1-1-2014

Regression Modeling of Longitudinal Outcomes With Outcome-Dependent Observation Times

Kay See Tan
University of Pennsylvania, kayseetan@gmail.com

Regression Modeling of Longitudinal Outcomes With Outcome-Dependent Observation Times

Abstract
Conventional longitudinal data analysis methods typically assume that outcomes are independent of the data-collection schedule. However, the independence assumption may be violated when an event triggers outcome assessment in between prescheduled follow-up visits. For example, patients initiating warfarin therapy who experience poor anticoagulation control may have extra physician visits to monitor the impact of necessary dose changes. Observation times may therefore be associated with outcome values, which may introduce bias when estimating the effect of covariates on outcomes using standard longitudinal regression methods. We consider a joint model approach with two components: a semi-parametric regression model for longitudinal outcomes and a recurrent event model for observation times. The semi-parametric model includes a parametric specification for covariate effects, but allows the effect of time to be unspecified. We formulate a framework of outcome-observation dependence mechanisms to describe conditional independence between the outcome and observation-time processes given observed covariates or shared latent variables.

We generalize existing methods for continuous outcomes by accommodating any combination of mechanisms through the use of observation-level weights and/or patient-level latent variables. We develop new methods for binary outcomes, while retaining the flexibility of a semi-parametric approach. We extend these methods to account for discontinuous risk intervals in which patients enter and leave the at-risk set multiple times during the study. Our methods are based on counting process approaches, rather than relying on possibly intractable likelihood-based or pseudo-likelihood-based approaches, and provide marginal, population-level inference. In simulations, we evaluate the statistical properties of our proposed methods. Comparisons are made to `naive' approaches that do not account for outcome-dependent observation times. We illustrate the utility of our proposed methods using data from a randomized trial of interventions designed to improve adherence to warfarin therapy and a randomized trial of malaria vaccines among children in Mali.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Epidemiology & Biostatistics

First Advisor
Andrea B. Troxel

Second Advisor
Benjamin French

Keywords
Informative observation times, Joint models, Observation-time process, Outcome process, Recurrent events, Semi-parametric regression

Subject Categories
Biostatistics

REGRESSION MODELING OF LONGITUDINAL OUTCOMES WITH OUTCOME-DEPENDENT OBSERVATION TIMES

Kay See Tan

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

Andrea B. Troxel

Professor of Biostatistics

Co-Supervisor of Dissertation

Benjamin French

Assistant Professor of Biostatistics

Graduate Group Chairperson

John H. Holmes, Associate Professor of Medical Informatics in Epidemiology

Dissertation Committee

Sarah J. Ratcliffe, Associate Professor of Biostatistics

Dylan Small, Associate Professor of Statistics

Kevin Volpp, Professor of Medicine and Health Care Management

REGRESSION MODELING OF LONGITUDINAL OUTCOMES WITH OUTCOME-DEPENDENT OBSERVATION TIMES

ACKNOWLEDGEMENT

I wish to express my heartfelt appreciation to my dissertation advisors: Dr. Benjamin French for his devotion and mentorship in developing me into an effective researcher and writer, and Dr. Andrea Troxel for providing tremendous support throughout the process of completing this thesis. I would also like to thank my committee members, Dr. Sarah Ratcliffe, Dr. Dylan Small and Dr. Kevin Volpp, for their constructive feedback that greatly improved my dissertation. I would like to acknowledge Dr. Stephen Kimmel for the use of the warfarin data, and Dr. Michael Fay for the use of the malaria vaccine data.

My sincere thanks go to the faculty in the Division of Biostatistics, including Dr. Wei-Ting Hwang for the opportunity to collaborate and publish, Dr. Sharon Xie for the invaluable research experience with the Department of Neurology, and Dr. Daniel Heitjan for his insightful advice and towering support throughout my research career. My journey would not be complete without the friendship and support of my fellow graduate students at the University of Pennsylvania.

I would like to acknowledge those who have provided personal support. I am grateful to Michael and Gail True for their kindness and opening their home to me for holiday dinners. I owe a special thanks to Dr. Marlin Eby who encouraged me to pursue graduate school, and Dennis Abline who convinced me to give college a try.

I would like to thank my sister and brother for their positive attitudes and words of inspiration. Lastly, my most appreciative and loving gratitude go to my parents. Without their unfailing support this dissertation would have been impossible. This dissertation is dedicated to them.

ABSTRACT

REGRESSION MODELING OF LONGITUDINAL OUTCOMES WITH OUTCOME-DEPENDENT OBSERVATION TIMES

Kay See Tan

Andrea B. Troxel

Benjamin French

Conventional longitudinal data analysis methods typically assume that outcomes are independent of the data-collection schedule. However, the independence assumption may be violated when an event triggers outcome assessment in between prescheduled follow-up visits. For example, patients initiating warfarin therapy who experience poor anticoagulation control may have extra physician visits to monitor the impact of necessary dose changes. Observation times may therefore be associated with outcome values, which may introduce bias when estimating the effect of covariates on outcomes using standard longitudinal regression methods. We consider a joint model approach with two components: a semi-parametric regression model for longitudinal outcomes and a recurrent event model for observation times. The semi-parametric model includes a parametric specification for covariate effects, but allows the effect of time to be unspecified. We formulate a framework of outcome-observation dependence mechanisms to describe conditional independence between the outcome and observation-time processes given observed covariates or shared latent variables.

We generalize existing methods for continuous outcomes by accommodating any combination of mechanisms through the use of observation-level weights and/or patient-level latent variables. We develop new methods for binary outcomes, while retaining the flexibility of a semi-parametric approach. We extend these methods to account for discontinuous risk intervals in which patients enter and leave the at-risk set multiple times during the study. Our methods are based on counting process approaches, rather than relying on possibly intractable likelihood-based or pseudo-likelihood-based approaches, and provide marginal, population-level inference. In simulations, we evaluate the statistical properties of our proposed methods. Comparisons are made to 'naïve' approaches that do not account for outcome-dependent observation times. We illustrate the utility of our pro-

posed methods using data from a randomized trial of interventions designed to improve adherence to warfarin therapy and a randomized trial of malaria vaccines among children in Mali.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

1.1. Background

1.1.1. Introduction

Conventional longitudinal data analysis methods assume that outcomes are independent of the data-collection schedule. However, the independence assumption may be violated, for example, when a specific treatment necessitates a different follow-up schedule than the control arm, or when adverse events trigger additional physician visits in between prescheduled follow-ups. If the probability of having a follow-up visit depends upon previous outcomes and measured or unmeasured covariates, then the outcome and observation-time processes are dependent and conventional longitudinal data analysis methods such as generalized estimating equations (GEE, Liang and Zeger, 1986) that ignore the observation-time process may provide biased estimates of covariate-outcome associations (French and Heagerty, 2009; Sun et al., 2005). We designate the longitudinal outcomes as the outcome process and the occurrence of visits over time as the observation-time process.

There has been considerable interest in the topic of addressing potential dependence between the outcome and observation-time processes. Lipsitz, Fitzmaurice, and Ibrahim (2002) developed a likelihood-based procedure for continuous outcomes, and Fitzmaurice et al. (2006) proposed a pseudo-likelihood estimation procedure for binary outcomes. Several other authors have adopted an estimating equations approach with observation-level inverse weights derived from explicit specification of the observation-time models (Bůžková and Lumley, 2007; Lin, Scharfstein, and Rosenheck, 2004). Others focused on estimation procedures based on joint likelihood approaches: Ryu et al. (2007) developed a Bayesian fully parametric regression model; Liu, Huang, and O'Quigley (2008) considered a joint mixed-effects model in which the outcomes, observation times, and censoring times were correlated through latent variables.

We consider a joint modeling approach with two components: (i) a semi-parametric regression model for longitudinal outcomes and (ii) a recurrent events model for observation times. The semi-

parametric approach is flexible as it assumes a non-parametric structure for the mean trajectory of the longitudinal outcomes (Bůžková and Lumley, 2009; Liang, Lu, and Ying, 2009; Lin and Ying, 2001; Sun, Song, and Zhou, 2011). In addition, our methods are based on counting process approaches, rather than relying on possibly intractable likelihood-based or pseudo-likelihood-based approaches, and provide marginal, population-level inference. We describe the two main components of the joint model below.

1.1.2. Semi-parametric regression model for outcome process

Our primary scientific interest lies in a semi-parametric regression model for the longitudinal outcomes $Y_i(t)$:

$$\mathrm{E}[Y_i(t) \mid X_i(t)] = g\{\mu(t) + \beta' X_i(t)\}, \tag{1.1}$$

in which the function $g(\cdot)$ links the expected outcome to the linear predictors, $\mu(t)$ is an arbitrary function of time, β is a $p \times 1$ vector of regression parameters of interest, and $X_i(t)$ is a p-dimensional covariate process for subject i. The semi-parametric outcome model assumes a parametric structure for the effect of $X_i(t)$ and a non-parametric structure for $\mu(t)$ (Lin and Carroll, 2001; Sun et al., 2005). Model (1.1) describes the marginal mean of $Y_i(\cdot)$ without specifying its correlation structure and distributional form; hence, it is appealing if the effects of $X_i(t)$ are of interest, but the effect of time is considered a nuisance. The primary target of inference is the marginal association between a set of covariates at t and an outcome of interest among a population of individuals, represented by the parameter β in Model (1.1). Model (1.1) does not condition on the entire covariate process or on past outcomes. Instead, it includes covariate information available at t, such as baseline covariates, covariates measured at or before t, and summaries of the covariate history; that is, Model (1.1) is a partly conditional mean regression model (Pepe and Couper, 1997).

1.1.3. Recurrent events model for observation-time process

The observation-time process describes the timing and intensity of follow-up visits and is characterized by a standard recurrent events model. We introduce a non-negative latent variable η_i with mean 1 and unknown variance σ^2. Given observation-time model covariates $Z_i(t)$ and η_i, the recurrent event process $N_i(\cdot)$ is a non-homogeneous Poisson process with intensity function (Lin et al.,

2000; Pepe and Cai, 1993):

$$\lambda_i(t) = \eta_i \lambda_0(t) \exp\{\gamma' Z_i(t)\}, \quad t \in [0, \tau] \tag{1.2}$$

in which γ is a vector of unknown parameters and $\lambda_0(t)$ is an arbitrary baseline intensity function. Model (1.2) implies that the occurrence of observations follows a proportional intensity model, in which η_i inflates or deflates the visit intensity. We assume the censoring time, C_i is independent of the observation-time process. The indicator $dN_i(t)$ equals 1 if a follow-up visit occurs at t and equals 0 otherwise. The parameter γ provides information regarding the observation-time model, but is considered a nuisance because our interest is in estimating the association parameter β from the outcome model. However, incorporating the observation-time process into estimation of β in a joint model facilitates reliable estimation under outcome-observation dependence, which we detail below.

1.1.4. Framework of outcome-observation mechanisms

Even though there has been great interest in the topic of longitudinal data with dependent observation times, there does not appear to be a unified framework to describe the various relationships that exist between the outcome and observation-time processes. We propose the following framework of outcome-observation dependence mechanism that we will refer to in the following chapters. Based on the two main components described in the previous section, we describe three sources of dependence between the outcome and observation-time processes, either through observed covariates or shared latent variables:

(M1) Conditional independence given past outcome-model covariates;

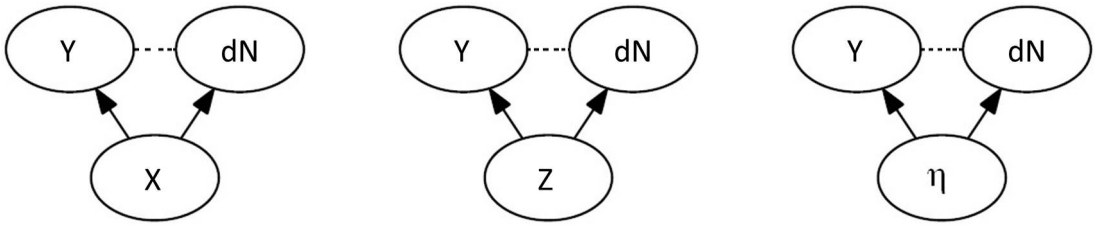
(M2) Conditional independence given past observation-time model covariates;

(M3) Conditional independence given shared latent variables.

In the remaining chapters, conditional independence given covariates implies conditional independence given past observed covariates.

(M1) Conditional independence given outcome-model covariates

The first mechanism assumes that the outcome process is conditionally independent of the obser-



(a) M1 (b) M2 (c) M3

Figure 1.1: Framework of outcome-observation mechanisms: (M1), (M2) and (M3)

vation-time process given outcome-model covariates $X_i(t)$, or a subset of $X_i(t)$:

$$\mathrm{E}[dN_i(t) \mid X_i(t), Y_i(t), C_i \geq t] = \mathrm{E}[dN_i(t) \mid X_i(t)].$$

The probability of observation at time t depends on $X_i(t)$, $Y_i(t)$, and C_i only through outcome-model covariates $X_i(t)$, hence is plausible if the occurrence of a follow-up visit is due to the features of the study design, rather than subject-specific behaviors (Figure 1.1a).

(M2) Conditional independence given observation-time model covariates

The second mechanism assumes that the occurrence of a follow-up visit depends on observation-time model covariates $Z_i(t)$:

$$\mathrm{E}[dN_i(t) \mid X_i(t), Z_i(t), Y_i(t), C_i \geq t] = \mathrm{E}[dN_i(t) \mid Z_i(t)].$$

The probability of observation at time t depends on $X_i(t), Z_i(t), Y_i(t)$, and C_i only through observation-time model covariates $Z_i(t)$ (Figure 1.1b). The set of covariates $Z_i(t)$ includes the full or partial subset of the outcome-model covariates and any additional measured covariates at or before time t, as well as previous outcomes. Note that (M1) $\subset$ (M2) because $X_i(t) \subset Z_i(t)$.

(M3) Conditional independence given shared latent variables

The third mechanism assumes that the outcome process is conditionally independent of the observation-time process given outcome-model covariates $X_i(t)$, and an unmeasured mean-one subject-specific latent variable η_i:

$$\mathrm{E}[dN_i(t) \mid X_i(t), Y_i(t), \eta_i, C_i \geq t] = \mathrm{E}[dN_i(t) \mid X_i(t), \eta_i].$$

The parameter η_i conveys information regarding subject-specific unmeasured confounders and propensity for physician visits (Figure 1.1c).

Our typology of outcome-observation dependence mechanisms provides a framework for reliable estimation of covariate-outcome associations. Mechanisms (M2) and (M3) allow the probability of an observation to depend on unmeasured patient characteristics in addition to measured observation-time model covariates, and hence place fewer restrictions than (M1) on the probability of having a visit; these are reasonable assumptions in most observational studies. However, (M2) and (M3) may require more advanced analytic methods to provide valid inference; we introduce these in subsequent chapters.

1.2. Motivating examples

We present several examples of longitudinal studies with outcome-dependent observation times and provide graphical and numerical descriptions to illustrate unique features of the data. The examples presented here will be investigated further in subsequent chapters.

1.2.1. Bladder tumor recurrence study

Andrews and Herzberg (1985) presented data from a bladder cancer study conducted by the Veterans Administration Cooperative Urological Research Group. Eighty-five patients with superficial bladder tumors were randomly assigned to control $(n = 47)$ or thiotepa treatment $(n = 38)$ and followed up to 53 months. Patients in the control group had a physician visit once every three months, while patients in the treatment group had a visit almost once a month due to the invasive nature of the treatment, which had to be directly distilled into the bladder.

At each follow-up visit, new tumors were counted before being removed transurethrally. There was notable heterogeneity in visit patterns across patients. The median (25^{th}, 75^{th} percentile) number of visits in the placebo group and treatment group was 9 (5, 12) and 9 (4, 23), respectively. The average time between visits for the placebo group was 3.7 months, compared to 2.3 months for the treatment group. These differences suggested that the patients in the treatment group visited the clinic more often (Figure 1.2). Hence the observation-time process must be considered in order to properly estimate the effect of treatment on tumor recurrence. We revisit this example in Section 2, where we apply various methods developed for continuous outcomes with outcome-dependent

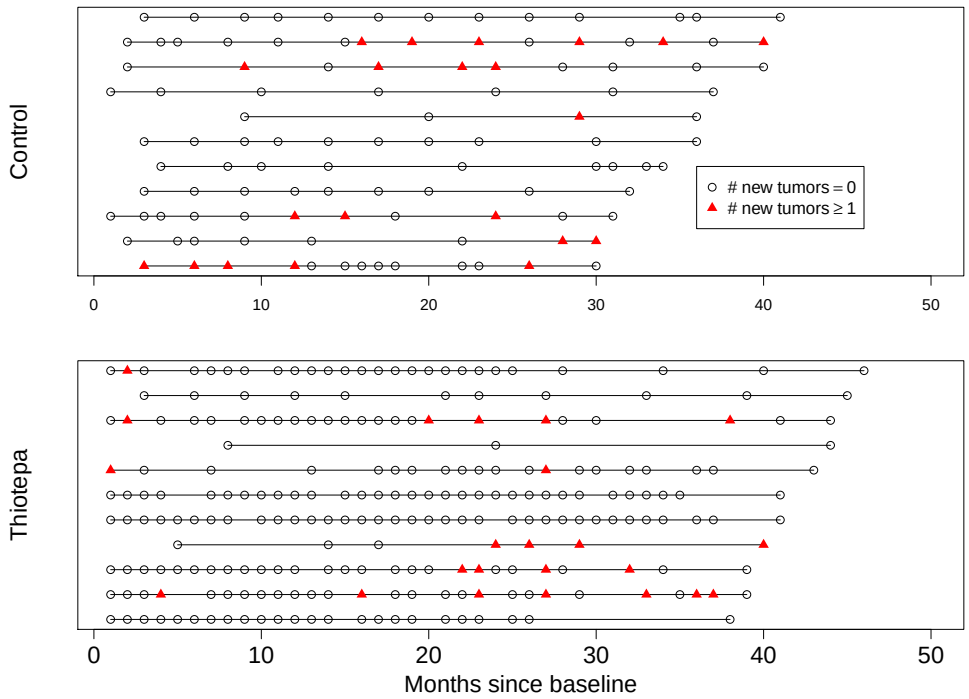


Figure 1.2: Observation times for four selected patients in the bladder tumor recurrence study and the corresponding observed outcomes: no new tumors or one or more new tumors found.

observation-times.

1.2.2. Warfarin adherence study

Thromboembolism is the formation of a clot that obstructs blood flow in a vein or artery and can lead to deep vein thrombosis and even stroke if untreated. Patients at risk of thromboembolism are commonly administered warfarin, an oral anticoagulant. Though highly efficacious, warfarin has a narrow therapeutic range: over-anticoagulation can lead to increased risk of bleeding complications, while under-anticoagulation can lead to thromboembolic events. A patient's international normalized ratio (INR) gives a measure of blood coagulation. The INR of a healthy individual is usually between 0.8 and 1.2, and the target range for patients on warfarin is between 2 and 3. Due to the drug's narrow therapeutic range, a patient on warfarin requires frequent monitoring. A clinician usually focuses on whether the patient is in or out of therapeutic range; an out-of-range INR typically triggers a dose change (Brigden et al., 1998).

We consider data from a randomized controlled trial among patients on warfarin therapy. The

main objective of the trial was to determine the effectiveness of interventions designed to increase adherence to therapy, and thus improve anticoagulation control (Kimmel et al., 2007). The study randomized 362 subjects into four treatment arms. The study protocol specified monthly follow-up visits, during which the INRs were measured. Physicians also scheduled as-needed visits in between protocol-required visits based on the patient's INR response (Figure 1.3).

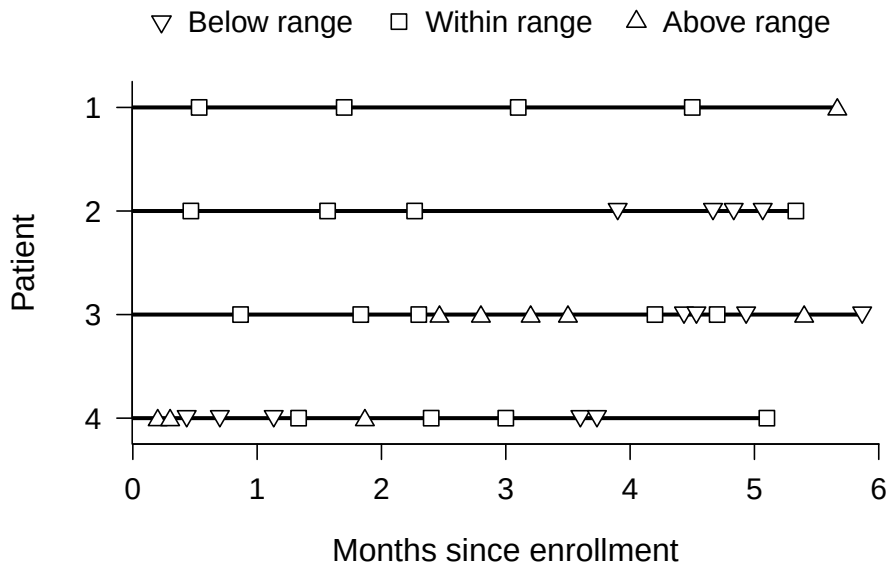


Figure 1.3: Observation times for four selected patients on warfarin and the corresponding observed outcomes: INR below, within, or above the therapeutic range.

The primary interest lies in estimating the effect of observed covariates on the probability of being out of therapeutic range, based on the binary outcome defined as either "in-range" or "out-of-range". At the analysis stage, a naïve approach considers only the outcomes from monthly protocol-scheduled visits at the expense of discarding data that can be potentially informative. Conversely, the analyst can also apply conventional longitudinal methods to the full data. However, doing so can over-represent certain types of patients and hence bias the estimates. One can also apply a 'cluster-weighted GEE' model to the full data, in which participant-level weights are equal to the inverse of the number of INRs (Williamson, Datta, and Satten, 2003). However, the cluster-weighted GEE approach does not account for the timing of the physician visits. We revisit this example in Chapter 3, in which we extend methods for continuous outcomes to binary outcomes.

