i]=0$ and $\mathbb{E}[1-\xi_i/\pi_i]=0$. So combining all results above we have shown that $\mathbb{V}ar[A_{1N}]=o_p(1)$. Also by Assumption C the O_p and o_p above are all uniform in u,θ,σ^2. Therefore it follows the uniform convergence in probability of A_N to zero.

□

Chapter 3

SIMULATION STUDIES FOR JOINT MODELING WITH CONTINUOUS BIOMARKERS

In this chapter, we evaluate the method developed in Chapter 2 via simulation studies. The event time data (Δ, V) are generated from Cox proportional hazards models with different combinations of the time-varying biomarker and vaccination indicator. We carry out 500 simulation runs for each scenario. Our primary goal is to evaluate the IPW and AIPW conditional score methods in two-phase sampling design cohort studies. For the purpose of comparison, we also calculate the conditional score estimator based on full cohort data (Full), the unobtainable benchmark in real vaccine trail studies, and conduct the naive complete-case (CC) analysis using only subjects being selected in the second phase and without any weighting.

We conduct mainly three simulation studies. Simulation Study I studies the Cox model with only one immune biomarker as the covariate. It is aimed to compare the IPW and AIPW estimators using various sets of auxiliary variables to estimate the augmentation terms. Simulation Study II considers the Cox model with both the immune biomarker variable and the vaccination indicator as covariates. We study the impact of the number of measurements and misspecified measurement error models on the performance of the proposed methods. Simulation Study III demonstrates the performance on the Cox model with immune biomarker and vaccination interaction term. All IPW and AIPW methods are implemented using both the pre-specified true sampling probabilities (π) and the estimated sampling probabilities ($\hat{\pi}$). In all simulation studies, we summarize the bias, Monte Carlo standard deviation (SD) and the average of estimated standard errors (ASE) for obtained estimates.

3.1 Simulation Study I

3.1.1 Data generation

We first consider the model with only one time-varying biomarker $\lambda(u) = \lambda_0(u)\exp\{\beta X(u)\}$, where the time-varying immune biomarker process $X(u)$ is characterized by a linear random effects model $X(u) = \alpha_0 + \alpha_1 u$. We simulate $X(u)$ to imitate the $\log_{10}$ CD4 cell counts in ACTG 175 study as described in [Song et al., 2002]. The random effects $(\alpha_0, \alpha_1)^T$ are generated from a bivariate Normal distribution with $\mathbb{E}(\alpha) = (2.5915, -0.00315)^T$, and $\mathbb{C}ov(\alpha) = D$ with elements $(D_{11}, D_{12}, D_{22})^T = (0.02408, -0.0008, 0.000014)^T$. The variance of measurement error e is $\sigma^2 = 0.01$. It represents a noise-to-signal ratio of $\mathbb{V}ar(e)/\mathbb{V}ar(X(0)) \approx 42\%$, which is approximately the same to that in the ACTG 175 data. The scheduled time points for a visit to measure $X(u)$ are at baseline and within a series of time windows: 2, 4, 8, 20, 32, 44, 56, 68, 80 ± 0.5. The event time data are generated from the Cox model with hazard ratios $e^\beta = \{1, 0.5, 0.25\}$. The censoring time follows the exponential distribution $Exp(1/180)$ and is subject to an administrative censoring at $u = 85$. We chose the baseline hazards to yield an event rate of around 10%. Specifically, the proportions of subjects dropping off the study during the follow-up and completing the study free of events are (35.5%, 54.1%), (36.0%, 54.0%) and (36.2%, 53.3%) when the hazard ratios are 1, 0.5 and 0.25, respectively. This high censoring rate reflects an HIV-1 vaccine efficacy trial where a typical infection rate is about 10%. The average number of immune biomarker measurements per subject is around 8.

The sample size for the full cohort data is $N = 1500$. The phase II sample is taken from the full cohort data from the case-control sampling $(S1)$:

$$(S1): \quad \mathbb{P}(\xi = 1|\Delta = 1) = 1, \qquad \mathbb{P}(\xi = 1|\Delta = 0) = 0.33$$

This results in around 60% missingness. Table 3.1 shows the average sample sizes for Phase I and Phase II samples.

For the AIPW method, we evaluate several sets of predictor variables which serve to estimate the augmentation terms in the non-parametric kernel regression (Nadaraya-Watson).

Table 3.1: The sample size for Phase I (N) and Phase II (n) sample under case-control sampling ($S1$) for Simulation Study I.

	Case ($\Delta = 1$)	Control ($\Delta = 0$)
β	$N(n)$	$N(n)$
0	156 (156)	1344 (443)
$-\ln 2$	151 (151)	1349 (445)
$-\ln 4$	158 (158)	1342 (443)

We generate three sets of auxiliary variables $A = (A_1, A_2)^T$ with null, moderate, and strong correlation with the immune biomarker values. The correlation is quantified by R^2. Specifically, we consider three sets of correlations with $R^2 = \{0, 0.5, 0.95\}$. For $R^2 = 0$, A is generated independently from α and (Δ, V), from $A_1 \sim N(\mathbb{E}(\alpha_1), 1)$ and $A_2 \sim N(\mathbb{E}(\alpha_1 + 40\alpha_2), 1)$. For $R^2 = 0.5, 0.95$, $A_1 = \alpha_1 + e_1$, $A_2 = \alpha_1 + 40\alpha_2 + e_2$, $e_1 \sim N(0, d_1^2)$, $e_2 \sim N(\alpha_1 + 40\alpha_2, d_2^2)$, with

$$d_1^2 = (\frac{1}{R^2} - 1)\mathbb{V}ar(\alpha_1), \qquad d_2^2 = (\frac{1}{R^2} - 1)\mathbb{V}ar(\alpha_1 + 40\alpha_2)$$

3.1.2 Methods

In this simulation study, the goal is to evaluate and compare different methods in making inference on the regression coefficient β: Full, CC, IPW and AIPW. We are particularly interested to do extensive exploration on the AIPW method. Specifically, eleven sets of predictor variables are used in the kernel regression: Δ, (Δ, V), A, (Δ, A) and (Δ, V, A). For each set of variables including A, there are also three choices for A with $R^2 = 0, 0.5, 0.95$ as described above. By comparing these eleven AIPW estimators, we look in how the predictor variables in the augmentation terms influence the performance of the AIPW method.

3.1.3 Results

From Table 3.2, we see that the results based on pre-specified sampling probabilities (π) and estimated sampling probabilities ($\hat{\pi}$) are very similar. In Figure 3.1 we plot the biases, 95% coverage probabilities and the relative efficiencies (calculated as the Monte Carlo variance

of $\hat{\beta}$ compared to that from Full). We only show the estimates obtained using π. The x axis lists the methods we are comparing with. The variables enclosed in the parentheses of AIPW() indicate which variables are used in estimating the augmentation terms. The numbers enclosed in the parentheses of AIPW() indicate which set of auxiliary variables A (corresponding to $R^2 = 0, 0.5, 0.95$) are used in AIPW(A), AIPW(Δ, A) or AIPW(Δ, V, A).

As expected the CC analysis generates very biased estimates when the β is large. We observe slightly large biases for AIPW(Δ, V, A) estimators when $\beta = -\ln 4$, even though the theory of double robustness guarantees its consistency as the sampling probabilities are correctly specified. This suggests a concern in terms of bias when including too many variables in the non-parametric kernel method. Actually we can see from the plot of relative efficiency that, as long as we include strong predictor A, adding additional variables Δ or V does not help to improve the efficiency. Also, if we cannot find auxiliary variables that are highly correlated with the immune biomarker, we suggest just using IPW or AIPW(Δ). However, we hesitate to use AIPW(A), even though the figure shows that it is more efficient than IPW and AIPW(Δ). Actually, the observed efficiency gain in this situation could be just due to that A is continuous and has great variability. In additional simulation studies with A being discrete or less variable, the efficiency gain disappears and even the efficiency loss shows up (Table 3.3). When the auxiliary variables are highly correlated with the $X(u)$, including them and using AIPW method provides less variable estimates than other methods, especially when $X(u)$ has strong effect on the event time. The 95% coverage probabilities are slightly below the nominal level when $\beta = -\ln 4$, suggesting SE($\hat{\beta}$) is underestimated by the sandwich variance estimation.

3.2 Simulation Study II

3.2.1 Data generation

We next consider the Cox model including both the immune biomarker and the vaccination indicator $\lambda(u) = \lambda_0(u) \exp\{\beta X(u) + \eta Z\}$. When there is variability of immune response levels in the placebo arm, we can use this model to assess the Prentice's surrogate by examining if η is plausibly closed to zero (assuming no dual predictors for

Table 3.2: Simulation results for Simulation Study I: $\lambda(u) = \lambda_0(u)\exp\{\beta X(u)\}$

	R^2	Sampling	$\beta = 0$			$\beta = -\ln 2$			$\beta = -\ln 4$		
Method		Prob.	Bias	SD($\hat{\beta}$)	ASE($\hat{\beta}$)	Bias	SD($\hat{\beta}$)	ASE($\hat{\beta}$)	Bias	SD($\hat{\beta}$)	ASE($\hat{\beta}$)
Full			-0.012	0.404	0.404	-0.013	0.377	0.367	-0.007	0.354	0.330
CC			-0.001	0.401	0.406	0.062	0.372	0.365	0.165	0.343	0.328
IPW		π	-0.001	0.439	0.446	-0.020	0.415	0.408	-0.020	0.402	0.383
AIPW(Δ)		π	-0.001	0.439	0.446	-0.018	0.413	0.408	-0.009	0.395	0.382
AIPW(Δ, V)		π	-0.001	0.440	0.445	-0.017	0.413	0.407	-0.016	0.399	0.382
AIPW(A)	0	π	-0.003	0.439	0.438	-0.004	0.407	0.400	0.020	0.388	0.372
AIPW(A, Δ)	0	π	-0.006	0.441	0.437	-0.017	0.414	0.399	-0.016	0.402	0.373
AIPW(A, Δ, V)	0	π	-0.006	0.445	0.433	-0.020	0.419	0.397	-0.014	0.399	0.370
AIPW(A)	0.5	π	-0.011	0.435	0.435	-0.012	0.401	0.398	-0.004	0.384	0.370
AIPW(A, Δ)	0.5	π	-0.010	0.437	0.436	-0.013	0.401	0.398	-0.019	0.391	0.371
AIPW(A, Δ, V)	0.5	π	-0.014	0.440	0.436	-0.020	0.405	0.397	-0.023	0.394	0.370
AIPW(A)	0.95	π	-0.010	0.418	0.419	-0.021	0.387	0.381	-0.026	0.368	0.351
AIPW(A, Δ)	0.95	π	-0.011	0.419	0.420	-0.023	0.389	0.382	-0.025	0.366	0.353
AIPW(A, Δ, V)	0.95	π	-0.011	0.428	0.426	-0.031	0.395	0.387	-0.039	0.368	0.356
IPW		$\hat{\pi}$	-0.001	0.439	0.446	-0.020	0.415	0.408	-0.021	0.402	0.383
AIPW(Δ)		$\hat{\pi}$	-0.001	0.439	0.446	-0.017	0.412	0.408	-0.010	0.395	0.382
AIPW(Δ, V)		$\hat{\pi}$	-0.001	0.440	0.445	-0.017	0.413	0.407	-0.017	0.397	0.382
AIPW(A)	0	$\hat{\pi}$	-0.003	0.439	0.438	-0.010	0.411	0.400	0.018	0.386	0.372
AIPW(A, Δ)	0	$\hat{\pi}$	-0.006	0.442	0.438	-0.017	0.415	0.400	-0.017	0.401	0.373
AIPW(A, Δ, V)	0	$\hat{\pi}$	-0.006	0.445	0.433	-0.020	0.419	0.397	-0.011	0.398	0.370
AIPW(A)	0.5	$\hat{\pi}$	-0.011	0.435	0.436	-0.013	0.401	0.398	-0.006	0.381	0.370
AIPW(A, Δ)	0.5	$\hat{\pi}$	-0.011	0.437	0.436	-0.013	0.401	0.398	-0.017	0.390	0.371
AIPW(A, Δ, V)	0.5	$\hat{\pi}$	-0.014	0.440	0.436	-0.020	0.405	0.397	-0.024	0.395	0.371
AIPW(A)	0.95	$\hat{\pi}$	-0.010	0.418	0.419	-0.021	0.387	0.381	-0.025	0.367	0.351
AIPW(A, Δ)	0.95	$\hat{\pi}$	-0.011	0.419	0.420	-0.024	0.389	0.382	-0.030	0.368	0.353
AIPW(A, Δ, V)	0.95	$\hat{\pi}$	-0.011	0.428	0.426	-0.031	0.395	0.387	-0.039	0.370	0.356

[1] π: The IPW and AIPW methods are implemented using pre-specified true sampling probabilities. $\hat{\pi}$: The IPW and AIPW methods are implemented using estimated sampling probabilities.
[2] R^2 quantifies the correlation of A and $X(u)$.
[3] For the kernel regression, the standard Normal kernel is used, and the bandwidth for predictor variable P is $0.75N^{-1/3}sd(P)$.
[4] The sampling probability model is $\pi(\Delta) = \Delta + 0.33(1 - \Delta)$.

Figure 3.1: Simulation results for Simulation Study I: $\lambda(u) = \lambda_0(u)\exp\{\beta X(u)\}$. The IPW and AIPW estimates are based on pre-specified true sampling probabilities π.

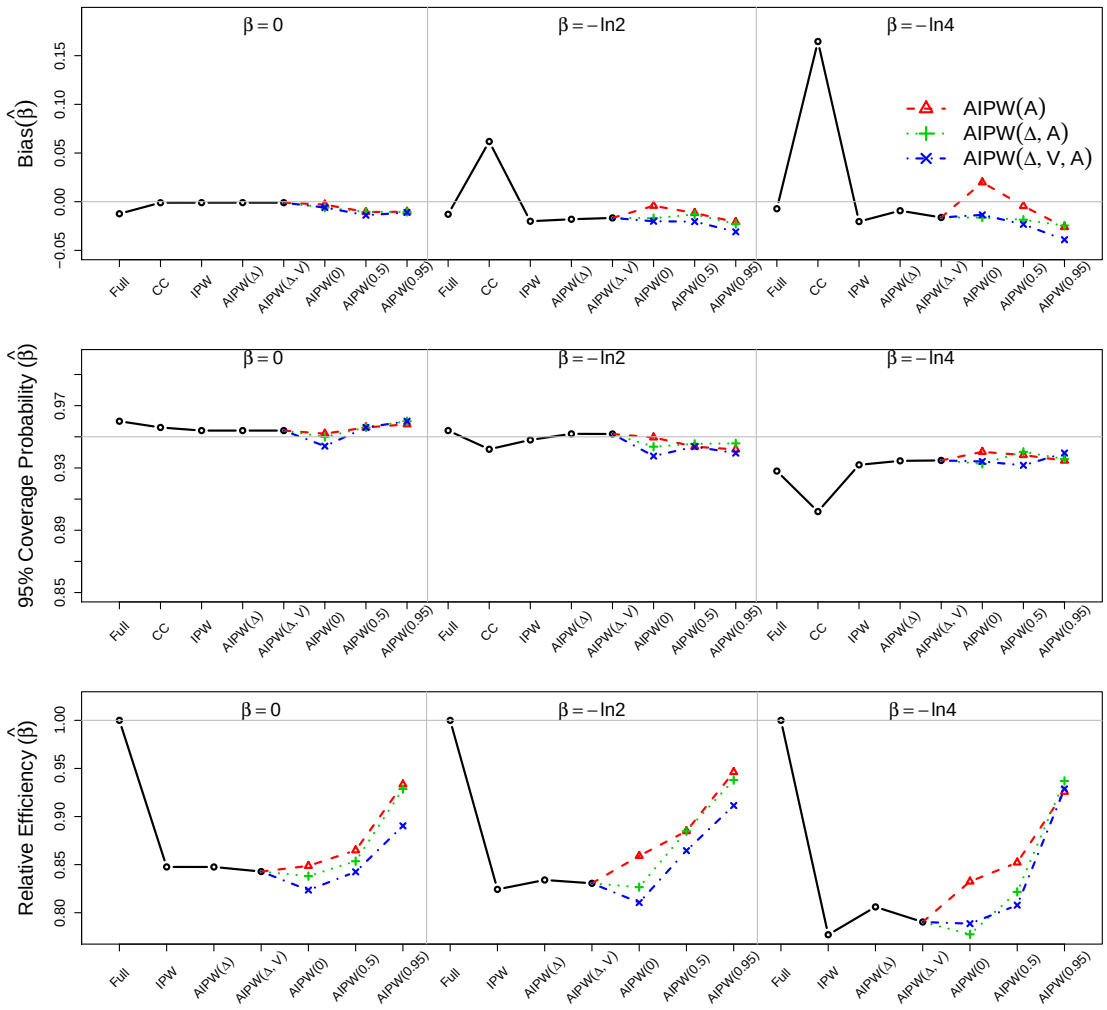


Table 3.3: Simulation results of AIPW(A) for Simulation Study I, with different sets of auxiliary variables.

	R^2	Sampling	$\beta = 0$			$\beta = -\ln 2$			$\beta = -\ln 4$		
Method		Prob.	Bias	SD($\hat{\beta}$)	ASE($\hat{\beta}$)	Bias	SD($\hat{\beta}$)	ASE($\hat{\beta}$)	Bias	SD($\hat{\beta}$)	ASE($\hat{\beta}$)
IPW		π	-0.001	0.439	0.446	-0.020	0.415	0.408	-0.020	0.402	0.383
AIPW(Δ)		π	-0.001	0.439	0.446	-0.018	0.413	0.408	-0.009	0.395	0.382
AIPW(A)	0	$\hat{\pi}$	-0.003	0.439	0.438	-0.004	0.407	0.400	0.020	0.388	0.372
AIPW(A^*)	0	π	0.001	0.434	0.444	-0.014	0.411	0.406	-0.002	0.394	0.380
AIPW($\tilde{A}$)	0	π	0.000	0.438	0.445	-0.014	0.412	0.407	-0.011	0.400	0.381
IPW		$\hat{\pi}$	-0.001	0.439	0.446	-0.020	0.415	0.408	-0.021	0.402	0.383
AIPW(Δ)		$\hat{\pi}$	-0.001	0.439	0.446	-0.017	0.412	0.408	-0.010	0.395	0.382
AIPW(A)	0	$\hat{\pi}$	-0.003	0.439	0.438	-0.010	0.411	0.400	0.018	0.386	0.372
AIPW(A^*)	0	$\hat{\pi}$	0.001	0.434	0.444	-0.014	0.411	0.407	-0.005	0.395	0.381
AIPW($\tilde{A}$)	0	$\hat{\pi}$	0.000	0.438	0.445	-0.015	0.412	0.407	-0.013	0.402	0.382

[1] $A = (A_1, A_2)^T$ are generated independently from α and (Δ, V): $A_1 \sim N(\mathbb{E}(\alpha_1), 1)$ and $A_2 \sim N(\mathbb{E}(\alpha_1 + 40\alpha_2), 1)$ (See Table 3.2)
[2] $A^* = (A_1^*, A_2^*)^T$: A_j^* is discrete variable generated based on quartiles of A_j, $j = 1, 2$.
[3] $\tilde{A} = (\tilde{A}_1, \tilde{A}_2)^T$ are generated independently from α and (Δ, V): $\tilde{A}_1 \sim N(\mathbb{E}(\alpha_1), 0.01)$ and $\tilde{A}_2 \sim N(\mathbb{E}(\alpha_1 + 40\alpha_2), 0.01)$

$X(u)$ and T). The time-varying immune biomarker process $X(u)$ is still characterized by a linear random effects model $X(u) = \alpha_0 + \alpha_1 u$, and $Z \sim Bernoulli(0.5)$ is the 1:1 treatment arm assignment with $Z = 1$ for vaccine and $Z = 0$ for placebo. The random effects α are simulated from bivariate Normal distribution with $\mathbb{E}(\alpha|Z = 0) = (2.5915, -0.00145)^T$, $\mathbb{E}(\alpha|Z = 1) = (2.5915, -0.00315)^T$ and $\mathbb{C}ov(\alpha|Z) = D$ with elements $(D_{11}, D_{12}, D_{22})^T = (0.02408, -0.0008, 0.000014)^T$. We consider two sets of hazard ratios $(e^\beta, e^\eta)^T = \{(0.5, 1)^T, (0.5, 0.5)^T\}$. The censoring time follows the exponential distribution $Exp(1/180)$ and is subject to an administrative censoring at $u = 85$.

In Simulation Study I, we have especially explored different AIPW estimators. In this simulation study, we would also like to compare several AIPW estimators to IPW and Full estimators in a similar setting as that in Simulation Study I. Beyond that however we also aim to evaluate the influence of 1) the number of measuring time points; and 2) misspecified measurement error model. We consider the following four scenarios in terms of different measuring schedules or measurement error distributions.

1. Simulation Study II(a): The scheduled visits for measuring $X(u)$ are at baseline and within a series of time windows: 2, 4, 8, 20, 32, 44, 56, 68, 80 $\pm$0.5. Measurement

error $e \sim N(0, 0.01)$.

2. Simulation Study II(b): The scheduled visits for measuring $X(u)$ are at baseline and within a series of time windows: 8, 44, 80 ± 0.5. Measurement error $e \sim N(0, 0.01)$.
3. Simulation Study II(c): The scheduled visits for measuring $X(u)$ are at baseline and within a series of time windows: 2, 4, 8, 20, 32, 44, 56, 68, 80 ± 0.5. Measurement error $e \sim Exp(10) - 0.1$.
4. Simulation Study II(d): The scheduled visits for measuring $X(u)$ are at baseline and within a series of time windows: 2, 4, 8, 20, 32, 44, 56, 68, 80 ± 0.5. Measurement error depends on the value of the immune biomarker: $e \sim N(0, 0.01)$ if $X(u) > 2.5$; $e \sim N(0, 0.05)$ if $X(u) \leq 2.5$.

The conditional score estimator, though does not require any distributional assumption on the random effects α, does assume the measurement errors are random and Normal. Simulation Study II(a) is the setting with random Normal measurement errors as required by the model assumptions. Also the visit schedule provides decent number of immune biomarker measurements for the inferential analysis. Simulation Study II(b), compared to Simulation Study II(a), reduces 60% of the number of available immune biomarker measurements. [Tsiatis and Davidian, 2001] and [Song et al., 2002] conducted simulation studies investigating different distributions for random effects and different levels of variance for the measurement errors. Here we set up Simulation Study II(c) and Simulation Study II(d) to assess the influence when the measurement error is not Normal or even worse, depends on the biomarker values.

For all models, the proportions of subjects dropping off the study during the follow-up and completing the study free of events are (35.9%, 53.7%) and (36.1%, 53.4%) when the hazard ratios $(e^{\beta}, e^{\eta})^T$ are $(0.5, 1)^T$, $(0.5, 0.5)^T$, respectively. The average number of immune biomarker measurements per subject in Simulation Study II(a),(c), and (d) is around 8.3, and in Simulation Study II(b) is around 3.3. For all studies, the full cohort data consists of $N = 1500$ subjects, and the second phase sample is still taken by case-control sampling $(S1)$ with sampling probabilities $\mathbb{P}(\xi = 1|\Delta = 1) = 1$ and $\mathbb{P}(\xi = 1|\Delta = 0) = 0.33$. Table 3.4 shows the average sample sizes for Phase I and Phase II samples.

Table 3.4: The sample size for Phase I (N) and Phase II (n) sample under case-control sampling ($S1$) for simulation Simulation Study II(a).

	Case ($\Delta = 1$)	Control ($\Delta = 0$)
(β, η)	$N(n)$	$N(n)$
$(-\ln 2, 0)$	156 (156)	1344 (443)
$(-\ln 2, -\ln 2)$	157 (157)	1343 (443)

3.2.2 Results for Simulation Study II(a)

Table 3.5 summarizes the fitting for Simulation Study II(a). Still the results from using π and $\hat{\pi}$ are very similar so we only plot the results from using π in Figure 3.2. Since Z is always included to estimate the augmentation terms in AIPW estimators, the corresponding estimates for η are almost as efficient as that from the full cohort data. On the other hand, $\hat{\beta}$ from AIPW(Δ, Z) has slightly smaller variance than that from AIPW(Δ, Z, A) when A is independent of $X(u)$ and (Δ, V) ($R^2 = 0$). From Figure 3.2, we still observe increasing efficiency for $\hat{\beta}$ from AIPW(Δ, Z, A) when A has higher correlation with $X(u)$. The coverage probabilities are closed to the nominal value.

3.2.3 Results for Simulation Study II(b)

We now reduce the frequency to measure the immune biomarker and summarize the results in Table 3.6. The average number of measurements per subject is around 3.3. By direct comparison of SD$(\hat{\beta})$ in this table to that from Simulation Study II(a), we observe dramatic reduction in the efficiency. Still the results from using π and $\hat{\pi}$ are very similar so we only plot the results from using π in Figure 3.3. There is a concern on the bias for AIPW estimators with $(\beta, \eta)^T = (-\ln 2, 0)$. Also only adding A that has extremely strong correlation with the biomarker in AIPW method can yield slightly efficiency gain. It suggests that when the event time depends on the biomarker only, but only very limited number of measurements are available for the biomarker variable, using more complex AIPW method can give poorer results than using the simple IPW method. We also found in this simulation study that for around 1%$\sim$2% of the runs the program failed.

Table 3.5: Simulation results for Simulation Study II(a): $\lambda(u) = \lambda_0(u)\exp\{\beta X(u) + \eta Z\}$

	R^2	Samp. Prob.	$(\beta,\eta) = (-\ln 2, 0)$						$(\beta,\eta) = (-\ln 2, -\ln 2)$					
			$\hat{\beta}$			$\hat{\eta}$			$\hat{\beta}$			$\hat{\eta}$		
Method			Bias	SD	ASE	Bias	SD	ASE	Bias	SD	ASE	Bias	SD	ASE
Full			0.008	0.350	0.361	0.009	0.165	0.164	0.019	0.348	0.338	-0.000	0.178	0.171
CC			0.079	0.351	0.361	0.005	0.166	0.165	0.113	0.335	0.338	0.080	0.172	0.171
IPW		π	-0.002	0.395	0.405	0.008	0.188	0.185	0.023	0.384	0.385	-0.001	0.195	0.192
AIPW(Δ, Z)		π	-0.001	0.395	0.405	0.008	0.166	0.166	0.025	0.382	0.385	-0.000	0.181	0.173
AIPW(Δ, Z, A)	0	π	-0.006	0.399	0.392	0.008	0.166	0.166	0.017	0.388	0.370	-0.001	0.182	0.173
AIPW(Δ, Z, A)	0.5	π	-0.007	0.390	0.394	0.007	0.166	0.166	0.018	0.378	0.373	-0.000	0.181	0.173
AIPW(Δ, Z, A)	0.95	π	-0.013	0.368	0.380	0.007	0.166	0.165	0.005	0.366	0.358	-0.002	0.179	0.172
IPW		$\hat{\pi}$	-0.002	0.395	0.405	0.008	0.188	0.185	0.023	0.384	0.385	-0.001	0.195	0.192
AIPW(Δ, Z)		$\hat{\pi}$	-0.002	0.395	0.405	0.008	0.166	0.166	0.025	0.382	0.385	-0.000	0.181	0.173
AIPW(Δ, Z, A)	0	$\hat{\pi}$	-0.006	0.399	0.392	0.008	0.166	0.166	0.017	0.388	0.370	-0.001	0.182	0.173
AIPW(Δ, Z, A)	0.5	$\hat{\pi}$	-0.007	0.390	0.394	0.007	0.166	0.166	0.019	0.378	0.373	-0.000	0.181	0.173
AIPW(Δ, Z, A)	0.95	$\hat{\pi}$	-0.014	0.368	0.380	0.007	0.166	0.165	0.005	0.366	0.358	-0.002	0.179	0.172

[1] π: The IPW and AIPW methods are implemented using pre-specified true sampling probabilities. $\hat{\pi}$: The IPW and AIPW methods are implemented using estimated sampling probabilities.
[2] R^2 quantifies the correlation of A and $X(u)$.
[3] For the kernel regression, the standard Normal kernel is used, and the bandwidth for predictor variable P is $0.75N^{-1/3}sd(P)$.
[4] The sampling probability model is $\pi(\Delta) = \Delta + 0.33(1-\Delta)$.
[5] The scheduled visits for measuring $X(u)$ are at baseline and within a series of time windows: 2, 4, 8, 20, 32, 44, 56, 68, 80 ± 0.5.
[6]: Measurement error $e \sim N(0, 0.01)$.

Table 3.6: Simulation results for Simulation Study II(b): $\lambda(u) = \lambda_0(u)\exp\{\beta X(u) + \eta Z\}$

	R^2	Samp. Prob.	$(\beta,\eta) = (-\ln 2, 0)$						$(\beta,\eta) = (-\ln 2, -\ln 2)$					
			$\hat{\beta}$			$\hat{\eta}$			$\hat{\beta}$			$\hat{\eta}$		
Method			Bias	SD	ASE	Bias	SD	ASE	Bias	SD	ASE	Bias	SD	ASE
Full			-0.008	0.552	0.535	0.008	0.167	0.169	0.006	0.487	0.477	-0.001	0.180	0.174
CC			0.058	0.570	0.548	0.003	0.169	0.170	0.098	0.477	0.480	0.078	0.174	0.175
IPW		π	-0.026	0.627	0.605	0.005	0.191	0.192	0.003	0.532	0.536	-0.003	0.198	0.197
AIPW(Δ, Z)		π	-0.062	0.640	0.606	0.003	0.171	0.174	-0.020	0.539	0.537	-0.004	0.184	0.179
AIPW(Δ, Z, A)	0	π	-0.080	0.670	0.599	0.001	0.172	0.173	-0.022	0.536	0.523	-0.008	0.186	0.179
AIPW(Δ, Z, A)	0.5	π	-0.067	0.638	0.600	0.002	0.171	0.174	-0.024	0.539	0.530	-0.006	0.185	0.178
AIPW(Δ, Z, A)	0.95	π	-0.074	0.609	0.597	0.001	0.170	0.172	-0.038	0.534	0.520	-0.006	0.183	0.177
IPW		$\hat{\pi}$	-0.026	0.627	0.605	0.005	0.191	0.192	0.003	0.532	0.536	-0.003	0.198	0.197
AIPW(Δ, Z)		$\hat{\pi}$	-0.062	0.640	0.606	0.002	0.171	0.174	-0.020	0.539	0.537	-0.004	0.184	0.179
AIPW(Δ, Z, A)	0	$\hat{\pi}$	-0.080	0.670	0.599	0.001	0.172	0.173	-0.023	0.536	0.523	-0.008	0.186	0.179
AIPW(Δ, Z, A)	0.5	$\hat{\pi}$	-0.067	0.638	0.600	0.002	0.171	0.174	-0.025	0.539	0.530	-0.006	0.185	0.178
AIPW(Δ, Z, A)	0.95	$\hat{\pi}$	-0.072	0.609	0.596	0.001	0.170	0.172	-0.038	0.534	0.520	-0.006	0.183	0.177

[1] π: The IPW and AIPW methods are implemented using pre-specified true sampling probabilities. $\hat{\pi}$: The IPW and AIPW methods are implemented using estimated sampling probabilities.
[2] R^2 quantifies the correlation of A and $X(u)$.
[3] For the kernel regression, the standard Normal kernel is used, and the bandwidth for predictor variable P is $0.75N^{-1/3}sd(P)$.
[4] The sampling probability model is $\pi(\Delta) = \Delta + 0.33(1-\Delta)$.
[5] The scheduled visits for measuring $X(u)$ are at baseline and within a series of time windows: 8, 44, 80 ± 0.5.
[6]: Measurement error $e \sim N(0, 0.01)$.

Figure 3.2: Simulation results for Simulation Study II(a): $\lambda(u) = \lambda_0(u)\exp\{\beta X(u) + \eta Z\}$. The IPW and AIPW estimates are based on pre-specified true sampling probabilities π.

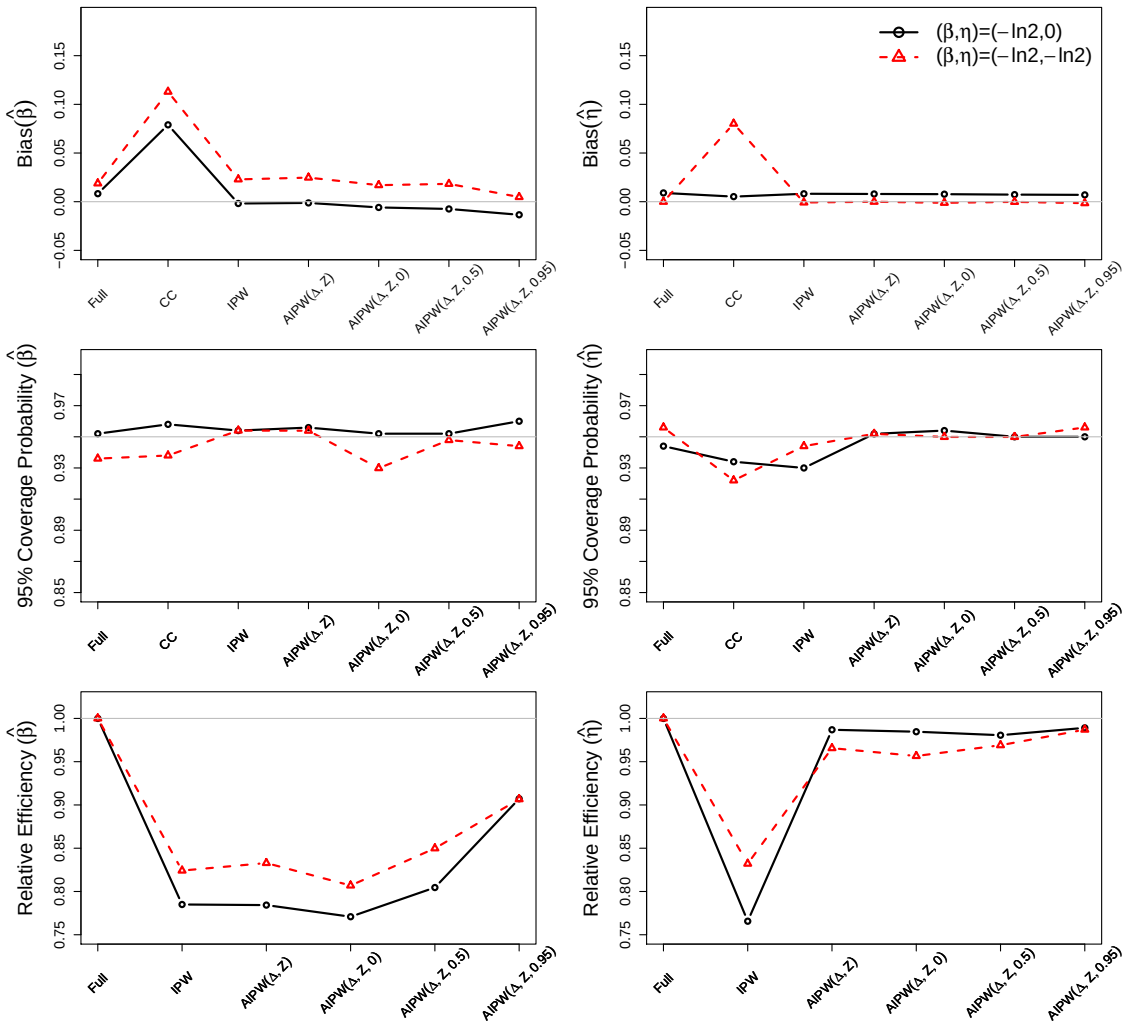
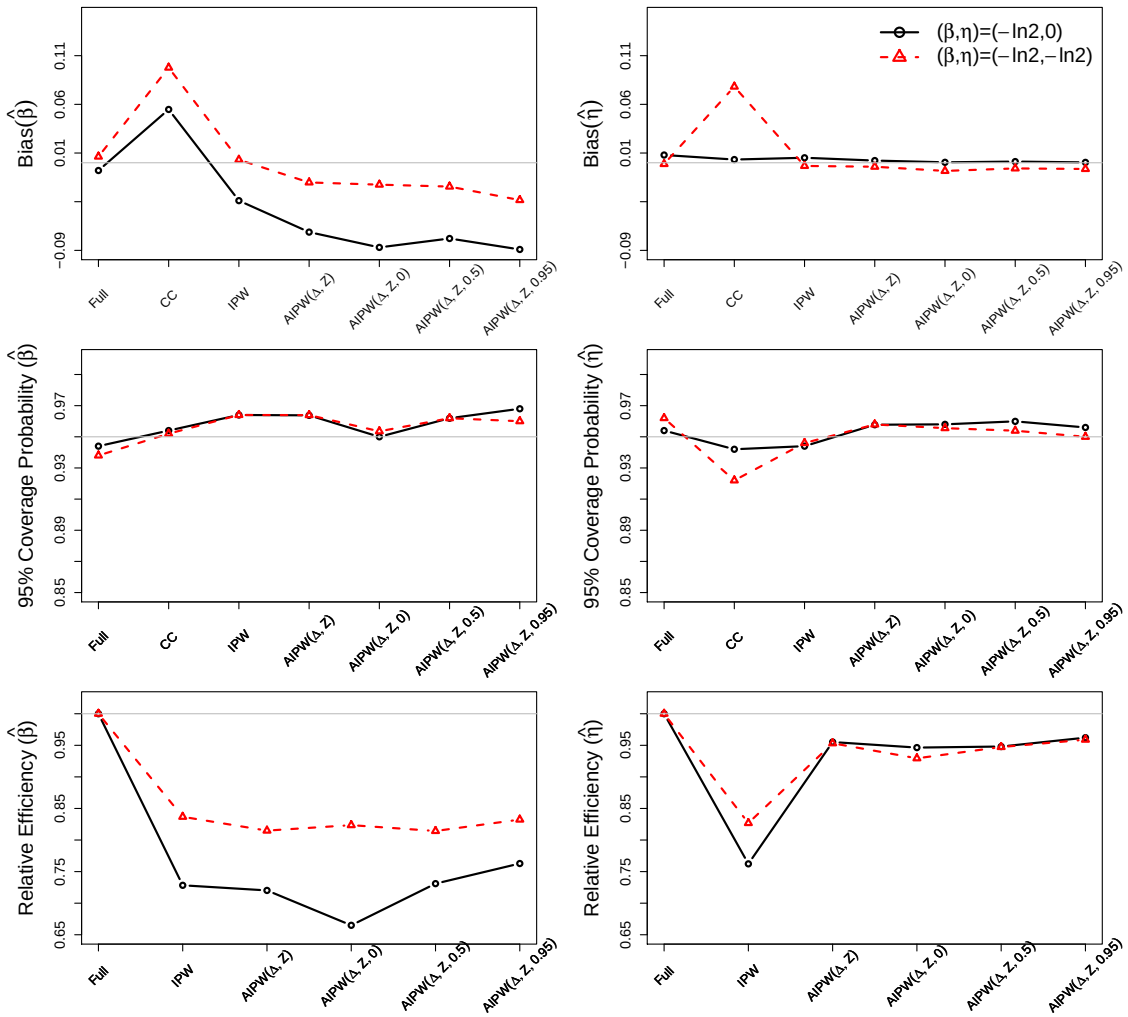


Figure 3.3: Simulation results for Simulation Study II(b): $\lambda(u) = \lambda_0(u)\exp\{\beta X(u) + \eta Z\}$. The IPW and AIPW estimates are based on pre-specified true sampling probabilities π.



3.2.4 Results for Simulation Study II(c)

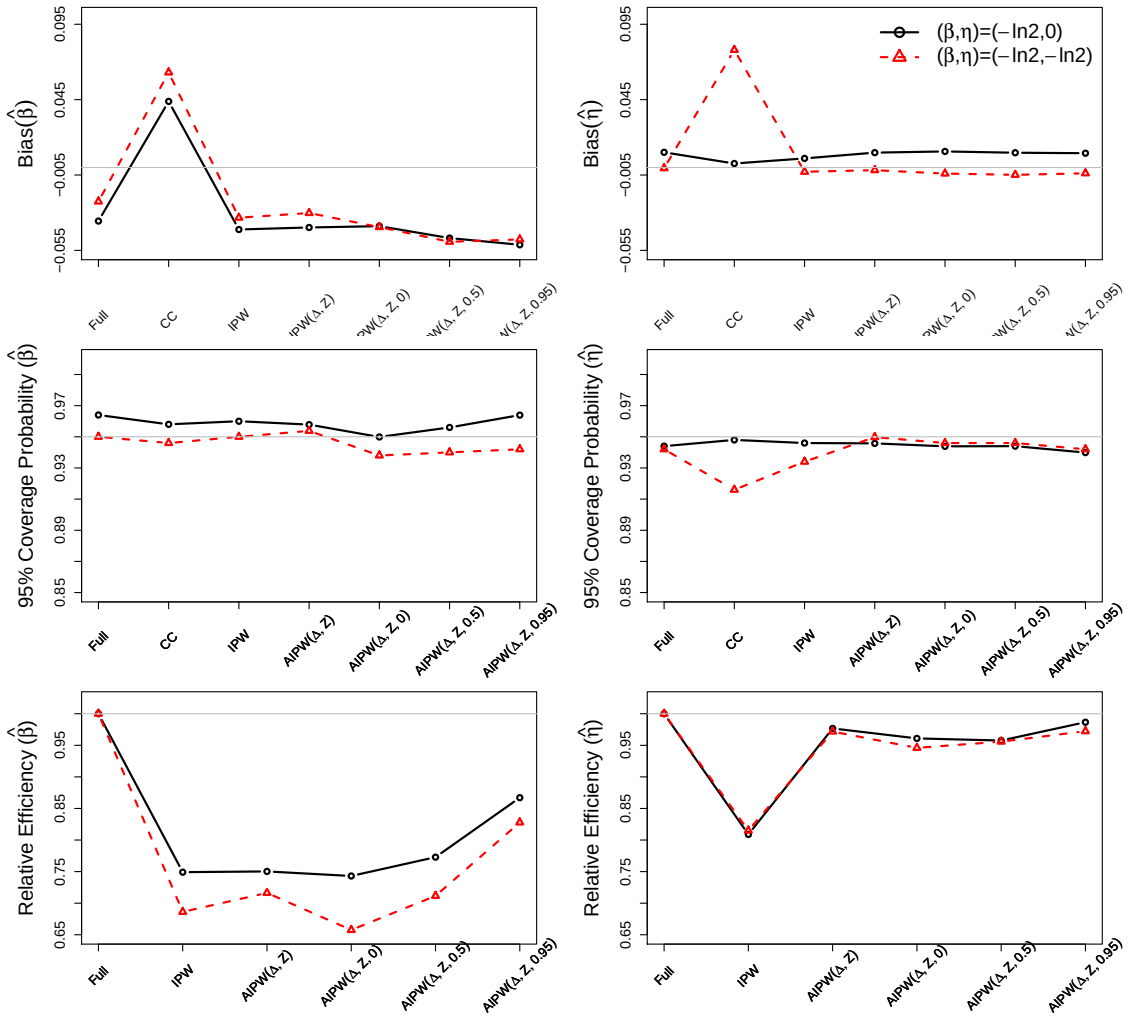
Here we intend to assess the influences on the methods when the measurement errors are not Normal. We simulate the measurement error e from $Exp(10) - 0.1$ so that it still has $\mathbb{E}(e) = 0$ and $\mathbb{V}ar(e) = 0.01$. However the density of e is no longer in a bell shape and has heavy right tail. We compare the results showing in Table 3.7 and Figure 3.4 to those from Simulation Study II(a). From the plot we see slightly large bias on $\hat{\beta}$ for all methods, but not on $\hat{\eta}$. This might be because Z does not involve any measurement errors. The 95% coverage probabilities are very closed to the nominal level expect for the CC method. Similar patter in the relative efficiency as in Simulation Study II(a) is observed here, suggesting unless very strong predictors for the biomarker exist, using IPW method rather than the AIPW method is recommended. The influence of misspecified measurement error seems to center on the bias of $\hat{\beta}$, but the level of bias is acceptable: with relative bias up to 7.4%.

Table 3.7: Simulation results for Simulation Study II(c): $\lambda(u) = \lambda_0(u)\exp\{\beta X(u) + \eta Z\}$

	R^2	Samp. Prob.	$(\beta,\eta) = (-\ln 2, 0)$						$(\beta,\eta) = (-\ln 2, -\ln 2)$					
			$\hat{\beta}$			$\hat{\eta}$			$\hat{\beta}$			$\hat{\eta}$		
Method			Bias	SD	ASE	Bias	SD	ASE	Bias	SD	ASE	Bias	SD	ASE
Full			-0.036	0.354	0.370	0.010	0.170	0.165	-0.023	0.345	0.344	-0.000	0.183	0.171
CC			0.044	0.366	0.371	0.003	0.168	0.165	0.063	0.365	0.344	0.078	0.179	0.171
IPW		π	-0.041	0.409	0.416	0.006	0.189	0.186	-0.033	0.416	0.392	-0.003	0.202	0.193
AIPW(Δ, Z)		π	-0.040	0.409	0.416	0.010	0.172	0.167	-0.030	0.408	0.392	-0.002	0.185	0.174
AIPW(Δ, Z, A)	0	π	-0.039	0.411	0.403	0.011	0.173	0.166	-0.040	0.425	0.378	-0.004	0.188	0.173
AIPW(Δ, Z, A)	0.5	π	-0.047	0.403	0.405	0.010	0.174	0.166	-0.049	0.409	0.380	-0.005	0.187	0.173
AIPW(Δ, Z, A)	0.95	π	-0.051	0.381	0.391	0.009	0.171	0.165	-0.048	0.379	0.365	-0.004	0.185	0.172
IPW		$\hat{\pi}$	-0.041	0.409	0.416	0.006	0.189	0.186	-0.034	0.417	0.392	-0.003	0.202	0.193
AIPW(Δ, Z)		$\hat{\pi}$	-0.040	0.409	0.416	0.010	0.172	0.167	-0.030	0.408	0.392	-0.002	0.185	0.174
AIPW(Δ, Z, A)	0	$\hat{\pi}$	-0.039	0.411	0.403	0.011	0.173	0.166	-0.040	0.426	0.378	-0.004	0.188	0.173
AIPW(Δ, Z, A)	0.5	$\hat{\pi}$	-0.047	0.403	0.405	0.010	0.174	0.166	-0.049	0.409	0.380	-0.005	0.187	0.173
AIPW(Δ, Z, A)	0.95	$\hat{\pi}$	-0.050	0.381	0.391	0.009	0.170	0.165	-0.046	0.378	0.365	-0.003	0.184	0.172

[1] π: The IPW and AIPW methods are implemented using pre-specified true sampling probabilities. $\hat{\pi}$: The IPW and AIPW methods are implemented using estimated sampling probabilities.
[2] R^2 quantifies the correlation of A and $X(u)$.
[3] For the kernel regression, the standard Normal kernel is used, and the bandwidth for predictor variable P is $0.75N^{-1/3}sd(P)$.
[4] The sampling probability model is $\pi(\Delta) = \Delta + 0.33(1 - \Delta)$.
[5] The scheduled visits for measuring $X(u)$ are at baseline and within a series of time windows: 0, 2, 4, 8, 20, 32, 44, 56, 68, 80 ± 0.5.
[6]: Measurement error $e \sim Exp(10) - 0.1$.

Figure 3.4: Simulation results for Simulation Study II(c): $\lambda(u) = \lambda_0(u)\exp\{\beta X(u) + \eta Z\}$. The IPW and AIPW estimates are based on pre-specified true sampling probabilities π.



3.2.5 Results for Simulation Study II(d)

We further look into the impact of misspecified measurement error model by considering the situation where the measurement error depends on the level of underlying immune biomarker: if the underlying biomarker level is greater than 2.5, $e \sim N(0, 0.01)$; otherwise $e \sim N(0, 0.05)$. This represents the case when the measured immune biomarker level is more variable when its level is low. The results are shown in Table 3.8 and Figure 3.5. Here we have more serious issue of bias than that in Simulation Study II(c) for $\hat{\beta}$, especially in the setting with $(\beta, \eta)^T = (-\ln 2, -\ln 2)^T$. The relative bias can be as high as 10.7%.

Table 3.8: Simulation results for Simulation Study II(d): $\lambda(u) = \lambda_0(u) \exp\{\beta X(u) + \eta Z\}$

	R^2	Samp. Prob.	$(\beta, \eta) = (-\ln 2, 0)$						$(\beta, \eta) = (-\ln 2, -\ln 2)$					
			$\hat{\beta}$			$\hat{\eta}$			$\hat{\beta}$			$\hat{\eta}$		
Method			Bias	SD	ASE	Bias	SD	ASE	Bias	SD	ASE	Bias	SD	ASE
Full			0.061	0.359	0.368	0.013	0.165	0.165	0.071	0.349	0.343	0.004	0.178	0.171
CC			0.122	0.362	0.372	0.008	0.168	0.165	0.157	0.347	0.347	0.084	0.171	0.171
IPW		π	0.049	0.397	0.409	0.011	0.190	0.186	0.074	0.391	0.388	0.004	0.195	0.193
AIPW(Δ, Z)		π	0.049	0.398	0.409	0.011	0.167	0.167	0.074	0.391	0.388	0.004	0.181	0.174
AIPW(Δ, Z, A)	0	π	0.045	0.404	0.397	0.011	0.167	0.166	0.066	0.399	0.376	0.003	0.182	0.173
AIPW(Δ, Z, A)	0.5	π	0.047	0.396	0.400	0.009	0.168	0.166	0.067	0.393	0.380	0.001	0.182	0.173
AIPW(Δ, Z, A)	0.95	π	0.040	0.381	0.390	0.011	0.167	0.166	0.057	0.379	0.367	0.003	0.179	0.172
IPW		$\hat{\pi}$	0.049	0.397	0.409	0.011	0.190	0.186	0.074	0.391	0.389	0.004	0.195	0.193
AIPW(Δ, Z)		$\hat{\pi}$	0.049	0.398	0.409	0.011	0.167	0.167	0.074	0.391	0.389	0.005	0.181	0.174
AIPW(Δ, Z, A)	0	$\hat{\pi}$	0.045	0.404	0.397	0.011	0.167	0.166	0.065	0.399	0.376	0.003	0.182	0.173
AIPW(Δ, Z, A)	0.5	$\hat{\pi}$	0.047	0.396	0.400	0.009	0.168	0.166	0.067	0.393	0.380	0.001	0.182	0.173
AIPW(Δ, Z, A)	0.95	$\hat{\pi}$	0.039	0.382	0.390	0.011	0.166	0.166	0.057	0.379	0.367	0.003	0.179	0.172

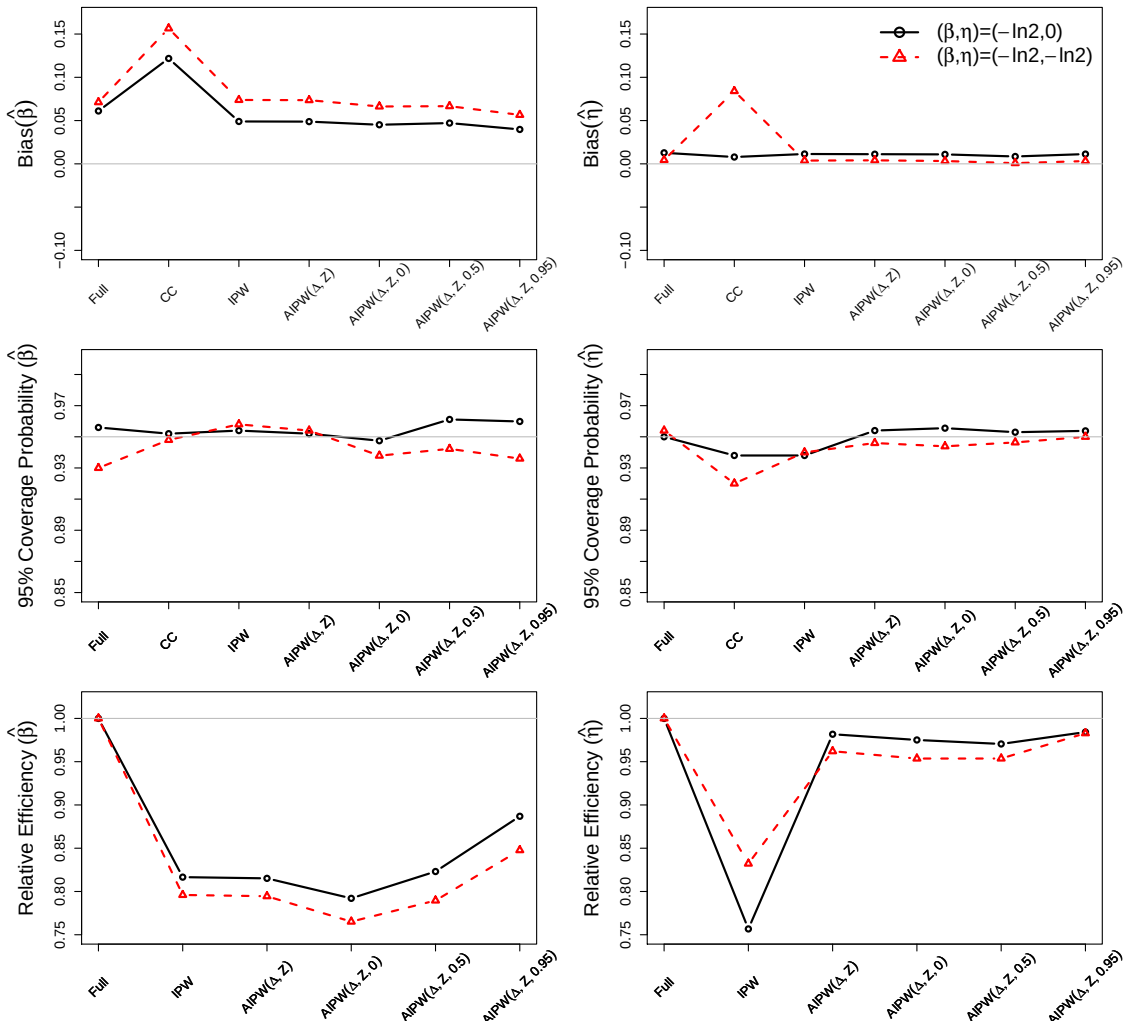
[1] π: The IPW and AIPW methods are implemented using pre-specified true sampling probabilities. $\hat{\pi}$: The IPW and AIPW methods are implemented using estimated sampling probabilities.
[2] R^2 quantifies the correlation of A and $X(u)$.
[3] For the kernel regression, the standard Normal kernel is used, and the bandwidth for predictor variable P is $0.75N^{-1/3}sd(P)$.
[4] The sampling probability model is $\pi(\Delta) = \Delta + 0.33(1 - \Delta)$.
[5] The scheduled visits for measuring $X(u)$ are at baseline and within a series of time windows: 0, 2, 4, 8, 20, 32, 44, 56, 68, 80 ± 0.5.
[6]: Measurement error $e \sim N(0, 0.01)$ if $X(u) > 2.5$; $e \sim N(0, 0.05)$ if $X(u) \leq 2.5$.

3.3 Simulation Study III

3.3.1 Data generation

We finally consider the Cox model including the interaction of the immune biomarker and the vaccination indicator $\lambda(u) = \lambda_0(u) \exp\{\beta X(u) + \eta Z + \gamma X(u) Z\}$. The simulation data set is

Figure 3.5: Simulation results for Simulation Study II(d): $\lambda(u) = \lambda_0(u)\exp\{\beta X(u) + \eta Z\}$. The IPW and AIPW estimates are based on pre-specified true sampling probabilities π.



exactly the same as that in Simulation Study II(a) with hazard ratio $(e^{\beta}, e^{\eta})^T = (0.5, 0.5)^T$. It means the true hazard ratios in this study are $(e^{\beta}, e^{\eta}, e^{\gamma})^T = (0.5, 0.5, 1)^T$.

3.3.2 Results

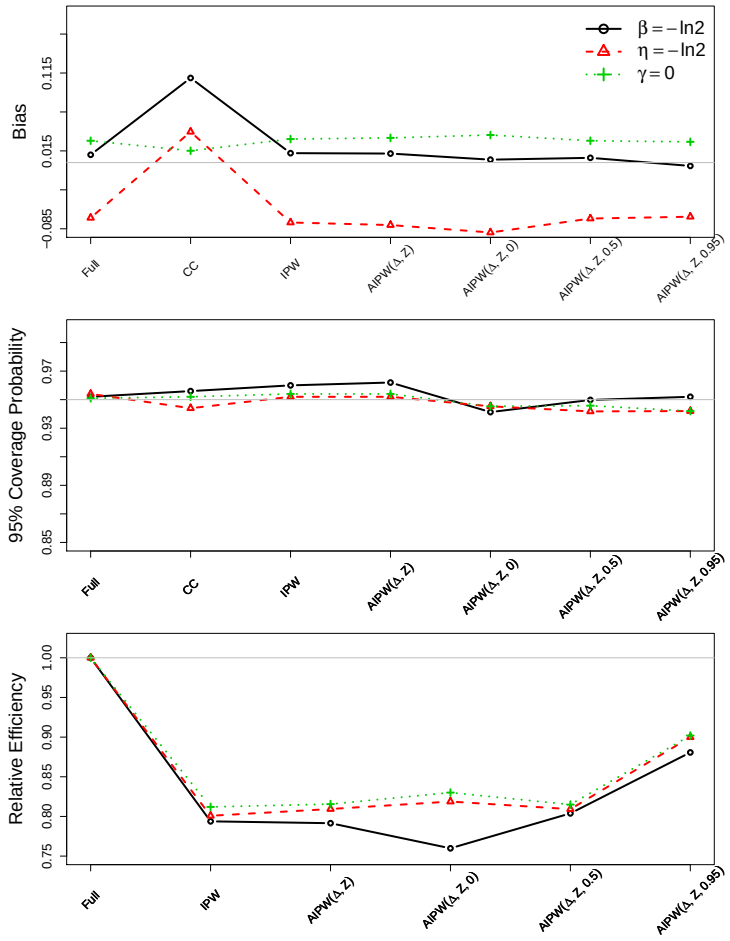
For this interaction model, we observe slightly larger bias and greater variability in $\hat{\eta}$. If the purpose of fitting the interaction model is to test and examine any effect modification of the immune biomarker on the treatment effect, the focus lies on the coefficient γ, which shows negligible bias. For further evaluation of the association between immune biomarker and the event endpoint, we suggest fitting the model by treatment subgroups or pooled vaccine and placebo groups and considering models in Simulation Study II. The efficiency gain from AIPW(Δ, Z, A) is not significant unless A has extremely strong correlation with the immune biomarker variable. The 95% coverage probabilities are very close to the nominal value.

Table 3.9: Simulation results for Simulation Study III: $\lambda(u) = \lambda_0(u)\exp\{\beta X(u) + \eta Z + \gamma X(u)Z)\}$

	R^2	Samp. Prob.	$\beta = -\ln 2$			$\eta = -\ln 2$			$\gamma = 0$		
Method			Bias	SD	ASE	Bias	SD	ASE	Bias	SD	ASE
Full			0.010	0.444	0.422	-0.071	1.713	1.654	0.028	0.704	0.682
CC			0.109	0.424	0.423	0.040	1.701	1.669	0.015	0.698	0.688
IPW		π	0.012	0.498	0.491	-0.077	1.914	1.874	0.030	0.781	0.769
AIPW(Δ, Z)		π	0.012	0.499	0.491	-0.080	1.904	1.871	0.032	0.780	0.769
AIPW(Δ, Z, A)	0	π	0.004	0.509	0.470	-0.090	1.893	1.811	0.035	0.773	0.745
AIPW(Δ, Z, A)	0.5	π	0.006	0.495	0.473	-0.072	1.904	1.819	0.028	0.780	0.749
AIPW(Δ, Z, A)	0.95	π	-0.004	0.473	0.450	-0.070	1.806	1.751	0.027	0.741	0.722
IPW		$\hat{\pi}$	0.012	0.499	0.491	-0.077	1.915	1.874	0.030	0.782	0.769
AIPW(Δ, Z)		$\hat{\pi}$	0.012	0.499	0.492	-0.080	1.904	1.871	0.032	0.780	0.769
AIPW(Δ, Z, A)	0	$\hat{\pi}$	0.002	0.511	0.471	-0.085	1.892	1.813	0.034	0.773	0.746
AIPW(Δ, Z, A)	0.5	$\hat{\pi}$	0.006	0.495	0.474	-0.078	1.908	1.820	0.031	0.781	0.749
AIPW(Δ, Z, A)	0.95	$\hat{\pi}$	-0.004	0.473	0.450	-0.070	1.805	1.751	0.027	0.741	0.722

[1] π: The IPW and AIPW methods are implemented using pre-specified true sampling probabilities. $\hat{\pi}$: The IPW and AIPW methods are implemented using estimated sampling probabilities.
[2] R^2 quantifies the correlation of A and $X(u)$.
[3] For the kernel regression, the standard Normal kernel is used, and the bandwidth for predictor variable P is $0.75N^{-1/3}sd(P)$.
[4] The sampling probability model is $\pi(\Delta) = \Delta + 0.33(1 - \Delta)$.

Figure 3.6: Simulation results for Simulation Study III: $\lambda(u) = \lambda_0(u)\exp\{\beta X(u) + \eta Z + \gamma X(u)Z\}$. The IPW and AIPW estimates are based on pre-specified true sampling probabilities π.



3.4 Discussion

Joint model of longitudinal data measured with error and the event time data has been studied extensively on full cohort data. Here we focus on the conditional score method generalized for the two-phase sampling design cohort studies. We also conduct the simulation studies evaluating the model with an interaction of the time-varying immune biomarker and the treatment indicator. We study the design where only subjects in the second phase sample have the longitudinal records of the immune response data, and subjects outside the second phase sample have their immune biomarker profiles completely missing. Since conditional score method is a semi-parametric method, using the technique of weighting the complete case by the inverse of the sampling probability is a natural way to deal with our problem of missingness. In order to obtain more efficient estimator, we also consider the AIPW technique and assess their performance in finite sample via simulation studies.

It is known that for non-parametric kernel regression, the choice of bandwidth is crucial. The R package `np` has functions that can handle multivariate kernel regression with mixed types of predictor variables (continuous, categorical, binary) and provide the algorithms such as cross-validation to determine the optimal bandwidth. However, it takes considerable computational time in finding the optimal bandwidth. In our AIPW method, it requires fitting the kernel regression across time points, so it becomes even more impractical to seek the optimal bandwidth for each regression. Therefore in our all simulation studies for AIPW method, we use fixed bandwidth determined from the variance of the predictor variables over time and do not explore the influence of different choices of bandwidth. Even so the AIPW method outperforms the IPW method when some auxiliary variables strongly correlated with the biomarker variable are included. However, when the correlation is very weak, including them could reduce the efficiency or even increase the bias in finite sample, especially when the number of immune biomarker measurements per subject is very limited. In that case, the IPW method is recommended.

Our simulation studies on misspecified measurement error models suggest that when the measurement error is not Normal but still random, the conditional score methods (Full, IPW, or AIPW) could lead to slightly biased estimates. More serious problem arises when

the measurement error violates the homoscedasticity assumption. These results suggest that checking if the assumption of random Normal measurement error should be necessary before applying the proposed methods. If evidence of violation of the assumption has been found, transforming the original biomarker variable to meet such an assumption is recommended. However it could be very hard to check this assumption based on observed biomarker data because of the complication in specifying a correct distribution for the inherent true biomarker level first. It might be able to obtain more information when using replicated samples where for every single subject there are repeated measurements at the same time point. Learning from the principles of the assays used to obtain the immune biomarker measurements is also a way to justify or disprove this assumption.

In additional to the simulation studies described above, we also explore different sampling design other than $S1$. We consider various stratified sampling designs among controls based on some strong predictor of the immune biomarker. In doing this we attempt to oversample controls having potentially higher variability in the inherent immune response profiles, and to construct somehow "more efficient" sampling design than $S1$. However, the resulting estimates are not as efficient as those from $S1$, as long as the sampling probabilities and the size of Phase II sample under different designs are controlled to be compatible. It might be because we already include all cases, and they dominate the variance of influence functions. So which controls are selected could provide only very limited influence on the efficiency as long as all cases are included. More exploration in this direction could be considered for a design where not all cases are sampled.

In Chapter 2, we also introduce the framework of natural direct/indirect effects for assessment of time-dependent CoPs. However, we do not conduct simulation studies for detailed evaluation of its performance. We calculate the proportion PCS defined in Section 2.3.2 for data analysis of ACTG 175 in Chapter 6.

Chapter 4

JOINT MODELING FOR DICHOTOMIZED TIME-DEPENDENT BIOMARKERS

4.1 Background

Our experience with the immune response data in HIV-1 and dengue trials shows that in some vaccinated participants the level of immune response declines below the lower quantification limit of the assay during a short period after the final immunization, while for others the level stays positive until the end of follow-up. The immune response could start to have an effect in protection only when its value is above some threshold. This suggests an interest in investigating the binary status of the immune response level (e.g. responder vs. non-responder, high vs. low) as an immune CoR/CoP. The threshold to determine the dichotomization of the immune response level could be obtained from the quantification limit of the assay itself, the study data and prior knowledge. In the vaccine trials we are considering, we do have observations of quantitative immune response levels, only that they are subject to measurement error. So this motivates the need for statistical methods starting with mis-measured quantitative immune biomarker level and ends up with modeling its underlying true dichotomized trajectory over time.

However, most existing joint modeling methods center on modeling continuous longitudinal biomarkers. Few papers have been found on joint modeling for binary longitudinal processes. [Faucett et al., 1998] published their work assuming the observed data were binary and used a Markov model to construct the correlation between two binary data points measured at adjacent time points. Likelihood methods are capable to solve such problems but they usually involve intense numerical integration. Our exploration on the likelihood method actually suggests a serious issue of convergence on this joint modeling framework.

The conditional score method [Tsiatis and Davidian, 2001] and corrected score method [Wang, 2006] developed for continuous biomarkers rely heavily on the properties of ordinary

least squares estimates of the subject-specific random effects. However, the least squares estimates are biased and inefficient for the binary data scenario. Thus there is no straightforward extension of those methods to our binary data model. We therefore explored this problem from the angle of measurement error methods. [Zucker and Spiegelman, 2008] proposed a corrected score method for mis-classified discrete covariates in Cox regression model. They identified a function of the mis-classified covariates whose conditional expectation given the true covariates were asymptotically equivalent to the desired partial likelihood equations. Their method in theory could be extended to our model since we are also interested in the true "binary" covariate. However, because we are dealing with time-varying covariates and the actually observed biomarker values are quantitative, adopting their idea of corrected score method is more complex and would majorly reduce the efficiency since we need to manually dichotomize the observed quantitative first. Another popular analysis approach for the Cox model with mis-measured covariates is the calibration method proposed by [Prentice, 1986]. With an objective of estimating the Cox regression coefficients for the true covariates which are not observed directly, Prentice defined an observed-covariate hazard function, which is obtained by taking the expectation of the true-covariate hazard function, conditioning on the observed biomarkers and being at risk. If the induced observed-covariate hazard function were analytically achievable, then by maximizing the corresponding partial likelihood function we would get the estimates of coefficients. [Wang et al., 2000] has utilized the regression calibration method for the joint modeling framework. However, usually the conditional expectation is in a complicated form which depends on the unknown baseline hazard and coefficient parameters. [Zucker, 2005] proposed a pseudo-partial-likelihood approach and utilized Expectation-Maximization (EM) algorithm to maximize the induced observe-covariate partial likelihood. Again, their model included only time-independent biomarkers and generalizing it to time-varying biomarkers and joint modeling framework is very complicated. So far, most regression calibration methods are conducted by seeking an approximation to the conditional expectations given the true covariates. It is relative simple to implement but at a cost of getting inconsistent estimates. To reduce the bias, a recalibration strategy can be adopted by evaluating the conditional expectations within each at-risk set [Dafni and Tsiatis, 1998, Tsiatis et al., 1995, Xie et al.,

2001]. Even though the bias cannot be eliminated, simulation studies show that the magnitude of the bias could be very small in the rare event setting. We are therefore motivated to use the risk set recalibration (RRC) method to solve our model, since the event rates in the vaccine trials are low. In this chapter, we propose our model in a general way with multiple biomarkers.