1.2.3. Malaria vaccine study

We consider data from a randomized malaria vaccine trial involving 289 children between two and five years old living in rural regions of Mali. The main aim of this double blind phase II clinical trial was to compare the efficacy and safety of a candidate vaccine and an active control, namely AMA1-C1 ($n = 139$) and Hiberix ($n = 140$), in the odds of developing malaria symptoms. The primary outcome was whether the parasite that leads to clinical malaria (*P. falciparum* parasitemia) exceeds a given threshold ($3000/\mu$L). Other outcomes were clinical malaria (parasite $> 3000/\mu$L with fever) and the continuous measures of hemoglobin and anemia.

Monthly visits were scheduled according to the protocol. However, there were many visits that occurred between two prescheduled visits. Approximately 65% of these as-needed visits were due to symptoms of clinical malaria or side-effects from the vaccine. There may be a relationship between the outcome and observation-time processes because physicians tend to schedule closely-spaced visits for patients who were more susceptible to a malaria episode.

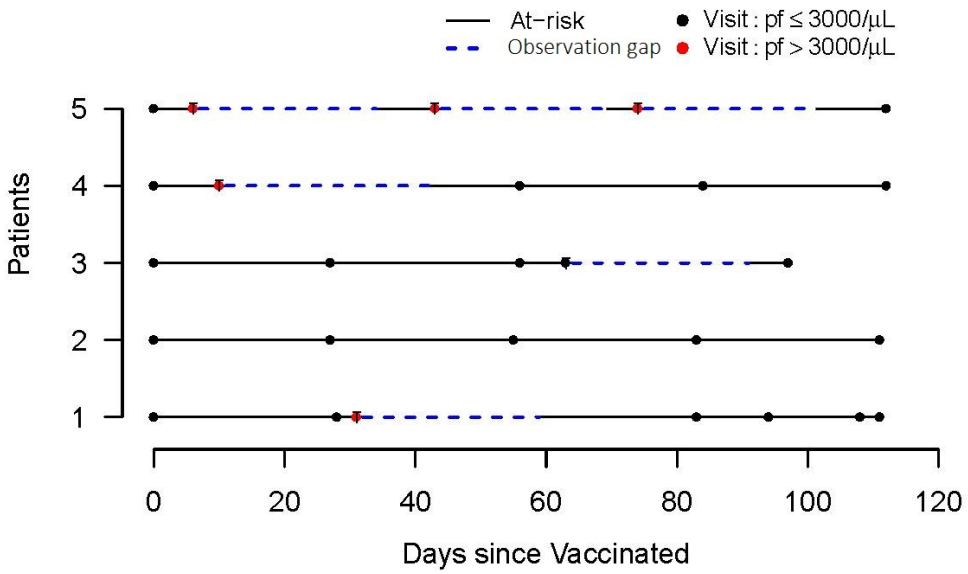


Figure 1.4: Observation times for four selected patients in the malaria vaccine study and the corresponding observed outcomes. A red dot indicates that parasite level $> 3000/\mu$L recorded at that visit, and the 28-day observation gaps are indicated by dashed blue lines.

A noteworthy point is that when a child was diagnosed with clinical malaria, the child received

malaria treatment and was not considered at risk for a new malaria episode (and another physician visit) until 28-days after the first day of treatment. Figure 1.4 shows the time in study for each subject, in which the at-risk process is interrupted by the observation gaps, leading to discontinuous risk intervals. These observation gaps started on the day the malaria treatment was administered and lasted for 28 days. The issue of discontinuous risk intervals is easily addressed by careful consideration of when individuals are at risk of a physician visit. Methods accommodating discontinuous risk intervals are discussed in Chapter 4.

1.3. Outline of the dissertation

Chapter 2 compares and contrasts four recently developed semi-parametric methods for continuous outcomes that accommodate one of three outcome-observation dependence mechanisms. To allow greater flexibility, we propose a comprehensive method to accommodate any combination of mechanisms through the use of observation-level weights and patient-level latent variables. In simulation studies, we show how incorrectly specifying the outcome-observation dependence may yield biased estimates of covariate-outcome associations and how our proposed extensions can accommodate a greater number of dependence mechanisms. We illustrate the implications of different modeling strategies in an application to bladder cancer data.

In Chapter 3, we develop a new method for the analysis of binary outcomes with outcome-dependent observation times, while retaining the flexibility of a semi-parametric approach. Our methods are based on counting process approaches, rather than relying on possibly intractable likelihood-based or pseudo-likelihood-based approaches, and provide marginal, population-level inference. In simulations, we evaluate the statistical properties of our proposed methods. Comparisons are made to 'naïve' GEE approaches that either do not account for outcome-dependent observation times or incorporate weights based on the observation-time process. We illustrate the utility of our proposed methods using data from a randomized controlled trial of interventions designed to improve adherence to warfarin therapy.

Chapter 4 extends the comprehensive method from Chapter 3 to data with discontinuous risk intervals in which patients may enter and leave the at-risk set multiple times during the study. We show that discontinuous risk intervals can be addressed by careful specification of the at-risk indicator. The methodology is applied to a randomized trial of malaria vaccines among children in Mali.

We conclude with Chapter 5, which provides a summary of the work and presents several directions for future research.

CHAPTER 2

SEMI-PARAMETRIC METHODS FOR CONTINUOUS OUTCOMES

2.1. Introduction

Longitudinal studies commonly assume that the data-collection schedule is independent of a subject's outcomes and measured or unmeasured characteristics. However, this independence assumption may be violated if observed covariate or outcome values influence the occurrence or timing of subsequent visits. For example, in a study following patients with diabetes, routine visits are scheduled every six months. However, spikes in blood sugar levels, exacerbation of other symptoms, or underlying patient characteristics may trigger additional closely-spaced physician visits until the blood sugar level has stabilized. The intensity of events such as physician visits is dependent on previous outcomes and measured or unmeasured covariates. Less healthy patients may be over-represented in the analysis due to more frequent data collection. In the presence of the resultant selection bias, conventional methods such as generalized estimating equations (GEE, Liang and Zeger, 1986) may yield biased estimates of covariate-outcome associations (Huang, Wang, and Zhang, 2006; Sun et al., 2005). Proper estimation must account for such selection bias. We focus on a marginal mean regression model to evaluate the association between observed covariates and a continuous outcome of interest. We denote the longitudinal outcomes as the outcome process and the occurrence of visits over time as the observation-time process.

Several authors have proposed parametric models to account for the potential dependence between the outcome and observation-time processes. Lipsitz, Fitzmaurice, and Ibrahim (2002) developed a likelihood-based procedure for continuous outcomes, Fitzmaurice et al. (2006) proposed a pseudo-likelihood estimation procedure for binary outcomes, and Lin, Scharfstein, and Rosenheck (2004) and Bůžková and Lumley (2007) utilized inverse intensity weighted estimators with observation-level inverse weights. Others focused on estimation procedures based on joint likelihood approaches: Ryu et al. (2007) developed a Bayesian fully parametric regression model; Liu, Huang, and O'Quigley (2008) considered a joint mixed-effects model in which the outcomes, observation times, and censoring times were correlated through latent variables. The study of outcome-dependent observation times shares features of research regarding incomplete (Albert,

2000; Troxel et al., 2010) and recurrent marked point process data (French and Heagerty, 2009), but differs in that subjects do not share a common set of visit times, and outcomes (e.g., blood sugar level) exist even if an event (e.g., a physician visit) does not occur.

We introduce a framework of three outcome-observation dependence mechanisms. The first mechanism applies when the outcome and the observation-time processes are conditionally independent given outcome-model covariates. The second mechanism applies when the processes are conditionally independent given observation-time model covariates, which may include outcome-model covariates. The third applies when the processes are conditionally independent given shared, unobserved, latent variables. We consider four semi-parametric marginal regression methods that do not require estimation of the mean effect of time on the outcomes: the Lin method (Lin and Ying, 2001) accommodates the first mechanism, the Bůžková method (Bůžková and Lumley, 2009) accommodates the second mechanism, and the Liang (Liang, Lu, and Ying, 2009) and Sun (Sun, Song, and Zhou, 2011) methods accommodate the third mechanism. We extend both the Liang and the Sun methods to accommodate a combination of the three mechanisms, thereby increasing the flexibility of the models.

In this chapter, we compare currently available and newly extended methods that accommodate outcome-dependent observation times. Our goal is to provide much-needed clarification of the strengths and limitations of each estimation method under alternative outcome-observation dependence mechanisms. In Section 2.2, we elaborate on our framework of outcome-observation dependence mechanisms. We review existing methods under each of these mechanisms (Section 2.2.2) and detail our extensions to both the Liang and Sun methods to accommodate conditional independence through observation-time model covariates, and our extension to the Liang method to accommodate time-dependent covariates in the observation-time model (Section 2.2.3). We present simulation studies to evaluate the performance of the reviewed methods under alternative outcome-observation dependence mechanisms in Section 2.3, and illustrate their application to a bladder cancer study in Section 2.4. Section 2.5 provides guidance on the selection of estimation methods.

2.2. Estimation methods

Let $Y_i(t)$ denote a continuous outcome of interest at time t and $X_i(t)$ denote a $p \times 1$ vector of possibly time-dependent covariates for subject $i = 1, \ldots, n,$. We only consider external covariates, such that any time-dependent covariate process at time *t* is conditionally independent of all previous outcomes given the history of the covariate process (Kalbfleisch and Prentice, 2002). The outcome $Y_i(\cdot)$ is measured at m_i observation times $0 \leq T_{i1} < T_{i2} < \cdots < T_{im_i} \leq \tau$, for which m_i denotes the number of follow-up measurements on the ith individual, and τ denotes the maximum study duration. Using counting process notation, let $N_i(t) = \sum_{s \leq t} dN_i(s)$ denote the number of observations on the i^{th} subject by $t \leq C_i$. The censoring time $C_i \leq \tau$ is the time of last visit or an administrative end-of-study time. The indicator variable $dN_i(t)$ is 1 if a follow-up visit occurred at t and 0 otherwise. We assume non-informative censoring, such that $\mathrm{E}[Y_i(t) \mid X_i(t), C_i \geq t] = \mathrm{E}[Y_i(t) \mid X_i(t)]$. That is, the covariate-outcome associations are the same in subjects who are censored at C_i as those who are still in the study.

2.2.1. Models and assumptions

Semi-parametric outcome model

We assume that primary scientific interest lies in a semi-parametric regression model for the longitudinal continuous outcomes (Lin and Ying, 2001):

$$Y_i(t) = \mu(t) + \beta' X_i(t) + \epsilon_i(t), \tag{2.1}$$

for which $\mu(t)$ is an arbitrary function of time, β is a $p \times 1$ vector of regression parameters of interest, and $\epsilon_i(t)$ is a zero-mean process independent of $X_i(t)$. Model (2.1) specifies a parametric structure for the effect of $X_i(t)$ and a non-parametric structure for $\mu(t)$ (Brumback and Rice, 1998; Lin and Carroll, 2001; Sun et al., 2005).

Observation-time model

We use a standard recurrent events model to describe the observation-time process. Given observation-time model covariates $Z_i(t)$ and a non-negative latent variable η_i with mean 1 and unknown

variance σ^2, N_i is a non-homogeneous Poisson process with intensity function (Lin et al., 2000):

$$\lambda_i(t) = \eta_i \exp\{\gamma' Z_i(t)\}\lambda_0(t), \tag{2.2}$$

in which γ is a vector of unknown parameters, $\lambda_0(t)$ is an unspecified baseline intensity function with $\lambda_0(t) = \int_0^t \lambda(u)du$, and $X_i(t) \subseteq Z_i(t)$. Unless otherwise specified, we assume that η_i is independent of $Z_i(t)$.

A framework of outcome-observation dependence mechanisms

We distinguish three mechanisms that describe the dependence between the outcome and observation-time processes:

(M1) Conditional independence given past outcome-model covariates;

(M2) Conditional independence given past observation-time model covariates;

(M3) Conditional independence given shared latent variables.

Recall that $X_i(t)$ incorporates covariate information available at t, which may include baseline covariates, covariates measured at or before t, and summaries of the covariate history. Details of the framework of outcome-observation dependence mechanisms can be found in Section 1.1.4.

Our framework for outcome-observation dependence in the analysis of longitudinal data provides guidance for the selection of reliable methods. (M2) and (M3) place fewer restrictions on the probability of having a visit than (M1) and are reasonable assumptions in most observational studies. However, (M2) and (M3) are more restrictive because fewer analysis methods are available to provide valid inference, which we detail in the following section.

2.2.2. Existing methods

In this section, we describe four existing methods to estimate covariate-outcome associations in the presence of outcome-observation dependence. All of the methods require estimation of an observation-time model. If the observation-time process is conditionally independent of the censoring times, the parameter γ can be consistently estimated by $\hat{\gamma}$ from the estimating equation (Lin et al., 2000):

$$U(\gamma) = \sum_{i=1}^{n} \int_0^{\tau} \{Z_i(t) - \bar{Z}(t;\gamma)\} dN_i(t) = 0, \tag{2.3}$$

for which $\xi_i(t) = I(C_i > t)$ and

$$\bar{Z}(t;\gamma) = \frac{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' Z_i(t)\} Z_i(t)}{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' Z_i(t)\}}.$$

Method under (M1)

Lin and Ying (2001) assume that the observation-time process is conditionally independent of the outcome process given the outcome-model covariates, as in (M1). The Lin method specifies a marginal semi-parametric outcome model $\mathrm{E}[Y_i(t) \mid X_i(t)] = \mu(t) + \beta' X_i(t)$ and a proportional rate observation-time model $\mathrm{E}[dN_i(t) \mid V_i(t)] = \exp\{\gamma' V_i(t)\} d\lambda_0(t)$, in which $V_i(t)$ is a subset of $X_i(t)$. We define a zero-mean stochastic process (Lin and Ying, 2001):

$$M_i(t; \mathcal{A}, \beta, \gamma) = \int_0^t \{Y_i(s) - \beta' X_i(s)\} dN_i(s) - \int_0^t \exp\{\gamma' V_i(s)\} \xi_i(s) d\mathcal{A}(s), \tag{2.4}$$

in which $\mathcal{A}(t) = \int_0^t \mu(s) d\Lambda(s)$. Based on (2.4), one set of estimating equations to solve for $\mu(t)$ and β is:

$$\sum_{i=1}^{n} M_i(t; \beta, \gamma) = 0, \qquad 0 < t \leq \tau \tag{2.5}$$

$$\sum_{i=1}^{n} \int_0^{\tau} W(t) X_i(t) dM_i(t; \beta, \gamma) = 0. \tag{2.6}$$

The common weight $W(t)$ can improve efficiency and may be data-dependent, such as the proportion of subjects left in the study, i.e., $n^{-1} \sum_{i=1}^{n} \xi_i(t)$. The closed-form expression of $\mathcal{A}(t)$ in (2.5) yields $\tilde{\mathcal{A}}(t;\beta) = \sum_{i=1}^{n} \int_0^t \frac{\{Y_i(s) - \beta' X_i(s)\} dN_i(s)}{\sum_{j=1}^{n} \xi_j(s) \exp\{\hat{\gamma}' V_i(s)\}}$, which replaces $\mathcal{A}(t)$ in (2.6) to form the estimating equation:

$$\sum_{i=1}^{n} \int_0^{\tau} W(t) \{X_i(t) - \bar{X}(t;\gamma)\} \{Y_i(t) - \beta' X_i(t)\} dN_i(t) = 0. \tag{2.7}$$

The centering term is defined as:

$$\bar{X}(t;\gamma) = \frac{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' V_i(t)\} X_i(t)}{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' V_i(t)\}}.$$

We note that:

$$\mathrm{E}\Big[\sum_{i=1}^{n}\int_0^{\tau}\{X_i(t)-\bar{X}(t;\gamma)\}g(t)dN_i(t)\mid\{X_i(t),C_i;i=1,\ldots,n\}\Big] = \sum_{i=1}^{n}\int_0^{\tau}\{X_i(t)-\bar{X}(t;\gamma)\}g(t)dN_i(t)=0$$

for any function $g(\cdot)$, so we extend the left side of (2.7) to obtain the class of estimating functions for β:

$$U_g(\beta;\gamma)=\sum_{i=1}^{n}\int_0^{\tau}W(t)[X_i(t)-\bar{X}(t;\gamma)]\{Y_i(t)-\beta' X_i(t)-g(t;\gamma)\}dN_i(t).$$

One optimal choice of $g(\cdot)$ that minimizes the variance of $U_g(\beta,\gamma)$ is $g(t;\gamma)=\bar{Y}^*(t;\gamma)-\beta'\bar{X}(t;\gamma)$, in which:

$$\bar{Y}^*(t;\gamma)=\frac{\sum_{i=1}^{n}\xi_i(t)\exp\{\gamma' V_i(t)\}Y_i^*(t)}{\sum_{i=1}^{n}\xi_i(t)\exp\{\gamma' V_i(t)\}},$$

and $Y_i^*(t)$ is the measurement of Y_i at the observation nearest to t. Hence β can be consistently estimated from the estimating equation (Lin and Ying, 2001):

$$U(\beta;\gamma)=\sum_{i=1}^{n}\int_0^{\tau}W(t)[X_i(t)-\bar{X}(t;\gamma)]\{Y_i(t)-\bar{Y}^*(t;\gamma)-\beta'[X_i(t)-\bar{X}(t;\gamma)]\}dN_i(t)=0,$$

in which γ is estimated by (2.3) conditioning on the covariates $V_i(t)$. The inclusion of the centering term for covariates accounts for the probability of being observed at t and removes the need for estimation of $\mu(t)$. The centering of the outcome increases the efficiency of the estimation procedure. Note that $Y_i^*(t)$ is the nearest-neighbor approximation of $Y_i(t)$ if the true measurement is not evaluable or collected at t. Li and Ryan (2004) documented the potential issue of such mismeasured covariates. Discussion of other forms of $g(\cdot)$ and $Y_i^*(t)$ can be found in the comments and rejoinder section of Lin and Ying (2001).

Method under (M1) and (M2)

Bůžková and Lumley (2009) relax the assumption of (M1) by addressing the dependence between the outcome and observation-time processes through observation-time model covariates. The set of covariates $Z_i(t)$ may include the outcome-model covariates $X_i(t)$ and past outcomes.

The Bůžková method uses inverse intensity rate ratio (IIRR) weighted estimators to estimate β in the outcome model $\mathrm{E}[Y_i(t)\mid X_i(t)]=\mu(t)+\beta' X_i(t)$. The observation-level inverse weights

standardize the observed data to the time-specific underlying population under the proportional rate model for observation times $\mathrm{E}[dN_i(t) \mid Z_i(t)] = \exp\{\gamma' Z_i(t)\} d\lambda_0(t)$. Inverse weighting has also been shown to reduce bias when cluster size is informative (i.e., the outcomes measured among clustered units are not independent of cluster size) (Williamson, Datta, and Satten, 2003) and when missing data are missing at random (i.e., missingness depends only on observed covariates and outcomes) (Rotnitzky and Robins, 1995; Zhao, Rotnitzky, and Robins, 1995). One particular weight with variance-stabilizing properties is:

$$\rho_i(t; \gamma, \delta) = \frac{\exp\{\gamma' Z_i(t)\}}{\exp\{\delta' X_i(t)\}},$$

for which δ is estimated by $\hat{\delta}$ using (2.3) conditioning on $X_i(t)$ instead of $Z_i(t)$. The proposed estimating equation for β is:

$$U(\beta; \hat{\gamma}, \delta) = \sum_{i=1}^{n} \int_0^{\tau} \frac{W(t)}{\rho_i(t; \hat{\gamma}, \delta)} [X_i(t) - \bar{X}(t; \delta)] \{Y_i(t) - \bar{Y}^*(t; \delta) - \beta' [X_i(t) - \bar{X}(t; \delta)]\} dN_i(t) = 0,$$

in which:

$$\bar{X}(t; \delta) = \frac{\sum_{i=1}^{n} \xi_i(t) \exp\{\delta' X_i(t)\} X_i(t)}{\sum_{i=1}^{n} \xi_i(t) \exp\{\delta' X_i(t)\}},$$

and

$$\bar{Y}^*(t; \delta) = \frac{\sum_{i=1}^{n} \xi_i(t) \exp\{\delta' X_i(t)\} Y_i^*(t)}{\sum_{i=1}^{n} \xi_i(t) \exp\{\delta' X_i(t)\}}.$$

If $Z_i(t) = X_i(t)$ then $\rho_i(t; \gamma, \delta) = 1$, and the Bůžková method reduces to the Lin method. The IIRR-weighted estimates are asymptotically consistent and normal, but the validity of the proposed IIRR-weighted estimator is contingent upon correct specification of $Z_i(t)$ in the observation-time model (Bůžková and Lumley, 2009).

Methods under (M1) and (M3).

The following two methods accommodate subject-specific observation-time processes with arbitrary visit patterns through the use of latent variables.

The Liang method (Liang, Lu, and Ying, 2009) specifies a semi-parametric mixed-effects outcome model:

$$\mathrm{E}[Y_i(t) \mid X_i(t), Q_i(t)] = \mu(t) + \beta' X_i(t) + \eta_{i1}' Q_i(t), \tag{2.8}$$

in which η_{i1} is a vector of unobserved subject-specific latent variables, and $Q_i(t)$ is a subset of $X_i(t)$. The observation-time process is modeled as $\lambda_i(t) = \eta_{i2}\lambda_0(t)\exp\{\gamma' V_i\}$, and V_i is a subset of baseline covariates in $X_i(t)$. The Gamma-distributed latent variable η_{i2} is independent of V_i, $\mathrm{E}[\eta_{i2}] = 1$, and $\mathrm{Var}[\eta_{i2}] = \sigma^2$ is unknown. The relationship between η_{i1} and η_{i2} is defined by the conditional expectation $\mathrm{E}[\eta_{i1} \mid \eta_{i2}] = \theta(\eta_{i2} - 1)$, so θ describes the magnitude and direction of the association between the outcome and observation-time processes. Note that the marginal expectation of η_{i1} is 0. The linear link between η_{i1} and η_{i2} can also be extended to other specified link functions (Liang, Lu, and Ying, 2009). When $\eta_{i1} = 0$, the Liang method reduces to the Lin method.