4.2 Risk set recalibration method in full cohort studies

4.2.1 Notation and modeling

For each subject, let (V, Δ) denote the observed failure/censoring time and failure status. Assume the K time-varying biomarker processes $\{X_k(u), 0 \le u \le \tau\}$, $k = 1, \cdots, K$ are not observed directly. For each time-varying biomarker process, we observe the mismeasured values of $X_k(u)$ at discrete time points $0 \le T^m_{k1} < \cdots < T^m_{kJ_k} \le V$, denoted by $W_{kj} = X_k(T^m_{kj}) + e_{kj}$, $j = 1, \cdots, J_k$. We assume the errors e_{kj} are normal with zero mean. As in [Song et al., 2002], we also assume non-zero covariance between measurement errors for biomarkers measured at the same time point, i.e. $\mathbb{C}ov(e_{kj}, e_{k'j'}) = \sigma_{kk'} I(T^m_{kj} = T^m_{k'j'})$. Let $\tilde{\sigma} = \{\sigma_{kk'}, k, k' = 1, \cdots, K\}$.

We assume a random effects model for each time-varying biomarker process $X_k(u)$ such that $X_k(u) = \alpha_k^T f_k(u)$, with $f_k(u)$ being a q_k-dimension vector of u. For example $f_k(u) = (1, u)^T$ specifies a random effects model linear in time for the trajectory of $X_k(u)$. We define the design matrix, the vector of observed longitudinal biomarkers and the vector of measurement errors for each subject up to and including time u for $X_k(u)$ as

$$\widetilde{F}_k(u) = \begin{pmatrix} f_k^T(T^m_{k1}) \\ \vdots \\ f_k^T(T^m_{kJ_k(u)}) \end{pmatrix}, \quad \widetilde{W}_k(u) = \begin{pmatrix} W_{k1} \\ \vdots \\ W_{kJ_k(u)} \end{pmatrix}, \quad \widetilde{e}_k(u) = \begin{pmatrix} e_{k1} \\ \vdots \\ e_{kJ_k(u)} \end{pmatrix} \tag{4.1}$$

with $J_k(u)$ indicating the maximum number of measuring time points for $X_k(u)$ up to time u, and $T^m_k(u) = (T^m_{k1}, \cdots, T^m_{kJ_k(u)})^T$ being the vector of measuring time points for $X_k(u)$ up

to and including time u. Obviously for $k = 1, \cdots, K$

$$\widetilde{W}_k(u) = \widetilde{F}_k(u)\alpha_k + \widetilde{e}_k(u)$$

Let $\alpha = \left(\alpha_1^T, \cdots, \alpha_K^T\right)^T$, $\widetilde{W}(u) = \left(\widetilde{W}_1^T(u), \cdots, \widetilde{W}_K^T(u)\right)^T$, $\widetilde{e}(u) = \left(\widetilde{e}_1^T(u), \cdots, \widetilde{e}_K^T(u)\right)^T$, $T^m(u) = (T_1^{mT}(u), \cdots, T_K^{mT}(u))^T$, $J(u) = (J_1(u), \cdots, J_K(u))^T$, and

$$\widetilde{F}(u) = \begin{pmatrix} \widetilde{F}_1(u) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \widetilde{F}_K(u) \end{pmatrix} \tag{4.2}$$

Then

$$\widetilde{W}(u) = \widetilde{F}(u)\alpha + \widetilde{e}(u) \tag{4.3}$$

For a pre-specified cutoff value b_k, we define the binary indicator process $B_k(u)$ as

$$B_k(u) = I(X_k(u) \geq b_k) \tag{4.4}$$

and the vector of binary biomarker process as $B(u) = (B_1(u), \cdots, B_K(u))^T$. Then we study the proportional hazards model

$$\begin{aligned} \lambda(u; \alpha, \tilde{Z}) &= \lim_{du \to 0} \frac{1}{du} \mathbb{P}(u \leq T < u + du | \alpha, \tilde{Z}, T^m, C, T \geq u) \\ &= \lim_{du \to 0} \frac{1}{du} \mathbb{P}(u \leq T < u + du | \alpha, \tilde{Z}, T \geq u) \\ &= \lambda_0(u) \exp\{\beta^T B(u) + \eta^T \tilde{Z} + \gamma^T B(u) Z\} \end{aligned} \tag{4.5}$$

where $\tilde{Z} = (Z, L^T)^T$. Z is a binary indicator of vaccination arm (1 for vaccine and 0 for placebo), L is a vector of $p-1$ dimensional vector of potential baseline confounding variables. Here we restrict our attention to L being categorical variables, since we need to estimate the distribution of unobserved random effects α given $\tilde{Z}$. The hazard function implies that the measurement time points and censoring time are not informative to the hazard given α

and $\tilde{Z}$. It also says that only whether or not the biomarkers values $X(u)$ being above the thresholds matter for the hazard.

4.2.2 Ideal risk set recalibration estimator

Now we outline the risk set recalibration (RRC) method to estimate $\theta = (\beta^T, \eta^T, \gamma^T)^T$. At any given time u, define the at risk process as $Y(u) = I(V \geq u)$ and the increment of event process $dN(u) = I(V = u, \Delta = 1)$. Under the assumption of non-informative censoring given α and $\tilde{Z}$, the hazard for V is the same as that for T. Also, as discussed in [Prentice, 1982], we assume the hazard is independent of $\{W, T^m\}$ given α and $\tilde{Z}$. Then the induced hazard function from (4.5) conditional on the observed covariates and being at risk is

$$\begin{aligned}
\lambda(u; \widetilde{W}(u), T^m(u), \tilde{Z}) &= \lim_{du \to 0} \frac{1}{du} \mathbb{P}(u \leq V < u + du | \widetilde{W}(u), T^m(u), \tilde{Z}, Y(u) = 1) \\
&= \lim_{du \to 0} \frac{1}{du} \int \mathbb{P}(u \leq V < u + du | \widetilde{W}(u), T^m(u), \alpha, \tilde{Z}, Y(u) = 1) \\
&\qquad p(\alpha | \widetilde{W}(u), T^m(u), \tilde{Z}, Y(u) = 1) d\alpha \\
&= \int \lambda(u; | \widetilde{W}(u), T^m(u), \alpha, \tilde{Z}) p(\alpha | \widetilde{W}(u), T^m(u), \tilde{Z}, Y(u) = 1) d\alpha \\
&= \int \lambda(u; | \alpha, \tilde{Z}) p(\alpha | \widetilde{W}(u), T^m(u), \tilde{Z}, Y(u) = 1) d\alpha \\
&= \mathbb{E}\left[\lambda(u; \alpha, \tilde{Z}) | \widetilde{W}(u), T^m(u), \tilde{Z}, Y(u) = 1 \right] \\
&= \lambda_0(u) R^0(u, \theta) \qquad (4.6)
\end{aligned}$$

where

$$R^0(u, \theta) = \exp\{\eta^T \tilde{Z}\} \mathbb{E}\left[\exp\{\beta^T B(u) + \gamma^T B(u) Z\} | \widetilde{W}(u), T^m(u), \tilde{Z}, Y(u) = 1 \right] \qquad (4.7)$$

Let $\{V_i, \Delta_i, W_i, T_i^m, J_i, \tilde{Z}_i, e_i\}, i = 1, \cdots, N$ be a random sample. Then the corresponding induced partial-likelihood can be written as

$$L(\theta) = \prod_{i=1}^{N} \left[\frac{R_i^0(V_i, \theta)}{\sum_{j=1}^{N} Y_j(V_i) R_j^0(V_i, \theta)} \right]^{\Delta_i} \qquad (4.8)$$

and we can estimate θ by solving $\partial \log L(\theta)/\partial\theta = 0$, i.e.

$$\sum_{i=1}^{N} \int_0^{\tau} \left\{ \frac{\dot{R}_i^0(u,\theta)}{R_i^0(u,\theta)} - \frac{\sum_{j=1}^{N} Y_j(u)\dot{R}_j^0(u,\theta)}{\sum_{j=1}^{N} Y_j(u) R_j^0(u,\theta)} \right\} dN_i(u) = 0 \tag{4.9}$$

where $\dot{R}_i^0(u,\theta) = \partial R_i^0(u,\theta)/\partial\theta$.

In practice the analytical form of $R^0(u,\theta)$ is unobtainable so (4.9) are unobtainable ideal score equations for estimating θ. As a result, in the next subsection, we propose a working assumption to solve the problem.

4.2.3 RRC estimating equations

We assume α is independent of $\{T^m(u), J(u)\}$ given $\{Y(u) = 1, \tilde{Z}\}$ and is independent of the measurement errors $\widetilde{e}(u)$. Let α's mean and covariance matrix conditioning on $\{Y(u) = 1, \tilde{Z}\}$ be $\mu(u,\tilde{Z})$ and $\Sigma(u,\tilde{Z})$ where

$$\mu(u,\tilde{Z}) = \begin{pmatrix} \mu_1(u,\tilde{Z}) \\ \vdots \\ \mu_K(u,\tilde{Z}) \end{pmatrix}, \quad \Sigma(u,\tilde{Z}) = \begin{pmatrix} \Sigma_{11}(u,\tilde{Z}) & \cdots & \Sigma_{1K}(u,\tilde{Z}) \\ \vdots & \ddots & \vdots \\ \Sigma_{K1}(u,\tilde{Z}) & \cdots & \Sigma_{KK}(u,\tilde{Z}) \end{pmatrix} \tag{4.10}$$

$\mathbb{E}[\alpha_k | Y(u) = 1, \tilde{Z}] = \mu_k(u,\tilde{Z})$, and $\mathbb{C}ov[\alpha_k, \alpha_{k'} | Y(u) = 1, \tilde{Z}] = \Sigma_{kk'}(u,\tilde{Z})$. Then the mean and covariance matrix for $(\alpha_k^T, \widetilde{W}^T(u))^T$ given $\{Y(u) = 1, \tilde{Z}, T^m(u), J(u)\}$ are

$$\begin{pmatrix} \mu_k(u,\tilde{Z}) \\ \widetilde{F}_1(u)\mu_1(u,\tilde{Z}) \\ \\ \widetilde{F}_K(u)\mu_K(u,\tilde{Z}) \end{pmatrix}, \quad \begin{pmatrix} \Sigma_{kk}(u,\tilde{Z}) & \Sigma_{k1}(u,\tilde{Z})\widetilde{F}_1^T(u) & \cdots & \Sigma_{kK}(u,\tilde{Z})\widetilde{F}_K^T(u) \\ \widetilde{F}_1(u)\Sigma_{1k}(u,\tilde{Z}) & \Gamma_{11}(u,\tilde{Z}) & \cdots & \Gamma_{1K}(u,\tilde{Z}) \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{F}_K(u)\Sigma_{Kk}(u,\tilde{Z}) & \Gamma_{1K}(u,\tilde{Z}) & \cdots & \Gamma_{KK}(u,\tilde{Z}) \end{pmatrix} \tag{4.11}$$

where $\Gamma_{kk'}(u,\tilde{Z}) = \widetilde{F}_k(u)\Sigma_{kk'}(u,\tilde{Z})\widetilde{F}_{k'}^T(u) + A_{kk'}\sigma_{kk'}$ and $A_{kk'}$ is a $J_k(u) \times J_{k'}(u)$ matrix with $(a,b) - th$ element equal to 1 if $T_{ka}^m = T_{k'b}^m$, and 0 otherwise. Let

$$\Sigma_{k.}(u,\tilde{Z}) = \begin{pmatrix} \Sigma_{k1}(u,\tilde{Z}) \\ \vdots \\ \Sigma_{kK}(u,\tilde{Z}) \end{pmatrix}, \qquad \Sigma_{.k}(u,\tilde{Z}) = \begin{pmatrix} \Sigma_{1k}(u,\tilde{Z}) & \cdots & \Sigma_{Kk}(u,\tilde{Z}) \end{pmatrix}$$

$$\Gamma(u,\tilde{Z}) = \begin{pmatrix} \Gamma_{11}(u,\tilde{Z}) & \cdots & \Gamma_{1K}(u,\tilde{Z}) \\ \vdots & \ddots & \vdots \\ \Gamma_{1K}(u,\tilde{Z}) & \cdots & \Gamma_{KK}(u,\tilde{Z}) \end{pmatrix}$$

Write $\mathcal{C} = \{Y(u) = 1, \tilde{Z}, T^m(u)\}$, $\mathbb{P}_{\mathcal{C}}[\cdot] = \mathbb{P}[\cdot|\mathcal{C}]$, $\mathbb{E}_{\mathcal{C}}[\cdot] = \mathbb{E}[\cdot|\mathcal{C}]$, and $\mathbb{C}ov_{\mathcal{C}}[\cdot] = \mathbb{C}ov[\cdot|\mathcal{C}]$. If we further make the normality assumption, then it follows that

$$\mu_{\mathcal{C}}(u) \equiv \mathbb{E}_{\mathcal{C}}\left[\alpha|\widetilde{W}(u)\right] = \mu(u,\tilde{Z}) + \Sigma(u,\tilde{Z})\widetilde{F}^T(u)\Gamma^{-1}(u,\tilde{Z})\left(\widetilde{W}(u) - \widetilde{F}(u)\mu(u,\tilde{Z})\right) \tag{4.12}$$

$$\Sigma_{\mathcal{C}}(u) \equiv \mathbb{C}ov_{\mathcal{C}}\left[\alpha|\widetilde{W}(u)\right] = \Sigma(u,\tilde{Z}) - \Sigma(u,\tilde{Z})\widetilde{F}^T(u)\Gamma^{-1}(u,\tilde{Z})\widetilde{F}(u)\Sigma(u,\tilde{Z}) \tag{4.13}$$

In other words, $\mu_{\mathcal{C}}(u)$ and $\Sigma_{\mathcal{C}}(u)$ fully specify the conditional distribution of α given $\{Y(u) = 1, \tilde{Z}, T^m(u), \widetilde{W}(u)\}$. So we are able to calculate for $m = (m_1, \cdots, m_K)^T$, $m_1, \cdots, m_K = 0, 1$,

$$p_{\mathcal{C}}(m; u, \mu, \Sigma) \equiv \mathbb{P}_{\mathcal{C}}\left[B_1(u) = m_1, B_2(u) = m_2, \cdots, B_K(u) = m_K|\widetilde{W}(u)\right] \tag{4.14}$$

which further yields the risk function in (4.7) as

$$\begin{aligned} R(u,\theta,\mu,\Sigma) &= \exp\{\eta^T\tilde{Z}\}\mathbb{E}\left[\exp\{\beta^T B(u) + \gamma^T B(u)Z\}|\widetilde{W}(u), T^m(u), \tilde{Z}, Y(u) = 1\right] \\ &= \exp\{\eta^T\tilde{Z}\}\left[\sum_{m_1,\cdots,m_K=0,1} \exp\{\sum_{k=1}^K(\beta_k + \gamma_k Z)m_k\}p_{\mathcal{C}}(m; u, \mu, \Sigma)\right] \end{aligned} \tag{4.15}$$

Note unlike in [Dafni and Tsiatis, 1998, Wang et al., 2000, Xie et al., 2001] where $R^0(u,\theta)$

is approximated by $\exp\left\{\eta^T\tilde{Z}+(\beta^T+\gamma^T Z)^T\mathbb{E}\left[B(u)|\widetilde{W}(u),T^m(u),\tilde{Z},Y(u)=1\right]\right\}$, here we derive the expression of $R^0(u,\theta)$ directly from the assumed conditional distribution of α given observed covariates and being at risk. Since we have made the normality assumption, the form of $R^0(u,\theta)$ depends on the nuisance parameters μ and Σ. Therefore we use the notation $R(u,\theta,\mu,\Sigma)$ in our proposed method accordingly.

Moreover (4.15) contains unknown nuisance parameters $\mu(u,\tilde{Z})$, $\Sigma(u,\tilde{Z})$ and $\tilde{\sigma}=\{\sigma_{kk'};$ $k,k'=1,\cdots,K\}$. Therefore, one more step is needed to estimate them before we solve the estimating equations for θ.

To estimate the covariance $\sigma_{kk'}$, we can do it similarly to that in [Song et al., 2002] by solving

$$\begin{aligned} S_{e,F}(\sigma_{kk'}) &= \sum_{i=1}^{N} I(J_{ik}\geq q_k, J_{ik'}\geq q_{k'})\left\{\left(\widetilde{W}_{ik}(V_i)-\widetilde{F}_{ik}(V_i)\hat{\alpha}_{ik}(V_i)\right)^T A_{ikk'}\right. \\ &\quad \left.\left(\widetilde{W}_{ik'}(V_i)-\widetilde{F}_{ik'}(V_i)\hat{\alpha}_{ik'}(V_i)\right)-\sigma_{kk'}tr\left(P_{ik}A_{ikk'}P_{ik'}A_{kk'}^T\right)\right\}=0 \quad (4.16) \end{aligned}$$

where $P_{ik}=I_{J_{ik}}-\widetilde{F}_{ik}(V_i)\{\widetilde{F}_{ik}^T(V_i)\widetilde{F}_{ik}(V_i)\}^{-1}\widetilde{F}_{ik}^T(V_i)$. To simplify the notations and formula, we establish the estimating equations for θ from now on assuming $\tilde{\sigma}$ are known. All the asymptotic theories can be extended to the case with $\tilde{\sigma}$ unknown by considering the estimating equations for θ and for $\sigma_{kk'}$ simultaneously.

Since $\mu_k(u,\tilde{Z})$ and $\Sigma_{kk'}(u,\tilde{Z})$ are $\tilde{Z}$-dependent and $\tilde{Z}$ are categorical variables, we propose to estimate them separately within each category. Suppose there are ν discrete values of $\tilde{Z}$, $\tilde{z}_1,\cdots,\tilde{z}_\nu$. Let $\mu(u)=\{\mu_{(j)}(u), j=1,\cdots,\nu\}$ and $\Sigma(u)=\{\Sigma_{(j)}(u), j=1,\cdots,\nu\}$, where for each $\tilde{z}_j$, $\mu_{(j)}(u)=\mu(u,\tilde{z}_j)$ and $\Sigma_{(j)}(u)=\Sigma(u,\tilde{z}_j)$. We would like to obtain estimates that satisfy

$$\sup_{u\in[0,\tau]}|vec\{\hat{\mu}(u)-\mu(u)\}|\xrightarrow{p}0,\quad \sup_{u\in[0,\tau]}\left|vec\{\hat{\Sigma}(u)-\Sigma(u)\}\right|\xrightarrow{p}0 \quad (4.17)$$

$$N^{1/2}vec\{\hat{\mu}(u)-\mu(u)\}=N^{-1/2}\sum_{i=1}^{N}\phi_i(u)+o_p(1) \quad (4.18)$$

$$N^{1/2}vec\{\hat{\Sigma}(u)-\Sigma(u)\}=N^{-1/2}\sum_{i=1}^{N}\psi_i(u)+o_p(1) \quad (4.19)$$

as $N \to \infty$, with $\phi_i(u)$ and $\psi_i(u)$ the influence functions and *vec* the vectorization of the matrices.

The problem with estimating the distribution of α is that α itself is not observed directly. So the usual approach such as regressing α on the observed data or the kernel density method do not work out here. However since $\mu(u)$ and $\Sigma(u)$ are nuisance parameters related to the observed biomarker values $\widetilde{W}$, we could maximize the likelihood of the observed data based on the random effects model. Another way to estimate $\mu_{(j)}(u)$ and $\Sigma_{(j)}(u)$ is by the method of moments in a way similar to that in [Dafni and Tsiatis, 1998]. Let $\hat{\alpha}_i(u)$ denote the least squares estimates of subject-specific random effects α_i using data up to and including time u, i.e. $\hat{\alpha}_i(u) = \{\widetilde{F}_i^T(u)\widetilde{F}_i(u)\}^{-1}\widetilde{F}_i^T(u)\widetilde{W}_i(u)$. It can be easily verified that $\hat{\alpha}_i(u) = (\hat{\alpha}_{i1}^T(u), \cdots, \hat{\alpha}_{iK}^T(u))^T$, with $\hat{\alpha}_{ik}^T(u) = \{\widetilde{F}_{ik}^T(u)\widetilde{F}_{ik}(u)\}^{-1}\widetilde{F}_{ik}^T(u)\widetilde{W}_{ik}(u)$ being the least squares estimate for α_{ik}, the random effects governing the $k-th$ underlying biomarker process. Since $\hat{\alpha}_{ik}(u) = \alpha_{ik} + \{\widetilde{F}_{ik}^T(u)\widetilde{F}_{ik}(u)\}^{-1}\widetilde{F}_{ik}^T(u)\widetilde{e}_{ik}(u)$, we have $\mathbb{E}_{\mathcal{C}}[\hat{\alpha}_{ik}(u)] = \mu_k(u, \tilde{Z})$, and $\mathbb{C}ov_{\mathcal{C}}[\hat{\alpha}_{ik}(u), \hat{\alpha}_{ik'}(u)] = \Sigma_{kk'}(u, \tilde{Z}_i) + \Sigma_{R_{ikk'}}(u)$, with $\Sigma_{R_{ikk'}}(u) = \{\widetilde{F}_{ik}^T(u)\widetilde{F}_{ik}(u)\}^{-1}\widetilde{F}_{ik}^T(u)A_{ikk'}(u)\widetilde{F}_{ik'}(u)\{\widetilde{F}_{ik'}^T(u)\widetilde{F}_{ik'}(u)\}^{-1}\sigma_{kk'}$. Therefore naturally the estimating equations for $\mu_k(u, \tilde{z})$ and $\Sigma_{kk'}(u, \tilde{z})$ for the category with $\tilde{Z} = \tilde{z}$ are

$$\sum_{i=1}^{N} Y_i(u) I(J_{ik}(u) \geq q_k) I(\tilde{Z}_i = \tilde{z})(\hat{\alpha}_{ik}(u) - \mu_k(u, \tilde{z})) = 0 \tag{4.20}$$

$$\sum_{i=1}^{N} Y_i(u) I(J_{ik}(u) \geq q_k, J_{ik'}(u) \geq q_{k'}) I(\tilde{Z}_i = \tilde{z}) \Big\{ [\hat{\alpha}_{ik}(u) - \mu_k(u, \tilde{z})] [\hat{\alpha}_{ik'}(u) - \mu_{k'}(u, \tilde{z})]^T - \Sigma_{kk'}(u, \tilde{z}) - \Sigma_{R_{ikk'}}(u) \Big\} = 0 \tag{4.21}$$

Apparently, under certain regularity conditions, (4.17)(4.18)(4.19) hold for the estimates of $\mu(u)$ and $\Sigma(u)$ from the likelihood approach or the method of moments.

Plugging $\hat{\mu}(u)$ and $\hat{\Sigma}(u)$ back into (4.12)(4.13) and (4.14) we obtain

$$\hat{\mu}_{\mathcal{C}_i}(u) = \hat{\mu}(u, \tilde{Z}_i) + \hat{\Sigma}(u, \tilde{Z}_i)\widetilde{F}_i^T(u)\hat{\Gamma}^{-1}(u, \tilde{Z}_i)\left(\widetilde{W}_i(u) - \widetilde{F}_i(u)\hat{\mu}(u, \tilde{Z}_i)\right) \tag{4.22}$$

$$\hat{\Sigma}_{\mathcal{C}_i}(u) = \hat{\Sigma}(u, \tilde{Z}_i) - \hat{\Sigma}(u, \tilde{Z}_i)\widetilde{F}_i^T(u)\hat{\Gamma}^{-1}(u, \tilde{Z}_i)\widetilde{F}_i(u)\hat{\Sigma}(u, \tilde{Z}_i) \tag{4.23}$$

$$\hat{p}_{\mathcal{C}_i}(m; u, \hat{\mu}, \hat{\Sigma}) = \hat{p}_{\mathcal{C}_i}(m; u, \mu = \hat{\mu}(u), \Sigma = \hat{\Sigma}(u)) \tag{4.24}$$

And finally, solving the following estimating equations yields the RRC estimates $\hat{\theta}^R$ for θ

$$U_F^R(\theta) = \sum_{i=1}^{N} \int \left\{ \frac{\hat{\dot{R}}_i(u,\theta,\hat{\mu},\hat{\Sigma})}{\hat{R}_i(u,\theta,\hat{\mu},\hat{\Sigma})} - \frac{\sum_{j=1}^{N} Y_j(u)\hat{\dot{R}}_j(u,\theta,\hat{\mu},\hat{\Sigma})}{\sum_{j=1}^{N} Y_j(u)\hat{R}_j(u,\theta,\hat{\mu},\hat{\Sigma})} \right\} dN_i(u) = 0 \tag{4.25}$$

where

$$\begin{aligned}
\hat{R}_i(u,\theta,\hat{\mu},\hat{\Sigma}) &= \exp\{\eta^T \tilde{Z}_i\} \left[\sum_m \exp\{\beta^T m + \gamma^T m Z_i\} \hat{P}_{\mathcal{C}_i}(m;u,\hat{\mu},\hat{\Sigma}) \right] \\
\hat{\dot{R}}_i(u,\theta,\hat{\mu},\hat{\Sigma}) &= \frac{\partial \hat{\dot{R}}_i(u,\theta,\hat{\mu},\hat{\Sigma})}{\partial \theta} \\
&= \begin{pmatrix} \exp\{\eta^T \tilde{Z}_i\} \left[\sum_m m \exp\{\beta^T m + \gamma^T m Z_i\} \hat{P}_{\mathcal{C}_i}(m;u,\hat{\mu},\hat{\Sigma}) \right] \\ \tilde{Z}_i \exp\{\eta^T \tilde{Z}_i\} \left[\sum_m \exp\{\beta^T m + \gamma^T m Z_i\} \hat{P}_{\mathcal{C}_i}(m;u,\hat{\mu},\hat{\Sigma}) \right] \\ Z_i \exp\{\eta^T \tilde{Z}_i\} \left[\sum_m m \exp\{\beta^T m + \gamma^T m Z_i\} \hat{P}_{\mathcal{C}_i}(m;u,\hat{\mu},\hat{\Sigma}) \right] \end{pmatrix}
\end{aligned}$$

4.2.4 Asymptotic distribution theory

In this section, we assume the regularity conditions given in Assumption D hold. Recall $\theta = (\beta^T, \eta^T, \gamma^T)^T$ and $\tilde{\sigma} = \{\sigma_{kk'}, k, k' = 1, \cdots, K\}$. Further define

$$A_F^{(0)}(u,\theta,\hat{\mu},\hat{\Sigma}) = N^{-1} \sum_{i=1}^{N} Y_i(u) \hat{R}_i(u,\theta,\hat{\mu},\hat{\Sigma}), \qquad a^{(0)}(u,\theta,\mu,\Sigma) = \mathbb{E}\left[Y(u)R(u,\theta,\mu,\Sigma)\right]$$

$$A_F^{(1)}(u,\theta,\hat{\mu},\hat{\Sigma}) = N^{-1} \sum_{i=1}^{N} Y_i(u) \hat{\dot{R}}_i(u,\theta,\hat{\mu},\hat{\Sigma}), \qquad a^{(1)}(u,\theta,\mu,\Sigma) = \mathbb{E}\left[Y(u)\dot{R}(u,\theta,\mu,\Sigma)\right]$$

$$A_F^{(2)}(u,\theta,\hat{\mu},\hat{\Sigma}) = N^{-1} \sum_{i=1}^{N} Y_i(u) \hat{\ddot{R}}_i(u,\theta,\hat{\mu},\hat{\Sigma}), \qquad a^{(2)}(u,\theta,\mu,\Sigma) = \mathbb{E}\left[Y(u)\ddot{R}(u,\theta,\mu,\Sigma)\right]$$

$$\ddot{R}(u,\theta,\mu,\Sigma) = \frac{\partial \dot{R}(u,\theta,\mu,\Sigma)}{\partial \theta}$$

$$b^{(0)}(u,\theta) = \lambda_0(u)\mathbb{E}\left\{ Y(u)\mathbb{E}\left[\exp\{\beta^T B(u) + \eta^T \tilde{Z} + \gamma^T B(u) Z\} \,\middle|\, \widetilde{W}(u), \tilde{Z}, Y(u) = 1 \right] \right\}$$

$$b^{(1)}(u,\theta) = \lambda_0(u)\mathbb{E}\left\{ Y(u) \frac{\dot{R}(u,\theta,\mu,\Sigma)}{R(u,\theta,\mu,\Sigma)} \mathbb{E}\left[\exp\{\beta^T B(u) + \eta^T \tilde{Z} + \gamma^T B(u) Z\} \,\middle|\, \widetilde{W}(u), \tilde{Z}, Y(u) = 1 \right] \right\}$$

$$H_i(\theta) = \int_0^\tau \left\{ \frac{\dot{R}_i(u,\theta,\mu,\Sigma)}{R_i(u,\theta,\mu,\Sigma)} - \frac{a^{(1)}(u,\theta,\mu,\Sigma)}{a^{(0)}(u,\theta,\mu,\Sigma)} \right\} dN_i(u)$$

$$h(\theta) = \int_0^\tau \left\{ b^{(1)}(u,\theta) - \frac{a^{(1)}(u,\theta,\mu,\Sigma)}{a^{(0)}(u,\theta,\mu,\Sigma)} b^{(0)}(u,\theta) \right\} du$$

Assumption D

D1. $\Lambda_0(\tau) < \infty$, $P(Y(\tau) = 1) > 0$.

D2. T and $(C, T^m)^T$ are independent given $(\alpha^T, \tilde{Z}^T)^T$

D3. $\tilde{Z}$ and J are bounded. The support of discrete variable $\tilde{Z}$ has fixed and finite number of values and for each value $\tilde{z}$, $u \in [0,\tau]$ and $k, k' = 1, \cdots, K$, $\mathbb{P}(Y(u) = 1, I(J_k(u) \geq q_k, I(J_{k'}(u) \geq q_{k'}) | \tilde{Z} = \tilde{z}) > 0$.

D4. (4.17)(4.18)(4.19) hold for $\hat{\mu}(u)$ and $\hat{\Sigma}(u)$.

D5. $h(\theta) = 0$ if and only if $\theta = \theta^*$.

D6. There exists an compact set Θ where θ^* lies in the interior, such that $a^{(r)}(u,\theta,\mu,\Sigma)$, $b^{(r)}(u,\theta)$ and $\dot{b}^{(r)}(u,\theta)$ exists and continuous in $(u,\theta) \in \Theta \times [0,\tau]$, $r = 0, 1, 2$.

D7. $\mathbb{E}\left[\sup_{(u,\theta)\in[0,\tau]\times\Theta} |Y(u)R^0(u,\theta)|\right] < \infty$,$\mathbb{E}\left[\sup_{(u,\theta)\in[0,\tau]\times\Theta} |Y(u)R^0(u,\theta)R^0(u,\theta)^T|\right] < \infty$

D8. $\mathbb{E}[\dot{H}(\theta^*)]$ exists and invertible.

D9. $\mathbb{V}ar[M(\theta^*)]$ is finite and positive definite where $M(\theta^*)$ is defined in (4.26).

Lemma 4.2.1. *For $r = 0, 1$, $\sup_{(u,\theta)\in[0,\tau]\times\Theta} \left| A_F^{(r)}(u,\theta,\hat{\mu},\hat{\Sigma}) - a^{(r)}(u,\theta,\mu,\Sigma) \right| \xrightarrow{p} 0$, as $N \to \infty$.*

Proof. We only outline the proof for $r = 0$. The other one for $r = 1$ can be proved similarly. By Theorem III.1 in [Andersen and Gill, 1982] we have $A_F^{(0)}(u,\theta,\mu,\Sigma) = a^{(r)}(u,\theta,\mu,\Sigma) + o_p(1)$ uniformly in u, θ. Therefore we only need to prove that $A_F^{(0)}(u,\theta,\hat{\mu},\hat{\Sigma}) = A_F^{(0)}(u,\theta,\mu,\Sigma) + o_p(1)$ uniformly in u, θ. Actually

$$\begin{aligned}
&\sup_{(u,\theta)\in[0,\tau]\times\Theta} \left| A_F^{(0)}(u,\theta,\hat{\mu},\hat{\Sigma}) - A_F^{(0)}(u,\theta,\mu,\Sigma) \right| \\
= &\sup_{(u,\theta)\in[0,\tau]\times\Theta} \Bigg| N^{-1} \sum_{j=1}^{\nu} \sum_{i=1}^{N} I(\tilde{Z}_i = \tilde{z}_j) Y_i(u) \exp\{\eta^T \tilde{z}_j\} \sum_m \exp\{\beta^T m + \gamma^T m z_j\} \times \\
&\qquad \left[\hat{p}_{\mathcal{C}_i}(m; u, \hat{\mu}_{(j)}, \hat{\Sigma}_{(j)}) - p_{\mathcal{C}_i}(m; u, \mu_{(j)}, \Sigma_{(j)}) \right] \Bigg|
\end{aligned}$$

$$\leq \sum_{j=1}^{\nu}\sum_{m}\left\{\sup_{\theta\in\Theta}\left|\exp\left\{\beta^T m+\eta^T\tilde{z}_j+\gamma^T m z_j\right\}\right|\times N^{-1}\sum_{i=1}^{N}\sup_{u\in[0,\tau]}\left|\hat{p}_{\mathcal{C}_i}(m;u,\hat{\mu}_{(j)},\hat{\Sigma}_{(j)})-p_{\mathcal{C}_i}(m;u,\mu_{(j)},\Sigma_{(j)})\right|\right\}$$

The second term converges to zero because of the uniform convergence of $\hat{\mu}_{(j)}$, $\hat{\Sigma}_{(j)}$ and the continuity of $p_{\mathcal{C}}(m,u,\mu,\Sigma)$ (in the form of Normal cumulative density function) in μ and Σ. On the other hand, since ν and the number of all possible values of m are fixed finite, $A_F^{(0)}(u,\theta,\hat{\mu},\hat{\Sigma})=A_F^{(0)}(u,\theta,\mu,\Sigma)+o_p(1)$ uniformly in u,θ. □

Theorem 4.2.2. *Under Assumption D, as $N\to\infty$, (i) $\hat{\theta}^R\xrightarrow{p}\theta^*$; (ii) $N^{1/2}(\hat{\theta}^R-\theta^*)$ converges weakly to a Normal random variate with mean zero and covariance $A^{-1}B(A^{-1})^T$, where*

$$A=\mathbb{E}[\dot{H}(\theta^*)],\qquad B=\mathbb{E}[M(\theta^*)M(\theta^*)^T]$$

and $M_i(\theta^*)$ is defined in (4.26).