Conditioning on η_{i2}, the observation-time process is a non-homogeneous process such that m_i has a Poisson distribution with mean $\eta_{i2}\exp(\gamma' V_i)\Lambda(C_i)$. The cumulative baseline intensity function $\lambda_0(t)$ can be consistently estimated by the Aalen-Breslow type estimator $\hat{\Lambda}(t) = \hat{\Lambda}(t, \hat{\gamma})$:

$$\hat{\Lambda}(t, \hat{\gamma}) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{\sum_{j=1}^{n} \xi_j(s)\exp(\hat{\gamma}' V_i)},$$

for which γ is estimated by (2.3) conditioning on the baseline covariates V_i. Given (C_i, m_i, η_{i2}), the observation times $(T_{i1}, T_{i2}, \ldots, T_{im_i})$ are the order statistics of a set of independently and identically distributed random variables with the density function:

$$\frac{\exp\{\gamma' V_i\} d\lambda_0(t)}{\int_0^{C_i} \exp\{\gamma' V_i\} d\Lambda(s)} = \frac{d\lambda_0(t)}{\Lambda(C_i)}, \qquad 0 \le t \le C_i,$$

and the conditional likelihood function for all subjects can be derived as (Huang, Qin, and Wang, 2010):

$$\prod_{i=1}^{n} p(t_{i1}, t_{i2}, \ldots, t_{im_i} \mid C_i, m_i, \eta_{i2}) = \prod_{i=1}^{n} \left\{ m_i! \prod_{j=1}^{m_i} \frac{d\Lambda(t_{ij})}{\Lambda(C_i)} \right\} \propto \prod_{i=1}^{n} \prod_{j=1}^{m_i} \frac{d\Lambda(t_{ij})}{\Lambda(C_i)}$$

Hence $\mathrm{E}\{dN_i(t) \mid C_i, m_i, \eta_{i2}\} = \xi_i(t) m_i \frac{d\lambda_0(t)}{\Lambda(C_i)}$. It follows that:

$$\begin{aligned}
&\mathrm{E}\Big[\{Y_i(t) - \beta' X_i(t)\} dN_i(t) \mid C_i, m_i\Big] \\
&\qquad = \mathrm{E}(\mathrm{E}[\{\mu(t) + \eta_{i1}' Q_i(t) + \epsilon_i(t)\} dN_i(t) \mid C_i, m_i, \eta_{i2}] \mid C_i, m_i) \\
&\qquad = \mu(t)\xi_i(t) m_i \frac{d\lambda_0(t)}{\Lambda(C_i)} + \theta' Q_i(t) \mathrm{E}\{(\eta_{i2} - 1) \mid C_i, m_i\} \mathrm{E}\{dN_i(t) \mid C_i, m_i\}. \quad (2.9)
\end{aligned}$$

We define $B_i(t) = Q_i(t)\mathrm{E}[(\eta_{i2} - 1) \mid C_i, m_i]$ as a covariate based on the subject-specific propensity of visit and $\mathcal{A}(t) = \int_0^t \mu(s)d\Lambda(s)$. Then (2.9) can be expressed as:

$$\mathrm{E}\Big[\{Y_i(t) - \beta' X_i(t) - \theta' B_i(t)\}dN_i(t) \mid C_i, m_i\Big] = \xi(t)\frac{m_i}{\Lambda(C_i)}d\mathcal{A}(t).$$

We can then formulate the zero-mean process:

$$M_{i2}(t; \mathcal{A}, \beta, \theta, \gamma) = \int_0^t \{Y_i(s) - \beta' X_i(s) - \theta' B_i(t)\}dN_i(s) - \int_0^t \xi_i(s)\frac{m_i}{\Lambda(C_i)}d\mathcal{A}(s), \tag{2.10}$$

and define the set of estimating equations based on (2.10) to estimate $\mu(t)$, β, and θ simultaneously:

$$\sum_{i=1}^{n} M_{i2}(t; \beta, \theta, \gamma) = 0, \qquad 0 < t \leq \tau \tag{2.11}$$

$$\sum_{i=1}^{n} \int_0^\tau \begin{pmatrix} X_i(t) \\ \hat{B}_i(t) \end{pmatrix} dM_{i2}(t; \beta, \theta, \gamma) = 0. \tag{2.12}$$

The closed-form expression for $\mathcal{A}(t)$ in (2.11) replaces $\mathcal{A}(t)$ in (2.12), so β and θ can be consistently estimated using the class of estimating equations (Liang, Lu, and Ying, 2009):

$$U(\beta, \theta; \hat{\Lambda}, \hat{B}) = \sum_{i=1}^{n} \int_0^\tau \begin{pmatrix} X_i(t) - \bar{X}(t) \\ \hat{B}_i(t) - \bar{\hat{B}}(t) \end{pmatrix} \{Y_i(t) - \beta' X_i(t) - \theta' \hat{B}_i(t)\}dN_i(t) = 0,$$

for which:

$$\bar{X}(t) = \frac{\sum_{i=1}^n \xi_i(t) X_i(t) m_i / \hat{\Lambda}(C_i)}{\sum_{i=1}^n \xi_i(t) m_i / \hat{\Lambda}(C_i)},$$

and:

$$\bar{\hat{B}}(t) = \frac{\sum_{i=1}^n \xi_i(t) \hat{B}_i(t) m_i / \hat{\Lambda}(C_i)}{\sum_{i=1}^n \xi_i(t) m_i / \hat{\Lambda}(C_i)}.$$

To estimate $B_i(t)$, the conditional expectation of η_{i2} given (C_i, m_i) is required. If we assume that η_{i2} is Gamma distributed with mean 1 and variance σ^2, the expectation of η_{i2} can be expressed as:

$$\mathrm{E}(\eta_{i2} | C_i, m_i) = \frac{1 + m_i\sigma^2}{1 + \exp(\gamma' V_i)\Lambda(C_i)\sigma^2}.$$

The covariate $B_i(t)$ can thus be estimated by:

$$\hat{B}_i(t) = \left(\frac{1 + m_i \hat{\sigma}^2}{1 + \exp(\hat{\gamma}' V_i) \hat{\Lambda}(C_i) \hat{\sigma}^2} - 1 \right) Q_i(t),$$

for which $\hat{\sigma}^2$ is a consistent estimator of σ^2 defined as:

$$\hat{\sigma}^2 = \max \left\{ \frac{\sum_{i=1}^n \{m_i^2 - m_i - \exp(2\hat{\gamma}' V_i) \hat{\Lambda}^2(C_i)\}}{\sum_{i=1}^n \exp(2\hat{\gamma}' V_i) \hat{\Lambda}^2(C_i)}, 0 \right\}. \tag{2.13}$$

Similar to the Liang method, the Sun method (Sun, Song, and Zhou, 2011) accommodates (M3). In contrast to the Liang method, the distribution of the latent variable is completely unspecified, and the same latent variable η_i is shared between the outcome and observation-time models. The Sun method specifies the semi-parametric marginal model:

$$\mathrm{E}[Y_i(t) \mid X_i(t), \eta_i] = \mu(t) + \beta' X_i(t) + \alpha \eta_i. \tag{2.14}$$

Similar to θ in the Liang method, α parameterizes the correlation between the outcome and observation-time processes. If $\alpha = 0$, then the Sun method reduces to the Lin method.

Conditioning on η_i, $N_i(t)$ is a non-homogeneous Poisson process with intensity function $\lambda_i(t) = \eta_i \lambda_0(t) \exp\{\gamma' X_i(t)\}$. The distribution of η_i under the Sun method may depend on observed time-independent outcome-model covariates V_i with $\mathrm{E}[\eta_i \mid V_i] = 1$. Discussion regarding covariate-dependent latent variables or frailties can be found in recent literature (Heagerty and Kurland, 2001; Liu, Kalbfleisch, and Schaubel, 2011; McCulloch and Neuhaus, 2011; Neuhaus and McCulloch, 2006). Let $\hat{\pi}(t; X_i) = \int_0^t \exp\{\hat{\gamma}' X_i(u)\} d\hat{\Lambda}(u)$, $\hat{\eta}_i = (m_i - 1)/\hat{\pi}(C_i; X_i)$, and $\hat{\Omega}_i = (m_i - 1)(m_i - 2)/\hat{\pi}(C_i; X_i)^2$. The class of estimating equations for β and α has the form:

$$U_1(\beta, \alpha; \gamma) = \sum_{i=1}^n \int_0^\tau W(t) [\{X_i(t) - \bar{X}(t; \gamma)\} \{Y_i(t) - \beta' X_i(t) - \alpha \hat{\eta}_i\}] dN_i(t) = 0,$$

and:

$$U_2(\beta, \alpha; \gamma) = \sum_{i=1}^n \int_0^\tau W(t) [\{\hat{\eta}_i - \bar{\eta}(t; \gamma)\} \{Y_i(t) - \beta' X_i(t)\} - \alpha \{\hat{\Omega}_i - \hat{\eta}_i \bar{\eta}(t; \gamma)\}] dN_i(t) = 0,$$

for which:

$$\bar{X}(t;\gamma) = \frac{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' X_i(t)\} X_i(t) m_i / \hat{\pi}(C_i; X_i)}{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' X_i(t)\} m_i / \hat{\pi}(C_i; X_i)},$$

and:

$$\bar{\eta}(t;\gamma) = \frac{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' X_i(t)\} \hat{\eta}_i m_i / \hat{\pi}(C_i; X_i)}{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' X_i(t)\} m_i / \hat{\pi}(C_i; X_i)}.$$

2.2.3. Extensions

Extension to Liang method to accommodate time-dependent covariates.

The estimation procedure of Liang, Lu, and Ying (2009) allows adjustment for time-independent covariates in the observation-time model. Here, we extend the Liang method to accommodate time-dependent covariates. Let $\hat{\pi}(t; V_i) = \int_0^t \exp\{\hat{\gamma}' V_i(u)\} d\hat{\Lambda}(u)$. The class of estimating equations for β and θ permitting time-dependent covariates in the observation-time model has the form:

$$U(\beta, \theta; \hat{\Lambda}, \hat{B}) = \sum_{i=1}^{n} \int_0^{\tau} \begin{pmatrix} X_i(t) - \bar{X}(t) \\ \hat{B}_i(t) - \bar{\hat{B}}(t) \end{pmatrix} \{Y_i(t) - \beta' X_i(t) - \theta' \hat{B}_i(t)\} dN_i(t) = 0,$$

for which:

$$\bar{X}(t) = \frac{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' X_i(t)\} X_i(t) m_i / \hat{\pi}(C_i; X_i)}{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' X_i(t)\} m_i / \hat{\pi}(C_i; X_i)},$$

$$\bar{\hat{B}}(t) = \frac{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' X_i(t)\} \hat{B}_i(t) m_i / \hat{\pi}(C_i; X_i)}{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' X_i(t)\} m_i / \hat{\pi}(C_i; X_i)},$$

and $\hat{B}_i(t)$ can be estimated as before by replacing $\hat{\Lambda}(C_i)$ with $m_i / \hat{\pi}(C_i; V_i)$. We provide details on consistency and asymptotic normality of the estimators in Appendix A.

Weighted-Liang and Weighted-Sun methods.

We propose extensions to the Liang and Sun methods to offer additional flexibility when parameterizing outcome-observation dependence under both (M2) and (M3). Recall that we denote $X_i(t)$ as the outcome-model covariates and $Z_i(t)$ as the observation-time model covariates. With the inclusion of observation-level weights $\rho_i(t; \hat{\gamma}, \delta)$, the set of estimating equation for the Weighted-Liang method can be expressed as:

$$U(\beta, \theta; \hat{\Lambda}, \hat{B}) = \sum_{i=1}^{n} \int_0^{\tau} \frac{W(t)}{\rho_i(t; \hat{\gamma}, \delta)} \begin{pmatrix} X_i(t) - \bar{X}(t) \\ \hat{B}_i(t) - \bar{\hat{B}}(t) \end{pmatrix} \{Y_i(t) - \beta' X_i(t) - \theta' \hat{B}_i(t)\} dN_i(t) = 0,$$

for which:

$$\bar{X}(t) = \frac{\sum_{i=1}^{n} \xi_i(t) \exp\{\delta' X_i(t)\} X_i(t) m_i / \hat{\pi}(C_i; Z_i)}{\sum_{i=1}^{n} \xi_i(t) \exp\{\delta' X_i(t)\} m_i / \hat{\pi}(C_i; Z_i)},$$

and:

$$\bar{\hat{B}}(t) = \frac{\sum_{i=1}^{n} \xi_i(t) \exp\{\delta' X_i(t)\} \hat{B}_i(t) m_i / \hat{\pi}(C_i; Z_i)}{\sum_{i=1}^{n} \xi_i(t) \exp\{\delta' X_i(t)\} m_i / \hat{\pi}(C_i; Z_i)}.$$

Similarly, the set of estimating functions for the Weighted-Sun method is:

$$U_1(\beta, \alpha; \hat{\gamma}) = \sum_{i=1}^{n} \int_0^{\tau} \frac{W(t)}{\rho_i(t; \hat{\gamma}, \delta)} [\{X_i(t) - \bar{X}(t)\}\{Y_i(t) - \beta' X_i(t) - \alpha \hat{\eta}_i\}] dN_i(t) = 0,$$

and:

$$U_2(\beta, \alpha; \hat{\gamma}) = \sum_{i=1}^{n} \int_0^{\tau} \frac{W(t)}{\rho_i(t; \hat{\gamma}, \delta)} [\{\hat{\eta}_i - \bar{\eta}(t)\}\{Y_i(t) - \beta' X_i(t)\} - \alpha\{\hat{\Omega}_i - \hat{\eta}_i \bar{\eta}(t)\}] dN_i(t) = 0,$$

for which $\hat{\eta}_i = \frac{m_i - 1}{\hat{\pi}(C_i; Z_i)}$,

$$\bar{\eta}(t) = \frac{\sum_{i=1}^{n} \xi_i(t) \exp\{\delta' X_i(t)\} \hat{\eta}_i m_i / \hat{\pi}(C_i; Z_i)}{\sum_{i=1}^{n} \xi_i(t) \exp\{\delta' X_i(t)\} m_i / \hat{\pi}(C_i; Z_i)},$$

and:

$$\bar{X}(t) = \frac{\sum_{i=1}^{n} \xi_i(t) \exp\{\delta' X_i(t)\} X_i(t) m_i / \hat{\pi}(C_i; Z_i)}{\sum_{i=1}^{n} \xi_i(t) \exp\{\delta' X_i(t)\} m_i / \hat{\pi}(C_i; Z_i)}.$$

We provide details on consistency and asymptotic normality of our extensions in Appendices A and B.

2.2.4. Summary

In this section, we formulated a semi-parametric linear regression model to evaluate the marginal association between covariates and a continuous outcome of interest in the presence of outcome-dependent observation times. We presented a framework of outcome-observation dependence mechanisms. The Lin method is the most restrictive of the reviewed methods, because it is suitable only for the stronger assumption of (M1); the Bůžková method accommodates (M2) and reduces to (M1) when the additional covariates in the observation-time model are not required; the Liang and Sun methods accommodate (M3), with (M1) as a special case. We proposed two methods, the Weighted-Liang and Weighted-Sun methods, which offer considerable flexibility in that they can ac-

commodate all (or any combination) of the three outcome-observation dependence mechanisms. We note that standard error estimation for all methods is most easily obtained using bootstrap procedures; in this setting, a cluster bootstrap, in which subjects are sampled with replacement, is required (Efron and Tibshirani, 1993; Field and Welsh, 2007). The resampling of subjects assumes that the correlation structure within each subject is retained (Cheng, Yu, and Huang, 2013; Chernick, 2011). R code for the cluster-bootstrap procedure is included in Appendix A.

2.3. Simulation study

We evaluated the statistical properties of the reviewed methods through simulation studies under two outcome-observation dependence settings: (i) (M2) and (ii) (M2) and (M3). All simulations were conducted in R 2.13.1 (R Development Core Team, Vienna, Austria).

2.3.1. Setting 1: Simulations under (M2)

Parameters

In this setting, we used covariates to induce correlation between the outcome and observation-time processes. Following the simulation procedure of Bůžková and Lumley (2009), we generated continuous outcomes at each of 1000 iterations using the linear mixed-effects model:

$$Y_i(t) = \mu(t) + \beta_1 X_{i1}(t) + \beta_2(X_{i2} - \mathrm{E}[X_{i2} \mid X_{i1}]) + \epsilon_i(t), \tag{2.15}$$

for which $\mu(t) = t$, $\epsilon_i(t) \sim$ Normal(0,1), and β_1 was the target of inference. The time-dependent covariate $X_{i1}(t) = X_{i1}\log(t)$ was a known function of time, in which X_{i1} followed a Uniform[0,1] distribution. The time-independent covariate X_{i2} was drawn from a mixture distribution, for which $X_{i2} \sim$ Normal(2,1) if $X_{i1} \leq 0.5$ and $X_{i2} \sim$ Normal(0,4) if $X_{i1} > 0.5$. Hence X_{i2} in model (2.15) influenced the covariate-outcome association of $X_{i1}(t)$. To ensure proper marginalization of model (2.15), X_{i2} was centered by its conditional mean given X_{i1}, resulting in the marginal semi-parametric outcome model:

$$\mathrm{E}[Y_i(t) \mid X_i(t)] = \mu(t) + \beta_1 X_{i1}(t). \tag{2.16}$$

We generated observation times T_{ik} from a non-homogeneous Poisson process with intensity func-

tion $\lambda_i(t) = \eta_i \lambda_0(t) \exp\{\gamma_1 X_{i1}(t) + \gamma_2 X_{i2}\}$. Note that X_{i2} induced additional correlation between the outcome and observation-time processes. We set $\lambda_0(t) = \frac{\sqrt{t}}{2}$ and generated the latent variable η_i from a Gamma distribution with mean 1 and variance $\sigma^2_\eta = 0.5$. The independent censoring time C_i was generated from Uniform[5,10]. We considered various combinations of outcome parameters $(\beta_1 = 1, \beta_2 = \{0, 0.3, 1\})$ and intensity parameters $(\gamma_1 = 0.5, \gamma_2 = \{0, -0.2, 0.5\})$. When $\beta_2 = 0$ and $\gamma_2 = 0$, the outcome-observation dependence model satisfied (M1); when $\gamma_2 \neq 0$, the outcome-observation dependence model satisfied (M2).

Results

Table 2.1 provides the estimated bias, empirical standard error estimates, and mean squared error estimates for estimation of β_1 in model (2.16). Recall that the Lin, Liang, and Sun methods estimate β_1 without accounting for X_{i2} in any way, whereas the Bůžková, Weighted-Liang and Weighted-Sun methods incorporate the effect of X_{i2} through observation-level weights. As anticipated, all six methods yielded approximately unbiased parameter estimates for β_1 if (M1) was satisfied ($\gamma_2 = 0$), i.e., when the outcome process was conditionally independent of the observation-time process given outcome-model covariates. The Lin, Bůžková, Weighted-Liang and Weighted-Sun estimates of β_1 were comparable in bias and efficiency to both the Liang and Sun estimators. However, if (M1) was violated ($\gamma_2 \neq 0$), i.e., the source of additional correlation between the two processes was induced by an additional covariate X_{i2}, then only the Bůžková, Weighted-Liang, and Weighted-Sun methods performed well, with negligible biases in all settings. When $\beta_2 = 0$ and $\gamma_2 \neq 0$, all methods performed well because X_{i2} was not associated with the outcome. As β_2 increased, the biases of Lin, Liang and Sun estimates for β_1 increased. A positive value of γ_2 with positive values of X_{i2} led to more observations per subject, which increased efficiency in the estimation of β_1 in most cases.

In this setting, we also quantified the price of assuming (M3) when the latent variable was unnecessary. We calculated the estimated relative efficiency (ERE) of unbiased estimators with the estimated variance of the Weighted-Liang and Weighted-Sun methods in the numerator, and the estimated variance of the Bůžková method in the denominator. The ERE indicated that the loss of efficiency was reasonable and comparable between the Weighted-Liang and Weighted-Sun methods. As β_2 increased (i.e., the dependence between the outcome and observation-time models increased), the ERE decreased. In addition, we also calculated the ERE of IIRR-weighted versus unweighted methods to investigate the loss of efficiency due to inclusion of the additional covariate

Table 2.1: Simulation results for β_1 under (M2): Bias, $\hat{\beta}_1 - \beta_1, \beta_1 = 1$; ESE, empirical sample error; MSE, mean-squared error; ERE, estimated relative efficiency

n	β_2	γ_2	**Lin**			**Bůžková**			**Liang** (extension)			**Weighted-Liang** (extension)				**Sun**			**Weighted-Sun** (extension)			
			Bias	ESE	MSE	Bias	ESE	MSE	Bias	ESE	MSE	Bias	ESE	MSE	ERE[a]	Bias	ESE	MSE	Bias	ESE	MSE	ERE[a]
100	0	0	0.003	0.284	0.081	0.003	0.284	0.080	0.004	0.280	0.078	−0.006	0.393	0.155	1.915	0.004	0.282	0.079	−0.006	0.393	0.155	1.915
		−0.2	−0.001	0.292	0.085	−0.002	0.289	0.084	0.003	0.287	0.082	−0.012	0.428	0.183	2.193	0.003	0.288	0.083	−0.013	0.428	0.183	2.193
		0.5	−0.020	0.342	0.118	−0.009	0.289	0.084	−0.018	0.324	0.105	−0.007	0.379	0.143	1.720	−0.018	0.329	0.109	−0.008	0.379	0.144	1.724
	0.3	0	0.003	0.318	0.101	0.002	0.313	0.098	0.004	0.313	0.098	−0.008	0.411	0.169	1.724	0.004	0.315	0.099	−0.007	0.411	0.169	1.313
		−0.2	−0.136	0.336	0.132	−0.010	0.332	0.110	−0.098	0.323	0.114	−0.023	0.467	0.219	1.979	−0.119	0.328	0.121	−0.024	0.467	0.219	1.979
		0.5	0.336	0.399	0.273	−0.003	0.323	0.104	0.231	0.352	0.177	0.002	0.385	0.149	1.421	0.300	0.367	0.225	0.011	0.393	0.155	1.480
	1	0	0.003	0.543	0.294	−0.001	0.521	0.271	0.004	0.536	0.288	−0.010	0.577	0.333	1.227	0.004	0.539	0.290	−0.010	0.577	0.333	1.227
		−0.2	−0.452	0.624	0.594	−0.029	0.587	0.345	−0.333	0.555	0.419	−0.048	0.687	0.475	1.370	−0.402	0.582	0.500	−0.050	0.686	0.473	1.366
		0.5	1.167	0.769	1.954	0.011	0.558	0.312	0.811	0.549	0.960	0.022	0.560	0.314	1.007	1.043	0.629	1.484	0.054	0.599	0.362	1.152
200	0	0	−0.004	0.203	0.041	−0.003	0.203	0.041	−0.002	0.202	0.041	−0.010	0.284	0.081	1.957	−0.003	0.202	0.041	−0.010	0.285	0.081	1.971
		−0.2	−0.004	0.213	0.045	−0.004	0.209	0.044	−0.001	0.210	0.044	−0.009	0.304	0.092	2.116	−0.001	0.211	0.044	−0.009	0.304	0.092	2.116
		0.5	0.001	0.258	0.067	0.004	0.202	0.041	−0.003	0.237	0.056	0.002	0.269	0.072	1.773	−0.001	0.243	0.059	0.001	0.269	0.072	1.773
	0.3	0	−0.007	0.227	0.052	−0.007	0.224	0.050	−0.006	0.225	0.051	−0.014	0.299	0.090	1.782	−0.006	0.226	0.051	−0.014	0.299	0.090	1.782
		−0.2	−0.140	0.246	0.080	−0.008	0.237	0.056	−0.099	0.235	0.065	−0.013	0.329	0.108	1.927	−0.121	0.239	0.072	−0.013	0.329	0.108	1.927
		0.5	0.378	0.305	0.236	0.006	0.220	0.048	0.255	0.255	0.130	0.007	0.271	0.073	1.517	0.326	0.267	0.178	0.010	0.270	0.073	1.506
	1	0	−0.014	0.389	0.152	−0.015	0.374	0.140	−0.015	0.386	0.149	−0.023	0.421	0.178	1.267	−0.014	0.388	0.150	−0.023	0.421	0.178	1.267
		−0.2	−0.457	0.462	0.422	−0.016	0.420	0.177	−0.328	0.405	0.272	−0.023	0.486	0.236	1.339	−0.401	0.428	0.344	−0.023	0.485	0.236	1.333
		0.5	1.257	0.602	1.943	0.012	0.369	0.136	0.856	0.389	0.884	0.018	0.379	0.144	1.055	1.089	0.449	1.386	0.031	0.377	0.143	1.044

[a] Estimated relative efficiency was calculated for unbiased estimators with the variance of the Bůžková parameter estimate in the denominator.