Proof. (i) We give a sketch of proof similarly to that in [Xie et al., 2001]. Note that

$$\begin{aligned}
&\mathbb{E}\left[\frac{\dot{R}(u,\theta,\mu,\Sigma)}{R(u,\theta,\mu,\Sigma)}dN(u)\right]\\
=\;&\mathbb{E}\left\{\mathbb{E}\left[\frac{\dot{R}(u,\theta,\mu,\Sigma)}{R(u,\theta,\mu,\Sigma)}dN(u)\,\middle|\,\widetilde{W}(u),\tilde{Z}\right]\right\}\\
=\;&\mathbb{E}\left\{\mathbb{E}\left[\mathbb{E}\left(\frac{\dot{R}(u,\theta,\mu,\Sigma)}{R(u,\theta,\mu,\Sigma)}dN(u)\,\middle|\,\widetilde{W}(u),\tilde{Z},Y(u)\right)\middle|\,\widetilde{W}(u),\tilde{Z}\right]\right\}\\
=\;&\mathbb{E}\left\{\frac{\dot{R}(u,\theta,\mu,\Sigma)}{R(u,\theta,\mu,\Sigma)}\mathbb{E}\left[dN(u)\,\middle|\,\widetilde{W}(u),\tilde{Z},Y(u)=1\right]\mathbb{E}\left[Y(u)\,\middle|\,\widetilde{W}(u),\tilde{Z}\right]\right\}\\
=\;&\lambda_0(u)du\mathbb{E}\left\{Y(u)\frac{\dot{R}(u,\theta,\mu,\Sigma)}{R(u,\theta,\mu,\Sigma)}\mathbb{E}\left[\exp\{\beta^T B(u)+\eta^T\tilde{Z}+\gamma^T B(u)Z\}\mid\widetilde{W}(u),\right.\right.\\
&\left.\left.\tilde{Z},Y(u)=1\right]\right\}\\
\equiv\;&b^{(1)}(u,\theta)du
\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[dN(u)\right] &= \lambda_0(u)du\mathbb{E}\left\{Y(u)\mathbb{E}\left[\exp\{\beta^T B(u)+\eta^T\tilde{Z}+\gamma^T B(u)Z\}\left|\widetilde{W}(u),\tilde{Z},Y(u)=1\right.\right]\right\}\\ &\equiv b^{(0)}(u,\theta)du\end{aligned}$$

Let $\bar{N}(u)=N^{-1}\sum_{i=1}^{N}N_i(u)$. Then $N^{-1}U_F^R(\theta)=N^{-1}U_{1N}(\theta)+N^{-1}U_{2N}(\theta)-N^{-1}U_{3N}(\theta)$ where

$$\begin{aligned}N^{-1}U_{1N}(\theta)&=N^{-1}\sum_{i=1}^{N}H_i(\theta)\\ N^{-1}U_{2N}(\theta)&=N^{-1}\sum_{i=1}^{N}\int_0^\tau\left(\frac{\hat{\dot{R}}_i(u,\theta,\hat{\mu},\hat{\Sigma})}{\hat{R}_i(u,\theta,\hat{\mu},\hat{\Sigma})}-\frac{\dot{R}_i(u,\theta,\mu,\Sigma)}{R_i(u,\theta,\mu,\Sigma)}\right)dN_i(u)\\ N^{-1}U_{3N}(\theta)&=\int_0^\tau\left(\frac{A_F^{(1)}(u,\theta,\hat{\mu},\hat{\Sigma})}{A_F^{(0)}(u,\theta,\hat{\mu},\hat{\Sigma})}-\frac{a^{(1)}(u,\theta,\mu,\Sigma)}{a^{(0)}(u,\theta,\mu,\Sigma)}\right)d\bar{N}(u)\end{aligned}$$

We investigate each of the three terms. For the second term, we can prove as in Lemma 4.2.1 that

$$\sup_{(u,\theta)\in[0,\tau]\times\Theta}\left|N^{-1}\sum_{i=1}^{N}\frac{\hat{\dot{R}}_i(u,\theta,\hat{\mu},\hat{\Sigma})}{\hat{R}_i(u,\theta,\hat{\mu},\hat{\Sigma})}-N^{-1}\sum_{i=1}^{N}\frac{\dot{R}_i(u,\theta,\mu,\Sigma)}{R_i(u,\theta,\mu,\Sigma)}\right|=o_p(1)$$

This implies that $\sup_{\theta\in\Theta}|N^{-1}U_{2N}(\theta)|=o_p(1)$.

By the arguments similar to those in [Fleming and Harrington, 1991] (page 305-306), we can prove that $a^{(0)}(u,\theta,\mu,\Sigma)$ is bounded away from zero on $[0,\tau]\times\Theta$. Together with Lemma 4.2.1, we have $\sup_{(u,\theta)\in[0,\tau]\times\Theta}\left|\frac{A_F^{(1)}(u,\theta,\hat{\mu},\hat{\Sigma})}{A_F^{(0)}(u,\theta,\hat{\mu},\hat{\Sigma})}-\frac{a^{(1)}(u,\theta,\mu,\Sigma)}{a^{(0)}(u,\theta,\mu,\Sigma)}\right|\xrightarrow{p}0$. Therefore

$$\sup_{\theta\in\Theta}|N^{-1}U_{3N}(\theta)| \leq \sup_{(u,\theta)\in[0,\tau]\times\Theta}\left|\frac{A_F^{(1)}(u,\theta,\hat{\mu},\hat{\Sigma})}{A_F^{(0)}(u,\theta,\hat{\mu},\hat{\Sigma})}-\frac{a^{(1)}(u,\theta,\mu,\Sigma)}{a^{(0)}(u,\theta,\mu,\Sigma)}\right|\int_0^\tau d\bar{N}(u)=o_p(1)$$

Therefore, $N^{-1}U_F^R(\theta)=N^{-1}\sum_{i=1}^{N}H_i(\theta)+o_p(1)$ uniformly in θ. On the other hand, since $R(u,\theta,\mu,\Sigma)$ and $\dot{R}(u,\theta,\mu,\Sigma)$ are continuous in θ, we can show that $N^{-1}U_{1N}(\theta)=h(\theta)+o_p(1)$ uniformly in θ. Therefore $\hat{\theta}^R$ converges to θ^* as $N\to\infty$.

(ii) Let $dM_i(u)=dN_i(u)-\lambda_0(u)Y_i(u)R_i^0(u,\theta)du$, $d\bar{M}(u)=N^{-1}\sum_{i=1}^{N}dM_i(u)$. Then we

have

$$\begin{aligned} N^{-1/2}U_F^R(\theta^*) &= N^{-1/2}\sum_{i=1}^{N}\int_0^\tau \frac{\hat{\dot{R}}_i(u,\theta^*,\hat{\mu},\hat{\Sigma})}{\hat{R}_i(u,\theta^*,\hat{\mu},\hat{\Sigma})}dN_i(u) - N^{1/2}\int_0^\tau \frac{A_F^{(1)}(u,\theta^*,\hat{\mu},\hat{\Sigma})}{A_F^{(0)}(u,\theta^*,\hat{\mu},\hat{\Sigma})}d\bar{N}(u) \\ &= U_{4N}(\theta^*) - U_{5N}(\theta^*) - U_{6N}(\theta^*) \end{aligned}$$

where

$$\begin{aligned} U_{4N}(\theta^*) &= N^{-1/2}\sum_{i=1}^{N}\int_0^\tau \frac{\hat{\dot{R}}_i(u,\theta^*,\hat{\mu},\hat{\Sigma})}{\hat{R}_i(u,\theta^*,\hat{\mu},\hat{\Sigma})}dN_i(u) - N^{1/2}\int_0^\tau \frac{a^{(1)}(u,\theta^*,\mu,\Sigma)}{a^{(0)}(u,\theta^*,\mu,\Sigma)}d\bar{M}(u) \\ U_{5N}(\theta^*) &= N^{1/2}\int_0^\tau \frac{A_F^{(1)}(u,\theta^*,\hat{\mu},\hat{\Sigma})}{A_F^{(0)}(u,\theta^*,\hat{\mu},\hat{\Sigma})}\left(N^{-1}\sum_{i=1}^{N}\lambda_0(u)Y_i(u)R_i^0(u,\theta)\right)du \\ U_{6N}(\theta^*) &= N^{1/2}\int_0^\tau \left(\frac{A_F^{(1)}(u,\theta^*,\hat{\mu},\hat{\Sigma})}{A_F^{(0)}(u,\theta^*,\hat{\mu},\hat{\Sigma})} - \frac{a^{(1)}(u,\theta^*,\mu,\Sigma)}{a^{(0)}(u,\theta^*,\mu,\Sigma)}\right)d\bar{M}(u) \end{aligned}$$

The Proposition A.1 in [Kulich and Lin, 2004] implies that $N^{1/2}d\bar{M}(u)$ converges weakly in $l^\infty[0,\tau]$ to a mean-zero Gaussian process uniformly in u. Then the convergence in probability to zero of $U_{6N}(\theta^*)$ follows from Lemma 4.2.1 and Lemma 4.2 in [Kosorok, 2008]. We can also approximate $U_{5N}(\theta^*)$ with the arguments similarly to those in Theorem 2.1 in [Lin and Wei, 1989] and Appendix A in [Xie et al., 2001]

$$\begin{aligned} U_{5N}(\theta^*) &= N^{1/2}\int_0^\tau \frac{1}{a^{(0)}(u,\theta^*,\mu,\Sigma)}\left\{A_F^{(1)}(u,\theta^*,\hat{\mu},\hat{\Sigma}) - \frac{a^{(1)}(u,\theta^*,\mu,\Sigma)}{a^{(0)}(u,\theta^*,\mu,\Sigma)}\right. \\ &\quad \left.\left(A_F^{(0)}(u,\theta^*,\hat{\mu},\hat{\Sigma}) - a^{(0)}(u,\theta^*,\mu,\Sigma)\right)\right\}\mathbb{E}(dN(u)) + o_p(1) \end{aligned}$$

Therefore by direct calculation

$$\begin{aligned} & N^{-1/2}U_F^R(\theta^*) \\ = \; & N^{-1/2}\sum_{i=1}^{N}\int_0^\tau \left(\frac{\hat{\dot{R}}_i(u,\theta^*,\hat{\mu},\hat{\Sigma})}{\hat{R}_i(u,\theta^*,\hat{\mu},\hat{\Sigma})} - \frac{a^{(1)}(u,\theta^*,\mu,\Sigma)}{a^{(0)}(u,\theta^*,\mu,\Sigma)}\right)dN_i(u) \\ & -N^{-1/2}\sum_{i=1}^{N}\int_0^\tau \frac{Y_i(u)}{a^{(0)}(u,\theta^*,\mu,\Sigma)}\left\{\hat{\dot{R}}_i(u,\theta^*,\hat{\mu},\hat{\Sigma}) - \frac{a^{(1)}(u,\theta^*,\mu,\Sigma)}{a^{(0)}(u,\theta^*,\mu,\Sigma)}\hat{R}_i(u,\theta^*,\hat{\mu},\hat{\Sigma})\right\} \\ & \mathbb{E}(dN(u)) + o_p(1) \end{aligned}$$

We need to further expand the above expression at $\hat{\mu}$ and $\hat{\Sigma}$. Let $N^{-1/2}M_{N1}(\theta^*)$ and $N^{-1/2}M_{N2}(\theta^*)$ denote the two terms in $N^{-1/2}U_F^R(\theta^*)$. Then first by Taylor expansion of x/y we have

$$\begin{aligned}
& N^{-1/2}M_{N1}(\theta^*) \\
= \; & N^{-1/2}\sum_{i=1}^{N}\int_0^\tau \left(\frac{\dot{R}_i(u,\theta^*,\mu,\Sigma)}{R_i(u,\theta^*,\mu,\Sigma)} - \frac{a^{(1)}(u,\theta^*,\mu,\Sigma)}{a^{(0)}(u,\theta^*,\mu,\Sigma)}\right) dN_i(u) \\
& + N^{-1/2}\sum_{i=1}^{N}\int_0^\tau \frac{\dot{R}_i(u,\theta^*,\mu,\Sigma)}{R_i^2(u,\theta^*,\mu,\Sigma)}\left\{\hat{R}_i(u,\theta^*,\hat{\mu},\hat{\Sigma}) - R_i(u,\theta^*,\mu,\Sigma)\right\} dN_i(u) \\
& + N^{-1/2}\sum_{i=1}^{N}\int_0^\tau \frac{1}{R_i^2(u,\theta^*,\mu,\Sigma)}\left\{\hat{\dot{R}}_i(u,\theta^*,\hat{\mu},\hat{\Sigma}) - \dot{R}_i(u,\theta^*,\mu,\Sigma)\right\} dN_i(u) + o_p(1) \\
= \; & N^{-1/2}\sum_{i=1}^{N}\int_0^\tau \left(\frac{\dot{R}_i(u,\theta^*,\mu,\Sigma)}{R_i(u,\theta^*,\mu,\Sigma)} - \frac{a^{(1)}(u,\theta^*,\mu,\Sigma)}{a^{(0)}(u,\theta^*,\mu,\Sigma)} + c(u)\phi_i(u) + d(u)\psi_i(u)\right) dN_i(u) + o_p(1)
\end{aligned}$$

where

$$\begin{aligned}
c(u) &= N^{-1}\sum_{i=1}^{N}\left\{\frac{\dot{R}_i(u,\theta^*,\mu,\Sigma)}{R_i^2(u,\theta^*,\mu,\Sigma)}\frac{\partial R_i(u,\theta^*,\mu,\Sigma)}{\partial vec\{\mu\}} + \frac{1}{R_i^2(u,\theta^*,\mu,\Sigma)}\frac{\partial \dot{R}_i(u,\theta^*,\mu,\Sigma)}{\partial vec\{\mu\}}\right\} \\
d(u) &= N^{-1}\sum_{i=1}^{N}\left\{\frac{\dot{R}_i(u,\theta^*,\mu,\Sigma)}{R_i^2(u,\theta^*,\mu,\Sigma)}\frac{\partial R_i(u,\theta^*,\mu,\Sigma)}{\partial vec\{\Sigma\}} + \frac{1}{R_i^2(u,\theta^*,\mu,\Sigma)}\frac{\partial \dot{R}_i(u,\theta^*,\mu,\Sigma)}{\partial vec\{\Sigma\}}\right\}
\end{aligned}$$

Similarly

$$\begin{aligned}
& N^{-1/2}M_{N2}(\theta^*) \\
= \; & -N^{-1/2}\sum_{i=1}^{N}\int_0^\tau \frac{Y_i(u)}{a^{(0)}(u,\theta^*,\mu,\Sigma)}\left\{\dot{R}_i(u,\theta^*,\mu,\Sigma) - \frac{a^{(1)}(u,\theta^*,\mu,\Sigma)}{a^{(0)}(u,\theta^*,\mu,\Sigma)}R_i(u,\theta^*,\mu,\Sigma)\right\}\mathbb{E}(dN(u)) \\
& -N^{-1/2}\sum_{i=1}^{N}\int_0^\tau \{e(u)\phi_i(u) + f(u)\psi_i(u)\}\,\mathbb{E}(dN(u)) + o_p(1)
\end{aligned}$$

where

$$e(u) = N^{-1}\sum_{i=1}^{N}\frac{Y_i(u)}{a^{(0)}(u,\theta^*,\mu,\Sigma)}\left\{\frac{\partial \dot{R}_i(u,\theta^*,\mu,\Sigma)}{\partial vec\{\mu\}} - \frac{a^{(1)}(u,\theta^*,\mu,\Sigma)}{a^{(0)}(u,\theta^*,\mu,\Sigma)}\frac{\partial R_i(u,\theta^*,\mu,\Sigma)}{\partial vec\{\mu\}}\right\}$$

$$f(u) = N^{-1} \sum_{i=1}^{N} \frac{Y_i(u)}{a^{(0)}(u, \theta^*, \mu, \Sigma)} \left\{ \frac{\partial \dot{R}_i(u, \theta^*, \mu, \Sigma)}{\partial vec\{\Sigma\}} - \frac{a^{(1)}(u, \theta^*, \mu, \Sigma)}{a^{(0)}(u, \theta^*, \mu, \Sigma)} \frac{\partial R_i(u, \theta^*, \mu, \Sigma)}{\partial vec\{\Sigma\}} \right\}$$

We combine all the results above and define

$$\begin{aligned} & M_i(\theta^*) \\ = & \int_0^\tau \left\{ \frac{\dot{R}_i(u, \theta^*, \mu, \Sigma)}{R_i(u, \theta^*, \mu, \Sigma)} - \frac{a^{(1)}(u, \theta^*, \mu, \Sigma)}{a^{(0)}(u, \theta^*, \mu, \Sigma)} + c(u)\phi_i(u) + d(u)\psi_i(u) \right\} dN_i(u) \\ & - \int_0^\tau \frac{Y_i(u)}{a^{(0)}(u, \theta^*, \mu, \Sigma)} \left\{ \dot{R}_i(u, \theta^*, \mu, \Sigma) - \frac{a^{(1)}(u, \theta^*, \mu, \Sigma)}{a^{(0)}(u, \theta^*, \mu, \Sigma)} R_i(u, \theta^*, \mu, \Sigma) \right\} \mathbb{E}(dN(u)) \\ & - \int_0^\tau \{e(u)\phi_i(u) + f(u)\psi_i(u)\} \mathbb{E}(dN(u)) \end{aligned} \tag{4.26}$$

Thus we have proved that $N^{-1/2} U_F^R(\theta^*) = N^{-1/2} \sum_{i=1}^N M_i(\theta^*) + o_p(1)$. On the other hand we can also show as in (i) that $N^{-1} \partial U_F^R(\theta)/\partial\theta = N^{-1} \sum_{i=1}^N \dot{H}_i(\theta) + o_p(1)$ uniformly in θ. The asymptotic distribution of $N^{1/2}(\hat{\theta}^R - \theta^*)$ is therefore $A^{-1}B(A^{-1})^T$.

□

Note in this calibration method, we need to estimate the nuisance parameters $\mu(u)$ and $\Sigma(u)$ at each observed event time point using subjects at risk and having enough measurements (i.e. we need at least q_k measurements to estimate the trajectory of $X_k(u)$). This is ensured by Assumption D3. For example, in the dengue vaccine trial, the primary endpoint is the symptomatic virologically confirmed dengue disease occurred 28 days after the third vaccination and during the active phase from Month 13 to Month 25. By the time of the third injection, three visits for measurements of antibody titers have already been made. This implies that almost all subjects at risk by that time have three measurements available. However, if we are interested in the dengue disease occurring during the active phase since the start of the study, then it is possible that by the time of the occurrence of a disease in the early stage of the study, only one or two scheduled visits for measurements have been made. That means estimating the trajectories with more than two parameters at that time point is infeasible. And those subjects who drop out of the study or develop disease before that time would make no contribution to the estimating equations. Therefore we will lose

considerate number of early cases in the analysis if a large proportion of the disease events occur before enough visits for measurements have been made. The ultimate inferential results could be misleading. Looking back at the conditional score method for continuous immune response level, even though it has the same issue of losing a number of cases in the analysis, can still provide consistent estimates because the number of measurements is incorporated in the estimating equations.

4.3 Risk set recalibration method in two-phase sampling design cohort studies

Since this dissertation is aimed to evaluate the time-dependent CoR and CoP in vaccine trial studies where the immune response data are collected on a two-phase sample, we need to further develop the method of RRC for estimating θ in (4.5) when the immune response data are only available on the second phase sample. The sampling design here we considered is the same as those for the continuous biomarker model in Chapter 2. We briefly restate the notations and sampling model described in Section 2.2.2.

In the first phase, we take a random sample with size N from the study population with measurements $(V_i, \Delta_i, Z_i, L_i^T)^T, i = 1, \cdots, N$. In the second phase, a random Bernoulli sample is taken from the N subjects, with sampling probabilities given by $\pi(O_i, \rho)$, where O_i are (a subset of) the variables collected at the first phase and ρ is a finite-dimensional vector of parameters. Let ξ be the binary indicator of being sampled ($\xi = 1$) with probability

$$\mathbb{P}(\xi = 1|O, \alpha, W, T^m, J) = \mathbb{P}(\xi = 1|O) = \pi(O; \rho)$$

Then the longitudinal immune biomarkers $\{W_i, T_i^m, J_i\}$ are assessed only on subjects with $\xi_i = 1$, i.e. the observed data for $i = 1, \cdots, N$ are $\{V_i, \Delta_i, Z_i, L_i, \xi_i, \xi_i W_i, \xi_i T_i^m, \xi_i J_i\}$. With such data with missing immune response data, we propose to estimate θ by solving an IPW version of the RRC estimating equations, defined as

$$U_{IPW}^R(\theta) = \sum_{i=1}^{N} \int \frac{\xi_i}{\pi_i} \left\{ \frac{\hat{\dot{R}}_i(u, \theta, \hat{\mu}, \hat{\Sigma})}{\hat{R}_i(u, \theta, \hat{\mu}, \hat{\Sigma})} - \frac{A_{IPW}^{(1)}(u, \theta, \hat{\mu}, \hat{\Sigma})}{A_{IPW}^{(0)}(u, \theta, \hat{\mu}, \hat{\Sigma})} \right\} dN_i(u) = 0 \tag{4.27}$$

where

$$A^{(0)}_{IPW}(u,\theta,\hat{\mu},\hat{\Sigma}) = N^{-1}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}Y_i(u)\hat{R}_i(u,\theta,\hat{\mu},\hat{\Sigma})$$
$$A^{(1)}_{IPW}(u,\theta,\hat{\mu},\hat{\Sigma}) = N^{-1}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}Y_i(u)\hat{\dot{R}}_i(u,\theta,\hat{\mu},\hat{\Sigma})$$

As in Chapter 2, we consider both situations with correctly and fully specified sampling probabilities and estimated sampling probabilities by solving $S_{\pi,F}(\rho) = 0$ below. The corresponding estimates of θ are denoted by $\hat{\theta}^R_{IPW}(\pi)$ and $\hat{\theta}^R_{IPW}(\hat{\pi})$.

$$\begin{aligned} S_{\pi,F}(\rho) = \sum_{i=1}^{N} S_{\pi,i}(\rho) &= \sum_{i=1}^{N}\frac{\partial}{\partial\rho}\log\left\{\pi(O_i;\rho)^{\xi_i}\left(1-\pi(O_i;\rho)\right)^{1-\xi_i}\right\} \\ &= \sum_{i=1}^{N}\frac{\xi_i-\pi(O_i;\rho)}{\pi(O_i;\rho)(1-\pi(O_i;\rho))}\frac{\partial\pi(O_i;\rho)}{\partial\rho} = 0 \end{aligned}$$

To establish the asymptotic theories for $\hat{\theta}^R_{IPW}(\pi)$ and $\hat{\theta}^R_{IPW}(\hat{\pi})$, in additional to Assumption D, we also assume the sampling probabilities are positive: $1 \geq \pi(O;\rho) > \delta > 0$, for all ρ and some constant $\delta > 0$. The following theories can be proved based on Theorem 4.2.2 and in a similar way as that in the proof of the theories for IPW estimators for continuous biomarker model in Theorem 2.6.3 and Theorem 2.6.5.

Theorem 4.3.1. *As $N \to \infty$, (i) $\hat{\theta}^R_{IPW}(\pi) \xrightarrow{p} \theta^*$; and (ii) $\sqrt{N}\left(\hat{\theta}^R_{IPW}(\pi) - \theta^*\right)$ converges weakly to a Normal random variate with mean zero and covariance $A^{-1}B^*(A^{-1})^T$, where*

$$A = \mathbb{E}\left[\dot{H}(\theta^*)\right] \quad B^* = \mathbb{E}\left[\frac{1}{\pi}M(\theta^*)M(\theta^*)^T\right]$$

where $M(\theta^*)$ is defined in (4.26).

Theorem 4.3.2. *As $N \to \infty$, (i) $\hat{\theta}^R_{IPW}(\hat{\pi}) \xrightarrow{p} \theta^*$; and (ii) $\sqrt{N}\left(\hat{\theta}^R_{IPW}(\hat{\pi}) - \theta^*\right)$ converges weakly to a Normal random variate with mean zero and covariance $A^{-1}B^{**}(A^{-1})^T$, where*

$$A = \mathbb{E}\left[\dot{H}(\theta^*)\right] \quad B^{**} = B^* - \mathbb{E}[M(\theta^*)\frac{\dot{\pi}}{\pi}]\left\{\mathbb{E}[S_\pi S_\pi^T]\right\}^{-1}\mathbb{E}[M(\theta^*)\frac{\dot{\pi}}{\pi}]^T$$

where $M(\theta^*)$ *is defined in* (4.26).

Chapter 5

SIMULATION STUDIES FOR JOINT MODELING WITH DICHOTOMIZED BIOMARKERS

5.1 Simulation for full cohort studies

In this section, we evaluate the RRC estimator developed for dichotomized immune biomarkers in Chapter 4. We compare the RRC estimator with the ideal estimator (Ideal) where the random effects for each subject are assumed known and the two-stage estimator (TS) where the subject-specific biomarker trajectories are first fitted by least squares estimates to predict the binary status of immune biomarker at each time point and then use them to fit the Cox regression model. We evaluate the three estimators through simulations studies in terms of the bias, relative bias to the true parameter (Bias %), Monte Carlo standard deviation (MCSD), and the relative mean squared error (RMSE) to the ideal estimator.

For the RRC estimator, we would also like to evaluate an estimate for its theoretical sandwich variance given in Theorem 4.2.2. However, direct programming of the estimate of $A = \mathbb{E}[M(\theta^*)M(\theta^*)^T]$ is very hard as can be seen from the expression of $M(\theta^*)$ defined in (4.26). First it requires the derivatives of $R(u, \theta, \mu, \Sigma)$ and $\dot{R}(u, \theta, \mu, \Sigma)$ with respect to $\mu(u)$ and $\Sigma(u)$ over time, which are apparently nonlinear. Also it depends on the "unknown" truth $R^0(u, \theta)$ which we propose to approximate based on the Normal working assumption. The same problem occurred in [Dafni and Tsiatis, 1998] where they dealt with a similar model but for the continuous biomarker $X(u)$. In their simulation studies (Table 3), they used the variance estimate obtained from maximizing the induced partial likelihood function simply with estimated nuisance parameters plugged in. In their simulation studies, such approximated variance estimates performed quite well compared with the empirical standard derivation of the parameter estimates. In our simulation studies, we approximate the standard error (ASE) estimates in a similar way. For a valid estimate of the theoretical variance we also investigate the bootstrap method as in [Wang et al., 2001].

5.1.1 Model 1

We first consider the model with one dichotomized biomarker $\lambda(u) = \lambda_0(u)\exp\{\beta B(u)\}$. The biomarker $X(u)$ is generated from $X(u) = \alpha_1 + \alpha_2 u$, where $(\alpha_1, \alpha_2)^T$ is from bivariate Normal distribution with mean $\mu = (2.575, -0.009)^T$ and covariance Σ with elements $(\Sigma_{11}, \Sigma_{12}, \Sigma_{22})^T = (0.0191, 0.00007, 0.0002)^T$. The parameters μ and Σ are estimated from the ACTG 175 data [Hammer et al., 1996]. The threshold l for $B(u) = I(X(u) \geq l)$ is 2.393 with $\mathbb{P}(B(13.5) = 1) = 0.6$. The censoring time is simulated from $Exp(1/80)$ and is subject to administrative censoring at time 30.

We consider four simulation scenarios (a)(b)(c) and (d) for Model 1. For each scenario, we simulate the event time data with hazard ratios $e^{\beta} = \{1, 0.75, 0.50, 0.25\}$. Also we consider three settings of measurement errors (low, moderate and high) with the variance of measurement error $\sigma^2 = \{0.01, 0.08, 0.15\}$. They represent a noise-to-signal ratio of $\mathbb{V}ar(e)/\mathbb{V}ar(X(0)) \approx 1/2$, 4, and 8, and $\mathbb{V}ar(e)/\mathbb{V}ar(X(10)) \approx 1/4$, 2, and 4 respectively.

(a) Low event rate setting with a moderate number of measurement time points. Sample size N=800. The longitudinal observations W are made at baseline and a random time point uniformly sampled from each of these 9 time windows $3, 6, 9, \cdots, 27 \pm 0.3$.

(b) Low event rate setting with a large number of measurement time points. Sample size N=800. The longitudinal observations W are made at baseline and a random time point uniformly sampled from each of these 56 time windows $0.5, 1.0, 1.5, \cdots, 28 \pm 0.05$.

(c) High event rate setting with a moderate number of measurement time points. Sample size N=160. The longitudinal observations W are made at baseline and a random time point uniformly sampled from each of these 9 time windows $3, 6, 9, \cdots, 27 \pm 0.3$.

(d) High event rate setting with a large number of measurement time points. Sample size N=160. The longitudinal observations W are made at baseline and a random time point uniformly sampled from each of these 56 time windows $0.5, 1.0, 1.5, \cdots, 28 \pm 0.05$.

For the two scenarios in Model 1(a) and Model 1(b) with low event rates, the average event rates are 11.7%, 11.4%, 12.6%, and 13.1% when the hazard ratios are 1, 0.75, 0.50 and 0.25, respectively. For the two scenarios in Model 1(c) and Model 1(d) with high event rates, the average event rates are 81.2%, 81.2%, 74.8% and 80.2% when the hazard ratios are 1,

0.75, 0.50 and 0.25, respectively. We only include the measurements before the occurrence of an event for cases. The average number of measurements available in each of the four scenarios are 8.0, 45.1, 3.9 and 20.1. See Table 5.1-5.4 and Figure 5.1-5.2 for the simulation results.

From the results we see that in low-event-rate settings, the RRC estimates provide small bias with relative bias less than 12% for all hazard ratios. As the number of measurements collected become large, the relative bias could be controlled below 6%. The TS estimates, however, can lead to a relative bias as high as 48.3%. Even though we enlarge the number of measurements, the relative bias with large measurement errors cannot be reduced to below 20% in several settings. In the high-event-rate settings, the RRC method could still give reasonably small bias except for the setting with e^{β}=0.25 and large measurement errors. This is what commonly expected for the calibration methods when the effect size of the covariate is large and event rate is high. The biases from TS estimates generally show a similar pattern as that in low event rate settings.

Generally speaking the TS method produces smaller MCSD than the RRC method, especially when the measurement errors are relative large. This is also reflected in the RMSE such that the TS estimator could give smaller RMSE than the RRC estimator in some settings even if the TS estimator is more biased. This implies a bias-variance trade-off when making the choice of which methods to be used. If it is of interest to estimate the effect of the biomarker on the event endpoint, we suggest using the RRC method instead of the TS method due to the high biases yielded by the latter.

The approximated standard errors (ASE) are very closed to the MCSD of $\hat{\beta}$ as it has been seen in [Dafni and Tsiatis, 1998]. It suggests using this simplified approximation of the SE($\hat{\beta}$) in such model with one dichotomized biomarker only is an acceptable choice.

Table 5.1: Simulation results for Model 1(a), with low event rates and moderate numbers of longitudinal immune response measurements.

	σ^2	Method	Bias(% Bias)	ASE	MCSD	RMSE
$\beta = 0$		Ideal	0.029	0.229	0.226	1.000
	0.01	TS	-0.058	0.226	0.235	1.124
	0.01	RRC	0.028	0.279	0.269	1.409
	0.08	TS	-0.141	0.218	0.230	1.395
	0.08	RRC	0.030	0.356	0.335	2.176
	0.15	TS	-0.156	0.215	0.226	1.450
	0.15	RRC	0.035	0.408	0.396	3.028
$\beta = -\ln 4/3$		Ideal	0.031 (-10.687)	0.226	0.217	1.000
	0.01	TS	0.019 (-6.466)	0.225	0.227	1.082
	0.01	RRC	0.032 (-11.054)	0.275	0.264	1.473
	0.08	TS	0.006 (-1.947)	0.219	0.227	1.081
	0.08	RRC	0.031 (-10.624)	0.354	0.333	2.340
	0.15	TS	0.018 (-6.350)	0.217	0.234	1.146
	0.15	RRC	0.032 (-11.277)	0.405	0.401	3.388
$\beta = -\ln 2$		Ideal	0.020 (-2.925)	0.213	0.207	1.000
	0.01	TS	0.106 (-15.293)	0.212	0.220	1.384
	0.01	RRC	0.015 (-2.227)	0.261	0.251	1.464
	0.08	TS	0.191 (-27.564)	0.207	0.210	1.872
	0.08	RRC	0.001 (-0.119)	0.338	0.319	2.366
	0.15	TS	0.255 (-36.806)	0.206	0.214	2.571
	0.15	RRC	0.009 (-1.235)	0.390	0.377	3.299
$\beta = -\ln 4$		Ideal	0.005 (-0.366)	0.223	0.230	1.000
	0.01	TS	0.277 (-20.008)	0.214	0.232	2.459
	0.01	RRC	-0.008 (0.585)	0.281	0.290	1.589
	0.08	TS	0.560 (-40.412)	0.206	0.209	6.730
	0.08	RRC	-0.028 (1.990)	0.380	0.392	2.912
	0.15	TS	0.669 (-48.266)	0.204	0.209	9.253
	0.15	RRC	-0.043 (3.082)	0.451	0.459	3.998

Sample size is N=800. The longitudinal measurements of W are made at baseline and randomly from time windows $3, 6, 9, \cdots, 27 \pm 0.3$, resulting on average 8.0 measurements available per subject.

Table 5.2: Simulation results for Model 1(b), with low event rates and large numbers of longitudinal immune response measurements.