X_{i2} when none was needed (Appendix A.3). The EREs between the Bůžková and Lin methods were close to 1 under all scenarios. The loss of efficiency was greater for the Weighted-Liang and Weighted-Sun methods, but decreased as the number of observations increased (i.e., greater γ_2) and when β_2 increased.

2.3.2. Setting 2: Simulation under (M2) and (M3)

Parameters

In the previous setting, we focused on outcome-observation dependence induced through covariates. In this setting, we focus on estimation of β_1 under various forms of latent variable structures. To simulate data under both (M2) and (M3), we generated outcomes at each of 1000 iterations using the linear mixed-effects model:

$$Y_i(t) = \mu(t) + \beta_1 X_{i1}(t) + \beta_2(X_{i2} - \mathrm{E}[X_{i2} \mid X_{i1}]) + \alpha\eta_{i1}Q_i(t) + \epsilon_i(t), \tag{2.17}$$

in which $\mu(t)$, $\epsilon_i(t)$, $X_{i1}(t)$ and X_{i2} were as defined in Section 2.3.1.

The observation times T_{ik} were generated from a non-homogeneous Poisson process with intensity function $\lambda_i(t) = \eta_{i2}\lambda_0(t)\exp\{\gamma_1 X_{i1}(t) + \gamma_2 X_{i2}\}$, for which $\lambda_0(t) = \frac{\sqrt{t}}{2}$. The independent censoring time C_i was generated from Uniform[7,10]. The coefficients were set at $\beta_1 = 1$, $\beta_2 = 0.3$, $\gamma_1 = 0.5$, $\gamma_2 = -0.2$, and $\alpha = 1$. Because $\alpha \neq 0$ in model (2.17), correlation was introduced between the outcome and the observation-time processes through latent variables. We generated the latent variable η_{i2} under two scenarios:

1. η_{i2} from Gamma distribution with mean 1 and variance 0.5; hereby $\eta_{i2}^{(1)}$.
2. η_{i2} from a mixture distribution, following Uniform[0.5,1.5] if $X_{i1} \leq 0.5$ and Gamma distribution with mean 1 and variance 0.7 if $X_{i1} > 0.5$; hereby $\eta_{i2}^{(2)}$.

The latent variable η_{i1} was generated under two scenarios:

1. $\eta_{i1} = \eta_{i2}$, hereby $\eta_{i1}^{(1)}$.
2. $\mathrm{E}[\eta_{i1} \mid \eta_{i2}] = \theta(\eta_{i2} - 1), \theta = 1$, hereby $\eta_{i1}^{(2)}$.

We let $Q_i(t) = 1$ or $Q_i(t) = X_{i1}$. When $Q_i(t) = X_{i1}$, model (2.17) can be considered a random

coefficient model. The latent variables were dependent on the outcome process either through $Q_i(t) = X_{i1}$ or $\eta_{i2}^{(2)}$. The simulation setup mirrored the setup of Sun, Song, and Zhou (2011) if $\eta_{i1} = \eta_{i2}$ and $Q_i(t) = 1$ and mirrored the setup of Liang, Lu, and Ying (2009) if $\alpha = 1$, η_{i2} was Gamma distributed with mean 1, and η_{i1} and η_{i2} were linearly linked through $\mathrm{E}[\eta_{i1} \mid \eta_{i2}] = \theta(\eta_{i2} - 1)$.

Results

Table 2.2 provides the estimated bias, empirical standard error estimates, and mean squared error estimates for estimation of β_1 in (2.17). The inclusion of X_{i2} in the observation-time model satisfied (M2) and induced additional correlation between the outcome and observation-time processes, so the IIRR-weighted methods (Bůžková, Weighted-Liang, and Weighted-Sun) performed better than their unweighted counterparts, reflecting the results of Setting 1.

Under the Sun setup (i.e., $\eta_{i1}^{(1)} : \eta_{i1} = \eta_{i2}$ and $Q_i(t) = 1$), all IIRR-weighted methods yielded approximately unbiased parameter estimates for β_1 under $\eta_{i2}^{(1)}$. Under the Liang setup (i.e., $\eta_{i1}^{(2)} : \mathrm{E}[\eta_{i1} \mid \eta_{i2}] = \theta(\eta_{i2} - 1)$ and $Q_i(t) = 1$), all methods yielded approximately unbiased estimates under $\eta_{i2}^{(1)}$. Under $\eta_{i2}^{(2)}$, in which the distribution of the latent variable depended on X_{i1}, only the Weighted-Sun method yielded approximately unbiased estimates under $Q_i(t) = 1$, though the bias under the Weighted-Liang method was smaller in magnitude than the Bůžková method. If the effect of the latent variable η_{i1} on the outcomes was associated with the value of X_{i1} (i.e., $Q_i(t) = X_{i1}$), then the bias of β_1 was small under the Weighted-Liang method but large under all other methods.

2.3.3. Setting 3: Simulations under (M2) and (M3) with additional covariates

Parameters

In Setting 2, we generated data using the same set of covariates in settings with and without latent variables; this setting is thus unfair to methods that do not explicitly incorporate latent variables, such as the Bůžková method. In this setting, we generated both outcomes and observation times according to Setting 2, but fit the observation-time model with additional covariates that were correlated with the latent variables: $Z_3 = I(\eta_{i2} > 1)$; or $Z_4 = 2(\eta_{i2} - 1)$. We fit three observation-time models:

Model 1: $\lambda_i(t) = \eta_{i2}\lambda_0(t)\exp\{\gamma_1 X_{i1}(t) + \gamma_2 X_{i2}\}$

Model 2: $\lambda_i(t) = \eta_{i2}\lambda_0(t)\exp\{\gamma_1 X_{i1}(t) + \gamma_2 X_{i2} + \gamma_3 Z_{i3}\}$

Table 2.2: Simulation results for β_1 under (M2) and (M3): Bias, $\hat{\beta}_1 - \beta_1$, $\beta_1 = 1$; ESE, empirical sample error; MSE, mean-squared error

n	η_{i1}[a]	Q_i	η_{i2}[b]	**Lin**			**Bůžková**			**Liang** (extension)			**Weighted-Liang** (extension)			**Sun**			**Weighted-Sun** (extension)		
				Bias	ESE	MSE	Bias	ESE	MSE	Bias	ESE	MSE	Bias	ESE	MSE	Bias	ESE	MSE	Bias	ESE	MSE
100	$\eta_{i1}^{(1)}$	1	$\eta_{i2}^{(1)}$	-0.155	0.430	0.209	-0.030	0.426	0.182	-0.603	0.496	0.611	0.007	0.421	0.177	-0.595	0.494	0.598	0.016	0.407	0.166
			$\eta_{i2}^{(2)}$	0.335	0.426	0.294	0.469	0.418	0.394	-0.281	0.464	0.294	0.116	0.405	0.178	-0.305	0.462	0.306	0.008	0.398	0.158
	$\eta_{i1}^{(2)}$	1	$\eta_{i2}^{(1)}$	-0.171	0.488	0.267	-0.041	0.484	0.236	-0.592	0.539	0.641	0.002	0.481	0.231	-0.585	0.536	0.630	0.014	0.472	0.223
			$\eta_{i2}^{(2)}$	0.356	0.528	0.406	0.490	0.528	0.519	-0.223	0.541	0.342	0.134	0.470	0.239	-0.247	0.540	0.353	0.025	0.465	0.217
		X_{i1}	$\eta_{i2}^{(1)}$	0.104	0.415	0.183	0.232	0.411	0.223	-0.358	0.481	0.359	0.004	0.445	0.198	-0.284	0.493	0.324	0.250	0.442	0.258
			$\eta_{i2}^{(2)}$	0.348	0.483	0.354	0.486	0.485	0.472	-0.233	0.504	0.309	0.045	0.443	0.198	-0.163	0.506	0.283	0.127	0.438	0.208
200	$\eta_{i1}^{(1)}$	1	$\eta_{i2}^{(1)}$	-0.166	0.307	0.122	-0.025	0.310	0.097	-0.621	0.351	0.509	0.012	0.302	0.092	-0.615	0.349	0.500	0.018	0.292	0.085
			$\eta_{i2}^{(2)}$	0.336	0.311	0.209	0.484	0.301	0.325	-0.285	0.342	0.198	0.113	0.294	0.100	-0.304	0.341	0.209	0.007	0.288	0.083
	$\eta_{i1}^{(2)}$	1	$\eta_{i2}^{(1)}$	-0.179	0.342	0.149	-0.043	0.350	0.124	-0.607	0.387	0.518	-0.004	0.353	0.124	-0.601	0.385	0.510	0.004	0.346	0.120
			$\eta_{i2}^{(2)}$	0.338	0.372	0.253	0.492	0.377	0.384	-0.240	0.390	0.210	0.104	0.346	0.131	-0.258	0.391	0.220	-0.002	0.343	0.118
		X_{i1}	$\eta_{i2}^{(1)}$	0.100	0.297	0.098	0.235	0.303	0.147	-0.361	0.352	0.254	0.003	0.324	0.105	-0.289	0.358	0.211	0.257	0.329	0.174
			$\eta_{i2}^{(2)}$	0.335	0.339	0.227	0.490	0.344	0.358	-0.237	0.367	0.190	0.024	0.323	0.105	-0.168	0.365	0.162	0.110	0.320	0.114

[a] Two possible links: $\eta_{i1}^{(1)} : \eta_{i1} = \eta_{i2}$; $\eta_{i1}^{(2)} : \mathrm{E}[\eta_{i1} \mid \eta_{i2}] = \theta(\eta_{i2} - 1), \theta = 1$

[b] Latent variable distributions: $\eta_{i2}^{(1)} : \eta_{i2} \sim$ Gamma(mean $= 1, \sigma^2 = 0.5$); $\eta_{i2}^{(2)} : \eta_{i2} \sim I(X_{i1} \leq 0.5)$Uniform$[0.5, 1.5] + I(X_{i1} > 0.5)$Gamma$(1, 0.7)$

Model 3: $\lambda_i(t) = \eta_{i2}\lambda_0(t)\exp\{\gamma_1 X_{i1}(t) + \gamma_2 X_{i2} + \gamma_4 Z_{i4}\}$.

The coefficients were set at $(\beta_1, \beta_2) = (1, 0.3)$, $(\gamma_1, \gamma_2) = (0.5, -0.2)$, and $\alpha = 1$. We considered combinations of $\eta_{i1}^{(1)}$, $\eta_{i1}^{(2)}$, $\eta_{i2}^{(1)}$, and $\eta_{i2}^{(2)}$, and let $Q_i(t) = 1$ or X_{i1}.

Results

Table 2.3 provides the estimated bias, empirical standard error estimates, and mean squared error estimates for estimation of β_1 in (2.17) under the Bůžková method. The results under Model 1 are reflective of results from Setting 2; the bias was small only when the latent variables were commonly distributed across all subjects (i.e., $\eta_{i2}^{(1)}$) and $Q_i(t) = 1$. When we fit the observation-time model covariates with additional covariates Z_3 or Z_4 that were correlated with the latent variable, as in Models 2 and 3, the bias under the Bůžková method for all cases was reduced. The estimates were unbiased only when the latent variable was commonly distributed across all subjects. The estimated observation-level weights under Models 2 and 3 capture information regarding subject-specific visit intensities, resulting in smaller biases under the Bůžková method compared to Model 1.

Table 2.3: Simulation results for β_1 under (M2) and (M3) with additional covariates: Bias, $\hat{\beta}_1 - \beta_1$, $\beta_1 = 1$; ESE, empirical sample error; MSE, mean-squared error

n	η_{i1}[a]	Q_i	η_{i2}[b]	**Bůžková** Model 1			**Bůžková** Model 2			**Bůžková** Model 3		
				Bias	ESE	MSE	Bias	ESE	MSE	Bias	ESE	MSE
100	$\eta_{i1}^{(1)}$	1	$\eta_{i2}^{(1)}$	-0.030	0.426	0.182	0.001	0.358	0.128	0.002	0.310	0.096
			$\eta_{i2}^{(2)}$	0.469	0.417	0.394	0.252	0.340	0.179	0.107	0.303	0.103
	$\eta_{i1}^{(2)}$	1	$\eta_{i2}^{(1)}$	-0.041	0.484	0.236	-0.018	0.419	0.176	0.002	0.381	0.145
			$\eta_{i2}^{(2)}$	0.490	0.528	0.519	0.272	0.437	0.265	0.123	0.402	0.176
		X_{i1}	$\eta_{i2}^{(1)}$	0.232	0.411	0.223	0.074	0.361	0.136	0.056	0.336	0.116
			$\eta_{i2}^{(2)}$	0.486	0.485	0.472	0.215	0.392	0.200	0.124	0.368	0.151

[a] Two possible links: $\eta_{i1}^{(1)} : \eta_{i1} = \eta_{i2}$; $\eta_{i1}^{(2)} : \mathrm{E}[\eta_{i1} \mid \eta_{i2}] = \theta(\eta_{i2} - 1), \theta = 1$

[b] Latent variable distributions: $\eta_{i2}^{(1)} : \eta_{i2} \sim$ Gamma(mean $= 1, \sigma^2 = 0.5$); $\eta_{i2}^{(2)} : \eta_{i2} \sim I(X_{i1} \leq 0.5)$Uniform$[0.5, 1.5] + I(X_{i1} > 0.5)$Gamma$(1, 0.7)$

We fit three observation-time models:

Model 1: $\lambda_i(t) = \eta_{i2}\lambda_0(t)\exp\{\gamma_1 X_{i1}(t) + \gamma_2 X_{i2}\}$

Model 2: $\lambda_i(t) = \eta_{i2}\lambda_0(t)\exp\{\gamma_1 X_{i1}(t) + \gamma_2 X_{i2} + \gamma_3 Z_{i3}\}$, $Z_{i3} = I(Z_{i2} > 1)$.

Model 3: $\lambda_i(t) = \eta_{i2}\lambda_0(t)\exp\{\gamma_1 X_{i1}(t) + \gamma_2 X_{i2} + \gamma_4 Z_{i4}\}$, $Z_{i4} = 2 * (Z_{i2} - 1)$.

2.3.4. Summary

Our simulation results quantified the potential for bias in estimated covariate-outcome associations under various outcome-observation dependence mechanisms. The Bůžková, Weighted-Liang, and Weighted-Sun methods performed better when (M2) is satisfied. In Setting 1, we examined the

robustness of the methods that included latent variables when they were not needed. We showed that the potential loss of efficiency was moderate and decreased when the dependence between the outcome and observation-time models increased. We also examined the relative efficiency between IIRR-weighted and unweighted methods to examine potential loss of efficiency due to including an unnecessary additional covariate in the observation-time model. The results indicated that the loss of efficiency was moderate and decreased with greater number of observations or increased dependence between the outcome and observation-time processes. The Weighted-Liang and Weighted-Sun methods were the most flexible in that they could accommodate a combination of outcome-observation dependence mechanisms. They also provided estimates with negligible bias depending on the relationship between the latent variable and the outcome model covariates. In practice, ensuring unbiased estimates through a more complex dependence model may be more important than a potential loss in efficiency. In Setting 3, we performed simulation studies in which the observation-time model under the Bůžková approach included additional covariates that were correlated with the latent variables. Results indicated that adjustment for these covariates reduced the bias under the Bůžková method. Hence in practice, observed covariates that are correlated with the unobserved latent variable may be used to capture information regarding subject-specific visit intensities.

In practice, we rely on exploratory data analysis, model diagnostics, and sensitivity analyses to investigate the relationship between the outcome and observation-time processes and to ensure selection of an appropriate analysis method. We illustrate and discuss strategies for model selection in Sections 2.4 and 2.5.

2.4. Case study

2.4.1. Background

We compared the reviewed methods using a subset of data from a bladder cancer study conducted by the Veterans Administration Cooperative Urological Research Group (Andrews and Herzberg, 1985). Eighty-five patients with superficial bladder tumors were randomly assigned to placebo $(n = 47)$ or thiotepa treatment $(n = 38)$. At each follow-up visit, new tumors were counted before being removed transurethrally. The maximum study duration was 53 months. There was notable heterogeneity in visit patterns across patients. The median (25^{th}, 75^{th} percentile) number of visits in

the placebo group and treatment group was 9 (5, 12) and 9 (4, 23), respectively. The average time between visits for the placebo group was 3.7 months, compared to 2.3 months for the treatment group. These differences suggested that the patients in the treatment group visited the clinic more often. Hence the observation-time process must account for this difference to estimate properly the effect of treatment on tumor recurrence.

Our analysis focused on the natural logarithm of the cumulative number of new tumors observed up to t plus 1, to retain a marginal response. We included a treatment indicator (X_1) and the natural logarithm of the initial number of tumors plus 1 (X_2) in the outcome model. We considered the following outcome models:

Lin and Bůžková methods: $\mathrm{E}[Y_i(t) \mid X_i(t)] = \mu(t) + \beta_1 X_{i1} + \beta_2 X_{i2}$;

Liang and Weighted-Liang methods: $\mathrm{E}[Y_i(t) \mid X_i(t), \eta_{i1}] = \mu(t) + \beta_1 X_{i1} + \beta_2 X_{i2} + \eta_{i1} Q_i$, $Q_i = X_{i1}$;

Sun and Weighted-Sun methods: $\mathrm{E}[Y_i(t) \mid X_i(t), \eta_{i1}] = \mu(t) + \beta_1 X_{i1} + \beta_2 X_{i2} + \alpha \eta_{i1}$.

The consensus of previous analyses was that the tumor recurrence and observation-time processes were dependent (Hu, Sun, and Wei, 2003; Liang, Lu, and Ying, 2009; Sun and Wei, 2000). We note that the outcome may be intrinsically dependent upon the measurement process, such that larger intervals between visits allows for more tumors to grow. The outcome is undoubtedly expected to increase with longer time between visits. We considered two observation-time models:

Case 1: $\lambda_i(t) = \eta_{i2} \exp\{\gamma_1 X_{i1} + \gamma_2 X_{i2}\} \lambda_0(t)$

Case 2: $\lambda_i(t) = \eta_{i2} \exp\{\gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 \log(\text{\# new tumors since baseline} + 1)\} \lambda_0(t)$

Case 1 specified the same set of covariates in both the outcome and observation-time models. Case 2 specified an additional covariate based on number of tumors since baseline because it is common for the physician to schedule a patient's next visit based on the outcomes so far. Recall that $\eta_{i1} = \eta_{i2}$ in the Sun and Weighted-Sun methods and $\mathrm{E}[\eta_{i1} \mid \eta_{i2}] = \theta(\eta_{i2} - 1)$ in the Liang and Weighted-Liang methods.

2.4.2. Results

Table 2.4 provides estimates for β and γ under the Lin, Liang, and Sun methods in Case 1. We obtained $\hat{\gamma}_1 = 0.444$ (SE, 0.093) and $\hat{\gamma}_2 = -0.001\,(0.115)$, which suggested that treatment assignment was significantly associated with the observation-time process. We specified $Q_i = X_{i1}$ because X_{i1} had a significant effect in the observation-time model, and the results in Table 2.4 mirrored the conclusion from Liang, Lu, and Ying (2009).

Table 2.4: Parameter estimates and estimated standard errors (SE) under Case 1

Method	$\hat{\beta}_1$ (SE)	$\hat{\beta}_2$ (SE)	$\hat{\theta}$ (SE)*	$\hat{\alpha}$ (SE)*
Lin	$-0.701\,(0.172)$	$0.657\,(0.165)$		
Liang	$-0.588\,(0.175)$	$0.682\,(0.147)$	$-0.235\,(0.243)$	
Sun	$-0.751\,(0.188)$	$0.680\,(0.159)$		$-0.043\,(0.398)$

$\hat{\gamma}_1 = 0.444\,(0.093)$, $\hat{\gamma}_2 = -0.001\,(0.115)$

* The parameters θ and α represent the association between the outcome and observation-time processes for the Liang and Sun methods respectively.

Next, we examined the importance of the additional covariate in Case 2. Table 2.5 provides estimates for β and γ for IIRR-weighted methods under Case 2. We found that the cumulative number of tumors since baseline was significantly related to the observation-time process. The Wald test of $\gamma_3 = 0$ in the observation-time model provided a p-value < 0.001, implying that the inclusion of the additional covariate was appropriate. Hence the IIRR-weighted methods were more appropriate than the unweighted methods, and we focused on the results in Table 2.5. The observation-level weights applied to the Bůžková, Weighted-Liang, and Weighted-Sun methods ranged from 0.50 to 1.26, with median (25^{th}, 75^{th} percentile) = 0.93 (0.84, 1.06). With the incorporation of observation-levels weights, the treatment effect under the Bůžková method was attenuated compared to the Lin method. The treatment effect estimates under Weighted-Liang and Weighted-Sun methods were lower than those under the Liang and Sun methods. Because the initial number of tumors was not significantly related to the observation-time process, the corresponding estimates $\hat{\beta}_2$ were comparable under all methods.

Table 2.5: Parameter estimates and their estimated standard errors (SE) under Case 2

Method	$\hat{\beta}_1$ (SE)	$\hat{\beta}_2$ (SE)	$\hat{\theta}$ (SE)*	$\hat{\alpha}$ (SE)*
Bůžková	$-0.565\,(0.170)$	$0.572\,(0.165)$		
Weighted-Liang	$-0.395\,(0.166)$	$0.584\,(0.147)$	$-0.266\,(0.229)$	
Weighted-Sun	$-0.423\,(0.182)$	$0.580\,(0.156)$		$-0.247\,(0.247)$

$\hat{\gamma}_1 = 0.536\,(0.090)$, $\hat{\gamma}_2 = -0.105\,(0.128)$, $\hat{\gamma}_3 = 0.227\,(0.076)$

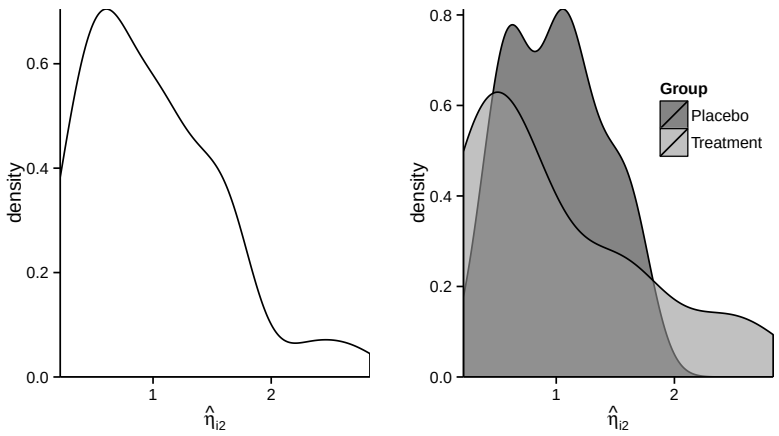
Stabilized weights: median (25^{th}, 75^{th} percentile)=0.93 (0.84, 1.06)

* The parameters θ and α represent the association between the outcome and observation-time processes for the Weighted-Liang and Weighted-Sun methods respectively.

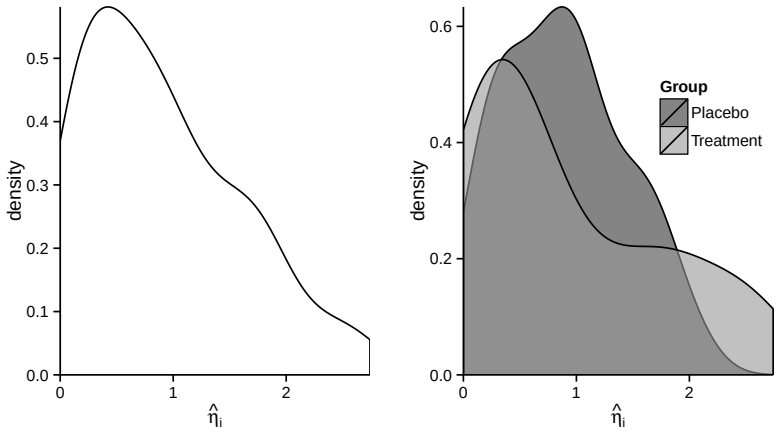
To determine the necessity of latent variables in the outcome models, we focused on the variance of the latent variable η_{i2} in the observation-time model. Under Case 2, the estimated variance based on (2.13) was 0.448, indicating that the latent variable approaches were appropriate. Based on the variance property of the Gamma distribution, we partitioned the variance of η_{i2} as the contribution from the placebo group (0.059) and the thiotepa group (0.417). The difference in variance estimates indicated the possibility of covariate-dependent η_{i2}, in that the distribution of η_{i2} was different between the treatment groups. Next, we used the density curve of $\hat{\eta}_{i2}$ to graphically check if η_{i2} was covariate dependent. The density curves were indeed different between the treatment groups (Figure 2.1).

Given the evidence of (M2) and (M3), we focused on the results under the Weighted-Liang and Weighted-Sun methods. We note that the same $Z_i(t)$ was used for the methods on Table 2.5. As in the simulation study, correct specification of covariates in $Z_i(t)$ may recover the effect of the latent variable under the Bůžková method. We did not have access to other measured covariates in this data set; if those were available, it may have been possible to find candidates for $Z_i(t)$ such that the treatment estimate under the Bůžková method were closer to those under the Weighted-Liang and Weighted-Sun methods.

The choice between Weighted-Liang and Weighted-Sun methods relied on the distribution of η_{i2}. The Weighted-Liang method assumes that η_{i2} is derived from a Gamma distribution with a common variance for all subjects, whereas the Weighted-Sun method places no distributional assumption on η_{i2}. Considering the evidence of covariate dependence based on the density curves, the results from the Weighted-Sun method best described the data, although the estimates for β_1 were similar between the Weighted-Liang and Weighted-Sun methods. Overall, the results indicated that



(a) Density plot of estimated η_{i2} under Weighted-Liang method



(b) Density plot of estimated η_i under Weighted-Sun method

Figure 2.1: Bladder data: Density plots of estimated latent variables

treatment and the initial number of tumors had significant effects on tumor recurrence. We also observed a negative correlation between tumor recurrence and the observation-time processes ($\hat{\alpha} = -0.247$).

Lastly, we evaluated the fit of the outcome model based on the procedure presented in Liang, Lu, and Ying (2009). We derived residuals $\hat{\epsilon}_i(t) = Y_i(t) - \hat{y}_i(t)$ using parameter estimates from Table 2.5. Denote $0 \leq t_1 < t_2 < \cdots < t_M$ as the M total observation times among all subjects. The estimate of $\mu(t)$ is a step function with jumps at unique observation times: $\hat{\mu}(t_k) = \frac{d\hat{\mathcal{A}}(t_k)}{d\hat{\Lambda}(t_k)} = \frac{\hat{\mathcal{A}}(t_k) - \hat{\mathcal{A}}(t_k-)}{\hat{\Lambda}(t_k) - \hat{\Lambda}(t_k-)}, 1 \leq k \leq M$. Based on the residual plots of $\hat{\epsilon}_i(t)$ against the observation times (Figure 2.2), there was some evidence of lack of fit for large outcome values, but it was not systematic

with respect to time and was similar across all weighted methods.

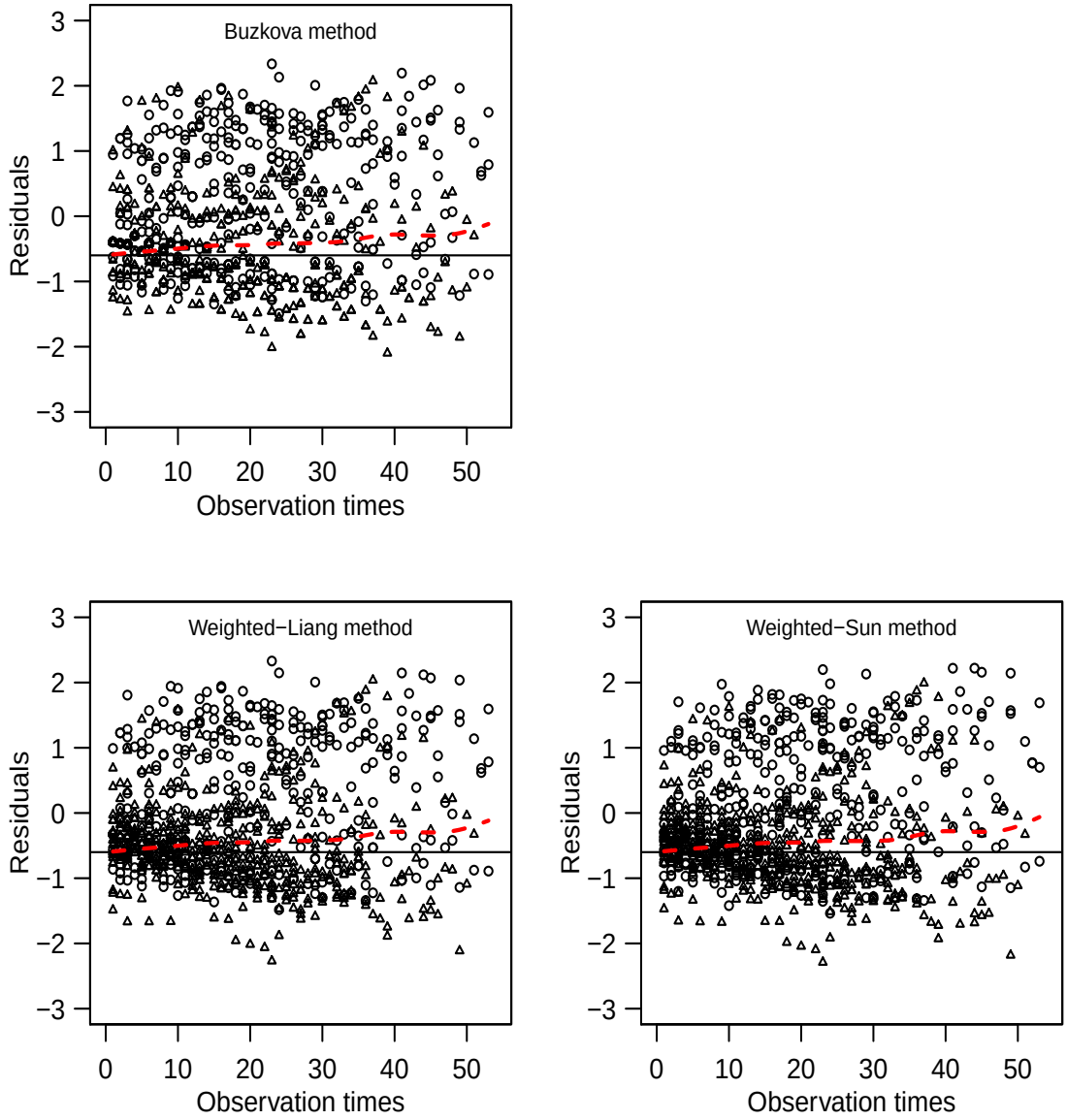


Figure 2.2: Bladder data: Residual plots by observation times

2.5. Discussion

In this chapter, we evaluated the statistical properties of currently available and newly extended semi-parametric methods for the analysis of longitudinal data with outcome-dependent observation times. Table 2.6 summarizes the strengths and limitations of each method under various outcome-

observation dependence mechanisms. The performance of each method hinges on the assumed mechanism of dependence between the outcome and observation-time processes. For conditional independence given covariates in the outcome model only (M1), all reviewed methods are appropriate. For conditional independence given observation-time model covariates only (M2), the Bůžková method is preferred. For conditional independence given unobserved latent variables only (M3), all methods perform well when the latent variables are independent of outcome-model covariates. However, if the distribution of the latent variables is covariate dependent, then the Sun method is preferred; if the effect of the latent variable in the outcome model is modified by any outcome-model covariates, then the Liang method is preferred. Under both (M2) and (M3), our extensions, the Weighted-Liang and Weighted-Sun methods, are the most flexible and remove the bias otherwise associated with the original Liang and Sun methods under (M2). In addition, our extension of the method by Liang, Lu, and Ying (2009) allows time-dependent covariates in the observation-time process, which would otherwise not be possible.

In practice, empirical model checking can be useful to decide which method is most appropriate. First, to decide between (M1) and (M2), one can focus on the observation-time model and perform a Wald test of the additional $q - p$ covariates (Lin et al., 2000). If the Wald test yields a significant result, the data suggest (M2). Next, one can determine the necessity of latent variables in the outcome model using the variance of the latent variable η_{i2}. If the estimated variance of the latent variables is small (i.e., close to 0), latent variables may not be required. One method to estimate $\mathrm{Var}[\eta_{i2}]$ is to assume a parametric distribution for the latent variables, such as using equation (2.13) if we can assume η_{i2} is Gamma distributed. The distribution of the latent variable in the observation-time model is unspecified in the Lin, Bůžková, and Sun methods, but is assumed to be Gamma distributed in the Liang method. There is a lack of formal techniques to check the Gamma distribution assumption of the unobserved latent variable. A series of sensitivity analyses is recommended. Liang, Lu, and Ying (2009) showed that the Liang method provided reasonable estimates for covariate-outcome association even if the distribution of the latent variable η_{i2} was misspecified, especially when the variance of the distribution was small. Robustness of the Liang and Weighted-Liang methods to misspecification of the distribution of η_{i2} can be improved by replacing the estimate of η_{i2} by $\hat{\eta}_{i2} = m_i \Big/ \int_0^{C_i} \exp\{\hat{\gamma}' X_i(t)\} d\lambda_0(t)$, removing any distributional assumption. The choice between the Liang and Sun methods rests upon whether the distribution of the latent variable is covariate-dependent. An informal check is to partition the estimated variance

of η_{i2} by the covariate values to determine if the partitioned variances are similar across levels of $X_i(t)$. We can also graphically display the density curves of $\hat{\eta}_{i2}$ to check for covariate-dependent latent variables. Lastly, we can evaluate the overall fit of the models based on residuals. Formal model selection is an area of future research.

Several features of the methods discussed here deserve comment. First, the semi-parametric outcome model does not require the estimation of $\mu(t)$. However, the potential gain from the flexibility of the form of $\mu(t)$ is countered by the potential loss in efficiency of estimation of the parameters of interest. Second, we assume that censoring times are independent of the outcome and observation-time model processes, i.e., non-informative censoring. This assumption may be relaxed to allow censoring to depend on the outcome and observation-time processes by estimating γ and Λ using the method proposed by Huang, Qin, and Wang (2010). In addition, the parameters in the outcome model are time-independent, which may not be appropriate in some cases. We refer readers to the procedure in Sun, Song, and Zhou (2011) to derive time-dependent regression coefficients. Third, our goal is to generate inference regarding the marginal association between a set of covariates and the outcome of interest, rather than to conduct formal causal inference. If we allow intervention on $X_i(t)$, modification of the exposure may influence not only the outcome of interest, but also occurrence of a visit. Hence the quantification of the causal effect of the exposure on the outcome of interest requires techniques that establish the temporal association between exposure and outcome. A g-computation algorithm (Robins, Greenland, and Hu, 1999) or inverse-probability-of-treatment weighted estimators (Robins, Hernán, and Brumback, 2000) may provide insight into estimation of causal effects. Lastly, the observation-time process can be modeled on two time scales: total time scale (i.e., time-to-events model) in which each recurrent event is measured from a time of origin, and gap time scale (i.e., time-between-events model) in which the measure of interest is time between successive events (Cook and Lawless, 2007). The methods in this paper adopt the total time scale, but it may be appropriate to consider the alternative parameterization. The time-between-events approach is well-studied within the recurrent events field (Huang and Liu, 2007), but the use of the gap time scale in the regression modeling of longitudinal data with outcome-dependent observation times warrants future research.

It is of interest to note that in the framework of incomplete data, GEE is able to accommodate missing completely at random (MCAR) data and the special case of covariate-dependent missing-

ness (Little, 1995; Rubin, 1976). Similarly, in the current focus on outcome-dependent observation times, GEE does provide reliable estimates of β under (M1), assuming a correctly specified function of time in the outcome model. With the inclusion of observation-level inverse intensity weights, a weighted-GEE model may also provide reasonable estimates of β under (M2) with the ease of currently available software packages (Bůžková and Lumley, 2007). However, the advantage of the methods in Section 2.2 is the flexibility provided by the non-parametric specification of the effect of time.

The methods we described are currently limited to linear models for continuous outcomes. Recent research has focused on the development of log-linear models for count outcomes (Bůžková and Lumley, 2008; Sun, Tong, and He, 2007). In the next chapter, we introduce a new semi-parametric method that can accommodate binary outcomes.

Table 2.6: Summary of methods for various outcome-observation dependence mechanisms:
$+$ Appropriate; $-$ Not appropriate; $+/-$ Appropriate under certain situations; N/A Not applicable

				Lin	**Bůžková**	**Liang**	**Liang** (extension)	**Weighted-Liang** (extension)	**Sun**	**Weighted-Sun** (extension)
Time-independent covariates in the observation-time model	Mechanism (M1)	Conditional independence given outcome-models covariates		$+$	$+$	$+$	$+$	$+$	$+$	$+$
	Mechanism (M2)	Conditional independence given observation-time model covariates		$-$	$+$	$-$	$-$	$+$	$-$	$+$
	Mechanism (M3) Conditional independence given latent variables (LV)	LV not associated with outcome-model covariates	LV distribution specified Gamma	$+$	$+$	$+$	$+$	$+$	$+$	$+$
			LV distribution unspecified	$+$	$+$	$+$	$+$	$+$	$+$	$+$
		LV associated with outcome-model covariates	Effect of LV modified by $Q(t)^*$	$-$	$-$	$-$	$+$	$+$	$-$	$-$
			Distribution of LV based on $Q(t)^*$	$-$	$-$	$-$	$-$	$-$	$+$	$+$
	Mechanisms (M2) + (M3)			$-$	$-$	$-$	$-$	$+/-$	$-$	$+/-$
Time-dependent covariates in the observation-time model				$+/-$	$+/-$	N/A	$+/-$	$+/-$	$+/-$	$+/-$

* $Q(t)$ is a subset of $X(t)$

CHAPTER 3

SEMI-PARAMETRIC METHOD FOR BINARY OUTCOMES

3.1. Introduction

Longitudinal studies typically focus on an explicit outcome of interest collected over time. The data-collection schedule may constitute an implicit outcome (Rizopoulos, 2012), in that the timing or frequency of data collection may communicate information regarding features of the study design or patient-level characteristics. Consider the use of warfarin, a commonly prescribed oral anticoagulant. A patient on warfarin requires frequent monitoring, based on the international normalized ratio (INR), due to the drug's narrow therapeutic range. Anticoagulation levels above or below the therapeutic range increase the risk of bleeding or thromboembolism, respectively (Hylek et al., 1996). An out-of-range INR typically triggers a dose change (Brigden et al., 1998); a physician may request multiple closely spaced follow-up visits to monitor the impact of the dose change on INR response (Figure 3.1). In such settings, the intensity of events such as follow-up visits may depend on previous outcomes and measured or unmeasured covariates. If interest lies in estimating the effect of observed covariates on the probability of being out of therapeutic range, then it is necessary to incorporate the data-collection schedule in the estimation procedure. We focus on a marginal mean regression model to estimate the association between observed covariates and a binary outcome of interest. We refer to the longitudinal outcomes as the outcome process and the occurrence of data collection over time as the observation-time process.

If the probability of having a follow-up visit depends upon previous outcomes and measured or unmeasured covariates, then the outcome and observation-time processes are dependent and conventional longitudinal data analysis methods such as generalized estimating equations (GEE, Liang and Zeger, 1986) that ignore the observation-time process may provide biased estimates of covariate-outcome associations (French and Heagerty, 2009; Sun et al., 2005). We have introduced a framework to describe the potential relationship between the outcome and observation-time processes based on assumptions regarding conditional independence (Section 1.1.4). Specifically, we assume that the outcome and observation-time processes are conditionally independent given past observed covariates in the outcome model, past observed covariates in the observation-

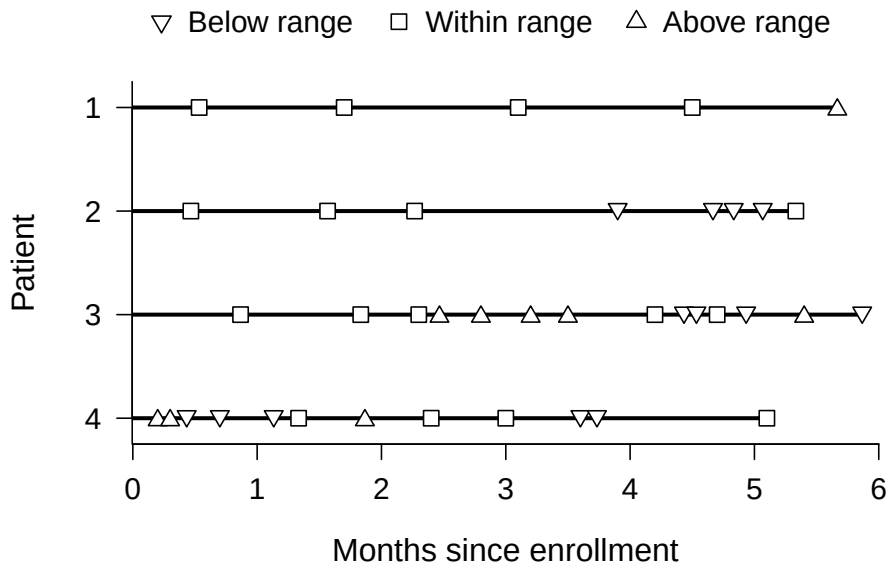


Figure 3.1: Observation times for four selected patients on warfarin and the corresponding observed outcomes: INR below, within, or above the therapeutic range.

time model, and/or shared, unobserved latent variables.

Various methods have been proposed to account for potential dependence between the observation-time process and longitudinal binary outcomes. Fitzmaurice et al. (2006) proposed a pseudo-likelihood estimator utilizing a linear approximation of the conditional distribution of the binary outcomes. The estimator requires strong assumptions about the observation-time process (e.g., that the conditional distribution of the outcome process at time t is independent of the observation-time process given the most recent observed value of response prior to t) and does not allow for explicit specification of the observation-time model. Other authors have adopted an estimating equations approach, explicitly specifying the observation-time models to be incorporated into the estimation of the outcome model (Bůžková and Lumley, 2007; Lin, Scharfstein, and Rosenheck, 2004). Although these estimating equations approaches allow weaker assumptions about the observation-time process, they require a parametric structure for the mean trajectory of the outcomes over time.