	σ^2	Method	Bias(% Bias)	ASE	MCSD	MSE
$\beta = 0$		Ideal	0.018	0.232	0.229	1.000
	0.01	TS	-0.030	0.230	0.233	1.041
	0.01	RRC	0.022	0.252	0.250	1.191
	0.08	TS	-0.144	0.225	0.227	1.367
	0.08	RRC	0.011	0.285	0.278	1.468
	0.15	TS	-0.196	0.222	0.229	1.720
	0.15	RRC	0.013	0.303	0.301	1.712
$\beta = -\ln 4/3$		Ideal	0.014 (-4.960)	0.228	0.235	1.000
	0.01	TS	0.007 (-2.321)	0.228	0.241	1.043
	0.01	RRC	0.017 (-5.794)	0.248	0.251	1.138
	0.08	TS	-0.057 (19.913)	0.225	0.237	1.067
	0.08	RRC	0.006 (-2.122)	0.281	0.279	1.404
	0.15	TS	-0.070 (24.313)	0.223	0.235	1.084
	0.15	RRC	0.003 (-1.064)	0.299	0.305	1.676
$\beta = -\ln 2$		Ideal	0.008 (-1.185)	0.214	0.219	1.000
	0.01	TS	0.057 (-8.160)	0.214	0.219	1.070
	0.01	RRC	0.016 (-2.259)	0.233	0.235	1.161
	0.08	TS	0.057 (-8.212)	0.212	0.211	1.001
	0.08	RRC	-0.004 (0.594)	0.264	0.264	1.456
	0.15	TS	0.079 (-11.397)	0.211	0.225	1.186
	0.15	RRC	-0.005 (0.690)	0.283	0.288	1.731
$\beta = -\ln 4$		Ideal	0.011 (-0.782)	0.223	0.222	1.000
	0.01	TS	0.161 (-11.578)	0.219	0.227	1.565
	0.01	RRC	0.013 (-0.958)	0.244	0.245	1.224
	0.08	TS	0.303 (-21.861)	0.214	0.209	2.746
	0.08	RRC	0.007 (-0.473)	0.280	0.277	1.561
	0.15	TS	0.366 (-26.424)	0.212	0.226	3.752
	0.15	RRC	-0.001 (0.059)	0.303	0.306	1.893

Sample size is N=800. The longitudinal measurements of W are made at baseline and randomly from time windows $0.5, 1.0, 1.5, \cdots, 28 \pm 0.05$, resulting on average 45.1 measurements available per subject.

Table 5.3: Simulation results for Model 1(c), with high event rates and moderate numbers of longitudinal immune response measurements.

	σ^2	Method	Bias(% Bias)	ASE	MCSD	MSE
$\beta = 0$		Ideal	-0.001	0.210	0.211	1.000
	0.01	TS	-0.133	0.201	0.222	1.513
	0.01	RRC	-0.032	0.279	0.281	1.803
	0.08	TS	-0.219	0.187	0.215	2.118
	0.08	RRC	-0.016	0.399	0.424	4.055
	0.15	TS	-0.218	0.183	0.218	2.132
	0.15	RRC	-0.012	0.464	0.478	5.149
$\beta = -\ln 4/3$		Ideal	0.005 (-1.617)	0.207	0.205	1.000
	0.01	TS	-0.030 (10.453)	0.202	0.221	1.186
	0.01	RRC	-0.017 (5.861)	0.267	0.276	1.830
	0.08	TS	-0.066 (22.987)	0.192	0.215	1.207
	0.08	RRC	-0.016 (5.649)	0.378	0.382	3.488
	0.15	TS	-0.063 (21.865)	0.189	0.221	1.263
	0.15	RRC	0.011 (-3.939)	0.450	0.473	5.332
$\beta = -\ln 2$		Ideal	0.000 (-0.044)	0.193	0.190	1.000
	0.01	TS	0.057 (-8.245)	0.191	0.206	1.261
	0.01	RRC	-0.009 (1.265)	0.261	0.267	1.962
	0.08	TS	0.132 (-19.068)	0.182	0.196	1.538
	0.08	RRC	0.038 (-5.482)	0.387	0.404	4.534
	0.15	TS	0.169 (-24.393)	0.180	0.210	2.010
	0.15	RRC	0.067 (-9.645)	0.446	0.439	5.447
$\beta = -\ln 4$		Ideal	0.002 (-0.156)	0.186	0.184	1.000
	0.01	TS	0.278 (-20.066)	0.189	0.196	3.429
	0.01	RRC	0.053 (-3.840)	0.260	0.270	2.241
	0.08	TS	0.511 (-36.888)	0.182	0.195	8.874
	0.08	RRC	0.202 (-14.605)	0.409	0.417	6.365
	0.15	TS	0.609 (-43.902)	0.181	0.196	12.119
	0.15	RRC	0.345 (-24.881)	0.476	0.494	10.768

Sample size is N=160. The longitudinal measurements of W are made at baseline and randomly from time windows $3, 6, 9, \cdots, 27 \pm 0.3$, resulting on average 3.9 measurements available per subject.

Table 5.4: Simulation results for Model 1(d), with high event rates and large numbers of longitudinal immune response measurements.

	σ^2	Method	Bias(% Bias)	ASE	MCSD	MSE
$\beta = 0$		Ideal	0.012	0.211	0.201	1.000
	0.01	TS	-0.077	0.206	0.203	1.163
	0.01	RRC	0.010	0.235	0.229	1.292
	0.08	TS	-0.270	0.194	0.206	2.842
	0.08	RRC	0.010	0.281	0.279	1.927
	0.15	TS	-0.300	0.190	0.210	3.311
	0.15	RRC	0.011	0.306	0.319	2.512
$\beta = -\ln 4/3$		Ideal	0.018 (-6.204)	0.207	0.202	1.000
	0.01	TS	-0.013 (4.640)	0.205	0.207	1.041
	0.01	RRC	0.018 (-6.169)	0.229	0.233	1.323
	0.08	TS	-0.123 (42.693)	0.197	0.210	1.439
	0.08	RRC	0.016 (-5.388)	0.272	0.283	1.956
	0.15	TS	-0.144 (50.211)	0.194	0.214	1.617
	0.15	RRC	0.025 (-8.531)	0.298	0.313	2.386
$\beta = -\ln 2$		Ideal	0.009 (-1.294)	0.193	0.189	1.000
	0.01	TS	0.046 (-6.612)	0.193	0.197	1.133
	0.01	RRC	0.008 (-1.189)	0.215	0.221	1.366
	0.08	TS	0.004 (-0.580)	0.187	0.196	1.065
	0.08	RRC	0.018 (-2.534)	0.260	0.277	2.146
	0.15	TS	0.016 (-2.254)	0.184	0.197	1.083
	0.15	RRC	0.039 (-5.573)	0.287	0.308	2.674
$\beta = -\ln 4$		Ideal	0.003 (-0.218)	0.185	0.186	1.000
	0.01	TS	0.172 (-12.386)	0.188	0.184	1.832
	0.01	RRC	0.017 (-1.198)	0.206	0.216	1.353
	0.08	TS	0.285 (-20.561)	0.185	0.190	3.398
	0.08	RRC	0.083 (-6.016)	0.258	0.278	2.437
	0.15	TS	0.353 (-25.459)	0.183	0.192	4.676
	0.15	RRC	0.144 (-10.366)	0.291	0.296	3.135

Sample size is N=800. The longitudinal measurements of W are made at baseline and randomly from time windows $0.5, 1.0, 1.5, \cdots, 28 \pm 0.05$, resulting on average 20.1 measurements available per subject.

Figure 5.1: Summary of simulation results for Model 1(a) and Model 1(b) with low event rates.

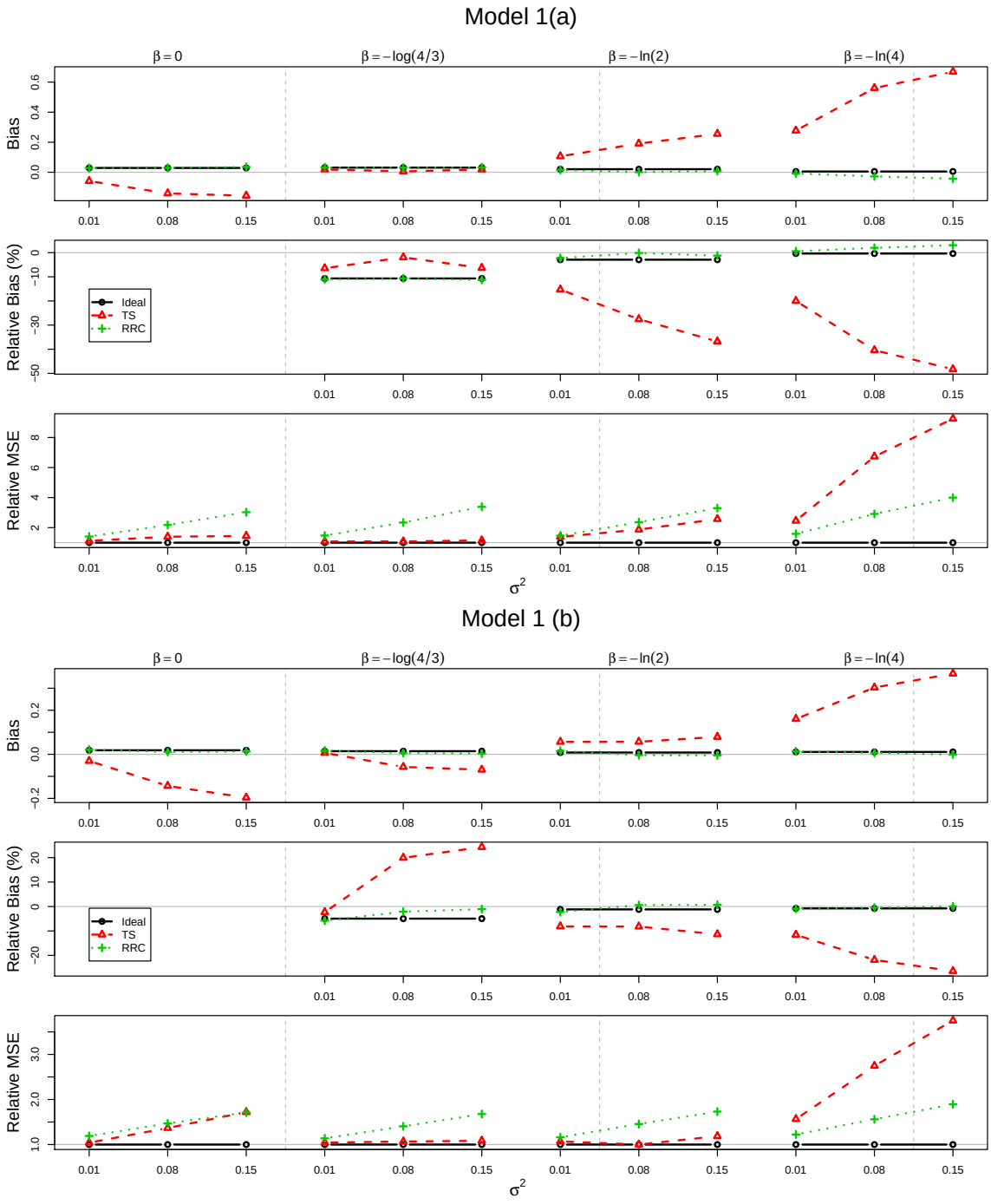
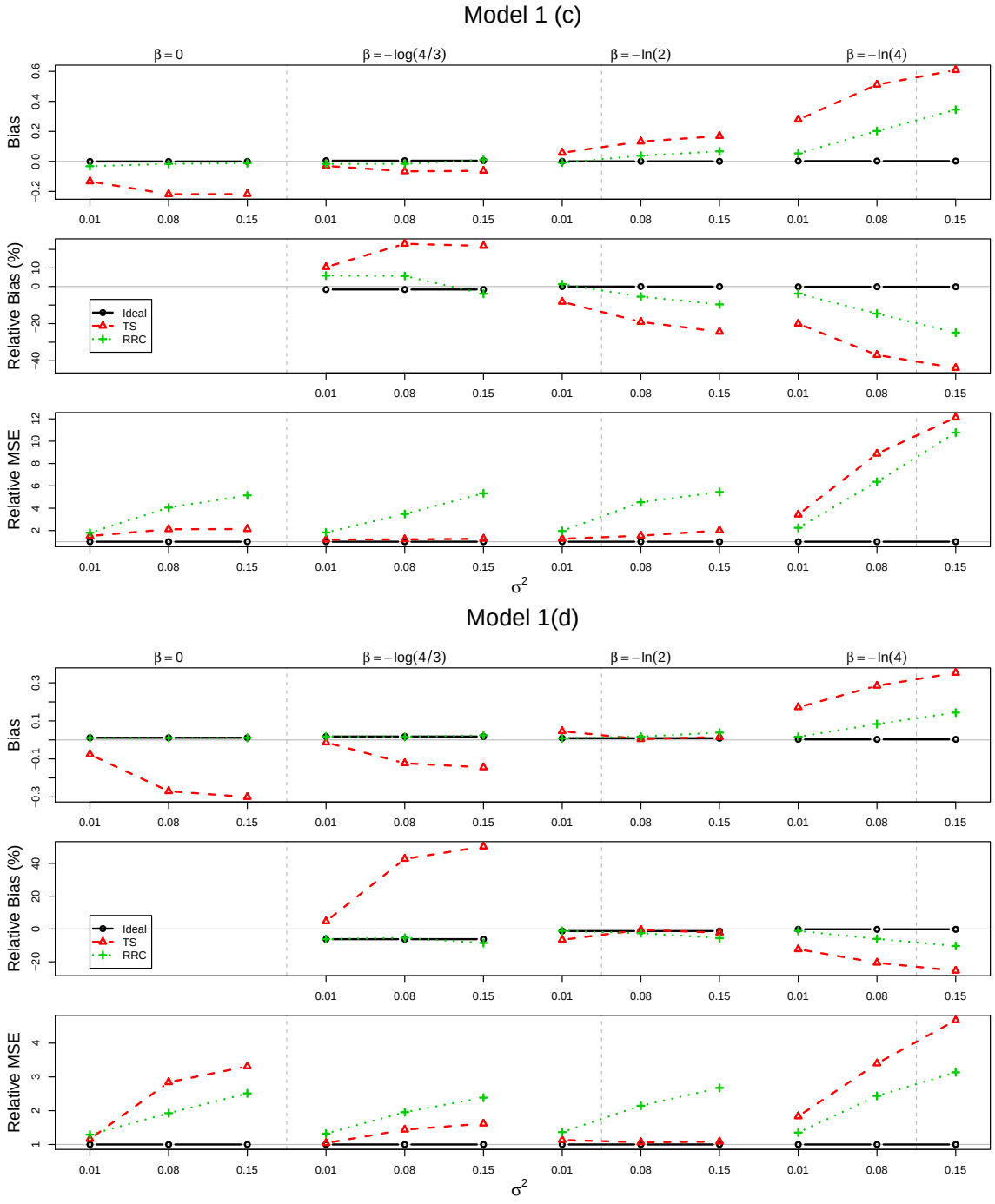


Figure 5.2: Summary of simulation results for Model 1(c) and Model 1(d) with high event rates.



5.1.2 Model 2

We next consider the model with an adjustment of vaccination indicator $\lambda(u) = \lambda_0(u) \exp\{\beta B(u) + \eta Z\}$. This model is useful in Prentice's framework to assess the vaccine effect on the disease endpoint adjusting for the dichotomized biomarker. The biomarker $X(u)$ is generated from $X(u) = \alpha_1 + \alpha_2 u$, where $\alpha = (\alpha_1, \alpha_2)^T$ given Z is generated from a bivariate Normal distribution with mean $\mathbb{E}[\alpha|Z=1] = (2.570, -0.009)^T$, $\mathbb{E}[\alpha|Z=0] = (2.577, -0.007)^T$, and covariance $\mathbb{C}ov[\alpha|Z] = \Sigma$ with elements $(\Sigma_{11}, \Sigma_{12}, \Sigma_{22}) = (0.0191, 0.00007, 0.0002)^T$. The threshold l for $B(u) = I(X(u) \geq l)$ is 2.4047 with $\mathbb{P}(B(13.5) = 1) = 0.6$. The vaccination indicator Z is generated from $Bernoulli(0.5)$. The censoring time is simulated from $Exp(1/80)$ with an administrative censoring time at 30. We still consider four simulation scenarios (a)(b)(c)(d) as in Model 1. For each scenario, we simulate the event time data with hazard ratios $(e^\beta, e^\eta)^T = \{(0.5, 1)^T, (0.5, 0.5)^T\}$.

For the two settings Model 2(a) and Model 2(b) with low event rates, the average event rates are 12.6% and 13.0% when the hazard ratios are $(0.5, 1)^T$ and $(0.5, 0.5)^T$, respectively. For the two settings Model 2(c) and Model 2(d) with high event rates, the average event rates are 79.5%, 81.7% when the hazard ratios are $(0.5, 1)^T$ and $(0.5, 0.5)^T$, respectively. The average number of measurements available per subject for each of the four scenarios are 8.0, 45.0, 4.1 and 22.0. See Table 5.5-5.8 and Figure 5.3-5.4 for the simulation results.

From the simulation results we can see that both TS and RRC methods provide very small biases for $\hat{\eta}$. Moreover, the TS estimator could be around 15% more efficient than the RRC estimator. This might suggest that if the objective only centers on evaluating the adjusted vaccine effect on the disease endpoint, TS estimator could also be considered as a reasonable method. However, if it is also interesting to look at the effect of the dichotomized biomarker on the disease endpoint, the RRC method is the only one recommended because the biases from the TS estimator of β can be very large, unless there are extremely large number of immune response measurements. In this model with both $B(u)$ and Z, the ASE is still very closed to the MCSD of $\hat{\beta}$ when the measurement error is not high. However, for η, the coefficient for Z, the ASE could underestimate the SE($\hat{\eta}$) up to 45% on average in Table 5.5 . Considering such high bias, the ASE is no longer considered as a valid estimate

for SE($\hat{\eta}$), so we would suggest using the bootstrap method for the standard error estimates. From Table 5.9 we see the bootstrap method provides reasonable variance estimates.

Table 5.5: Simulation results for Model 2(a), with low event rates and moderate numbers of longitudinal immune response measurements.

		σ^2	Method	Bias(% Bias)	ASE	MCSD	MSE
(β, η)	$\beta = -\ln 2$		Ideal	-0.009 (1.365)	0.211	0.213	1.000
=		0.01	TS	0.091 (-13.143)	0.210	0.222	1.262
$(-\ln 2, 0)$		0.01	RRC	-0.015 (2.132)	0.253	0.265	1.544
		0.08	TS	0.192 (-27.751)	0.206	0.210	1.780
		0.08	RRC	-0.003 (0.479)	0.328	0.332	2.420
		0.15	TS	0.242 (-34.869)	0.205	0.228	2.425
		0.15	RRC	-0.003 (0.413)	0.378	0.404	3.588
	$\eta = 0$		Ideal	0.007	0.201	0.202	1.000
		0.01	TS	0.005	0.201	0.202	1.001
		0.01	RRC	0.010	0.179	0.216	1.138
		0.08	TS	0.001	0.201	0.203	1.005
		0.08	RRC	0.010	0.142	0.216	1.137
		0.15	TS	-0.000	0.201	0.203	1.007
		0.15	RRC	0.013	0.130	0.223	1.220
(β, η)	$\beta = -\ln 2$		Ideal	-0.016 (2.369)	0.209	0.211	1.000
=		0.01	TS	0.076 (-10.981)	0.207	0.220	1.213
$(-\ln 2, -\ln 2)$		0.01	RRC	-0.024 (3.425)	0.252	0.261	1.532
		0.08	TS	0.184 (-26.499)	0.203	0.212	1.756
		0.08	RRC	-0.026 (3.747)	0.327	0.336	2.535
		0.15	TS	0.233 (-33.569)	0.202	0.221	2.303
		0.15	RRC	-0.030 (4.268)	0.379	0.417	3.900
	$\eta = -\ln 2$		Ideal	0.002 (-0.292)	0.209	0.224	1.000
		0.01	TS	-0.000 (0.048)	0.209	0.224	1.002
		0.01	RRC	0.004 (-0.575)	0.202	0.242	1.168
		0.08	TS	-0.003 (0.404)	0.209	0.224	1.001
		0.08	RRC	0.004 (-0.641)	0.163	0.242	1.168
		0.15	TS	-0.004 (0.550)	0.209	0.224	1.003
		0.15	RRC	0.006 (-0.886)	0.149	0.246	1.200

Sample size is N=800. The longitudinal measurements of W are made at baseline and randomly from time windows $3, 6, 9, \cdots, 27 \pm 0.3$, resulting on average 8.0 measurements available per subject.

Table 5.6: Simulation results for Model 2(b), with low event rates and large numbers of longitudinal immune response measurements.

		σ^2	Method	Bias(% Bias)	ASE	MCSD	MSE
(β, η)	$\beta = -\ln 2$		Ideal	0.011 (-1.545)	0.213	0.219	1.000
=		0.01	TS	0.053 (-7.670)	0.212	0.218	1.041
$(-\ln 2, 0)$		0.01	RRC	0.003 (-0.467)	0.225	0.240	1.189
		0.08	TS	0.071 (-10.185)	0.211	0.221	1.118
		0.08	RRC	-0.004 (0.548)	0.257	0.264	1.440
		0.15	TS	0.100 (-14.454)	0.209	0.216	1.174
		0.15	RRC	0.019 (-2.715)	0.273	0.286	1.705
	$\eta = 0$		Ideal	-0.002	0.202	0.201	1.000
		0.01	TS	-0.003	0.202	0.201	0.999
		0.01	RRC	0.000	0.185	0.204	1.029
		0.08	TS	-0.004	0.202	0.200	0.995
		0.08	RRC	0.001	0.165	0.203	1.027
		0.15	TS	-0.005	0.202	0.201	1.000
		0.15	RRC	-0.002	0.158	0.207	1.068
(β, η)	$\beta = -\ln 2$		Ideal	0.015 (-2.111)	0.210	0.213	1.000
=		0.01	TS	0.060 (-8.684)	0.210	0.215	1.101
$(-\ln 2, -\ln 2)$		0.01	RRC	0.012 (-1.728)	0.224	0.235	1.220
		0.08	TS	0.080 (-11.567)	0.208	0.214	1.152
		0.08	RRC	0.013 (-1.805)	0.254	0.254	1.420
		0.15	TS	0.099 (-14.333)	0.207	0.215	1.235
		0.15	RRC	0.013 (-1.894)	0.272	0.279	1.710
	$\eta = -\ln 2$		Ideal	0.002 (-0.262)	0.210	0.215	1.000
		0.01	TS	0.001 (-0.176)	0.210	0.215	1.003
		0.01	RRC	0.005 (-0.713)	0.216	0.218	1.024
		0.08	TS	0.000 (-0.070)	0.210	0.215	0.998
		0.08	RRC	0.005 (-0.705)	0.195	0.218	1.024
		0.15	TS	0.000 (-0.018)	0.210	0.216	1.007
		0.15	RRC	0.005 (-0.738)	0.187	0.219	1.034

Sample size is N=800. The longitudinal measurements of W are made at baseline and randomly from time windows $0.5, 1.0, 1.5, \cdots, 28 \pm 0.05$, resulting on average 45.0 measurements available per subject.

Table 5.7: Simulation results for Model 2(c), with high event rates and moderate numbers of longitudinal immune response measurements.

		σ^2	Method	Bias(% Bias)	ASE	MCSD	MSE
(β, η)	$\beta = -\ln 2$		Ideal	-0.002 (0.355)	0.193	0.200	1.000
=		0.01	TS	0.075 (-10.802)	0.192	0.219	1.336
$(-\ln 2, 0)$		0.01	RRC	0.007 (-0.956)	0.260	0.268	1.795
		0.08	TS	0.138 (-19.959)	0.184	0.208	1.553
		0.08	RRC	0.116 (-16.785)	0.429	0.410	4.527
		0.15	TS	0.176 (-25.450)	0.182	0.218	1.958
		0.15	RRC	0.187 (-26.983)	0.495	0.481	6.638
	$\eta = 0$		Ideal	-0.005	0.180	0.188	1.000
		0.01	TS	-0.005	0.180	0.196	1.080
		0.01	RRC	-0.000	0.171	0.202	1.149
		0.08	TS	-0.007	0.180	0.195	1.068
		0.08	RRC	-0.006	0.277	0.248	1.737
		0.15	TS	-0.009	0.180	0.192	1.036
		0.15	RRC	-0.013	0.391	0.313	2.759
(β, η)	$\beta = -\ln 2$		Ideal	-0.008 (1.201)	0.194	0.203	1.000
=		0.01	TS	0.078 (-11.206)	0.193	0.225	1.374
$(-\ln 2, -\ln 2)$		0.01	RRC	0.017 (-2.406)	0.289	0.286	1.981
		0.08	TS	0.132 (-19.045)	0.182	0.217	1.565
		0.08	RRC	0.144 (-20.717)	0.470	0.419	4.743
		0.15	TS	0.190 (-27.430)	0.180	0.217	2.010
		0.15	RRC	0.186 (-26.898)	0.527	0.464	6.049
	$\eta = -\ln 2$		Ideal	-0.012 (1.669)	0.183	0.191	1.000
		0.01	TS	-0.029 (4.191)	0.182	0.197	1.090
		0.01	RRC	-0.002 (0.341)	0.216	0.211	1.214
		0.08	TS	-0.019 (2.785)	0.182	0.196	1.067
		0.08	RRC	0.012 (-1.671)	0.360	0.287	2.254
		0.15	TS	-0.015 (2.149)	0.182	0.195	1.047
		0.15	RRC	0.031 (-4.410)	0.511	0.356	3.491

Sample size is N=160. The longitudinal measurements of W are made at baseline and randomly from time windows $3, 6, 9, \cdots, 27 \pm 0.3$, resulting on average 4.1 measurements available per subject.

Table 5.8: Simulation results for Model 2(d), with high event rates and large numbers of longitudinal immune response measurements.

		σ^2	Method	Bias(% Bias)	ASE	MCSD	MSE
(β, η)	$\beta = -\ln 2$		Ideal	0.014 (-2.059)	0.193	0.191	1.000
=		0.01	TS	0.057 (-8.208)	0.193	0.191	1.077
$(-\ln 2, 0)$		0.01	RRC	0.019 (-2.779)	0.216	0.222	1.351
		0.08	TS	0.022 (-3.148)	0.188	0.186	0.957
		0.08	RRC	0.061 (-8.810)	0.264	0.259	1.927
		0.15	TS	0.040 (-5.712)	0.185	0.197	1.098
		0.15	RRC	0.105 (-15.132)	0.287	0.299	2.725
	$\eta = 0$		Ideal	-0.004	0.180	0.187	1.000
		0.01	TS	-0.004	0.180	0.188	1.011
		0.01	RRC	-0.003	0.167	0.193	1.065
		0.08	TS	-0.005	0.180	0.188	1.011
		0.08	RRC	-0.002	0.176	0.205	1.202
		0.15	TS	-0.004	0.180	0.189	1.012
		0.15	RRC	-0.008	0.181	0.213	1.297
(β, η)	$\beta = -\ln 2$		Ideal	0.015 (-2.231)	0.194	0.194	1.000
=		0.01	TS	0.046 (-6.629)	0.193	0.195	1.052
$(-\ln 2, -\ln 2)$		0.01	RRC	0.029 (-4.184)	0.232	0.222	1.316
		0.08	TS	0.010 (-1.461)	0.186	0.194	0.997
		0.08	RRC	0.088 (-12.668)	0.283	0.272	2.152
		0.15	TS	0.032 (-4.685)	0.184	0.198	1.055
		0.15	RRC	0.127 (-18.289)	0.306	0.304	2.855
	$\eta = -\ln 2$		Ideal	-0.019 (2.712)	0.183	0.190	1.000
		0.01	TS	-0.016 (2.284)	0.183	0.191	1.004
		0.01	RRC	-0.015 (2.133)	0.206	0.196	1.059
		0.08	TS	-0.012 (1.764)	0.183	0.192	1.017
		0.08	RRC	-0.005 (0.750)	0.222	0.208	1.191
		0.15	TS	-0.006 (0.903)	0.183	0.192	1.015
		0.15	RRC	-0.002 (0.221)	0.225	0.226	1.399

Sample size is N=160. The longitudinal measurements of W are made at baseline and randomly from time windows $0.5, 1.0, 1.5, \cdots, 28 \pm 0.05$, resulting on average 22.0 measurements available per subject.

Figure 5.3: Summary of simulation results for Model 2(a) and Model 2(b) with low event rates.

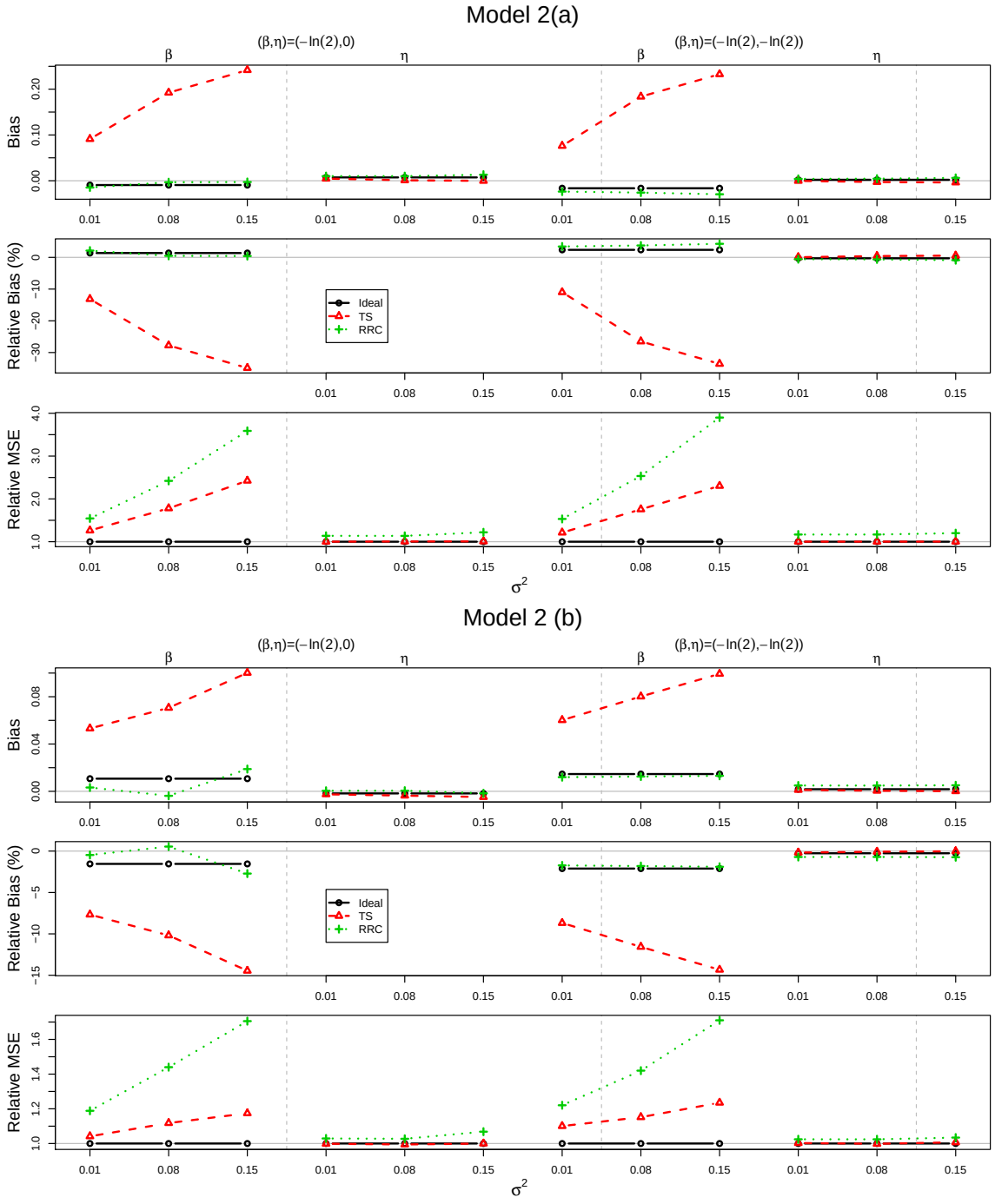


Figure 5.4: Summary of simulation results for Model 2(c) and Model 2(d) with high event rates.

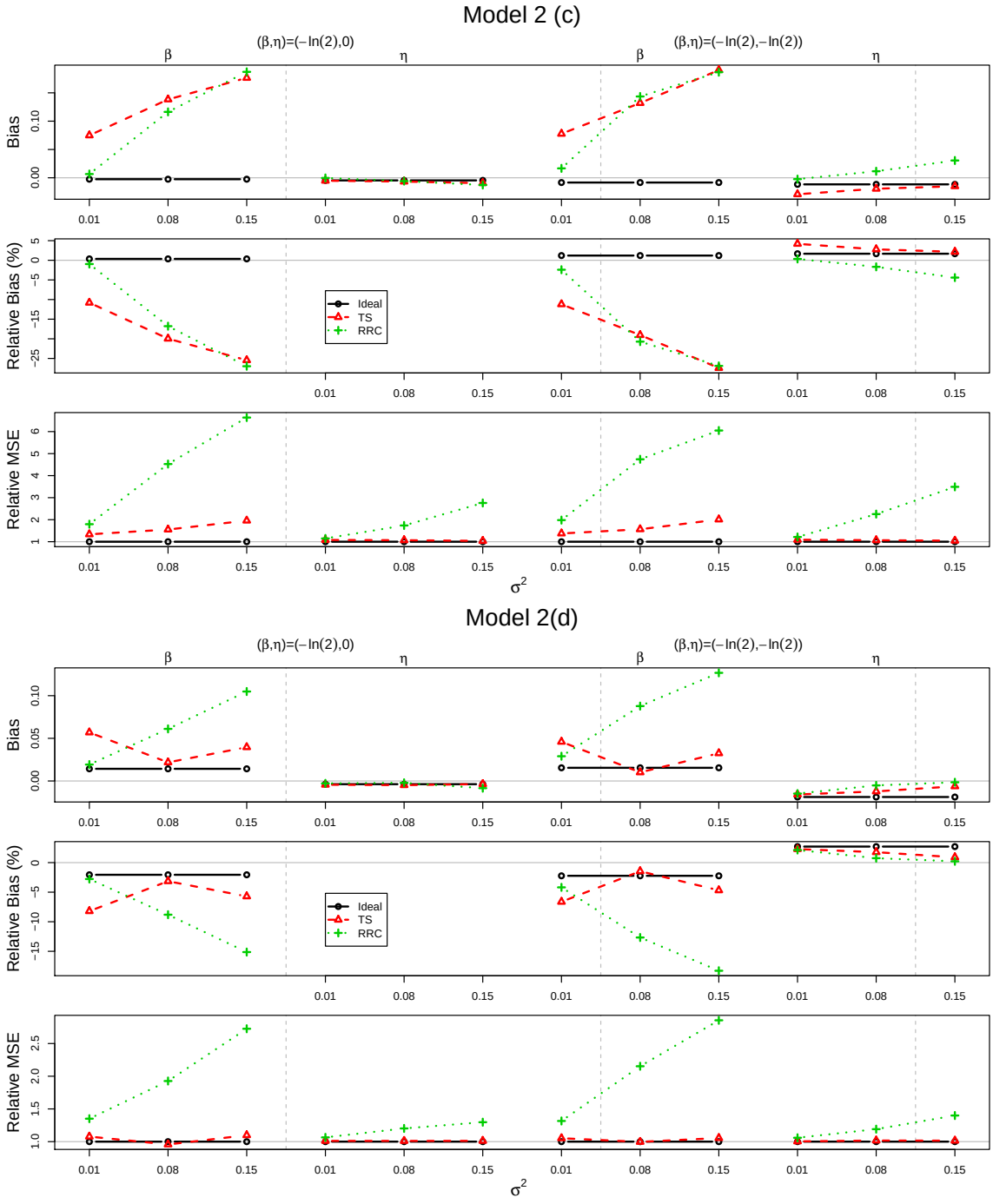


Table 5.9: Simulation results on RRC bootstrap standard error estimates for Model 2(a) with $(\beta, \eta)^T = (-\ln 2, 0)^T$.