Several authors have proposed semi-parametric estimation procedures that assume a non-parametric structure for the mean trajectory of the longitudinal outcomes and a parametric structure for covariate effects (Bůžková and Lumley, 2009; Liang, Lu, and Ying, 2009; Lin and Ying, 2001; Sun, Song, and Zhou, 2011). These models provide flexibility when the focus is on the effect of

a particular covariate of interest, while the effect of time is considered a nuisance. Closed-form solutions for the mean trajectory and the parameters of interest are derived using the properties of mean-zero processes. These proposed semi-parametric estimation procedures allow for more flexible modeling of the longitudinal outcomes, but are currently limited to continuous and count outcomes.

We consider a joint model approach to semi-parametric marginal regression to accommodate outcome-observation dependence in longitudinal studies with binary outcomes. Through the incorporation of observation-level visit-intensity weights and shared latent variables, our proposed joint model approach provides flexibility to accommodate the assumptions of conditional independence given observed covariates and/or subject-level latent variables, while not imposing a parametric assumption on the mean trajectory of the longitudinal binary outcomes.

In Section 3.2, we detail assumptions regarding conditional independence between the outcome and observation-time processes. In Section 3.3, we introduce a comprehensive estimation procedure for regression modeling of binary outcomes in the presence of outcome-dependent observation times. We present simulation studies to evaluate the performance of our proposed procedure under alternative outcome-observation dependence mechanisms in Section 3.4, and illustrate its application to data from a warfarin study in Section 3.5. Section 3.6 provides discussion and concluding remarks. R code to implement our proposed method is available in Appendix B.

3.2. Model formulation and assumptions

We consider a longitudinal study with *n* independent subjects in the study interval $[0, \tau]$, for which τ is the maximum study duration. For subject i, $i = 1, \ldots, n$, let $Y_i(t)$ denote a binary outcome of interest at time t, and $X_i(t)$ denote a $p \times 1$ vector of possibly time-dependent covariates. Unless otherwise specified, we consider only external covariates, such that any time-dependent covariate process at time *t* is conditionally independent of all previous outcomes, given the history of the covariate process (Kalbfleisch and Prentice, 2002). $Y_i(\cdot)$ is measured at m_i observation times $0 \leq T_{i1} < T_{i2} < \cdots < T_{im_i} \leq \tau$, for which m_i denotes the number of follow-up measurements on the i^{th} individual. Using counting process notation, let $N_i(t) = \sum_{s \leq t} dN_i(s)$ denote the number of observations on the i^{th} subject by time $t \leq C_i$, in which C_i is the censoring time. The indicator variable $dN_i(t)$ equals 1 if a follow-up visit occurred on the i^{th} individual at time t and equals 0 otherwise. We

assume non-informative censoring, such that $\Pr[Y_i(t) = 1 \mid X_i(t), C_i \geq t] = \Pr[Y_i(t) = 1 \mid X_i(t)]$. That is, the covariate-outcome associations are the same in those who are censored at C_i as those who have survived beyond C_i.

3.2.1. Semi-parametric outcome model

We assume that primary scientific interest lies in a semi-parametric regression model for the longitudinal binary outcomes. We extend the semi-parametric linear regression model for continuous outcomes proposed by Lin and Ying (2001) to binary outcomes $Y_i(t)$ under independent or dependent observation times:

$$\Pr[Y_i(t) = 1 \mid X_i(t)] = \text{expit}\{\mu(t) + \beta' X_i(t)\} = \frac{\mu(t) + \beta' X_i(t)}{1 + \exp\{\mu(t) + \beta' X_i(t)\}}, \tag{3.1}$$

for which $\mu(t)$ is an arbitrary function of time and β is a $p \times 1$ vector of regression parameters of interest.

3.2.2. Observation-time model

The observation-time process describes the timing and intensity of follow-up visits and is characterized by a standard recurrent events model. We introduce a non-negative latent variable η_i with mean 1 and unknown variance σ^2. Given observation-time model covariates $Z_i(t)$ and η_i, the recurrent event process $N_i(\cdot)$ is a non-homogeneous Poisson process with intensity function (Lin et al., 2000; Pepe and Cai, 1993):

$$\lambda_i(t) = \eta_i \lambda_0(t) \exp\{\gamma' Z_i(t)\}, \quad t \in [0, \tau] \tag{3.2}$$

for which γ is a vector of unknown parameters and $\lambda_0(t)$ is an arbitrary baseline intensity function with $\lambda_0(t) = \int_0^t \lambda(u) du$. If the censoring time is independent of the observation-time process, then the parameter γ can be consistently estimated by $\hat{\gamma}$ from the following estimating function (Lin et al., 2000):

$$U(\gamma) = \sum_{i=1}^{n} \int_0^{\tau} \{Z_i(t) - \bar{Z}(t; \gamma)\} dN_i(t), \tag{3.3}$$

for which:

$$\bar{Z}(t;\gamma) = \frac{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' Z_i(t)\} Z_i(t)}{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' Z_i(t)\}},$$

and $\xi_i(t) = I(C_i > t)$.

3.2.3. Assumptions regarding conditional independence

Recall the framework of outcome-observation dependence mechanisms that describes the dependence between the outcome and observation-time processes (Section 1.1.4):
(M1) Conditional independence given past outcome-model covariates;
(M2) Conditional independence given past observation-time model covariates;
(M3) Conditional independence given shared latent variables.

For the remainder of the chapter, conditional independence given covariates implies conditional independence given past observed covariates.

3.3. Estimation and inference

In this section, we detail a new estimation procedure to estimate covariate-outcome associations with binary outcomes in a joint modeling approach under any combination of the three outcome-observation dependence mechanisms described in the previous section.

3.3.1. Estimators

Estimator under M1

Given the semi-parametric outcome model (3.1), and the observation-time model $\mathrm{E}[\, dN_i(t) \mid X_i(t)] = \exp\{\gamma' X_i(t)\}\, d\lambda_0(t)$, we can define the zero-mean stochastic process for binary outcomes as:

$$M_i(t;\beta,\gamma) = \int_0^t \left[Y_i(s)\xi_i(s)\, dN_i(s) - \text{expit}\{\mu(s) + \beta' X_i(s)\}\xi_i(s) \exp\{\gamma' X_i(s)\}\, d\Lambda(s) \right]. \qquad (3.4)$$

$M_i(t;\beta,\gamma)$ is appropriate if it is assumed that the occurrence of a follow-up visit is a feature of the study design or known patient characteristics and not due to previous outcomes or unmeasured patient characteristics.

Estimator under M2

For continuous outcomes, Bůžková and Lumley (2009) proposed a method that relaxes the as-

sumption of (M1) and accommodates (M2) by applying observation-level weights to the estimating equation to account for dependence through covariates in the observation-time model, $Z_i(t)$. Recall that $Z_i(t)$ may include the outcome-model covariates $X_i(t)$ and summaries of past outcomes.

Given the marginal semi-parametric regression model (3.1), the observation-level weights standardize the observed data to the time-specific underlying population under the proportional rate model for observation times $\mathrm{E}[dN_i(t) \mid Z_i(t)] = \exp\{\gamma' Z_i(t)\} d\lambda_0(t)$. One particular observation-level weight with variance-stabilizing properties is:

$$\rho_i(t; \gamma, \delta) = \frac{\exp\{\gamma' Z_i(t)\}}{\exp\{\delta' X_i(t)\}},$$

for which δ is estimated by $\hat{\delta}$ using (3.3) conditioning on $X_i(t)$. The zero-mean process $M_i(t; \beta, \gamma)$ from (3.4) can then be extended as:

$$M_{i1}(t; \beta, \gamma, \delta) = \int_0^t \frac{1}{\rho_i(s, \gamma, \delta)} \Big[Y_i(s) \xi_i(s)\, dN_i(s) \\ - \text{expit}\{\mu(s) + \beta' X_i(s)\} \xi_i(s) \exp\{\gamma' Z_i(s)\}\, d\Lambda(s) \Big]. \tag{3.5}$$

Estimator under M2 and M3

To allow outcome-observation dependence through observed covariates and unobserved latent variables, the outcome model (3.1) can be extended to:

$$\Pr[Y_i(t) = 1 \mid X_i(t)] = \text{expit}\{\mu(t) + \beta' X_i(t) + \eta_{i1}' Q_i(t)\}, \tag{3.6}$$

in which $Q_i(t)$ is a $q \times 1$ subvector of $X_i(t)$ and η_{i1} is a q-dimensional vector of subject-specific latent variables that represent subject-level propensity for visit (Liang, Lu, and Ying, 2009). The observation-time model can be expressed as:

$$\mathrm{E}[d\Lambda_i(t) \mid Z_i(t)] = \eta_{i2} \exp\{\gamma' Z_i(t)\}\, d\lambda_0(t), \tag{3.7}$$

in which η_{i2} is a mean-one, non-negative latent variable. The distribution of η_{i2} may depend on observed time-independent outcome-model covariates V_i with $\mathrm{E}[\eta_{i2} \mid V_i] = 1$. Discussion regarding covariate-dependent latent variables or frailties can be found in recent literature (Heagerty and

Kurland, 2001; Liu, Kalbfleisch, and Schaubel, 2011; McCulloch and Neuhaus, 2011; Neuhaus and McCulloch, 2006). The latent variables from models (3.6) and (3.7) are assumed to be linearly linked through $\mathrm{E}[\eta_{i1} \mid \eta_{i2}] = \theta(\eta_{i2} - 1)$. The parameter θ describes the association between the outcome and observation-time processes. Thus, to ensure that β retains a marginal interpretation with the inclusion of the latent variable, we define $B_i(t) = \mathrm{E}[(\eta_{i2} - 1) \mid m_i, C_i]Q_i(t)$ as a fixed covariate that incorporates the subject-specific propensity for visit. The outcome model (3.6) can be re-expressed as:

$$\Pr[Y_i(t) = 1 \mid X_i(t), B_i(t)] = \text{expit}\{\mu(t) + \beta' X_i(t) + \theta' B_i(t)\}. \tag{3.8}$$

Next, we re-express the observation-time model. Let $\mathcal{Z}_i(t) = \{Z_i(s) : 0 \leq s < t\}$ denote the covariate history of Z_i up to t. Following the results from Huang, Qin, and Wang (2010), the event times ($t_{i1} < t_{i2} < \cdots < t_{im_i}$) of the i^{th} subject conditional on $\{C_i, m_i, \eta_{i2}, \mathcal{Z}(C_i)\}$ are order statistics of a set of independent and identically distributed random variables with the density function:

$$\frac{\exp\{\gamma' Z_i(t)\} d\lambda_0(t)}{\int_0^{C_i} \exp\{\gamma' Z_i(s)\} d\Lambda(s)}, \qquad 0 \leq t \leq C_i.$$

Define $\pi(t; Z_i) = \int_0^t \exp\{\gamma' Z_i(s)\} d\Lambda(s)$. The conditional likelihood function for all subjects can be derived as (Huang, Qin, and Wang, 2010):

$$\prod_{i=1}^{n} p(t_{i1}, t_{i2}, \ldots, t_{im_i} \mid C_i, m_i, \eta_{i2}, \mathcal{Z}(C_i)) = \prod_{i=1}^{n} \left\{ m_i! \prod_{j=1}^{m_i} \frac{d\pi(t; Z_i)}{\pi(C_i; Z_i)} \right\} \propto \prod_{i=1}^{n} \prod_{j=1}^{m_i} \frac{d\pi(t; Z_i)}{\pi(C_i; Z_i)}.$$

It follows that:

$$\mathrm{E}[dN_i(t) \mid C_i, m_i, \eta_{i2}, \mathcal{Z}(C_i)] = \xi_i(t) m_i \frac{d\pi(t; Z_i)}{\pi(C_i; Z_i)}.$$

Using both re-expressed outcome and observation-time models, the zero-mean process (3.5) can then be extended as:

$$M_{i2}(t; \beta, \theta, \gamma, \delta) = \int_0^t \frac{1}{\rho_i(s, \gamma, \delta)} \Big[Y_i(s) \xi_i(s)\, dN_i(s) - \text{expit}\{\mu(s) + \beta' X_i(s) + \theta' \hat{B}_i(s)\} \xi_i(s) m_i \frac{d\pi(s, Z_i)}{\pi(C_i, Z_i)} \Big] \tag{3.9}$$

in the presence of both (M2) and (M3).

To estimate $B_i(t)$, we first estimate η_{i2} from the observation-time model. We utilize the property that given $\{\eta_{i2}, C_i, \mathcal{Z}(C_i)\}$, m_i follows a Poisson distribution with mean $\eta_{i2}\pi(C_i, Z_i)$ to obtain $\hat{\eta}_{i2} = \{\frac{m_i}{\pi(C_i,Z_i)}\}$, so $\hat{B}_i(t) = \{\frac{m_i}{\pi(C_i,Z_i)} - 1\}Q_i(t)$. Equation (3.9) is the most general formulation of the joint model and can accommodate (M1), (M2), and (M3); that is, provide valid estimation of β under any combination of the three conditional independence mechanisms. Given a specific mechanism of (M1), (M2) or (M3), $M_{i2}(t; \beta, \theta, \gamma, \delta)$ can be reduced to (3.4) and (3.5). In subsequent sections, we proceed with estimation of β via the estimation equation (3.9), which we refer to as the 'proposed estimator.'

3.3.2. Estimation procedure

Unlike for continuous or count outcomes, there is no closed-form solution for $\mu(t)$ and β for binary outcomes based on the zero-mean process (3.9), by setting $M_{i2}(t; \beta, \theta, \gamma, \delta) = 0$. Computational issues may arise because $\mu(t)$ is infinite dimensional, and iterative procedures may be difficult with sparse data resulting from few subjects with visits at each unique observation time. To overcome these computational burdens, we impose a flexible structure on $\mu(t)$ using basis approximations. Generalizing the notation from Huang and Liu (2007), suppose the smooth function $\mu(\cdot)$ can be approximated by a spline function such that $\mu(t) \approx \sum_{k=1}^{K_n} \varphi_k G_k(t) = \varphi' G(t)$ in which $\{G_k(\cdot), k = 1, \ldots, K_n\}$ is a basis system of B-splines, $\varphi = (\tau_1, \ldots, \tau_{K_n})'$ and $G(t) = (G_1(t), \ldots, G_{K_n}(t))'$. Let $\tilde{H}_i(t) = G_i(t)$ or $\tilde{H}_{ij} = G(T_{ij})$. (3.8) can thus be approximated by $\Pr[Y_i(t) = 1 \mid X_i(t), B_i(t)] =$ expit$\{\varphi\tilde{H}_i(t) + \beta' X_i(t) + \theta' B_i(t)\}$. Let $s_1 < s_2 < \cdots < s_J$ denote the J distinct ordered observation times from all subjects $\{t_{ik}, i = 1, \ldots, n; k = 1, \ldots, m_i\}$. We propose to estimate β from (3.9) by the estimating equation:

$$\sum_{i=1}^{n}\sum_{k=1}^{m_i}\begin{pmatrix}\tilde{H}_i(t_{ik})\\ X_i(t_{ik})\\ \hat{B}_i(t_{ik})\end{pmatrix}\frac{Y_i(t_{ik})}{\rho_i(t_{ik},\gamma,\delta)}\xi_i(t_{ik})\,dN_i(t_{ik}) \tag{3.10}$$
$$-\sum_{j=1}^{J}\sum_{i=1}^{n}\begin{pmatrix}\tilde{H}_i(s_j)\\ X_i(s_j)\\ \hat{B}_i(s_j)\end{pmatrix}\frac{1}{\rho_i(s_j,\gamma,\delta)}\text{expit}\{\varphi'\tilde{H}_i(s_j) + \beta' X_i(s_j) + \theta'\hat{B}_i(s_j)\}\xi_i(s_j)m_i\frac{d\pi(s_j,Z_i)}{\pi(C_i,Z_i)} = 0.$$

γ and $\lambda_0(t)$ can be estimated by $\hat{\gamma}$ from (3.3) and $\hat{\Lambda}(t) = \sum_{i=1}^{n} \int_0^t dN_i(s) \Big/ \sum_{j=1}^{n} \xi_j(s) \exp\{\gamma' Z_j(s)\}$. The number of equations represented by $\tilde{H}_i(\cdot)$ reflects K_n, the number of knots selected. $\hat{\mu}(\cdot) = \hat{\varphi}' G(\cdot)$ estimates the non-parametric portion of the outcome model and $\hat{\beta}$ estimates the parametric portion of the outcome model, while $\hat{\theta}$ incorporates the effect of the visit process into the outcome model. Standard error estimation for our proposed estimation procedure can be obtained using a cluster bootstrap, in which subjects are sampled with replacement (Field and Welsh, 2007). Bootstrapping ensures that uncertainty from estimating μ and θ are accounted for in standard error estimate for $\hat{\beta}$.

3.3.3. Parameter interpretation

The inclusion of the fixed covariate $B_i(t)$ in (3.8) ensures that β retains a marginal interpretation. Consider a model with a binary treatment indicator X_{i1} and a confounder X_{i2}:

$$\Pr[Y_i(t) = 1 \mid X_i, \hat{B}_i] = \text{expit}\{\mu(t) + \beta_1 X_{i1} + \beta_2 X_{i2} + \theta \hat{B}_i\}, \tag{3.11}$$

such that $\hat{B}_i = (\hat{\eta}_{i2} - 1) Q_i(t)$ and β_1 is the parameter of interest. We examine four possible configurations of (3.11):

(i) $\Pr[Y_i(t) = 1 \mid X_i] = \text{expit}\{\mu(t) + \beta_1 X_{i1} + \beta_2 X_{i2}\}$;

(ii) $\Pr[Y_i(t) = 1 \mid X_i, \hat{B}_i] = \text{expit}\{\mu(t) + \beta_1 X_{i1} + \beta_2 X_{i2} + \theta(\hat{\eta}_{i2} - 1)\}$;

(iii) $\Pr[Y_i(t) = 1 \mid X_i, \hat{B}_i] = \text{expit}\{\mu(t) + \beta_1 X_{i1} + \beta_2 X_{i2} + \theta(\hat{\eta}_{i2} - 1) X_{i2}\}$;

(iv) $\Pr[Y_i(t) = 1 \mid X_i, \hat{B}_i] = \text{expit}\{\mu(t) + \beta_1 X_{i1} + \beta_2 X_{i2} + \theta(\hat{\eta}_{i2} - 1) X_{i1}\}$;

In (i), corresponding to the outcome models in the (M1)-only or (M1) and (M2)-only case, β_1 represents the difference in log odds of the response between two populations of treated and untreated individuals, regardless of their visit propensity. In (ii) and (iii), β_1 represents the difference in log odds of the response between two populations of treated and untreated individuals with the same value of X_{i2} and visit propensity. In (iv), the interpretation of β_1 is similar to the interpretation of the main effect in the presence of an interaction. The log odds for each treatment group can be expressed as:

$$\text{logit}\ \Pr[Y_i(t) = 1 \mid X_{i1} = 1, X_{i2}] = \beta_0 + \beta_1 + \beta_2 X_{i2} + \theta(\hat{\eta}_{i2} - 1);$$

$$\text{logit}\ \Pr[Y_i(t) = 1 \mid X_{i1} = 0, X_{i2}] = \beta_0 + \beta_2 X_{i2}.$$

Thus the comparison of the treatment groups results in the coefficient $\beta_1 + \theta(\hat{\eta}_{i2} - 1)$. Therefore, β_1 represents the difference in log odds of the response between two populations of treated and untreated individuals with the same value of X_{i2} and an average visit propensity (i.e., $\eta_{i2} = 0$).

3.4. Simulation study

We conducted simulation studies to evaluate the statistical properties of our proposed method under two outcome-observation dependence settings: (i) (M2) and (ii) (M2) and (M3). All simulations were conducted in R 2.13.1 (R Development Core Team, Vienna, Austria). For all simulations, we generated 1000 simulated datasets, each with $n = 100$ or 200 independent subjects. For comparison, we fit a GEE with a working independence correlation structure (IEE). The IEE should be used in the presence of time-dependent covariates unless a key assumption can be verified (French and Heagerty, 2009; Pepe and Anderson, 1994). The IEE fitted here represents the most basic model without any covariance assumption while providing valid marginal coefficients. We also fit an IEE that incorporated observation-level weights $\rho_i(t; \gamma, \delta)$ and $\hat{B}_i(t)$ as a covariate (weighted-IEE). All outcome models used B-splines with four degrees of freedom to approximate $\mu(t)$.

3.4.1. Setting 1: Simulations under (M2)

Parameters

In setting 1, we used covariates to induce correlation between the outcome and observation-time processes to satisfy (M2). We specified the outcome model as:

$$\Pr[Y_i(t) = 1 \mid X_i(t)] = \text{expit}\{\mu(t) + \beta_1 X_{i1}(t) + \beta_2 X_{i2}\}, \tag{3.12}$$

for which $\mu(t) = -1 + 0.5t^{-1/2}$, $\epsilon_i(t) \sim$ Normal(0,1), and (β_1, β_2) were the parameters of interest. The time-dependent covariate of interest $X_{i1}(t)$ took the form $X_{i1}\log(t)$, in which $X_{i1} \sim$ Uniform[0,1], and $X_{i2} \sim$ Bernoulli(0.5).

Following the simulation procedure of Bůžková and Lumley (2009) based on a probit link approximation, we generated binary outcomes based on the following equation:

$$Y_i(t) = I\Big[f^*(t) + \beta_1^* X_{i1}(t) + \beta_2^* X_{i2} + \beta_3 X_{i3} + \phi_i + \epsilon_i(t) > 0\Big], \tag{3.13}$$

for which $f^*(t) = \mu(t)M - \beta_3 \mathrm{E}[X_{i3} \mid X_{i1}]$, $\beta_1^* = \beta_1 M$, $\beta_2^* = \beta_2 M$, and $M = \sqrt{\sigma_\epsilon^2 + \sigma_\phi^2 + \beta_3^2 \mathrm{Var}[X_{i3} \mid X_{i1}]}/1.7$. We included an additional covariate X_{i3} drawn from a mixture distribution, for which $X_{i3} \sim$ Normal(2,1) if $X_{i1} \leq 0.5$ and $X_{i3} \sim$ Normal(0,4) if $X_{i1} > 0.5$. The parameter ϕ_i was a subject-specific latent variable that induced an exchangeable correlation structure on the outcomes from the same subject. We assumed ϕ_i was normally distributed with mean 0 and variance $\sigma_\phi^2 = 0.25$.