	σ^2	Bootstrap SE	MCSD
$\beta = -\ln 2$	0.01	0.271	0.264
	0.08	0.357	0.349
	0.15	0.422	0.384
$\eta = 0$	0.01	0.212	0.201
	0.08	0.212	0.202
	0.15	0.214	0.206

The results are based on $B = 50$ bootstrap samples and 200 simulation runs.

5.1.3 Model 3

We also consider the model with the interaction effect of the vaccination and dichotomized biomarker $\lambda(u) = \lambda_0(u) \exp\{\beta B(u) + \eta Z + \gamma B(u) Z\}$. The simulation datasets are generated exactly the same as those in Model 2 with hazard ratios $(e^\beta, e^\eta)^T = (0.5, 0.5)^T$. This actually indicates a simulation setting with true $e^\gamma = 1$. See Table 5.10-5.13 and Figure 5.5-5.6 for simulation results.

For this interaction model, if the event rate is low, the RRC method produces much smaller biases for $\hat{\beta}$ than does the TS method. The biases from the TS method reduce as the number of measurements get very large. In the setting with high event rate, even the RRC method could give biases larger than 0.1. For γ, both RRC and TS methods provide estimates with very small bias in rare event rate setting. As the event rate gets high, the TS and RRC estimates can be quite biased. Also we found it could be very unstable to fit such an interaction model, with only less than 80% of the simulations runs converged. The interaction model is actually rarely examined in existing literatures on joint modeling methods. The ASE for the interaction model could overestimate the standard error to as high as 50%. Therefore we still recommend using the bootstrap method.

Table 5.10: Simulation results for Model 3(a), with low event rates and moderate numbers of longitudinal immune response measurements.

	σ^2	Method	Bias(% Bias)	ASE	MCSD	RMSE
$\beta = -\ln 2$		Ideal	-0.015 (2.121)	0.255	0.249	1.000
	0.01	TS	0.083 (-11.988)	0.254	0.274	1.321
	0.01	RRC	-0.015 (2.114)	0.454	0.293	1.386
	0.08	TS	0.185 (-26.736)	0.251	0.278	1.797
	0.08	RRC	-0.017 (2.460)	0.780	0.400	2.576
	0.15	TS	0.238 (-34.318)	0.250	0.275	2.125
	0.15	RRC	0.036 (-5.195)	1.081	0.459	3.412
$\eta = -\ln 2$		Ideal	-0.001 (0.162)	0.291	0.295	1.000
	0.01	TS	0.002 (-0.283)	0.296	0.327	1.233
	0.01	RRC	0.019 (-2.779)	0.397	0.276	0.883
	0.08	TS	-0.008 (1.143)	0.296	0.329	1.246
	0.08	RRC	0.013 (-1.804)	0.617	0.340	1.336
	0.15	TS	-0.003 (0.493)	0.295	0.324	1.207
	0.15	RRC	0.003 (-0.489)	0.812	0.397	1.819
$\gamma = 0$		Ideal	-0.011	0.425	0.424	1.000
	0.01	TS	-0.030	0.427	0.522	1.524
	0.01	RRC	-0.065	0.729	0.448	1.141
	0.08	TS	-0.009	0.426	0.501	1.397
	0.08	RRC	-0.052	1.236	0.623	2.175
	0.15	TS	-0.022	0.426	0.503	1.414
	0.15	RRC	-0.039	1.626	0.733	3.002

Sample size is N=800. The longitudinal measurements of W_{ij} are made at baseline and randomly from time windows $3, 6, 9, \cdots, 27 \pm 0.3$, resulting on average 8.0 measurements available per subject.

Table 5.11: Simulation results for Model 3(b), with low event rates and large numbers of longitudinal immune response measurements.

	σ^2	Method	Bias(% Bias)	ASE	MCSD	RMSE
$\beta = -\ln 2$		Ideal	0.018 (-2.566)	0.256	0.256	1.000
	0.01	TS	0.059 (-8.579)	0.256	0.266	1.126
	0.01	RRC	0.021 (-3.061)	0.388	0.275	1.153
	0.08	TS	0.081 (-11.693)	0.254	0.261	1.134
	0.08	RRC	0.024 (-3.472)	0.485	0.307	1.442
	0.15	TS	0.102 (-14.764)	0.253	0.260	1.181
	0.15	RRC	0.038 (-5.459)	0.537	0.319	1.566
$\eta = -\ln 2$		Ideal	-0.001 (0.138)	0.297	0.299	1.000
	0.01	TS	-0.008 (1.121)	0.301	0.308	1.060
	0.01	RRC	0.023 (-3.299)	0.357	0.283	0.899
	0.08	TS	-0.005 (0.714)	0.299	0.299	1.001
	0.08	RRC	-0.010 (1.483)	0.419	0.333	1.238
	0.15	TS	-0.004 (0.625)	0.298	0.313	1.090
	0.15	RRC	0.007 (-1.044)	0.456	0.347	1.341
$\gamma = 0$		Ideal	-0.011	0.426	0.431	1.000
	0.01	TS	0.002	0.426	0.439	1.037
	0.01	RRC	-0.041	0.583	0.402	0.879
	0.08	TS	-0.005	0.426	0.430	0.994
	0.08	RRC	-0.003	0.736	0.510	1.399
	0.15	TS	-0.009	0.426	0.448	1.081
	0.15	RRC	-0.043	0.836	0.554	1.665

Sample size is N=800. The longitudinal measurements of W_{ij} are made at baseline and randomly from time windows $0.5, 1.0, 1.5, \cdots, 28 \pm 0.05$, resulting on average 45.0 measurements available per subject.

Table 5.12: Simulation results for Model 3(c), with high event rates and moderate numbers of longitudinal immune response measurements.

	σ^2	Method	Bias(% Bias)	ASE	MCSD	RMSE
$\beta = -\ln 2$		Ideal	-0.027 (3.966)	0.268	0.276	1.000
	0.01	TS	0.022 (-3.155)	0.265	0.382	1.904
	0.01	RRC	0.072 (-10.416)	1.379	0.381	1.965
	0.08	TS	0.104 (-15.012)	0.249	0.350	1.741
	0.08	RRC	0.308 (-44.368)	1.888	0.592	5.794
	0.15	TS	0.150 (-21.605)	0.246	0.332	1.729
	0.15	RRC	0.373 (-53.871)	1.991	0.642	7.185
$\eta = -\ln 2$		Ideal	-0.040 (5.838)	0.319	0.321	1.000
	0.01	TS	-0.119 (17.121)	0.321	0.492	2.444
	0.01	RRC	0.037 (-5.389)	1.251	0.363	1.266
	0.08	TS	-0.063 (9.036)	0.289	0.423	1.741
	0.08	RRC	0.138 (-19.918)	1.685	0.586	3.455
	0.15	TS	-0.069 (9.996)	0.281	0.390	1.497
	0.15	RRC	0.176 (-25.405)	1.751	0.607	3.806
$\gamma = 0$		Ideal	0.038	0.384	0.376	1.000
	0.01	TS	0.125	0.387	0.618	2.778
	0.01	RRC	-0.053	1.507	0.432	1.321
	0.08	TS	0.065	0.369	0.558	2.204
	0.08	RRC	-0.169	2.252	0.760	4.240
	0.15	TS	0.086	0.365	0.542	2.103
	0.15	RRC	-0.194	2.475	0.867	5.515

Sample size is N=160. The longitudinal measurements of W_{ij} are made at baseline and randomly from time windows $3, 6, 9, \cdots, 27 \pm 0.3$, resulting on average 4.1 measurements available per subject.

Table 5.13: Simulation results for Model 3(d), with high event rates and large numbers of longitudinal immune response measurements.

	σ^2	Method	Bias(% Bias)	ASE	MCSD	RMSE
$\beta = -\ln 2$		Ideal	0.003 (-0.387)	0.270	0.292	1.000
	0.01	TS	0.032 (-4.559)	0.269	0.301	1.075
	0.01	RRC	0.031 (-4.500)	0.856	0.329	1.277
	0.08	TS	-0.017 (2.439)	0.256	0.293	1.007
	0.08	RRC	0.122 (-17.618)	1.164	0.399	2.044
	0.15	TS	0.002 (-0.324)	0.252	0.286	0.962
	0.15	RRC	0.163 (-23.492)	1.243	0.397	2.159
$\eta = -\ln 2$		Ideal	-0.037 (5.403)	0.320	0.331	1.000
	0.01	TS	-0.038 (5.498)	0.319	0.357	1.158
	0.01	RRC	-0.014 (2.044)	0.819	0.364	1.191
	0.08	TS	-0.050 (7.184)	0.301	0.339	1.053
	0.08	RRC	0.027 (-3.964)	1.058	0.410	1.514
	0.15	TS	-0.046 (6.638)	0.293	0.322	0.951
	0.15	RRC	0.041 (-5.948)	1.096	0.393	1.401
$\gamma = 0$		Ideal	0.023	0.385	0.402	1.000
	0.01	TS	0.028	0.385	0.439	1.192
	0.01	RRC	-0.006	0.941	0.453	1.263
	0.08	TS	0.054	0.373	0.428	1.145
	0.08	RRC	-0.057	1.326	0.543	1.834
	0.15	TS	0.061	0.369	0.403	1.022
	0.15	RRC	-0.063	1.445	0.534	1.782

Sample size is N=160. The longitudinal measurements of W_{ij} are made at baseline and randomly from time windows $0.5, 1.0, 1.5, \cdots, 28 \pm 0.05$, resulting on average 22.0 measurements available per subject.

Figure 5.5: Summary of simulation results for Model 3(a) and Model 3(b) with low event rates.

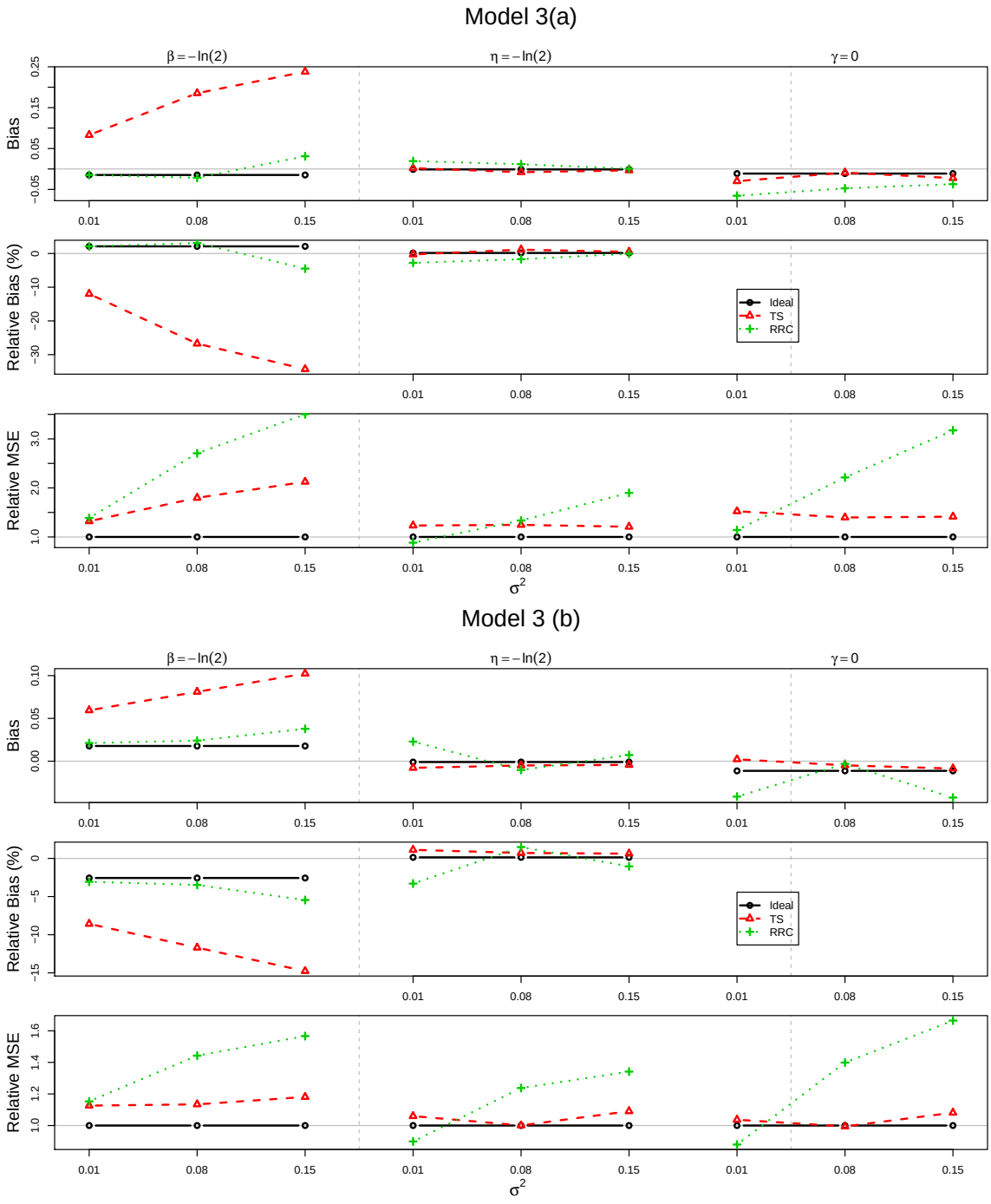
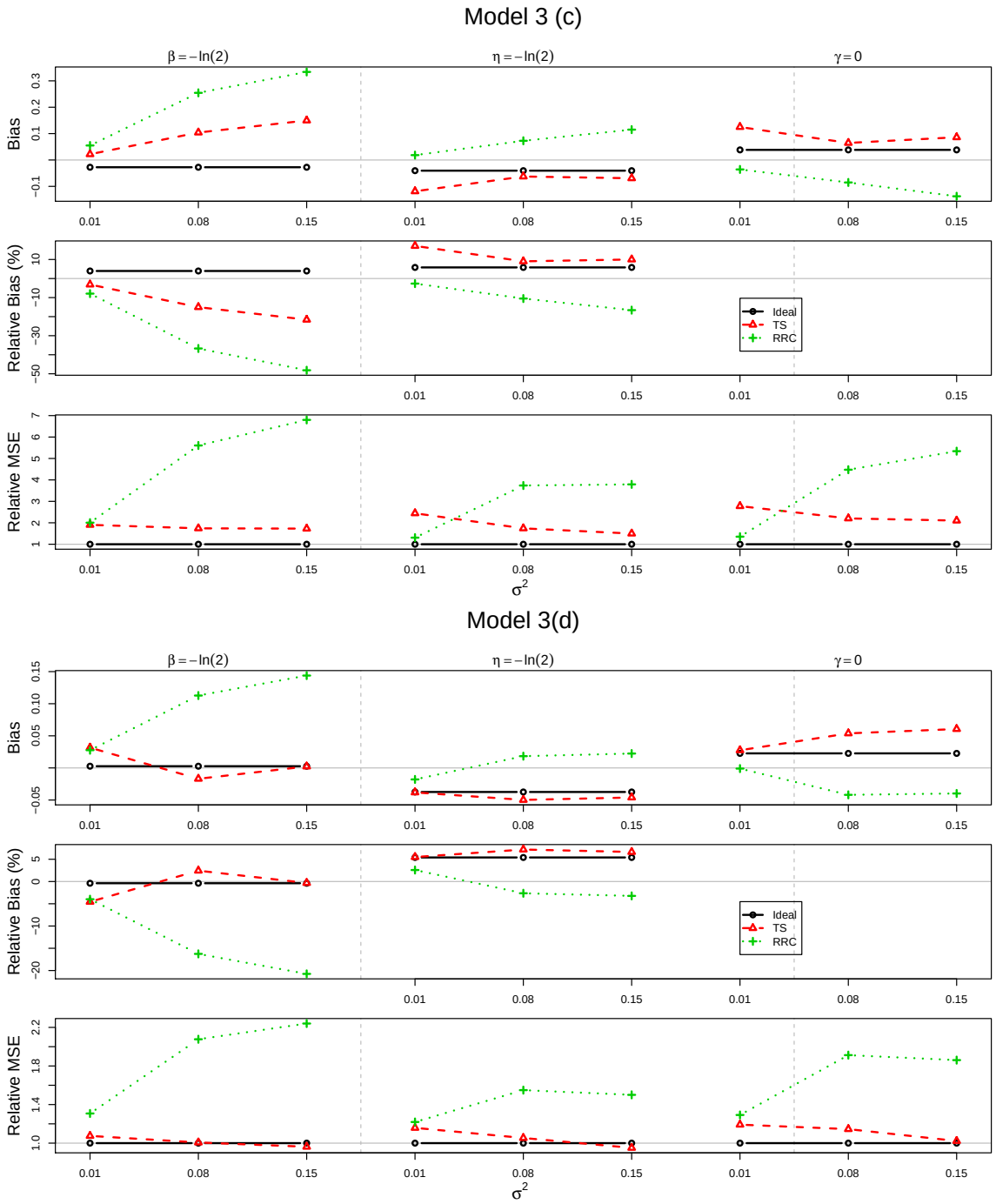


Figure 5.6: Summary of simulation results for Model 2(c) and Model 2(d) with high event rates.



5.1.4 Model 4

We finally consider the model with two dichotomized biomarkers $\lambda(u) = \lambda_0(u)\exp\{\beta_1 B_1(u) + \beta_2 B_2(u) + \eta Z\}$. The biomarkers $X_1(u)$ and $X_2(u)$ are generated as $X_1(u) = \alpha_1 + \alpha_2 u$, and $X_2(u) = \alpha_3 + \alpha_4 u$, where $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^T$ is multivariate Normal distribution with mean $\mathbb{E}[\alpha] = (2.575, -0.009, 2.925, -0.003)^T$ and covariance $\mathbb{C}ov[\alpha] = \Sigma$

$$\Sigma = \begin{pmatrix} 0.0191 & 7e-05 & 0.01036 & -6e-5 \\ 7e-05 & 0.0002 & 1e-5 & 1e-6 \\ 0.01036 & 1e-5 & 0.0354 & -0.00011 \\ -6e-5 & 1e-6 & -0.00011 & 0.0003 \end{pmatrix}$$

The vaccination indicator Z is generated from $Bernoulli(0.5)$. The censoring time is simulated from $Exp(1/80)$ with an administrative censoring time at 30. We consider only the scenario (a) with low event rate, moderate number of measurements and sample size $N = 800$. We simulate the survival data with hazard ratios $(e^{\beta_1}, e^{\beta_2}, e^{\eta}) = \{(0.5, 0.5, 1)^T, (0.5, 0.5, 0.5)^T\}$. The average event rates under those two sets of hazard ratios are 11.1% and 10.8% respectively. See Table 5.14 and Figure 5.7 for the simulation results.

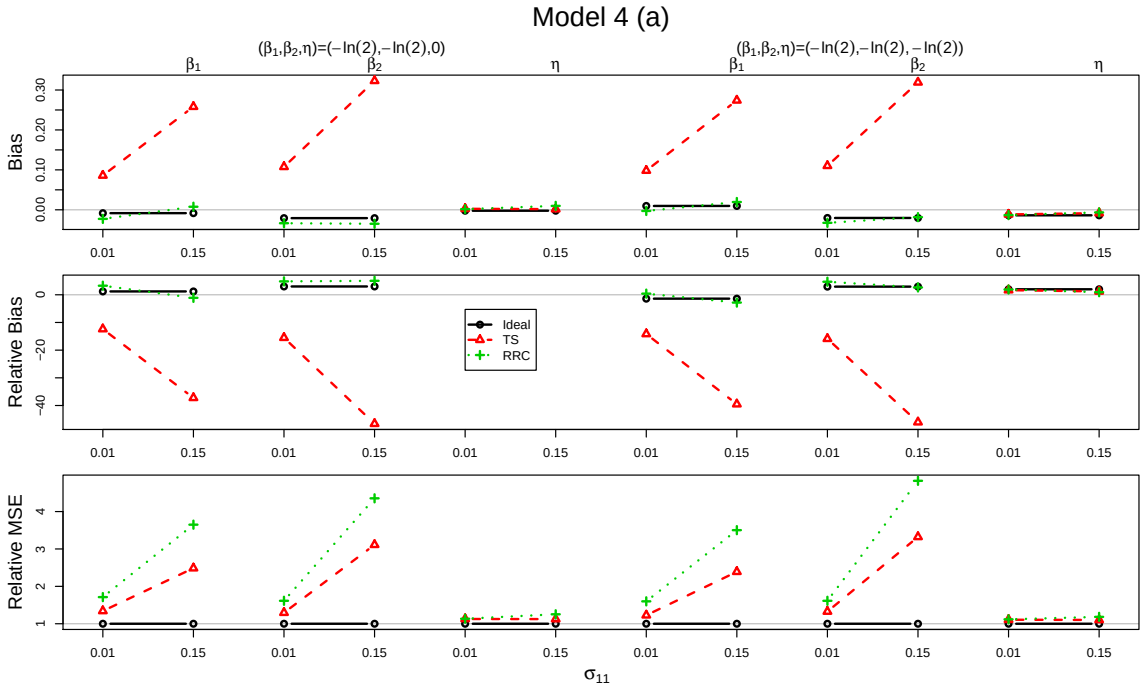
As in Model 2(a) with one biomarker, here with two biomarkers, we also see that both TS and RRC methods provide very small biases for $\hat{\eta}$, and the TS method tends to give slightly smaller MCSD for large measurement error setting. As for estimating β_1 and β_2, still only the RRC method performs well. The ASE estimates for all three parameters are very close to the corresponding MCSDs.

Table 5.14: Simulation results for Model 4(a), with low event rates and moderate numbers of longitudinal immune response measurements.

		σ_{11}	Method	Bias(% Bias)	ASE	MCSD	RMSE
(β_1, β_2)	$\beta_1 = -\ln 2$		Ideal	-0.008 (1.212)	0.227	0.220	1.000
$=$		0.01	TS	0.086 (-12.394)	0.235	0.240	1.341
$(-\ln 2, -\ln 2)$		0.01	RRC	-0.023 (3.280)	0.279	0.287	1.712
$\eta = 0$		0.15	TS	0.258 (-37.234)	0.232	0.232	2.485
		0.15	RRC	0.008 (-1.104)	0.433	0.420	3.649
	$\beta_2 = -\ln 2$		Ideal	-0.021 (3.006)	0.219	0.222	1.000
		0.01	TS	0.108 (-15.516)	0.229	0.230	1.295
		0.01	RRC	-0.034 (4.846)	0.276	0.281	1.614
		0.15	TS	0.323 (-46.591)	0.228	0.225	3.112
		0.15	RRC	-0.035 (5.044)	0.443	0.464	4.354
	$\eta = 0$		Ideal	-0.002	0.214	0.202	1.000
		0.01	TS	0.003	0.226	0.215	1.127
		0.01	RRC	0.002	0.225	0.215	1.134
		0.15	TS	0.002	0.226	0.214	1.124
		0.15	RRC	0.010	0.237	0.226	1.253
(β_1, β_2)	$\beta_1 = -\ln 2$		Ideal	0.010 (-1.400)	0.231	0.234	1.000
$=$		0.01	TS	0.098 (-14.161)	0.239	0.241	1.227
$(-\ln 2, -\ln 2)$		0.01	RRC	-0.003 (0.411)	0.283	0.297	1.599
$\eta = -\ln 2$		0.15	TS	0.274 (-39.525)	0.236	0.238	2.390
		0.15	RRC	0.020 (-2.815)	0.442	0.439	3.503
	$\beta_2 = -\ln 2$		Ideal	-0.020 (2.943)	0.223	0.219	1.000
		0.01	TS	0.110 (-15.889)	0.232	0.228	1.326
		0.01	RRC	-0.033 (4.708)	0.281	0.278	1.613
		0.15	TS	0.319 (-46.034)	0.232	0.243	3.321
		0.15	RRC	-0.018 (2.652)	0.451	0.483	4.823
	$\eta = -\ln 2$		Ideal	-0.014 (1.991)	0.230	0.227	1.000
		0.01	TS	-0.012 (1.669)	0.242	0.239	1.107
		0.01	RRC	-0.013 (1.906)	0.242	0.240	1.118
		0.15	TS	-0.008 (1.204)	0.242	0.239	1.101
		0.15	RRC	-0.007 (0.940)	0.254	0.247	1.183

Sample size is N=800. The longitudinal measurements of W are made at baseline and randomly from time windows $3, 6, 9, \cdots, 27 \pm 0.3$, resulting on average 8.0 measurements available per subject.

Figure 5.7: Summary of simulation results for Model 4(a).



5.1.5 Alternative fitting on Model 1

As discussed at the end of Chapter 4, subjects experiencing an event before enough visits for measuring biomarker data have been made make no contribution to the estimating equations $U_F^R(\theta) = 0$ (if the nuisance parameters $\mu(u)$ and $\Sigma(u)$ are estimated using only subjects with more than q measurements). Also from the simulation studies above we have seen that the number of measurements does have an influence on performance of the RRC and TS estimates, especially for the high-event-rate settings. The reason for this could be because more cases are getting involved in the estimation procedure. We now consider an alternative setting where we hope not to lose cases by considering the hazard of event occurring after some time point (for example, the third scheduled visits for biomarker measurements), and censoring the event occurring before that time point. So in such an analysis, the cases experiencing an event would have made at least the first several visits for measuring biomarker data.

We modified Model 1(a) and Model 1(c) accordingly in this way. We define the new event to occur after time $u = 8$. In this way, all cases would have at least three measurements. We call the new simulation models as Model 1(a^*) and Model 1(c^*). For Model 1(a^*), the average event rates under hazard ratios 1, 0.75, 0.5, and 0.25 are 8.0%, 9.0%, 8.0% and 10.0%, respectively. For Model 1(c^*), the average event rates are 35.8%, 38.2%, 38.0% and 38.1%. This indicates that in Model 1(c), around half of the events occur within $u \in [0, 8]$. Table 5.15,5.16 and Figure 5.8 summarize the simulation results. We see that in the setting with high event rate and moderate or unity hazard ratio, the TS estimator is majorly improved and could even perform better than the RRC estimator. The RRC method still performs as well as it does in Model 1(c).

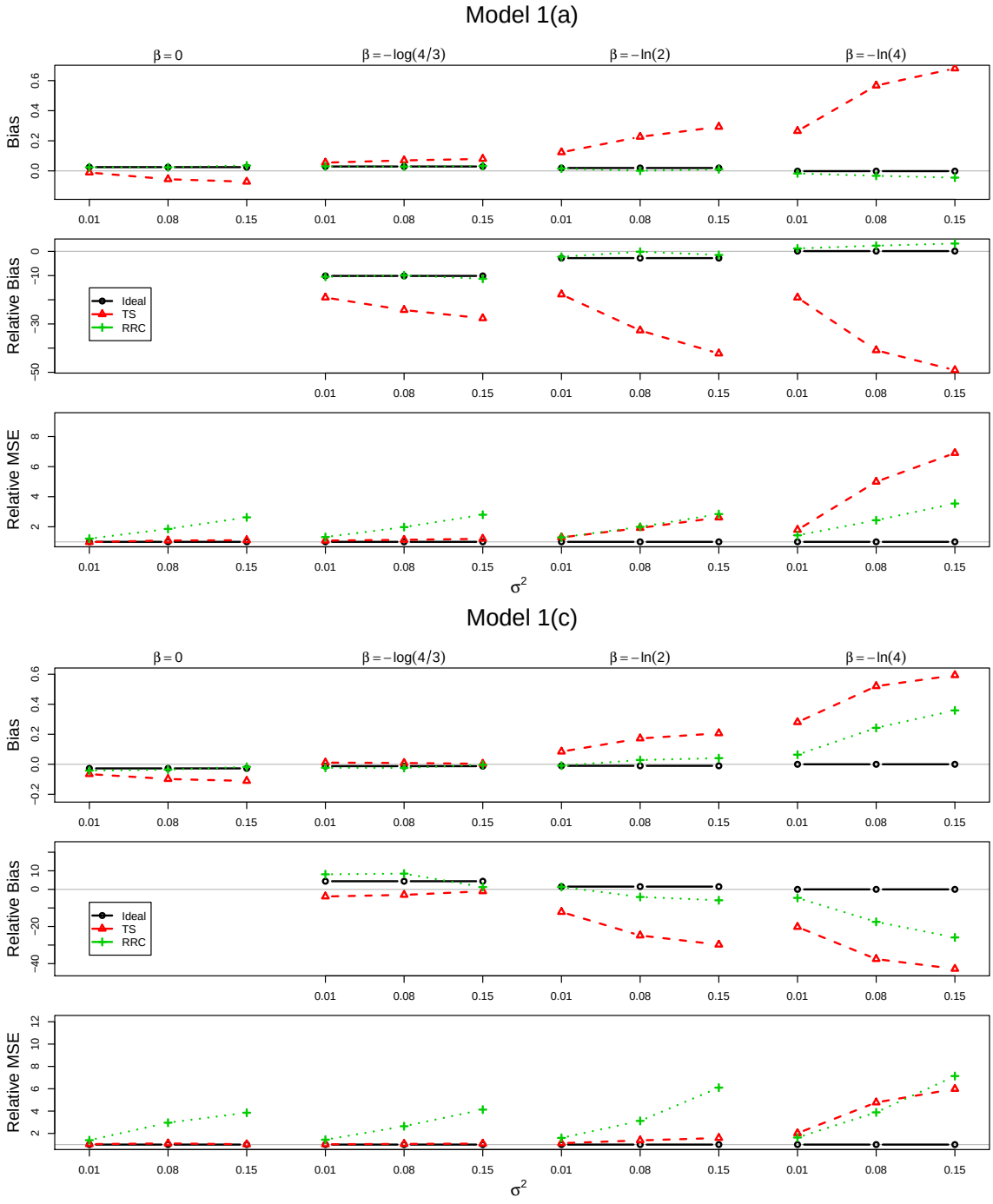
Table 5.15: Simulation results for Model 1(a^*).

	σ^2	Method	Bias(% Bias)	ASE	MCSD	RMSE
$\beta = 0$		Ideal	0.025	0.258	0.255	1.000
	0.01	TS	-0.011	0.258	0.254	0.983
	0.01	RRC	0.024	0.296	0.281	1.214
	0.08	TS	-0.055	0.257	0.262	1.090
	0.08	RRC	0.024	0.371	0.349	1.859
	0.15	TS	-0.072	0.256	0.260	1.106
	0.15	RRC	0.036	0.421	0.414	2.627
$\beta = -\ln 4/3$		Ideal	0.029 (-10.136)	0.257	0.248	1.000
	0.01	TS	0.055 (-19.068)	0.257	0.254	1.083
	0.01	RRC	0.030 (-10.509)	0.295	0.285	1.320
	0.08	TS	0.070 (-24.244)	0.256	0.256	1.134
	0.08	RRC	0.028 (-9.880)	0.370	0.350	1.979
	0.15	TS	0.080 (-27.660)	0.256	0.262	1.202
	0.15	RRC	0.033 (-11.330)	0.420	0.416	2.801
$\beta = -\ln 2$		Ideal	0.019 (-2.772)	0.246	0.234	1.000
	0.01	TS	0.123 (-17.804)	0.244	0.237	1.293
	0.01	RRC	0.015 (-2.225)	0.283	0.268	1.297
	0.08	TS	0.227 (-32.686)	0.241	0.235	1.922
	0.08	RRC	0.001 (-0.137)	0.357	0.332	1.996
	0.15	TS	0.293 (-42.229)	0.240	0.243	2.617
	0.15	RRC	0.010 (-1.470)	0.408	0.397	2.844
$\beta = -\ln 4$		Ideal	-0.002 (0.119)	0.266	0.275	1.000
	0.01	TS	0.265 (-19.104)	0.250	0.256	1.795
	0.01	RRC	-0.017 (1.227)	0.313	0.329	1.432
	0.08	TS	0.567 (-40.893)	0.238	0.237	4.994
	0.08	RRC	-0.033 (2.366)	0.412	0.428	2.436
	0.15	TS	0.681 (-49.127)	0.235	0.240	6.902
	0.15	RRC	-0.045 (3.243)	0.494	0.516	3.545

Table 5.16: Simulation results for Model 1(c^*).

	σ^2	Method	Bias(% Bias)	ASE	MCSD	RMSE
$\beta = 0$		Ideal	-0.027	0.278	0.276	1.000
	0.01	TS	-0.066	0.276	0.275	1.040
	0.01	RRC	-0.042	0.324	0.325	1.395
	0.08	TS	-0.098	0.274	0.274	1.102
	0.08	RRC	-0.036	0.445	0.476	2.959
	0.15	TS	-0.111	0.273	0.258	1.024
	0.15	RRC	-0.017	0.506	0.545	3.853
$\beta = -\ln 4/3$		Ideal	-0.013 (4.380)	0.264	0.257	1.000
	0.01	TS	0.011 (-3.836)	0.264	0.260	1.028
	0.01	RRC	-0.023 (8.118)	0.306	0.308	1.439
	0.08	TS	0.009 (-2.973)	0.263	0.264	1.057
	0.08	RRC	-0.024 (8.479)	0.414	0.417	2.643
	0.15	TS	0.003 (-1.021)	0.262	0.268	1.086
	0.15	RRC	-0.004 (1.247)	0.488	0.523	4.131
$\beta = -\ln 2$		Ideal	-0.010 (1.488)	0.265	0.263	1.000
	0.01	TS	0.084 (-12.178)	0.266	0.266	1.125
	0.01	RRC	-0.009 (1.306)	0.312	0.333	1.598
	0.08	TS	0.172 (-24.841)	0.264	0.256	1.375
	0.08	RRC	0.028 (-4.092)	0.439	0.464	3.120
	0.15	TS	0.206 (-29.783)	0.263	0.260	1.589
	0.15	RRC	0.041 (-5.873)	0.491	0.649	6.100
$\beta = -\ln 4$		Ideal	-0.000 (0.004)	0.265	0.264	1.000
	0.01	TS	0.281 (-20.239)	0.268	0.249	2.021
	0.01	RRC	0.064 (-4.612)	0.316	0.331	1.632
	0.08	TS	0.521 (-37.546)	0.266	0.249	4.772
	0.08	RRC	0.243 (-17.503)	0.464	0.461	3.886
	0.15	TS	0.594 (-42.832)	0.265	0.255	5.984
	0.15	RRC	0.359 (-25.900)	0.526	0.607	7.136

Figure 5.8: Summary of simulation results for Model 1(a^*) and Model 1(c^*).



5.1.6 Investigate the working distributional assumption

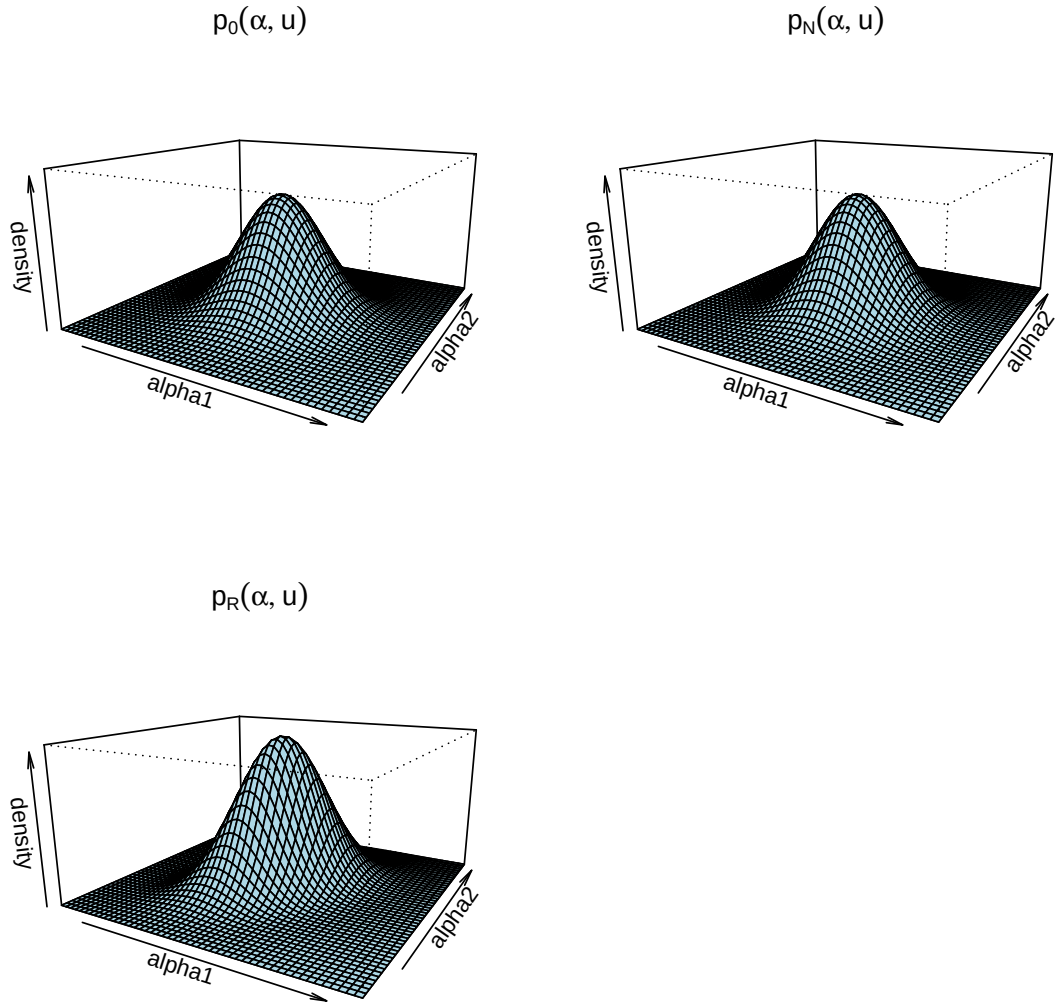
Note our working distributional assumption contains three elements: Normal, mean and variance. We explore which of the three elements in the assumption are most likely to be violated.

First consider the simple setting with $\lambda(u) = \lambda_0 \exp\{-\ln(4)X(u)\}$ in Model 1(a) with $X(u) = \alpha_1 + \alpha_2 u$. The random effects $(\alpha_1, \alpha_2)^T$ are from a bivariate Normal distribution with mean $\mu = (2.575, -0.009)^T$ and covariance Σ with elements $(\Sigma_{11}, \Sigma_{12}, \Sigma_{22})^T = (0.0191, 0.00007, 0.0002)^T$. The baseline hazard is $\lambda_0 = 0.011$. By direct calculation, it is easy to derive the expression of the true density function for $\alpha|T \geq u$. Our goal here is to investigate how much our proposed "Normal" working distribution departs from the truth. We generate one simulation data set and compare the following density functions. At each time when an event occurs, we calculate the parameters and plot the density functions.

1. Let $p_0(\alpha; u)$ denote the true density of $\alpha|T \geq u$, and $\mu_0(u)$, $\Sigma_0(u)$ denote the corresponding true mean and covariance.
2. Let $p_N(\alpha; u)$ denote the Normal density function with true mean and covariance, i.e. $N(\mu_0(u), \Sigma_0(u))$.
3. Let $p_R(\alpha; u)$ denote the Normal density function estimated in the RRC method, i.e. $N(\hat{\mu}(u), \hat{\Sigma}(u))$, where $\hat{\mu}(u)$ and $\hat{\Sigma}(u)$ are estimated from (4.20) and (4.21).

Note $p_0(\alpha; u)$ is the correct conditional probability density function of $\alpha|T \geq u$. For $p_N(\alpha; u)$, it leads to the correct mean and variance, but has incorrect distribution. And $p_R(\alpha; u)$ represents the working distributional assumption. Figure 5.9 shows the plots of density functions at one chosen time point. As shown in this figure, actually across all event time points, the shapes of $p_0(\alpha; u)$ and $p_N(\alpha; u)$ look very similar. This implies in such setting with α Normal, $\alpha|T \geq u$ could also be approximated by a Normal distribution across all time points. However, by comparing $p_0(\alpha; u)$ and $p_R(\alpha; u)$, we found that $p_R(\alpha; u)$ tends to have slightly higher peak, suggesting $\hat{\Sigma}(u)$ might not be a good estimate for $\Sigma_0(u)$.

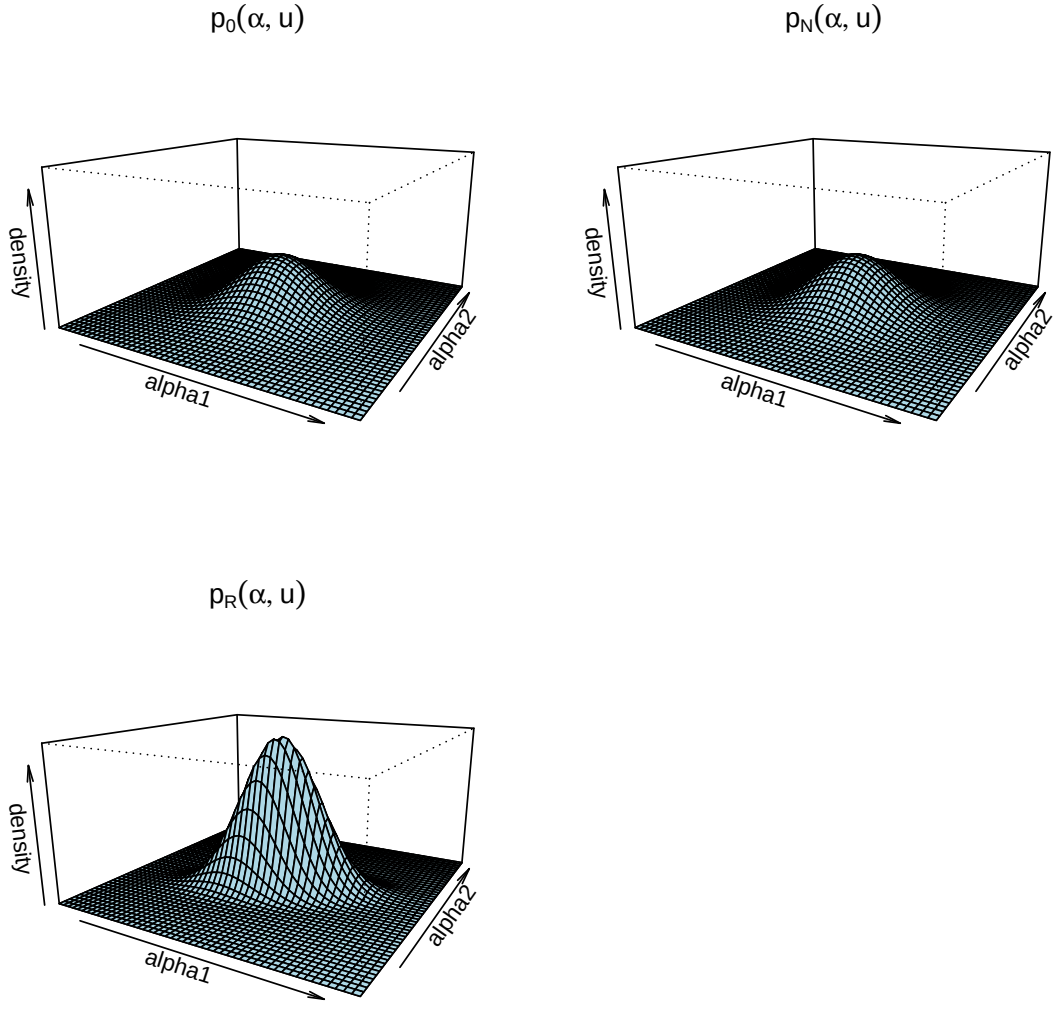
Figure 5.9: Examining the working distributional assumption based on density functions for Model 1(a) with low event rates.



Similar investigation was done for Model 1(c) with high event rate (Figure 5.10). Still $p_0(\alpha; u)$ could be approximated by a Normal distribution well. However, the shape of $p_R(\alpha; u)$ could be very different from the truth. We observed this discrepancy more frequently and more extremely than that in Model 1(a), probably due to a number of events

occurring at the early stage by when only two or three measurements were available.

Figure 5.10: Examining the working distributional assumption based on density functions for Model 1(c) with high event rates.



5.2 Simulation for two-phase sampling design cohort studies

In this section, we evaluate the RRC estimator for a dichotomous biomarker in two phase sampling design studies.The second phase sample is selected based on case-control sampling, including all cases ($\Delta = 1$) and 14.3% of the controls ($\Delta = 0$). We consider the IPW RRC method with pre-specified sampling probabilities (RRC(π)) and with estimated sampling probabilities (RRC($\hat{\pi}$)), as well as the complete-case RRC estimator (CC). We only consider the scenario (a) with low event rate. For the estimates of the standard errors, we consider both the ASE and the bootstrap SE. The RMSE is calculated as the Monte Carlo variance of $\hat{\beta}$ compared to the RRC estimates obtained based on full cohort data.

5.2.1 Model 1(a)

The full cohort data is generated from Model 1(a) in Section 5.1.1, from hazard function $\lambda(u) = \lambda_0(u)\exp\{\beta B(u)\}$. The number of subjects being sampled for measuring immune biomarker data is shown in Table 5.17. The simulation results are in Table5.19. From the results we see that due to only including around 25% of the sample in IPW analysis, we could lose around 50% to 80% of the efficiency. As expected, though, the IPW estimates still have very small biases with relative biases less than 6%. However, the CC estimates could generate a relative bias as high as 25%. The ASE still provides a good estimate of SE($\hat{\beta}$), as what we have seen in the full data analysis.

Table 5.17: The sample size for Phase I (N) and Phase II (n) sample for Model 1(a).

	Case ($\Delta = 1$)	Control ($\Delta = 0$)
β	$N(n)$	$N(n)$
0	94 (94)	706 (101)
$-\ln(2)$	91 (91)	709 (101)
$-\ln(4)$	101 (101)	699 (100)

Table 5.18: The sample size for Phase I (N) and Phase II (n) sample for Model 2(a).

	Case ($\Delta = 1$)	Control ($\Delta = 0$)
(β, η)	$N(n)$	$N(n)$
$(-\ln(2), 0)$	101 (101)	699 (100)
$(-\ln(2), -\ln(2))$	104 (104)	696 (99)

5.2.2 Model 2(a)

We next consider the two-phase sampling on the data generated from Model 2(a) in Section 5.1.2, with $\lambda(u) = \lambda_0(u)\exp\{\beta B(u) + \eta Z\}$. The number of subjects sampled with measurements of immune biomarker data is shown in Table 5.18. The simulation results are in Table5.21. Similar pattern in terms of biases and RMSE as that for Model 1(a) shows up here. The only concern is still that the ASE for $\hat{\eta}$ underestimates SE($\hat{\eta}$) to 30%$\sim$40%, as what we have seen in full data analysis. So we suggest running bootstrap method for the standard error estimates for such models with adjustments. We conduct the bootstrap fixing the Phase II sample. Suppose in the Phase II sample, the number of cases and controls are n_1 and n_0. Then within the Phase II case group, we sample with replacement n_1 subjects to form the bootstrap case subjects, and within the Phase II control group, we sample with replacement n_0 subjects to form the bootstrap control subjects. It is similar to the B2 procedure described in [Odile, 2007] by fixing the case and control group and perform bootstrap within each group independently. See Table 5.20 for the bootstrap SE. Again, the bootstrap yields accurate standard error estimates.

Table 5.19: Simulation results for Model 1(a) with two-phase sampled data.

	σ^2	Method	Bias(% Bias)	ASE	MCSD	RMSE
$\beta = 0$	0.01	RRCfull	0.028	0.279	0.269	1.000
	0.01	CC	0.027	0.297	0.277	1.058
	0.01	RRC(π)	0.032	0.351	0.346	1.649
	0.01	RRC($\hat{\pi}$)	0.032	0.351	0.346	1.649
	0.08	RRCfull	0.030	0.356	0.335	1.000
	0.08	CC	0.033	0.323	0.336	1.003
	0.08	RRC(π)	0.045	0.446	0.415	1.541
	0.08	RRC($\hat{\pi}$)	0.045	0.446	0.416	1.543
	0.15	RRCfull	0.035	0.408	0.396	1.000
	0.15	CC	0.037	0.432	0.418	1.118
	0.15	RRC(π)	0.056	0.519	0.508	1.656
	0.15	RRC($\hat{\pi}$)	0.055	0.521	0.509	1.663
$\beta = -\ln 2$	0.01	RRCfull	0.015 (-2.227)	0.261	0.251	1.000
	0.01	CC	0.172 (-24.807)	0.241	0.254	1.492
	0.01	RRC(π)	0.020 (-2.927)	0.334	0.326	1.691
	0.01	RRC($\hat{\pi}$)	0.020 (-2.878)	0.334	0.326	1.692
	0.08	RRCfull	0.001 (-0.119)	0.338	0.319	1.000
	0.08	CC	0.162 (-23.360)	0.351	0.332	1.339
	0.08	RRC(π)	0.015 (-2.136)	0.430	0.419	1.723
	0.08	RRC($\hat{\pi}$)	0.014 (-2.074)	0.430	0.420	1.727
	0.15	RRCfull	0.009 (-1.235)	0.390	0.377	1.000
	0.15	CC	0.163 (-23.550)	0.370	0.394	1.281
	0.15	RRC(π)	0.038 (-5.453)	0.500	0.505	1.804
	0.15	RRC($\hat{\pi}$)	0.038 (-5.478)	0.500	0.505	1.801
$\beta = -\ln 4$	0.01	RRCfull	-0.008 (0.585)	0.281	0.290	1.000
	0.01	CC	0.302 (-21.759)	0.285	0.282	2.022
	0.01	RRC(π)	-0.006 (0.443)	0.349	0.353	1.479
	0.01	RRC($\hat{\pi}$)	-0.007 (0.494)	0.350	0.353	1.481
	0.08	RRCfull	-0.028 (1.990)	0.380	0.392	1.000
	0.08	CC	0.336 (-24.252)	0.350	0.376	1.645
	0.08	RRC(π)	-0.023 (1.640)	0.472	0.478	1.482
	0.08	RRC($\hat{\pi}$)	-0.024 (1.706)	0.472	0.479	1.486
	0.15	RRCfull	-0.043 (3.082)	0.451	0.459	1.000
	0.15	CC	0.264 (-19.041)	0.422	0.459	1.320
	0.15	RRC(π)	-0.021 (1.503)	0.569	0.581	1.594
	0.15	RRC($\hat{\pi}$)	-0.021 (1.516)	0.569	0.582	1.595

Sample size is N=800. The longitudinal measurements of W_{ij} are made at baseline and randomly from time windows $0.5, 1.0, 1.5, \cdots, 28 \pm 0.05$. The two-phase sampling is conducted by sampling all cases ($\Delta = 1$) and 14.3% of controls ($\Delta = 0$).

Table 5.20: Simulation results on RRC($\hat{\pi}$) bootstrap standard error estimates for Model 2(a) with two-phase sampled data. Regression parameters are $(\beta, \eta)^T = (-\ln 2, 0)^T$.

	σ^2	Bootstrap SE	MCSD
$\beta = -\ln 2$	0.01	0.344	0.348
	0.08	0.480	0.473
	0.15	0.538	0.580
$\eta = 0$	0.01	0.296	0.281
	0.08	0.300	0.284
	0.15	0.301	0.281

The Bootstrap SE is evaluated for RRC($\hat{\pi}$). The results are based on $B = 100$ bootstrap samples and 200 simulation runs.

5.3 Discussion

We have evaluated the performance of the RRC and TS method for joint modeling of dichotomized biomarkers and event time data, as well as the RRC method in the two-phase sampling design studies. Generally speaking, in the rare-event setting the RRC method gives reasonably small bias regardless of the magnitude of measurement error and the size of hazard ratio. In high event rate setting, the RRC estimator tends to be very biased when the relative risk is very high. The TS estimator seems to give intolerable biases generally, even though it produces relatively small MCSD compared to the RRC estimator. The simulation results imply that the RRC estimator could be useful in vaccine efficacy trials where the infection or disease rate is small. The TS estimator, which is much simpler to be implemented, can be potentially used when the measurement errors are very small and the hazard ratios are moderate. We also explored the TS method where the subject-specific trajectory of biomarker was fitted from the mixed effect model by using the R package `lme4` and was fitted from shrinking least squares estimates. We observed only slight improvement with zero coefficient.

The simulation studies also demonstrate that increasing the number of measurements can significantly improve the performance of both RRC and TS methods, especially for the TS method. The TS estimates are more likely to be influenced by the noise-to-signal ratios.

Table 5.21: Simulation results for Model 2(a) with two-phase sampled data.

		σ^2	Method	Bias(% Bias)	ASE	MCSD	RMSE
(β, η)	$\beta = -\ln 2$	0.01	RRCfull	-0.015 (2.132)	0.253	0.265	1.000
=		0.01	CC	0.131 (-18.920)	0.251	0.254	1.161
$(-\ln 2, 0)$		0.01	RRC(π)	-0.034 (4.898)	0.327	0.337	1.636
		0.01	RRC($\hat{\pi}$)	-0.035 (5.002)	0.327	0.338	1.640
		0.08	RRCfull	-0.003 (0.479)	0.328	0.332	1.000
		0.08	CC	0.144 (-20.789)	0.327	0.335	1.210
		0.08	RRC(π)	-0.011 (1.568)	0.423	0.429	1.670
		0.08	RRC($\hat{\pi}$)	-0.011 (1.647)	0.423	0.429	1.674
		0.15	RRCfull	-0.003 (0.413)	0.378	0.404	1.000
		0.15	CC	0.141 (-20.295)	0.373	0.397	1.085
		0.15	RRC(π)	-0.001 (0.185)	0.495	0.521	1.661
		0.15	RRC($\hat{\pi}$)	-0.001 (0.154)	0.496	0.519	1.648
	$\eta = 0$	0.01	RRCfull	0.010	0.179	0.216	1.000
		0.01	CC	0.018	0.179	0.219	1.038
		0.01	RRC(π)	0.019	0.234	0.295	1.873
		0.01	RRC($\hat{\pi}$)	0.019	0.235	0.295	1.879
		0.08	RRCfull	0.010	0.142	0.216	1.000
		0.08	CC	0.017	0.144	0.223	1.077
		0.08	RRC(π)	0.023	0.185	0.299	1.927
		0.08	RRC($\hat{\pi}$)	0.023	0.185	0.299	1.936
		0.15	RRCfull	0.013	0.130	0.223	1.000
		0.15	CC	0.019	0.130	0.226	1.029
		0.15	RRC(π)	0.019	0.172	0.299	1.797
		0.15	RRC($\hat{\pi}$)	0.020	0.172	0.299	1.801
(β, η)	$\beta = -\ln 2$	0.01	RRCfull	-0.024 (3.425)	0.252	0.261	1.000
=		0.01	CC	0.139 (-20.083)	0.247	0.261	1.278
$(-\ln 2, -\ln 2)$		0.01	RRC(π)	-0.024 (3.426)	0.330	0.345	1.745
		0.01	RRC($\hat{\pi}$)	-0.025 (3.537)	0.331	0.346	1.753
		0.08	RRCfull	-0.026 (3.747)	0.327	0.336	1.000
		0.08	CC	0.135 (-19.492)	0.321	0.342	1.192
		0.08	RRC(π)	-0.021 (2.980)	0.430	0.458	1.856
		0.08	RRC($\hat{\pi}$)	-0.021 (3.038)	0.431	0.459	1.863
		0.15	RRCfull	-0.030 (4.268)	0.379	0.417	1.000
		0.15	CC	0.119 (-17.167)	0.374	0.398	0.988
		0.15	RRC(π)	-0.016 (2.301)	0.502	0.536	1.646
		0.15	RRC($\hat{\pi}$)	-0.017 (2.513)	0.503	0.537	1.653
	$\eta = -\ln 2$	0.01	RRCfull	0.004 (-0.575)	0.202	0.242	1.000
		0.01	CC	0.192 (-27.631)	0.203	0.246	1.653
		0.01	RRC(π)	0.003 (-0.431)	0.259	0.323	1.778
		0.01	RRC($\hat{\pi}$)	0.003 (-0.361)	0.259	0.323	1.777
		0.08	RRCfull	0.004 (-0.641)	0.163	0.242	1.000
		0.08	CC	0.196 (-28.316)	0.164	0.245	1.679
		0.08	RRC(π)	-0.000 (0.014)	0.210	0.323	1.775
		0.08	RRC($\hat{\pi}$)	-0.001 (0.091)	0.210	0.323	1.774
		0.15	RRCfull	0.006 (-0.886)	0.149	0.246	1.000
		0.15	CC	0.188 (-27.121)	0.151	0.253	1.646
		0.15	RRC(π)	0.003 (-0.493)	0.197	0.330	1.808
		0.15	RRC($\hat{\pi}$)	0.003 (-0.378)	0.197	0.330	1.805

Sample size is N=800. The longitudinal measurements of W_{ij} are made at baseline and randomly from time windows $0.5, 1.0, 1.5, \cdots, 28 \pm 0.05$. The two-phase sampling is conducted by sampling all cases ($\Delta = 1$) and 14.3% of controls ($\Delta = 0$).

Apparently the larger the measurement error, the more biased the TS estimates. For the RRC estimator, the bias is consistently small across different magnitudes of measurement error in the setting with rare event rate, but the precision does decrease. When the event rate is high, the bias tends to increase with the larger measurement error, though is still smaller than that from the TS method.

Also we explored the estimates for $SE(\hat{\theta})$ using the ASE and Bootstrap SE. The former was used in [Dafni and Tsiatis, 1998] as an approximation of $SE(\hat{\theta})$. In their paper, they investigated only the Cox model with one immune biomarker as the covariate, which is the Model 1 in this dissertation. We both found the ASE could provide a good approximation to $SE(\hat{\beta})$ in such a model. This dissertation also considers the Cox model with adjustment (Model 2) and with interaction of the immune biomarker and the treatment indicator (Model 3). However in Model 2 and Model 3 we did observe the ASE poorly estimates the SE, especially for $\hat{\eta}$ and $\hat{\gamma}$. So as suggested in [Wang et al., 2001], in the latter two settings, we should use the bootstrap SE estimates.

We also examined the working distributional assumption that α is Normal given $(\tilde{Z}, T \geq u)$. All above simulation studies generated α given $\tilde{Z}$ from Normal, where apparently the working assumption did not hold. We found it was most likely to have a poor estimate of the covariance matrix, especially when the size of at-risk set or the number of biomarker measurements were small. We also conducted simulation studies with α from a mixed-Normal distribution. For example for Model 1, α was generated a mixture of two Normal distributions with mean $(2.2295 - 0.009)^T$ and $(2.9205, -0.009)^T$ respectively. The results were similarly to those with Normal α, indicating RRC method could still give small biases.

Chapter 6

DATA ANALYSIS: ACTG 175

6.1 Background

In this chapter we apply the methods developed to AIDS Clinical Trials Group (ACTG) 175 dataset [Hammer et al., 1996]. ACTG 175 was a randomized clinical trial comparing four treatment regimens (zidovudine only, zidovudine+didanosine, zidovudine+zalcitabine and didanosine only) among HIV-infected subjects with CD4 cell counts 200-500 per cubic millimeter. Their primary study included 2,467 subjects. The study was designed to measure the CD4 cell counts on all study subjects at a schedule of every 12 weeks since Week 8. The study subjects were followed up for endpoint of $\geq$50% decline in the CD4 cell count, AIDS or death. The median follow-up time was 143 weeks. Taking the group with zidovudine only as the reference group, the original study on all study subjects found significant effect of each of the other three on the endpoint.

Now for this study, we assess the CD4 cell count as a time-dependent CoR and CoP. As in [Song et al., 2002], the treatment ($Z = 1$) of interest is combining the three of zidovudine+didanosine, zidovudine+zalcitabine and didanosine, and zidovudine only is considered as the control treatment ($Z = 0$). The primary clinical endpoint of interest in the correlates of protection analysis is the progression to AIDS or death. We refer any subject who experienced the primary endpoint during the study as a case and who never had the primary endpoint throughout the study as a control. A total of 308 cases were observed. Note this study was not originally designed in a two-phase manner to measure the CD4 cell counts. For a better demonstration of our proposed methods on this data, we created a case-control sample from the full study cohort and pretended the CD4 cell counts were measured only on subjects in the case-control sample. The case-control sampling was conducted to include all cases and 14.3% of the controls with CD4 cell counts available. This led to 306 cases and 306 controls. We only considered the CD4 cell counts measured before

the primary endpoint for cases. The average number of CD4 cell counts measurements per subject was 9.25.

6.2 Descriptive analysis

As in [Song et al., 2002], we analyzed the inherent trajectory of $\log_{10}$ cell CD4 counts. Figure 6.1 shows the observed trajectory of $\log_{10}$ cell CD4 counts by treatment groups and by subgroups cross-classified by treatment/control treatment and case/control status. The red lines are the smoothed mean curves and the shaded areas represent the pointwise 95% confidence intervals. Apparently cases tend to have steeper drop over time in their trajectories than controls, and subjects in the treatment group have on average higher $\log_{10}$ CD4 cell counts than those in the control treatment group.

Then we explored if the level of $\log_{10}$ CD4 cell count is predictive of the progression of AIDS or death. We plotted the cumulative incidence rate of the primary endpoint by subgroups with low, medium and high $\log_{10}$ CD4 cell counts measured at the visit of Week 8 (Figure 6.2). The subgroups were defined by the tertiles of $\log_{10}$ CD4 cell counts measured at the visit of Week 8 pooled control treatment and treatment group, taking into account the case-control sampling weights. The plot shows that higher log CD4 cell counts are associated with lower rates of the event. And the difference in the cumulative incidence rates comparing the low and medium group is greater than that comparing the medium and high group.

Figure 6.1: Spaghetti plot of observed $\log_{10}$ CD4 cell counts with smoothed mean curves and pointwise 95% confidence intervals by subgroups.

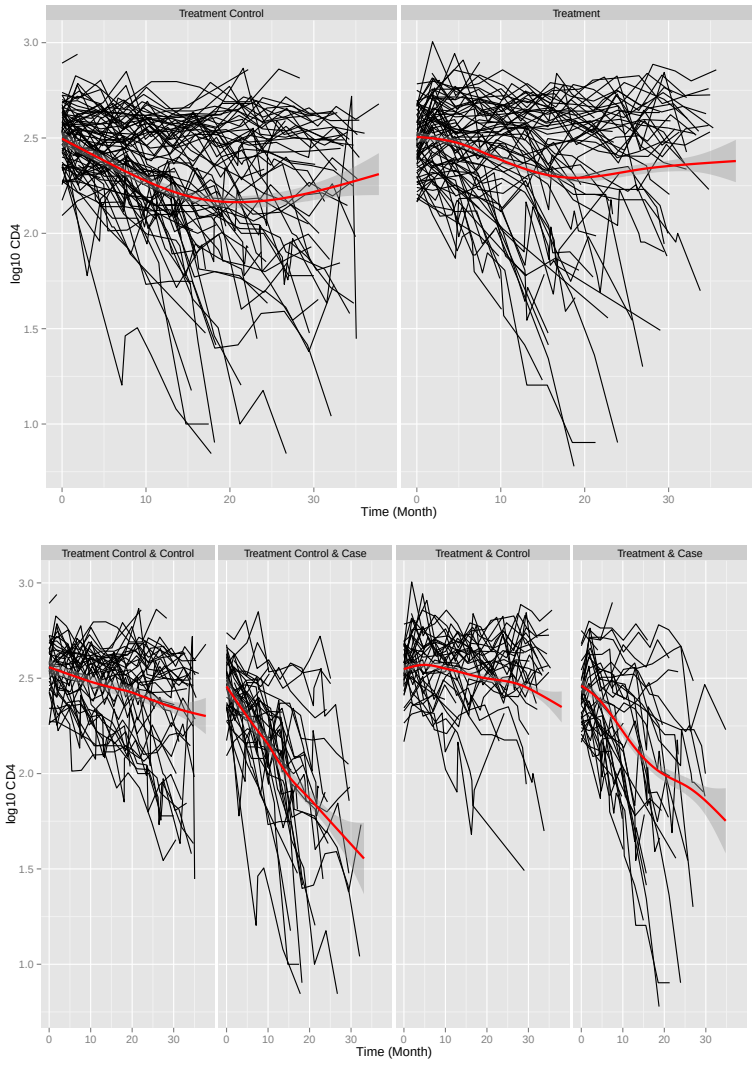
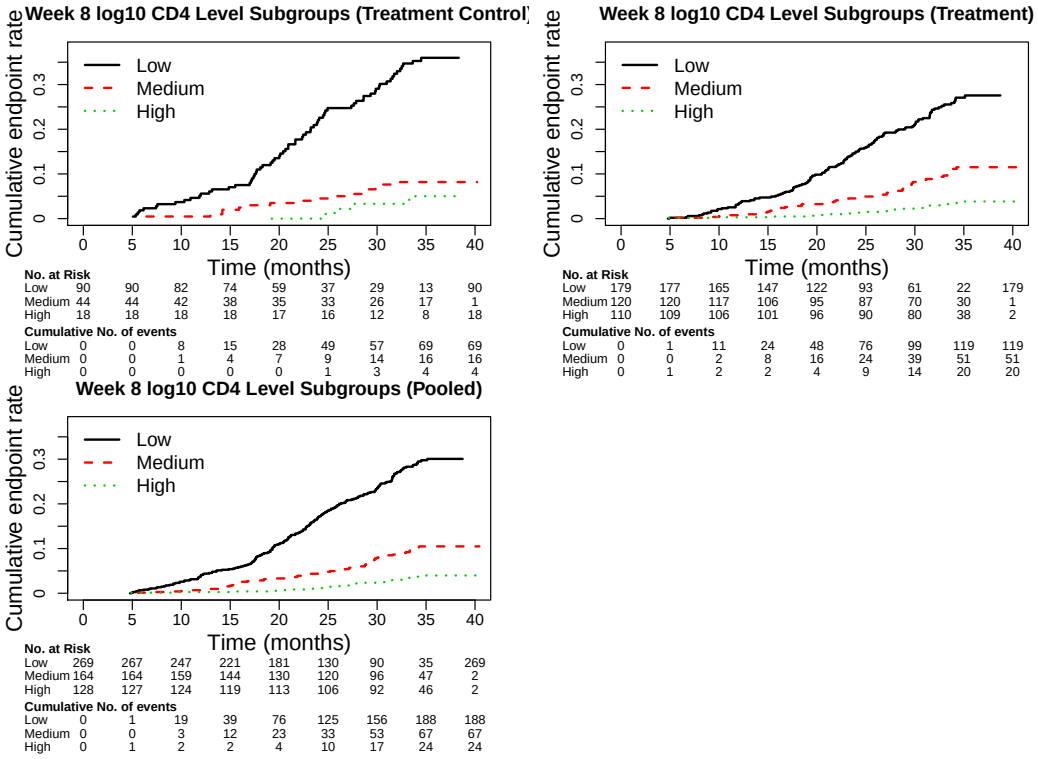


Figure 6.2: Cumulative incidence plot of the primary endpoint by subgroups of $\log_{10}$ CD4 cell count levels (low, medium and high) at the visit of Week 8.



6.3 Analysis of continuous trajectory of $\log_{10}$ *CD4 cell counts*

Before we conduct the correlates analysis on $\log_{10}$ CD4 cell counts, we explore the functional form of its trajectory $X(u)$ over time. We investigate the simple linear (L1) and quadratic (L2) weighted mixed effects models with fixed effects β's and mean-zero Normal random effects b's on all case-control samples.