Model (3.13) describes the case when X_{i3} affects the covariate-outcome association by $X_{i1}(t)$. Proper marginalization over the additional covariate X_{i3}, the random effect, and the error term in (3.13) results in the marginal semi-parametric outcome model (3.12).

We generated observation times T_{ik} from a non-homogeneous Poisson process with intensity function $\lambda_i(t) = \eta_i \lambda_0(t) \exp\{\gamma_1 X_{i1}(t) + \gamma_2 X_{i2} + \gamma_3 X_{i3}\}$, in which $\lambda_0(t) = \frac{\sqrt{t}}{2}$. Note that X_{i3} induced additional correlation between the outcome and observation-time processes, and X_{i3} was specified in the observation-time model but not in the marginal outcome model (3.12). The latent variable η_i was generated from a Gamma distribution with mean 1 and variance $\sigma_\eta^2 = 0.5$. The independent censoring time C_i was generated from Uniform[5,10]. To examine the performance of our proposed estimators under (M2), we considered various combinations of outcome parameters $\beta_1 = \log(1.5)$, $\beta_2 = \log(1.2)$, $\beta_3 = \{0, \log(0.5)\}$ and intensity parameters $\gamma_1 = 0.3$, $\gamma_2 = 0.2$, $\gamma_3 = (0, 0.2, 0.3)$. When $\gamma_3 = 0$, the outcome-observation dependence model satisfied (M1); when $\beta_3 \neq 0$ and $\gamma_3 \neq 0$, the outcome-observation dependence model satisfied (M2).

Results

Table 3.1 provides the estimated bias, empirical standard error estimates, and mean squared error estimates for estimation of β_1 in model (3.12) by the IEE, weighted-IEE, and our proposed method. If (M1) was satisfied ($\gamma_3 = 0$), i.e., the outcome and observation-time processes were conditionally independent given outcome-model covariates $X_{i1}(t)$ and X_{i2}, then all three methods performed well. Biases in the estimates of β_1 were negligible. However, if (M1) was violated ($\gamma_3 \neq 0$ and $\beta_3 \neq 0$), i.e., the two processes had additional correlation induced by X_{i3}, then the bias under the proposed method was smaller than the bias under IEE. IEE estimates β_1 without accounting for

Table 3.1: Simulation results for $\beta_1 = \log(1.5)$: Bias, $\hat{\beta}_1 - \beta_1$; ESE, empirical sample error; MSE, mean squared error

β_3	n	γ_3	IEE			Weighted-IEE			Proposed method		
			Bias	ESE	MSE	Bias	ESE	MSE	Bias	ESE	MSE
0	100	0	−0.02	0.25	0.06	−0.02	0.24	0.06	−0.01	0.26	0.07
		0.2	−0.04	0.24	0.06	−0.05	0.23	0.05	−0.04	0.24	0.06
		0.3	−0.04	0.24	0.06	−0.05	0.24	0.06	−0.05	0.27	0.08
	200	0	−0.02	0.18	0.03	−0.02	0.17	0.03	−0.02	0.18	0.03
		0.2	−0.04	0.18	0.03	−0.05	0.17	0.03	−0.05	0.18	0.03
		0.3	−0.04	0.17	0.03	−0.05	0.16	0.03	−0.05	0.20	0.04
$\log(0.5)$	100	0	0.01	0.38	0.14	0.01	0.37	0.14	0.02	0.38	0.14
		0.2	−0.32	0.38	0.25	−0.06	0.36	0.13	−0.06	0.37	0.14
		0.3	−0.42	0.39	0.33	−0.04	0.36	0.13	−0.04	0.39	0.15
	200	0	−0.02	0.25	0.06	−0.01	0.25	0.06	−0.01	0.25	0.06
		0.2	−0.33	0.26	0.18	−0.07	0.25	0.07	−0.07	0.26	0.07
		0.3	−0.45	0.27	0.28	−0.06	0.25	0.07	−0.06	0.28	0.08

* All outcome models were fitted with B-splines with 4 degrees of freedom.

the additional covariate X_{i3} in any manner, whereas the proposed method incorporates the effect of X_{i3} through observation-level weights. Thus, IEE provided biased estimates. The performance of the weighted-IEE was comparable to the proposed method; both methods provided comparable bias and mean squared errors of the covariate effects.

Because X_{i2} was independent of X_{i3}, the biases for β_2 were negligible under all three methods for all scenarios. This indicated that when the additional covariate is independent of an outcome-model covariate, the performance of IEE is comparable to the weighted-IEE and proposed methods.

3.4.2. Setting 2: Simulation under (M2) and (M3)

Parameters

In setting 1, we focused on (M2). In setting 2, we examined the performance of our proposed estimator under (M3) when (M2) was satisfied. Following (3.6), the model of interest for binary outcomes in the presence of a latent variable representing visit propensity was:

$$\mathrm{P}[Y_i(t) = 1 \mid X_i(t)] = \mathsf{expit}\{\mu(t) + \beta_1 X_{i1}(t) + \beta_2 X_{i2} + \eta_{i1} Q_i(t)\}, \tag{3.14}$$

in which $\mu(t)$, $\epsilon_i(t)$, $X_{i1}(t)$ and X_{i2} were as defined in Setting 1, and (β_1, β_2) were the parameters of interest. We introduced an additional covariate X_{i3}, defined as the mixture distribution as in Setting 1, which affected the covariate-outcome association of $X_{i1}(t)$ through (M2). Extending (3.14), we generated data under both (M2) and (M3) with the following equation:

$$Y_i(t) = I\bigg[f^*(t) + \beta_1^* X_{i1}(t) + \beta_2^* X_{i2} + \beta_3 X_{i3} + \eta_{i1}^* Q_i + \epsilon_i(t) > 0\bigg], \tag{3.15}$$

for which $f^*(t)$, β_1^* and β_2^* were as defined in Setting 1. With the inclusion of η_{i1}, we defined $\eta_{i1}^* = \eta_{i1} M$ and $M = \sqrt{\sigma_\epsilon^2 + Q_i^2\sigma_\phi^2 + \beta_2^2 \text{var}[X_{i2} \mid X_{i1}]}/1.7$.

The observation times T_{ik} were generated from a non-homogeneous Poisson process with intensity function $\lambda_i(t) = \eta_{i2}\lambda_0(t)\exp\{\gamma_1 X_{i1}(t) + \gamma_2 X_{i2} + \gamma_3 X_{i3}\}$, with $\lambda_0(t) = \frac{\sqrt{t}}{2}$. The independent censoring time C_i was generated from Uniform[5,10]. The latent variable η_{i1} was defined as $\text{E}[\eta_{i1} \mid \eta_{i2}] = \theta(\eta_{i2} - 1) + \phi_i$, for which $\phi_i \sim$ Normal$(0, \sigma_\phi^2)$ and $\sigma_\phi^2 = 1$.

We generated the latent variable η_{i2} in the observation-time model under two scenarios:

1. η_{i2} from Gamma distribution with mean 1 and variance 0.5; hereby $\eta_{i2}^{(1)}$.

2. η_{i2} from a mixture distribution, following Uniform[0.5,1.5] if $X_{i2} = 1$ and Gamma distribution with mean 1 and variance 0.7 if $X_{i2} = 0$; hereby $\eta_{i2}^{(2)}$.

$\eta_{i2}^{(2)}$ would imply a covariate-dependent latent variable.

The coefficients were defined as $(\beta_1, \beta_2, \beta_3) = \log(1.5, 1.2, 0.5)$, $(\gamma_1, \gamma_2, \gamma_3) = (0.3, 0.2, 0.3)$, and $\theta = 1$. $\theta \neq 0$ in model (3.14) introduced correlation between the outcome and the observation-time processes through latent variables.

We let $Q_i = 1$ or $Q_i = X_{i1}$. When $Q_i = 1$, the effect of the latent variable η_{i1} was not modified by any covariates in the outcomes process. When $Q_i = X_{i1}$, the effect of the latent variable η_{i1} was modified by the value of X_{i1}. By varying $Q_i = (1, X_{i1})$ and $\eta_{i2} = (\eta_{i2}^{(1)}, \eta_{i2}^{(2)})$, we considered different ways the latent variables induced a relationship between the outcome and observation-time processes.

Results

Table 3.2 provides the estimated bias, empirical standard error estimates, and mean squared error

estimates for the estimation of β_1 and β_2 in model (3.14). The inclusion of X_{i3} in the observation-time model satisfied (M2) and induced additional correlation and biases the covariate-outcome association of $X_{i1}(t)$. The inclusion of η_{i1} in the outcome model satisfied (M3).

We focus on the performance of the methods under various combinations of η_{i2} ($\eta_{i2}^{(1)}$ or $\eta_{i2}^{(2)}$) and Q_i (1 or X_i). If η_{i1} was unrelated to any of the outcome-model covariates, (i.e., $\eta_{i2}^{(1)}$ and $Q_i = 1$), then all three methods performed well for β_2, while IEE provided heavily biased estimates of β_1, consistent with the results from Setting 1. Under $\eta_{i2}^{(2)}$, in which the distribution of η_{i2}, and hence value of η_{i1}, depended on the status of X_{i2}, IEE provided biased estimates of β_2, while the weighted-IEE and the proposed method provided unbiased estimates. If the effect of η_{i1} was modified by the value of X_{i1} (i.e., $Q_i = X_{i1}$), then the bias for β_1 under the weighted-IEE and proposed method was smaller than IEE. Under the (M1) or (M2)-only assumption ($\theta = 0$, $\eta_{i1} = 0$), the proposed method is expected to be less efficient than IEE because the estimation procedure attempts to estimate θ, which results in loss of efficiency.

3.4.3. Summary

The preceding simulation results quantified the potential for bias in estimated covariate-outcome associations under various outcome-observation dependence mechanisms. Under (M1), all three methods performed well. Under (M2), only the weighted-IEE and our proposed method performed well. Under (M3) when (M2) was satisfied, both the weighted-IEE and our proposed method performed well in the presence of a latent variable representing visit propensity in the outcome model, especially when the latent variables were associated with outcome-model covariates either (i) if the distribution of the latent variables was covariate-dependent, or (ii) the effect of the latent variable was modified by an outcome-model covariate. In all simulations, the weighted-IEE and the proposed method were the most reliable and provided estimates with negligible biases under any combination of outcome-observation dependence mechanisms.

3.5. Application

3.5.1. Background

In this section, we apply our proposed joint model approach to data from a randomized controlled trial among patients on warfarin therapy. The goal of the trial was to determine the effectiveness of

Table 3.2: Simulation results for $\beta_1 = \log(1.5)$, $\beta_2 = \log(1.2)$, $\theta = 1$: Bias, $\hat{\beta} - \beta$; ESE, empirical sample error; MSE, mean squared error

n	η_2	Q_i		**IEE**			**Weighted-IEE**			**Proposed method**		
				Bias	ESE	MSE	Bias	ESE	MSE	Bias	ESE	MSE
100	$\eta_2^{(1)}$	1	β_1	-0.40	0.47	0.38	-0.05	0.42	0.18	-0.05	0.44	0.20
			β_2	-0.01	0.46	0.21	0.03	0.44	0.19	0.03	0.44	0.20
		X_{i1}	β_1	-0.11	0.41	0.18	-0.05	0.34	0.12	-0.05	0.39	0.15
			β_2	0.00	0.38	0.15	0.02	0.34	0.12	0.02	0.35	0.12
	$\eta_2^{(2)}$	1	β_1	-0.42	0.43	0.36	-0.08	0.39	0.16	-0.08	0.41	0.17
			β_2	-0.39	0.41	0.32	-0.09	0.38	0.15	-0.09	0.39	0.16
		X_{i1}	β_1	-0.24	0.38	0.20	-0.09	0.33	0.12	-0.09	0.36	0.14
			β_2	-0.19	0.34	0.15	-0.03	0.32	0.10	-0.03	0.33	0.11
200	$\eta_2^{(1)}$	1	β_1	-0.42	0.33	0.29	-0.06	0.30	0.09	-0.05	0.30	0.10
			β_2	-0.02	0.31	0.10	0.02	0.30	0.09	0.02	0.30	0.09
		X_{i1}	β_1	-0.12	0.29	0.10	-0.06	0.24	0.06	-0.06	0.26	0.07
			β_2	0.00	0.27	0.07	0.02	0.24	0.06	0.02	0.24	0.06
	$\eta_2^{(2)}$	1	β_1	-0.42	0.30	0.27	-0.07	0.27	0.08	-0.07	0.27	0.08
			β_2	-0.40	0.29	0.25	-0.08	0.26	0.08	-0.08	0.26	0.08
		X_{i1}	β_1	-0.24	0.27	0.13	-0.08	0.23	0.06	-0.08	0.24	0.06
			β_2	-0.20	0.24	0.10	-0.02	0.22	0.05	-0.02	0.22	0.05

* All outcome models were fitted with B-splines with 4 degrees of freedom.

[a] Latent variable distributions: $\eta_{i2}^{(1)} : \eta_{i2} \sim$ Gamma(mean=$1, \sigma^2 = 0.5$);
$\eta_{i2}^{(2)} : \eta_{i2} \sim I[X_2 = 1]$Uniform$[0.5, 1.5] + I[X_2 = 0]$Gamma$(1, 0.5)$

interventions designed to increase adherence to therapy, and thus improve anticoagulation control (Kimmel et al., 2007). The study randomized 362 subjects into four treatment arms, which we are unable to reveal in this preliminary analysis. The study protocol specified monthly follow-up visits, at which INR was measured. Physicians also scheduled as-needed visits in between protocol-required visits based on the patient's INR response.

The outcome $Y_i(t)$ in the outcome model was binary: 1 if the INR was outside the therapeutic range (out-of-range) at time t, and 0 otherwise. The primary exposure was treatment assignment. Descriptive analyses (data not shown) revealed several baseline covariates that were imbalanced across the four treatment groups $(P < 0.2)$, and were thus adjusted for in the outcome model: employment status (working, disabled, or retired/unemployed), baseline age, race, Medicare insurance, education, history of diabetes, target INR range, and sub-therapeutic INR at baseline.

We considered two outcome models:

Model 1: $\Pr[Y_i(t) = 1 \mid X_i, B_i]$ = expit$\{\mu(t) + \beta' X_i + \theta B_i\}$

Model 2: $\Pr[Y_i(t) = 1 \mid X_i, B_i]$ = expit$\{\mu(t) + \beta' X_i + \theta_1 B_{i,\text{disabled}} + \theta_2 B_{i,\text{retired/unemployed}}\}$.

Recall that B_i includes the latent variable from the observation-time model, and B_{i,Q_i}=$(\hat{\eta}_{i2} - 1)Q_i$. Model 1 assumed that the effect of the latent variable was not modified by any of the outcome-model covariates. Model 2 assumed that the effect of subject-specific latent variables was different based on employment status.

The observation-time model was defined as: $\mathrm{E}[d\Lambda_i(t) \mid Z_i(t)] = \eta_{i2} \exp\{\gamma' Z_i(t)\} d\lambda_0(t)$, in which $Z_i(t)$ included whether the INR was out-of-range at the previous visit. Thus, the observation-time model mirrored the clinical management of patients with suboptimal anticoagulation status who required additional follow-up. Outcome-model covariates were also screened for inclusion in the observation-time model. Univariable recurrent event models were used to assess unadjusted covariate associations with the observation times.

Censoring time C_i was defined as the time of the last follow-up visit at the study site. To estimate 95% confidence intervals (CI), we performed a cluster bootstrap in which subjects were sampled with replacement. The sampling procedure was repeated 1000 times and the 95% bootstrap CI was obtained from the 2.5^{th} and 97.5^{th} percentile of the empirical distribution produced from these 1000

Table 3.3: Parameter estimates and 95% CI of γ from the observation-time model

	$\hat{\gamma}$ (95% CI)
Employment Status	
Working	–
Disabled	$0.13\,(\,0.04, 0.29)$
Retired/Unemployed	$0.09\,(\,0.00, 0.27)$
Out-of-range INR at previous visit	$0.40\,(\,0.37, 0.49)$

estimates of β. The β estimates and 95% CI from the outcome models were exponentiated to obtain the odds ratios and corresponding 95% CI. Odds ratios less than 1 indicated decreased odds of out-of-range INR. For comparison, we fit a GEE with a working independence correlation structure (IEE) and an IEE that incorporated observation-level weights $\rho_i(t;\hat{\gamma},\hat{\delta})$ and $\hat{B}_i(t)$ as a covariate (weighted-IEE). All outcome models used B-splines with 4 degrees of freedom to approximate $\mu(t)$.

3.5.2. Results

The estimates of γ from the observation-time model indicated that employment status was significantly associated with the observation times (Table 3.3). Patients who were disabled or retired/unemployed were more likely to have a visit compared to patients who were working. The median number of visits for those in the 'working' group was 6 (range, 3–11), while the median number of visits in the 'disabled' and 'retired/unemployed' groups were 8 (range, 2–16) and 7 (range, 1–24), respectively. Employment status may be a proxy for other factors such as access to care and availability of time for physician visits. Patients were also significantly more likely to have a visit if the INR was out-of-range at the previous visit [$\hat{\gamma}$, 0.40; 95% CI:(0.37, 0.49)]. We found no significant interaction between employment status and out-of-range INR. The observation-level weights applied to the estimation procedures had a median of 1.21 and ranged from 0.88 to 1.71.

The odds ratios (ORs) and 95% CIs from the outcome models are presented in Table 3.4. Under Model 1, the estimate of B_i was positive, implying that the outcome and observation-time processes were positively associated, such that patients with greater odds of being out-of-range (i.e., poorer anticoagulation status) had more frequent visits. With both observation-level weights and a latent variable, the OR estimates from weighted-IEE and the proposed method shifted toward the null for those disabled and retired/unemployed. Here, β for each employment status represents the difference in log odds of an out-of-range INR between populations of 'disabled' or 'retired/unemployed'

individuals and 'employed' individuals with the same visit propensity.

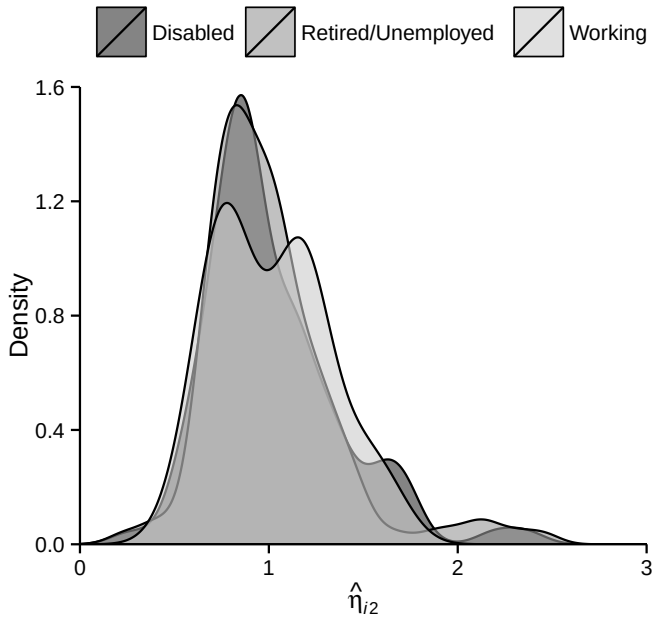


Figure 3.2: The empirical distribution of η_{i2} by employment status.

Model 2 investigated whether the latent variable was associated with employment status. Based on the empirical distribution of η_{i2} by employment status (Figure 3.2), there was little evidence that the distribution of η_{i2} from the observation-time model differed by employment status. The OR estimates for those in the 'disabled' group were similar between Model 1 and 2, but the estimates for those in the 'retired/unemployed' group further attenuated, even though the confidence intervals were wide. Both the 'disabled' and the 'retired/unemployed' groups had more frequent visits compared to the working group, hence the incorporation of the observation-level weights and effects of latent variables based on employment status adjusted for potential outcome-observation dependence. Here, β for each employment status represents the difference in log odds of an out-of-range INR between populations of 'disabled' or 'retired/unemployed' individuals and 'employed' individuals, all with average visit propensity.

3.6. Discussion

In this chapter, we presented a new approach to analyze longitudinal binary outcomes in the presence of outcome-dependent observation times. We introduced three mechanisms to describe the

Table 3.4: Odds ratios (OR) and 95% confidence intervals (CI) for out-of-range INR

			Model 1[1]				**Model 2**[2]			
	IEE		**Weighted-IEE**		**Proposed method**		**Weighted-IEE**		**Proposed method**	
	OR (95% CI)	$P^{\dagger}$	OR (95% CI)	$P^{\dagger}$	OR (95% CI)	$P^{\dagger}$	OR (95% CI)	$P^{\dagger}$	OR (95% CI)	$P^{\dagger}$
Employment Status		0.20		0.31		0.30		0.37		0.36
Working	–		–		–		–		–	
Disabled	$1.34\,(1.00, 2.12)$		$1.27\,(0.92, 1.98)$		$1.28\,(0.93, 2.00)$		$1.27\,(0.91, 1.95)$		$1.28\,(0.92, 2.03)$	
Retired/Unemployed	$1.22\,(0.93, 2.20)$		$1.16\,(0.85, 1.97)$		$1.17\,(0.84, 2.02)$		$1.11\,(0.80, 1.86)$		$1.12\,(0.80, 1.89)$	
θ	–		$1.17\,(0.98, 1.73)$		$1.15\,(0.95, 1.80)$		–		–	
θ_1	–		–		–		$1.00\,(0.58, 1.62)$		$0.98\,(0.57, 1.56)$	
θ_2	–		–		–		$1.29\,(1.05, 2.49)$		$1.26\,(1.02, 2.70)$	

* All outcome models were fitted with B-splines with 4 degrees of freedom.

† P-value corresponds to 2 degrees of freedom multivariate Wald test for β_{disabled}=$\beta_{\text{retired/unemployed}}$=0.

Outcome models:

[1] Model 1: $\mathrm{P}[Y_i(t) = 1 \mid X_i]$ = expit$\{\mu(t) + \beta' X_i + \theta B_i\}$

[2] Model 2: $\mathrm{P}[Y_i(t) = 1 \mid X_i]$ = expit$\{\mu(t) + \beta' X_i + \theta_1 B_{i,\text{disabled}} + \theta_2 B_{i,\text{retired/unemployed}}\}$.

dependence between the outcome and observation-time processes, and showed that our proposed method is applicable under any combination of the mechanisms. Our proposed method performed as well as the weighted-IEE that incorporates observation-level weights and latent variables. Both methods performed better than the naïve IEE when the dependence between outcomes and observation times is parameterized using observation-time model covariates and/or latent variables.