L1: $X(u) = \beta_0 + \beta_1 u + d_0 + d_1 u$

L2: $X(u) = \beta_0 + \beta_1 u + \beta_2 u^2 + d_0 + d_1 u + d_2 u^2$

The results on fitted model using maximum-likelihood estimation (MLE) from `lme()` are shown in Table 6.1. Figure 6.3 is the spaghetti plot of observed $\log_{10}$ cell CD4 counts on 10 randomly selected subjects as well as fitted lines from the linear and quadratic mixed effects models. Apparently, the results in Table 6.1 indicates the quadratic model provides a better fitting than does the simple linear model. The mixed effects models in L1 and L2 are only used to help to chose the functional form of the trajectory of $\log_{10}$ cell CD4 over time. To model it in the Cox regression model, we use the random effects model as discussed in (2.1). We model the trajectory of $\log_{10}$ CD4 $X(u)$ as being quadratic in time, i.e. $X(u) = \alpha_0 + \alpha_1 u + \alpha_2 u^2$.

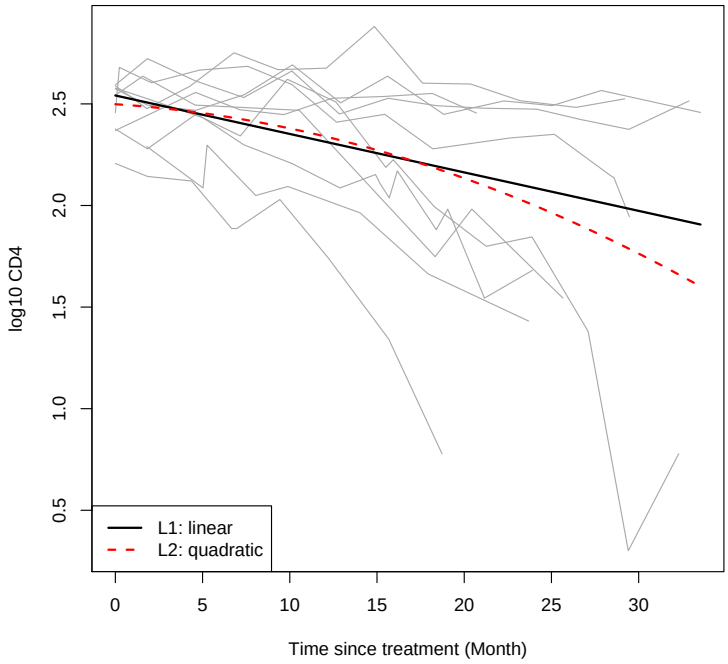
Table 6.1: Results on fixed effects from Model L1 and L2.

Model	Intercept	Time	Time^2	log-likelihood	P-value[1]
L1	2.5414	-0.0189		891.85	
L2	2.4984	-0.0056	0.0006	1397.50	<0.0001

[1] Likelihood-ratio test p-value.

We now evaluate the $\log_{10}$ CD4 cell count as time-dependent CoP in Prentice's framework by a set of Cox regression models. The Prentice's criteria are given in Section 2.3.1. The models are adjusted for the average baseline $\log_{10}$ CD4 cell counts measured at two time points before randomization. See Table 6.2 for the estimated coefficients from IPW and AIPW methods using estimated $\hat{\pi}$. The two methods yield similar results. The variables used in predicting the augmentation terms in the AIPW method are the treatment

Figure 6.3: Spaghetti plot of observed $\log_{10}$ CD4 cell counts on 10 randomly selected subjects and the fitted lines from linear and quadratic mixed effects models.



and event status indicators. No significant efficiency improvement is found on all coefficient estimates for this analysis. We therefore focus on the results from IPW method, since it is simpler to implement for further exploration as discussed below.

Model 0 shows that the treatment significantly affects the primary endpoint. Model 3 and Model 4 indicate that $\log_{10}$ CD4 cell count has no significant effect on the clinical endpoint in the treatment group, but has significant effect in the control treatment group. The effect of $\log_{10}$ CD4 cell count is significantly modified by the treatment group (Model 1). Therefore both the Prentice criteria (ii) and (iii) do not hold. It suggests that $\log_{10}$ CD4 cell fails to be a Prentice surrogate. Though there is a significant interaction effect, we still try to calculate the quantities defined below to quantify the surrogacy of $\log_{10}$ CD4 cell count (Table 6.2).

1. The proportion of treatment effect explained by the surrogate (PTE): $1 - \theta_{trt,2}/\theta_{trt,0}$,

where $\theta_{trt,j}$ is the coefficient for treatment indicator in Model j.

2. The proportion of treatment effect captured by the surrogate based on Prentice's framework (PCS) [Kobayashi and Kuroki, 2014]: $\frac{1}{1+(1/\text{PTE}-1)^2}$
3. The proportion of natural indirect effect of the treatment, assuming the identification assumptions I1-I4 given in Section 2.3.2 hold (pNIE): $\frac{\text{NIE}_0(26)}{\text{NIE}_0(26)+\text{NDE}_0(26)}$, where $\text{NDE}_0(26) = \mathbb{E}[I(T^{1X^0_\alpha} \geq 26) - I(T^{0X^0_\alpha} \geq 26)]$ and $\text{NIE}_0(26) = \mathbb{E}[I(T^{1X^1_\alpha} \geq 26) - I(T^{1X^0_\alpha} \geq 26)]$. Here we chose a time Month 26 which is a relative late stage of follow-up.

All three quantities are very close to zero. For pNIE, the CI was calculated from 500 bootstrap samples. For PTE and PCS, the CIs were determined from the asymptotic Normal distribution of $(\theta_{trt,0}, \theta_{trt,2})^T$.

Table 6.2: Estimates and 95% CIs for the coefficients in Cox regression models, to assess continuous $\log_{10}$ CD4 cell count as time-dependent CoP in Prentice's framework.

		logCD4	trt	logCD4 × trt	bl.logCD4
Pooled	Model 0		-0.37 (-0.61, -0.13)		
		IPW($\hat{\pi}$)			
Pooled	Model 1	-2.14 (-2.81, -1.47)	-3.33 (-5.31, -1.34)	1.42 (0.46, 2.38)	-3.60 (-5.16, -2.03)
	Model 2	-1.04 (-1.93, -0.15)	-0.44 (-0.79, -0.08)		-3.61 (-5.70, -1.51)
Treatment Group	Model 3	-0.73 (-1.68, 0.22)			-4.23 (-6.46, -2.00)
control treatment Group	Model 4	-2.89 (-3.60, -2.18)			-1.03 (-3.68, 1.62)
		AIPW($\hat{\pi}$)			
Pooled	Model 1	-2.35 (-3.22, -1.47)	-2.93 (-6.29, 0.43)	1.36 (-0.22, 2.94)	-1.55 (-2.90, -0.20)
	Model 2	-1.17 (-1.94, -0.39)	-0.35 (-0.67, -0.02)		-2.69 (-6.61, 1.23)
Trt Group	Model 3	-0.92 (-2.24, 0.41)			-3.14 (-7.91, 1.62)
Control Trt Group	Model 4	-3.03 (-3.84, -2.23)			-0.57 (-9.33, 8.20)

[1] For the AIPW method, the variables used in estimating the augmentation terms included treatment indicator and event indicator.
[2] PTE = -0.173 (95% CI: -1.026, 0.680).
[3] PCS = 0.021 (95% CI: 0.000, 0.586).
[4] pNIE = 0.079 (95% CI: -0.836, 0.999).

6.4 Analysis of dichotomized trajectory of $\log_{10}$ *CD4 cell counts*

We next evaluate the dichotomized $\log_{10}$ CD4 cell count as time-dependent CoP in Prentice's framework. We dichotomize the count level at $l = \log_{10} 200$, below which immediate

treatment has been recommended for HIV infected patients in treatment guidances for many years. Table 6.3 shows the fitted results. For the dichotomized $\log_{10}$ CD4 cell count, Model 3 and 4 suggest it significantly affects the clinical endpoint in both treatment and control treatment groups. Therefore the Prentice criterion (ii) holds. On the other hand, Model 1 suggests no significant effect modification of dichotomized $\log_{10}$ CD4 cell count by treatment group. Model 2 suggests that after controlling the dichotomized $\log_{10}$ CD4 cell count, the treatment group does not affect the clinical endpoint. It implies that Prentice criterion (iii) holds. The overall results show the evidence of the dichotomized $\log_{10}$ CD4 cell count being consistent with the Prentice's criteria as a surrogate for the treatment on the primary endpoint. This can also be seen from the three quantities (PTE, PCS, and pNIE), all of which demonstrate at least moderate or substantial level of surrogacy for the dichotomized $\log_{10}$ CD4 cell count. The CIs for all three quantities were determine from 500 bootstrap samples. We observe very wide CIs for all three quantities, especially for PCE. This is possibly because some bootstrap samples may have near zero treatment effect $\hat{\theta}_{trt,0}$ which leads to large variability in PTE. Compared the finding here to that for the continuous $\log_{10}$ CD4 cell count, we conjecture that there might exist a non-linear relationship between the $\log_{10}$ CD4 cell count and the hazard, especially in the treatment group. Chapter 7 discusses future research directions to examine the assumptions in the Cox proportional hazards model.

Table 6.3: Estimates and 95% CIs for the coefficients in Cox regression models, to assess dichotomized $\log_{10}$ CD4 cell count as time-dependent CoP in Prentice's framework.

		d.logCD4	trt	d.logCD4 × trt	bl.logCD4
Pooled	Model 1	-3.77 (-5.18, -2.35)	-0.14 (-0.59, 0.32)	0.51 (-0.96, 1.99)	-0.99 (-2.69, 0.72)
	Model 2	-3.35 (-3.93, -2.76)	-0.10 (-0.52, 0.33)		-1.01 (-2.69, 0.66)
Trt Group	Model 3	-3.11 (-3.70, -2.52)			-1.56 (-3.37, 0.24)
Control Trt Group	Model 4	-4.12 (-5.62, -2.62)			0.43 (-3.00, 3.85)

[1] The CIs were calculated based on 500 bootstrap samples.
[2] The nuisance parameters $\mu(u)$ and $\Sigma(u)$ for the distribution of random effects α were estimated by subgroups cross-classified by treatment group and dichotomized baseline $\log_{10}$ CD4 cell count ($\geq \log_{10} 300$, $< \log_{10} 300$).
[2] PTE = 0.739 (95% CI: -0.885, 3.721).
[3] PCS = 0.889 (95% CI: 0.002, 0.999).
[4] pNIE = 0.903 (95% CI: 0.042, 2.116).

Chapter 7

DISCUSSION AND FUTURE DIRECTION

7.1 Discussion

Assessing immune correlates of protection has always been an important objective in vaccine efficacy trials. This dissertation studies the quantitative and dichotomized inherent time-varying immune responses as immune correlates of risk and protection in two-phase sampling design cohort studies. The evaluation is based on Cox proportional hazards models with the continuous or dichotomized underlying immune response process, which is characterized by a random effects model. The model has the interpretation of association between the current value of the inherent immune biomarker and the instantaneous rate of the event. It provides straightforward assessment of the immune correlate of risk and also allows for the assessment of immune correlate of protection based on Prentice's framework. We also study the framework of causal effects by defining the natural direct and indirect effects of the vaccine on the probability of being free of event at some fixed time point. They are defined in terms of counterfactual outcomes that can only be partially observed under assigned treatment. We show that under certain sequential ignorability assumptions, the defined effects can be estimated from the Cox model mentioned above fitted on observed data. Since this dissertation works on the unobserved underlying immune response trajectory, it does not offer any guidance on the threshold of the protection level, even for the dichotomized model. The application is more to generate hypotheses about the biological mechanisms of protection.

The objective of this dissertation is to develop statistical methods to make inference on such joint models when the longitudinal immune biomarker data are only observed on study subjects selected in the second phase. In Chapter 2 and Chapter 3, we study the IPW and AIPW conditional score estimators for the continuous immune response trajectory via asymptotic theories and simulation studies. We explore the efficiency gain from

the AIPW method with various sets of predictor variables in estimating the augmentation terms compared to that from the IPW method. We find that when there exist auxiliary variables strongly correlated with immune biomarkers, including them to estimate the augmentation terms in AIPW method can provide significant efficiency gain for the parameter of the biomarker. Otherwise, our simulations studies show that the AIPW method does not improve the efficiency much, and can even reduce the efficiency if many weak auxiliary predictors are included. In addition, if the treatment effect is of interest in the Cox model, we recommend including the treatment indicator in the AIPW predictor in order to achieve greater precision close to that from the full cohort data analysis.

We also show that when the immune response data are only measured on a very limited number of time points, the resulting estimates on the biomarker effect can be very variable. Using the AIPW method in this situation can even lead to slightly rising bias and decreasing precision compared to the simple IPW method. This may be due to the fact that the conditional score method is constructed based on individual least squares estimates of the random effects. And such least squares estimates are also calculated at every observed event time so even fewer measurements are actually used (because we can only use the measurements taken up to and including that time point). If the number of measurements per subject is low, the fitting of individual trajectories can be very unreliable, especially when the measurement error is relatively large. The extra step in estimating the augmentation terms in AIPW method may lead to more variation and uncertainty in finite sample studies. Note that in our simulation studies, the inherent immune response level varies linearly over time. However, it is more often to observe curvilinear immune response trajectory over time in real trial studies, like for the CD4 cell counts in the ACTG 175 study and the antibody levels in the dengue vaccine trials. It implies more sophisticated models than linear should be used for a better fitting. However, that comes along with the need of more immune response measurements. So inadequacy in the number of measurements can also limit the choice of functional form of the immune response trajectory. Our simulation studies on misspecified measurement error distributions reveal that if the assumption of random and Normal measurement error is violated, we would expect slightly biased inferential results.

In Chapter 4, we propose the risk set recalibration method for the dichotomized immune

response model. Calibration is a commonly used approach in measurement error problems by deriving an induced hazard function of the observed but mismeasured immune response as a function of the parameters from the target Cox regression. Our induced hazard function is established based on a working assumption that the random effects are Normal given the status of being at risk. We reveal in Chapter 5 that the resulting estimator, even though theoretically inconsistent, has very small bias in studies with rare event rates even when the measurement error is relatively large. In contrast, the naive two-stage method is often very biased. Increasing the number of measurements per subject can majorly reduce the bias, especially for the two-stage method. Also the two-stage method is sensitive to the magnitude of measurement errors in that the bias increases with high measurement error. However, the proposed recalibration method is only slightly influenced by the measurement error in the rare event rate setting. We also encounter the common problem of the recalibration method that the variance of resulting estimates could be relatively high compared to the naive method. For random effects models, it is commonly assumed that the random effects follow the Normal distribution. We show that under Normal random effects, our working assumption is generally very close to the truth in the rare event setting as long as the number of at-risk subjects is moderate. When the number of at-risk subjects is very small, the inconsistency arises mostly because of poor estimation of the covariance matrix.

The joint modeling framework is intriguing in that it allows the modeling of subject-specific evolution of the immune response level over time and examining its relationship with the instantaneous rate of event simultaneously. However, the computation burden is always a concern for such a model. The semi-parametric methods considered in this dissertation are much less computationally intensive than approaches that require multidimensional integrals. But we still encounter an issue of convergence, especially when the measurement error is relatively large and the number of measurements per subject is small.

7.2 Future directions

Along with the research presented in this dissertation, below we describe several other interesting related questions.

7.2.1 Nonlinear models for immune response trajectory

In this dissertation, the immune response trajectory is modeled by a linear mixed effects model $X(u) = \alpha^T f(u)$. However, the trajectory over time for a single subject can be very variable and the linear models may not fully capture the entire evolution. Therefore, one interesting direction for future research is to study the nonlinear models. For the continuous trajectory, [Wu et al., 2008][Wu et al., 2010] proposed the nonlinear mixed effects model, like the exponential decay dynamic, to account for the biological understanding of the biomarker process in response to treatment. An interesting question is how to incorporate the nonlinear trajectory into the dichotomized model. Also a more complex pattern of the trajectory, the sawtooth pattern, due to booster doses has been observed in the VAX004 HIV-1 vaccine trial. One possible approach is to use the growing and decaying rates to characterize a single wave between two subsequent doses [Schlub et al., 2011]. To connect the waves over time, we may need assumptions on the period as well as on the time it takes after each dose for the immune response level to reach the peak.

7.2.2 The threshold for dichotomization

Our model on dichotomized immune response is aimed to assess if a "fixed" threshold of the level can predict the vaccine efficacy. It assumes that before we conduct the analysis there already exists some prior threshold of interest. However, a more natural question is what the protection level is if a specific immune response correlates with the protection. Even though this dissertation does not address the question of guiding the protection level, it still involves the issue of choosing a threshold. Actually, for our model, if no adequate evidence is found for a predetermined threshold to be an immune correlate of protection, it may just because we do not chose the right threshold. The threshold problem is of particular interest when the true relationship is nonlinear. When no particular preference exits for the threshold, people may tend to explore a range of candidate thresholds. In that case the type I error definitely needs to be controlled. Along this direction, one possible way out is to consider the "changepoint model" [Koziol and Wu, 1996][Vexler and Gurevich, 2009][Fong et al., 2014].

7.2.3 Model diagnosis on the proportional hazards assumption

The assessment of immune correlates of risk and protection in this dissertation relies on the assumption of proportional hazards (PH). If this assumption is strongly violated, applying the proposed methods can lead to misleading inferential results. In that case alternative models such as with time-varying coefficients or with interaction of time and the immune biomarker process may be considered [Song and Wang, 2008][Song et al., 2002]. [Fleming and Harrington, 1991](p173) proposed a list of methods to check the validity of PH assumption. The complication for the Cox model in the joint modeling framework is that it contains the unobserved random effects as the covariates. Thus any methods based on comparing the observed to the expected under the PH assumption do not apply directly here. However, we are able to estimate the random effects for each subject and perform the PH test by means of residuals obtained from the model with estimated random effects. The question is if by any way such test based on estimated random effects is linked to our target model on the unobserved true random effects. Another way possibly applies here is to fit Cox models for a series of time intervals separately and to see if the obtained coefficients are compatible across the intervals. It raises a question to develop some formal test statistic that can quantify the level of compatibility. Or we can add an additional term $X(u)g(u)$ to the Cox model and test if its coefficient is zero. The problem left is how to choose the functional form of $g(u)$. This method automatically provides a test p-value. However, even with an insignificant p-value, we cannot make a statement of inadequate evidence to reject the null of PH, because it only tests for the specific time-dependent pattern characterized by $g(u)$.

BIBLIOGRAPHY

P.K. Andersen and R.D. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

O. Borgan, B. Langholz, S.O. Samuelsen, L. Goldstein, and J. Pogoda. Exposure stratified case-cohort designs. *Lifetime data analysis*, 6:39–58, 2000.

N. Breslow and J. Wellner. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*, 34:86–102, 2007.

N. Breslow, T. Lumley, C.M. Ballantyne, L.E. Chambless, and M. Kulich. Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistics in Biosciences*, 1(1):32–49, 2009a.

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

M. Buyse and G. Molenberghs. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*, 54:1014–1029, 1998.

W. Cao, A.A. Tsiatis, and M. Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734, 2009.

M.R. Capeding, N.H. Tran, S.R. Hadinegoro, H.I. Ismail, T. Chotpitayasunondh, M.N. Chua, C.Q. Luong, K. Rusmil, D.N. Wirawan, R. Nallusamy, P. Pitisuttithum, U. Thisyakorn, I.K. Yoon, D. van der Vliet, E. Langevin, T. Laot, Y. Hutagalung, C. Frago, M. Boaz, T.A. Wartel, N.G. Tornieporth, M. Saville, A. Bouckenooghe, and CYD14 Study Group. Clinical efficacy and safety of a novel tetravalent dengue vaccine

in healthy children in Asia: a phase 3, randomised, observer-masked, placebo-controlled trial. *Lancet*, 10.1016/S0140-6736(14)61060-6, 2014.

N. Chatterjee and Y.H. Chen. A semiparametric pseudo-score method for analysis of two-phase studies with continuous phase-I covariates. *Lifetime data analysis*, 13:607–622, 2007.

N. Chatterjee, Y.H. Chen, and N. Breslow. A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98:158–168, 2003.

U.G. Dafni and A. Tsiatis. Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, 54(4):1445–1462, 1998.

V. DeGruttola and X.M. Tu. Modeling progression of CD4 lymphocyte count and its relationship to survival time. *Biometrics*, 50:1003–1014, 1994.

C.J. Faucett and D.C. Thomas. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, 15: 1663–1685, 1996.

C.L. Faucett, N. Schenker, and R.M. Elashoff. Analysis of censored survival data with intermittently observed time-dependent binary covariates. *Journal of American Statistical Association*, 93:427–437, 1998.

P. Flandre and Y. Saidi. Letters to the editor: Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine*, 18:107–109, 1999.

T.R. Fleming and D.P. Harrington. *Counting Processes and Survival Analyses*. New York: Wiley, second edition, 1991.

N.M. Flynn, D.N. Forthal, C.D. Harro, F.N. Judson, K.H. Mayer, and M.F. Para. Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *Journal of Infectious Disease*, 191:654–665, 2005.

Y. Fong, C. Di, and S. Permar. Change-point testing in logistic regression models with interaction term. *UW Biostatistics Working Paper Series*, page Working Paper 400, 2014.

D.N. Forthal, P. Gilbert, G. Landucci, and T. Phan. Recombinant gp120 vaccine-induced antibodies inhibit clinical strains of HIV-1 in the presence of Fc receptor-bearing effector cells and correlate inversely with HIV infection rate. *Journal of Immunology*, 178(10): 6596–6603, 2007.

C.E. Frangakis and D.B. Rubin. Principal stratification in causal inference. *Biometrics*, 58: 21–29, 2002.

L.S. Freedman, B.I. Graubard, and A. Schatzkin. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, 11:167–178, 1992.

P. Gilbert, M.L. Peterson, D. Follmann, M.G. Hudgens, D.P. Francis, M. Gurwith, W.L. Heyward, D.V. Jobes, V. Popovic, S. Self, F. Sinangil, D. Burke, and P.W. Berman. Correlation between immunologic responses to a recombinant glycoprotein 120 vaccine and incidence of HIV-1 infection in a phase 3 HIV-1 preventive vaccine trial. *The Journal of Infectious Disease*, 191:666–677, 2005.

P. Gilbert, L. Qin, and S. Self. Evaluating a surrogate endpoint at three levels, with application to vaccine development. *Statistics in Medicine*, 27:4758–4778, 2008.

X. Guo and B.P. Carlin. Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, 58(1):16–24, 2004.

P. Hall, J. Racine, and Q. Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004.

M.E. Halloran, I.M. Longini, and C.J. Struchiner. Estimability and interpretation of vaccine efficacy using frailty mixing models. *American Journal of Epidemiology*, 144(1):83–97, 1998.

S.M. Hammer, D.A. Katzenstein, M.D. Hughes, H. Gundacker, R.T. Schooley, R.H. Haubrich, W.K. Henry, M.M. Lederman, J.P. Phair, M. Niu, M.S. Hirsch, and T.C. Merigan. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *The New England Journal of Medicine*, 335(15):1081–1090, 1996.

P. Han. A note on improving the efficiency of inverse probability weighted estimator using the augmentation term. *Statistics and Probability Letters*, 82:2221–2228, 2012.

S. Haneuse, T. Saegusa, and T. Lumley. osdesign: An R package for the analysis, evaluation, and design of two-phase and case-control studies. *Journal of statistical software*, 43, 2011.

T.E. Hanson, A.J. Branscum, and W.O. Johnson. Predictive comparison of joint longitudinal-survival modeling: a case study illustrating competing approaches. *Lifetime Data Anal*, 17(1):3–28, 2011.

B.F. Haynes, G. Pantaleo, and A.S. Fauci. Toward an understanding of the correlates of protective immunity to hiv infection. *Science, New Series*, 271(5247):324–328, 1996.

B.F. Haynes, P. Gilbert, M.J. McElrath, S. Zolla-Pazner, G.D. Tomaras, S.M. Alam, D.T. Evans, D.C. Montefiori, C. Karnasuta, R. Sutthent, H.X. Liao, A.L. DeVico, G. Lewis, G.K. Williams, A. Pinter, Y. Fong, H. Janes, A. DeCamp, Y. Huang, M. Rao, E. Billings, and et al. Immune-correlates analysis of an HIV-1 vaccine efficacy trial. *The New England Journal of Medicine*, 366(14):1275–1286, 2012.

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

D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47:663–685, 1952.

M.M. Joffe and T. Greene. Related causal frameworks for surrgate outcomes. *Biometrics*, 65:530–538, 2009.

M.M. Joffe, D. Small, and C.Y. Hsu. Defining and estimating intervention effects for groups that will develop an ausiliary outcome. *Statistical Science*, 22(1):74–97, 2007.

J.D. Kalbfleisch and R.L. Prentice. *The statistical analysis of failure time data*. New York: Wiley-Interscience, 2002.

J.D.Y. Kang and J.L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22 (4):523–539, 2007.

F. Kobayashi and M. Kuroki. A new proportion measure of the treatment effect captured by candidate surrogate endpoints. *Statistics in Medicine*, 33:3338–3353, 2014.

M.R. Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer, 2008.

J.A. Koziol and S.H. Wu. Changepoint statistics for assessing a treatment-covariate interaction. *Biometrics*, pages 1147–1152, 1996.

M. Kulich and D.Y. Lin. Improving the efficiency of relative-risk estimation in case-cohort studies. *The Journal of the Acoustical Society of America*, 99(467):832–844, 2004.

T. Lange and J.V. Hansen. Direct and indirect effects in a survival context. *Epidemiology Cambridge Mass*, 22(4):575–581, 2011.

S.D. Lendle, M.S. Subbaraman, and M.J. van der Vaart. Identification and efficient estimation of the natural direct effect among the untreated. *Biostatistics*, 69(2):310–317, 2013.

N.L. Letvin. Progress toward an HIV vaccine. *Annual Review of Medicine*, 56:213–223, 2005.

Z. Li, P. Gilbert, and B. Nan. Weighted likelihood method for grouped survival data in case-cohort studies with application to HIV vaccine trials. *Biometrics*, 64(4):1247–1255, 2008.

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

D.Y. Lin, T.R. Fleming, and V. De Gruttola. Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine*, 16(13):1515–1527, 1997.

X. Luo, W.Y. Tsai, and Q. Xu. Pseudo-partial likelihood estimators for the Cox regression model with missing covariates. *Biometrika*, 96(3):617–633, 2009.

T. Martinussen, S. Vansteelandt, M. Gerster, and J. Hjelmborg. Estimation of direct effects for survival data by using the Aalen additive hazards model. *Journal of the Royal Statistical Society Series B*, 73:773–788, 2011.

L.M. McCrink, A.H. Marshall, and K.J. Cairns. Advances in joint modelling: a review of recent developments with application to the survival of end stage renal disease patients. *International Statistical Review*, 81(2):249–269, 2013.

B. Nan. Efficient estimation for case-cohort studies. *The Canadian Journal of Statistics*, 32:403–419, 2004.

B. Nan, M. Emond, and J. Wellner. Information bounds for Cox regression models with missing data. *The Annals of Statistics*, 32:723–753, 2004.

J. Neyman. Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):101–116, 1938.

P. Odile. Bootstrap of means under stratified sampling. *Electronic Journal of Statistics*, 1: 381–391, 2007.

G. Pantaleo and R.A. Koup. Correlates of immune protection in HIV-1 infection: what we know, what we don't know, what we should know. *Nature Medicine*, 10(8):806–810, 2004.

Y. Pawitan and S. Self. Modeling disease marker processes in AIDS. *Journal of the American Statistical Association*, 88(423):719–726, 1993.

J. Pearl. Direct and indirect effects. *In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, Francisco, CA: Morgan Kaufmann*, pages 411–420, 2001.

P. Philipson, I. Sousa, P. Diggle, P. Williamson, R. Kolamunnage-Dona, and R. Henderson. joineR: Joint modelling of repeated measurements and time-to-event data. *http://cran.r-project.org/web/packages/joineR/index.html*, 2012.

S.A. Plotkin and P. Gilbert. Nomenclature for immune correlates of protection after vaccination. *Clinical Infectious Disease*, 54:1615–1617, 2012.

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

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

R.L. Prentice. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8:431–440, 1989.

L. Qi, C.Y. Wang, and R.L. Prentice. Weighted estimators for proportional hazards regression with missing covariates. *Journal of American Statistical Association*, 100(472): 1250–1263, 2005.

L. Qin, P. Gilbert, L. Corey, M.J. McElrath, and S. Self. A framework for assessing immunological correlates of protection of vaccine trials. *The Journal of Infectious Disease*, 196:1304–1312, 2007.

S. Rerks-Ngarm, P. Pitisuttithum, S. Nitayaphan, J. Kaewkungwal, J. Chiu, R. Paris, N. Premsri, C. Namwat, M. de Souza, E. Adams, M. Benenson, S. Gurunathan, J. Tartaglia, J.G. McNeil, D.P. Francis, D. Stablein, D.L. Birx, S. Chunsuttiwat, C. Khamboonruang, P. Thongcharoen, M.L. Robb, N.L. Michael, P. Kunasol, J.H. Kim, and MOPH-TAVEG Investigators. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *The New England Journal of Medicine*, 361:2209–2220, 2009.

P.H. Rhodes, M.E. Halloran, and I.M. Longini. Counting process models for infectious disease data: distinguishing exposure to infection from susceptibility. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(4):751–762, 1996.

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

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

J.M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3:143–155, 1992.

J.M. Robins, A. Rotnitzky, and L. Zhao. Estimation of regression coefficients when some some regressors are not always observed. *Journal of American Statistical Association*, 89 (427):846–865, 1994.

T. Saegusa and J.A. Wellner. Weighted likelihood estimation under two-phase sampling. *The Annals of Statistics*, 41(1):269–295, 2013.

T.H. Scheike and T. Martinussen. Likelihood estimation for Cox's regression model under case-cohort sampling. *Scandinavian Journal of Statistics*, 31(2):283–293, 2004.

T.E. Schlub, J.C. Sun, S.M. Walton, S.H. Robbins, A.K. Pinto, M.W. Munks, A.B. Hill, L. Brossay, A. Oxenius, and M.P. Davenport. Comparing the kinetics of NK cells, CD4, and CD8 T cells in murine cytomegalovirus infection. *The journal of Immunology*, 187: 1385–1392, 2011.

S. Seaman and I.R. White. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295, 2011.

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

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

X. Song, M. Davidian, and A. Tsiatis. An estimator for the proprotional hazards model with multiple logntitudinal covariates measured with error. *Biostatistics*, 3(4):511–528, 2002.

E.J. Tchetgen Tchetgen. On causal mediation analysis with a survival outcome. *The International Journal of Biostatistics*, 7:Artical 33, 2011.

E.J. Tchetgen Tchetgen. Identification and estimation of survivor average causal effects. *Statistics in Medicine*, 33(21):3601–3628, 2014.

Y.K. Tseng, F. Hsieh, and J.L. Wnag. Joint modeling of accelerated failure time and longitudinal data. *Biometrika*, 92(3):587–603, 2005.

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

A. Tsiatis and M. Davidian. Joint modelin of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14:809–834, 2004.

A. Tsiatis, V. Degruttola, and M.S. Wulfsohn. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of American Statistical Association*, 90(429):27–37, 1995.

A.W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge Univ. Press, 1998.

T.J. VanderWeele. Causal mediation analysis with survival data. *Epidemiology Cambridge Mass*, 22(4):582–585, 2011.

A. Vexler and G. Gurevich. Average most powerful tests for a segmented regression. *Communications in Statistics - Theory and Methods*, 38:2214–2231, 2009.

L.V. Villar, G.H. Dayan, Arredondo-García J.L., D.M. Rivera, R. Cunha, C. Deseda, H. Reynales, M.S. Costa, J.O. Morales-Ramírez, G. Carrasquilla, L.C. Rey, R. Dietze, K. Luz, E. Rivas, M.C. Montoya, M.C. Supelano, B. Zambrano, and et al. Efficacy of a tetravalent dengue vaccine in children in Latin America. *The New England Journal of Medicine*, DOI: 10.1056/NEJMoa1411037, 2014.

C.Y. Wang. Corrected score estimator for joint modeling of longitudinal and failure time data. *Statistica Sinica*, 16(235-253), 2006.

C.Y. Wang and H. Chen. Augmented inverse probability weighted estimator for Cox missing covariate regression. *Biometrics*, 57:414–419, 2001.

C.Y. Wang, L. Hsu, Z. Feng, and R.L. Prentice. Regression calibration in failure time regression. *Biometrics*, 53:131–145, 1997.

C.Y. Wang, N. Wang, and S. Wang. Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. *Biometrics*, 56, 2000.

C.Y. Wang, S.X. Xie, and R.L. Prentice. Recalibration based on an approximate relative risk estimator in Cox regression with missing covariates. *Statistica Sinica*, 11:1081–1104, 2001.

S. Wang and C.Y. Wang. A note on kernel assisted estimators in missing covariate regression. *Statistics and Probability Letters*, 55(4):439–449, 2001.

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

Y. Wang and J.M.G. Taylor. Joint modelin of longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of American Statistical Association*, 96(895-905), 2001b.

C.J. Weir and R.J. Walley. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Statistics in Medicine*, 25:183–203, 2006.

L. Wu, X. J. Hu, and H. Wu. Joint inference for nonlinear mixed-effects models and time-to-event at the presence of missing data. *Biostatistics*, 9:308–320, 2008.

L. Wu, W. Liu, and X.J. Hu. Joint inference on hiv viral dynamics and immune suppression in presence of measurement errors. *Biometrics*, 66(2):327–335, 2010.

L. Wu, W. Liu, G.Y. Yi, and Y. Huang. Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, page dio:10.1155/2012/640153, 2012.

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

S.X. Xie, C.Y. Wang, and R.L. Prentice. A risk set calibration method for failure time regression by using a covariate eliability sample. *J. R. Statist. Soc. B*, 63:855–870, 2001.

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

W. Zheng and M.J. van der Laan. Causal mediation in a survival setting with time-dependent mediators. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2012.

D.M. Zucker. A pseudo-partial likelihood method for semiparametric survival regression with covariate errors. *Journal of American Statistical Association*, 100(472):1264–1277, 2005.

D.M. Zucker and D. Spiegelman. Corrected score estimation in the proportional hazards model with misclassified discrete covariates. *Statistics in Medicine*, 27:1911–1933, 2008.

VITA

Rong Fu was born in 1987 in Xinjiang, China. She received a Bachelor of Science in Statistics in 2009 at Peking University. In 2009 she entered the Doctoral program in Biostatistics at University of Washington, and earned a Master of Science in June 2014.