The advantage of our proposed method over the weighted-IEE would be apparent in the case of more complicated data-collection schedules, such as when the censoring times are not independent of the outcome and observation-time processes. In addition, our proposed method allows explicit specification of separate models for the outcome and the observation-time processes. Although not our primary target of inference, the parameters in the observation-time model provide relevant information to clinicians regarding the timing of care provided to patients. The ability of our proposed method to explicitly specify the secondary model for the observation-time process can be extended to accommodate more complex data-collection schedules.

Several key features of our approach are worth noting. First, we applied our proposed method to the analysis of binary outcomes, but our approach can be extended to other types of outcomes given an appropriate link function, such as the generalized logit link for a multinomial outcome. Second, we modeled the effect of time with B-splines instead of assuming a parametric structure. The potential gain in computational ease from the smooth spline approximation of $\mu(t)$ is countered by the potential loss in efficiency of estimation of the parameters of interest. Third, the validity of the proposed estimator is contingent upon correct specification of the observation-time model. One could utilize a Wald test for the importance of the additional covariates in $Z_i(t)$ to guide model building. Fourth, the model is able to accommodate censoring times that are dependent on the outcome and observation-time processes by estimating γ following the procedure in Huang, Qin, and Wang (2010). Finally, we note that in our application, patients were nested within physicians who made scheduling decisions based on the patient's anticoagulation status. Therefore, in addition to unmeasured patient characteristics, we may need to account for physician-level characteristics. Incorporating multiple sources of correlation, such as in a multi-level model, warrants future research.

CHAPTER 4

EXTENSION TO DISCONTINUOUS RISK INTERVALS

4.1. Introduction

In Chapters 2 and 3, subjects were considered at risk for a visit at any time until the time of censoring, i.e., the occurrence of a visit did not preclude the possibility of the next visit immediately thereafter. However, it is common in medical studies to encounter longitudinal data with observation gaps, during which subjects are considered not at risk for an event or physician visit; these gaps result in discontinuous risk intervals. Discontinuous risk intervals are a common feature of recurrent-event data, or more specifically, recurrent-episode data, such as recurrent infections and hospitalizations (Guo, Gill, and Allore, 2008; Kim, 2014; Zhao and Sun, 2006).

This chapter is motivated by a randomized trial of malaria vaccines conducted in Mali (Section 1.2.1). In this study, subjects were randomized to one of two vaccine arms, and the goal was to assess the impact of both vaccines on parasite and hemoglobin levels (Sagara et al., 2009). This study posed two main analysis issues. First, there was potential dependence between the outcome and observation-time processes. Subjects were scheduled for monthly follow-up visits according to the protocol. However, as-needed visits were also common between prescheduled visits, due to side effects or negative episodes such as clinical symptoms of malaria (Figure 4.1). Possible dependence between the longitudinal outcome and observation-time processes may require methods developed in Chapters 2 and 3 when evaluating covariate-outcome associations. The second analysis issue was the presence of discontinuous risk intervals, which is the main focus of this chapter. Upon confirmed diagnosis of clinical malaria during any visit, the patient received malaria treatment and was considered not at risk for a new malaria episode until 28 days after the first day of treatment. The patient was also not at risk for another physician visit during this 28-day period; we refer to this as an observation gap.

One possible approach to accommodate discontinuous risk intervals is to dichotomize the longitudinal outcomes and focus only on the outcomes that pass a certain threshold (i.e., episodes) and proceed with incidence rate analysis or recurrent episode analysis. To incorporate discontinuous risk intervals into the incidence rate (i.e., the number of events over total person time in study), the

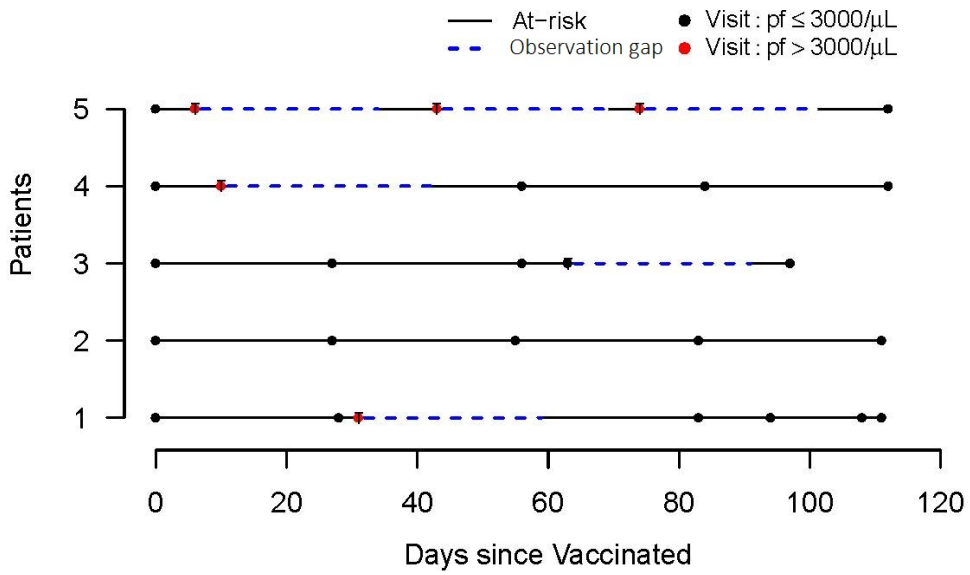


Figure 4.1: Observation times for four selected patients in the malaria vaccine study and the corresponding observed outcomes. A red dot indicates that parasite level $> 3000/\mu$L recorded at that visit, and the 28-day observation gaps are indicated by dashed blue lines.

person's time off the study (observation gaps) is removed from the total time in study (Guo, Gill, and Allore, 2008). However, this method does not easily accommodate covariate adjustments and does not account for the intensity of the episodes. To account for the intensity, several authors proposed methods within the recurrent episode framework by specifying when the subjects are not in the observation gaps. Hu et al. (2011) proposed a modified Cox regression model with time-dependent stratification and an adjusted risk set to accommodate non-negligible episode duration. Kim (2014) proposed a full likelihood approach to accommodate recurrent-episode data in which observation gaps are observed incompletely. However, the recurrent episode approach requires binary outcomes and ignores outcome measurements collected between episodes.

The topic of modeling longitudinal data with informative or outcome-dependent observation times has gained substantial interest in recent years. Lin, Scharfstein, and Rosenheck (2004) and Sun et al. (2005) proposed semi-parametric approaches; Lipsitz, Fitzmaurice, and Ibrahim (2002) considered a full likelihood approach for continuous outcomes while Fitzmaurice et al. (2006) proposed a pseudolikelihood model for binary outcomes. However, none of these methods can accommodate

discontinuous risk intervals.

Recently, Zhu et al. (2013) proposed a joint-likelihood approach motivated by the analysis of medical costs related to hospitalizations. The method addresses both the possible correlation between the outcome and observation-time process and the hospitalization duration. Their full likelihood approach can only accommodate continuous outcomes, and in their data example the medical cost data only exist when the hospitalization occurs, i.e., the data constitute a recurrent marked point process (French and Heagerty, 2009). In contrast, our malaria data consist of both binary and continuous outcomes, and the longitudinal outcomes exist regardless of the observation-time process.

We present an extension of the generalized model from Chapter 3 for the analysis of data with dependence between the outcome and observation-time processes in the presence of discontinuous risk intervals. In Section 4.2, we introduce the modified at-risk definition and the proposed estimation procedure for regression modeling in the presence of both outcome-dependent observation times and discontinuous risk intervals. We present simulation studies to evaluate the performance of our proposed procedure under various scenarios of discontinuous risk intervals in Section 4.3, and illustrate its application to data from a malaria vaccine trial in Section 4.4. Section 4.5 provides discussion and concluding remarks.

4.2. Model formulation

We consider a longitudinal study with *n* independent subjects in the study interval $[0, \tau]$, for which τ is the maximum study duration. For subject i, $i = 1, \ldots, n$, let $Y_i(t)$ denote an outcome of interest at time t, and $X_i(t)$ denote a $p \times 1$ vector of possibly time-dependent covariates. Unless otherwise specified, we consider only external covariates, such that any time-dependent covariate process at time *t* is conditionally independent of all previous outcomes, given the history of the covariate process (Kalbfleisch and Prentice, 2002). $Y_i(\cdot)$ is measured at m_i observation times $0 \leq T_{i1} < T_{i2} < \cdots < T_{im_i} \leq \tau$, for which m_i denotes the number of follow-up measurements on the i^{th} individual. Using counting process notation, let $N_i(t) = \sum_{s \leq t} dN_i(s)$ denote the number of observations on the i^{th} subject by time $t \leq C_i$, in which C_i is the censoring time. The indicator variable $dN_i(t)$ equals 1 if a follow-up visit occurred on the i^{th} individual at time t and equals 0 otherwise. We assume non-informative censoring, such that $\mathrm{E}[Y_i(t) \mid X_i(t), C_i \geq t] = \mathrm{E}[Y_i(t) \mid$

$X_i(t)]$. That is, the covariate-outcome associations are the same in those who are censored at C_i as those who are still in the study at C_i.

4.2.1. 'At risk' intervals

The complication of discontinuous risk intervals can be addressed by careful consideration of when individuals are at risk of a follow-up visit. We assume visits and outcomes occur and are recorded in continuous time over the study interval $[0, \tau]$. Let $I(\cdot)$ be an indicator function such that $I(A) = 1$ if A is true and $I(A) = 0$ otherwise. Without any observation gaps, the patients are always at risk for a visit prior to the time of censoring, i.e., $\xi_i(t) = I(0 \leq t \leq C_i)$. In the presence of discontinuous risk intervals, we specify $\Delta_i(t) = 1$ if an individual is under observation (i.e., not in an observation gap) and therefore at risk of a visit at t, and $\Delta_i(t) = 0$ otherwise. The 'at risk' indicator $\Delta_i(t)$ denotes which individuals provide information about visit occurrence at a given time and induces an adjusted risk set that includes only those not in an observation gap and who have not been censored at t.

4.2.2. Semi-parametric outcome model

As in previous chapters, we assume that primary scientific interest lies in a semi-parametric regression model for the longitudinal outcomes. We consider the generalized semi-parametric regression model for longitudinal outcomes $Y_i(t)$ under independent or dependent observation times (Lin and Ying, 2001):

$$\mathrm{E}[Y_i(t) \mid X_i(t)] = \mathrm{g}\{\mu(t) + \beta' X_i(t)\}, \tag{4.1}$$

for which $\mu(t)$ is an arbitrary function of time and β is a $p \times 1$ vector of regression parameters of interest. The function $\mathrm{g}(\cdot)$ links the expected outcome to the linear predictors; $\mathrm{g}(\cdot)$ is the identity function for continuous outcomes and the expit function for binary outcomes. The parameter β in model (4.1) remains the primary target of inference: β represents the marginal association between a set of covariates and an outcome of interest among a population of individuals.

4.2.3. Observation-time model

The observation-time process describes the timing and intensity of follow-up visits and is characterized by a standard recurrent events model. We introduce a non-negative latent variable η_i with mean 1 and unknown variance σ^2. Given observation-time model covariates $Z_i(t)$ and η_i, the recurrent event process $N_i(\cdot)$ is a non-homogeneous Poisson process with intensity function (Lin et al., 2000; Pepe and Cai, 1993):

$$\lambda_i(t) = \Delta_i(t)\eta_i\lambda_0(t)\exp\{\gamma' Z_i(t)\}, \quad t \in [0, \tau] \tag{4.2}$$

for which γ is a vector of unknown parameters and $\lambda_0(t)$ is an arbitrary baseline intensity function with $\lambda_0(t) = \int_0^t \lambda(u)du$. The inclusion of $\Delta_i(t)$ accommodates discontinuous risk intervals. The incorporation of the observation-time process into estimation of β in a joint model facilitates reliable estimation under outcome-observation dependence, as defined in Section 1.1.4.

4.3. Estimation and inference

In this section, we detail an extension to the general method proposed in Chapter 3 to accommodate any combination of the three outcome-observation dependence mechanisms in the presence of discontinuous risk intervals.

4.3.1. Estimators

To allow outcome-observation dependence through observed covariates and unobserved latent variables, the outcome model (4.1) can be extended to:

$$\mathrm{E}[Y_i(t) \mid X_i(t)] = \mathrm{g}\{\mu(t) + \beta' X_i(t) + \eta'_{i1} Q_i(t)\}, \tag{4.3}$$

in which $Q_i(t)$ is a $q \times 1$ subvector of $X_i(t)$ and η_{i1} is a q-dimensional vector of subject-specific latent variables that represent subject-level propensity for visits (Liang, Lu, and Ying, 2009). The observation-time model can be expressed as:

$$\mathrm{E}[d\Lambda_i(t) \mid Z_i(t)] = \eta_{i2}\exp\{\gamma' Z_i(t)\}\, d\lambda_0(t), \tag{4.4}$$

in which η_{i2} is a mean-one, non-negative latent variable. The distribution of η_{i2} may depend on observed time-independent outcome-model covariates V_i with $\mathrm{E}[\eta_{i2} \mid V_i] = 1$. The latent variables from models (4.3) and (4.4) are assumed to be linearly linked through $\mathrm{E}[\eta_{i1} \mid \eta_{i2}] = \theta(\eta_{i2} - 1)$. The parameter θ describes the association between the outcome and observation-time processes. Thus, to ensure that β retains a marginal interpretation with the inclusion of the latent variable, we define $B_i(t) = \mathrm{E}[(\eta_{i2} - 1) \mid m_i, C_i]Q_i(t)$ as a fixed covariate that incorporates the subject-specific propensity for visit. The outcome model (4.3) can be re-expressed as:

$$\mathrm{E}[Y_i(t) \mid X_i(t), B_i(t)] = \mathrm{g}\{\mu(t) + \beta' X_i(t) + \theta' B_i(t)\}. \tag{4.5}$$

Next, we re-express the observation-time model. Let $\mathcal{Z}_i(t) = \{Z_i(s) : 0 \leq s < t\}$ denote the covariate history of Z_i up to t. Without discontinuous risk intervals, the event times ($t_{i1} < t_{i2} < \cdots < t_{im_i}$) of the i^{th} subject conditional on $\{C_i, m_i, \eta_{i2}, \mathcal{Z}(C_i)\}$ are order statistics of a set of independent and identically distributed random variables with the density function (Huang, Qin, and Wang, 2010):

$$\frac{\exp\{\gamma' Z_i(t)\} d\lambda_0(t)}{\int_0^{C_i} \exp\{\gamma' Z_i(s)\} d\Lambda(s)}, \qquad 0 \leq t \leq C_i.$$

In the presence of discontinuous risk intervals, we incorporate the at risk indicator $\Delta_i(t)$ in the density function:

$$\frac{\Delta_i(t) \exp\{\gamma' Z_i(t)\} d\lambda_0(t)}{\int_0^{C_i} \Delta_i(t) \exp\{\gamma' Z_i(s)\} d\Lambda(s)}, \qquad 0 \leq t \leq C_i.$$

Define $\pi(t; Z_i) = \int_0^t \Delta_i(s) \exp\{\gamma' Z_i(s)\} d\Lambda(s)$. It follows that:

$$\mathrm{E}[dN_i(t) \mid C_i, m_i, \eta_{i2}, \mathcal{Z}(C_i)] = \Delta_i(t) m_i \frac{d\pi(t; Z_i)}{\pi(C_i; Z_i)}.$$

Using both re-expressed outcome and observation-time models, we can formulate the zero-mean process as:

$$\begin{aligned} M_i(t; \beta, \theta, \gamma, \delta) = \int_0^t \frac{1}{\rho_i(s, \gamma, \delta)} \Big[& Y_i(s) \Delta_i(s)\, dN_i(s) \\ & - \mathrm{g}\{\mu(s) + \beta' X_i(s) + \theta' \hat{B}_i(s)\} \Delta_i(s) m_i \frac{d\pi(s, Z_i)}{\pi(C_i, Z_i)} \Big]. \end{aligned} \tag{4.6}$$

The estimator accommodates outcome-observation dependence through observed covariates (M2) and unobserved latent variables (M3) through observation-level weights $\rho_i(t, \gamma, \delta)$ and $B_i(t)$. In addition, the inclusion of $\Delta_i(t)$ properly accounts for discontinuous risk intervals.

To estimate $B_i(t)$, we first estimate η_{i2} from the observation-time model (4.4). We utilize the property that given $\{\eta_{i2}, C_i, \mathcal{Z}(C_i)\}$, m_i follows a Poisson distribution with mean $\eta_{i2}\pi(C_i, Z_i)$ to obtain $\hat{\eta}_{i2} = \{\frac{m_i}{\pi(C_i, Z_i)}\}$, so $\hat{B}_i(t) = \{\frac{m_i}{\pi(C_i, Z_i)} - 1\}Q_i(t)$. The observation-level weights $\rho_i(t, \gamma, \delta)$ standardize the observed data to the time-specific underlying population under the observation-time model. One particular observation-level weight with variance-stabilizing properties is:

$$\rho_i(t; \gamma, \delta) = \frac{\exp\{\gamma' Z_i(t)\}}{\exp\{\delta' X_i(t)\}},$$

for which δ is from the observation-time model (4.2) conditioning on $X_i(t)$. (4.6) is the most general formulation of the joint model and can accommodate (M1), (M2), and (M3); that is, the equation (4.6) provides valid estimation of β under any combination of the three conditional independence mechanisms in the presence of discontinuous risk intervals. In subsequent sections, we proceed with estimation of β via the estimation equation (4.6), which we refer to as the 'proposed estimator.'

4.3.2. Estimation procedure

If the censoring time is independent of the observation-time process, then the parameter γ from the observation-time model (4.2) can be consistently estimated by $\hat{\gamma}$ with the estimating function (Lin et al., 2000; Zhao and Sun, 2006):

$$U(\gamma) = \sum_{i=1}^{n} \int_0^\tau \{Z_i(t) - \bar{Z}(t; \gamma)\} dN_i(t), \tag{4.7}$$

for which:

$$\bar{Z}(t; \gamma) = \frac{\sum_{i=1}^n \Delta_i(t) \exp\{\gamma' Z_i(t)\} Z_i(t)}{\sum_{i=1}^n \Delta_i(t) \exp\{\gamma' Z_i(t)\}},$$

and $\Delta_i(t)$ is as defined in Section 4.2.3 in the presence of discontinuous risk intervals.

To create the estimating equation for β in $M_i(t; \beta, \theta, \gamma, \delta)$, we impose a flexible structure on $\mu(t)$ using basis approximations. Generalizing the notation from Huang and Liu (2007), suppose the smooth function $\mu(\cdot)$ can be approximated by a spline function such that $\mu(t) \approx \sum_{k=1}^{K_n} \varphi_k G_k(t) =$

$\varphi' G(t)$ in which $\{G_k(\cdot), k = 1, \ldots, K_n\}$ is a basis system of B-splines, $\varphi = (\tau_1, \ldots, \tau_{K_n})'$ and $G(t) = (G_1(t), \ldots, G_{K_n}(t))'$. Let $\tilde{H}_i(t) = G_i(t)$ or $\tilde{H}_{ij} = G(T_{ij})$. The process (4.5) can thus be approximated by $\mathrm{E}[Y_i(t) \mid X_i(t), B_i(t)] = \mathrm{g}\{\varphi \tilde{H}_i(t) + \beta' X_i(t) + \theta' B_i(t)\}$. Let $s_1 < s_2 < \cdots < s_J$ denote the J distinct ordered observation times from all subjects $\{t_{ik}, i = 1, \ldots, n; k = 1, \ldots, m_i\}$. We propose to estimate β, φ, and θ from (4.6) by the estimating equation:

$$\sum_{i=1}^{n} \sum_{k=1}^{m_i} \begin{pmatrix} \tilde{H}_i(t_{ik}) \\ X_i(t_{ik}) \\ \hat{B}_i(t_{ik}) \end{pmatrix} \frac{Y_i(t_{ik})}{\rho_i(t_{ik}, \gamma, \delta)} \Delta_i(t_{ik})\, dN_i(t_{ik}) \tag{4.8}$$

$$- \sum_{j=1}^{J} \sum_{i=1}^{n} \begin{pmatrix} \tilde{H}_i(s_j) \\ X_i(s_j) \\ \hat{B}_i(s_j) \end{pmatrix} \frac{1}{\rho_i(s_j, \gamma, \delta)} \mathrm{g}\{\varphi' \tilde{H}_i(s_j) + \beta' X_i(s_j) + \theta' \hat{B}_i(s_j)\} \Delta_i(s_j) m_i \frac{d\pi(s_j, Z_i)}{\pi(C_i, Z_i)} = 0.$$

We estimate γ by (4.7) and

$$\hat{\Lambda}(t) = \sum_{i=1}^{n} \int_0^t \frac{\Delta_i(s) dN_i(s)}{\sum_{j=1}^{n} \Delta_j(s) \exp\{\gamma' Z_j(s)\}}.$$

4.4. Simulation study

We conducted simulation studies to evaluate the statistical properties of our proposed method when observation gaps occur after $Y_i(t) = 1$. All simulations were conducted in R 2.13.1 (R Development Core Team, Vienna, Austria). For all simulations, we generated 1000 simulated datasets, each with $n = 200$ independent subjects. For comparison, we fit a GEE with a working independence correlation structure (IEE). We also fit two weighted-IEEs that incorporated observation-level weights $\rho_i(t; \gamma, \delta)$ and $\hat{B}_i(t)$ as a covariate, one in which $\rho_i(t; \gamma, \delta)$ and $\hat{B}_i(t)$ were calculated from the observation-time model without incorporation of observation gaps (unadjusted risk set), and the other from the observation-time model with proper at-risk indicators (adjusted risk set). All outcome models used B-splines with three degrees of freedom to approximate $\mu(t)$.

Parameters

We simulated data according to the structure of the malaria trial, in which an observation gap with

fixed duration occurred after every malaria episode, i.e., after every occurrence of $Y_i(t) = 1$. We examined the performance of our proposed method in the presence of discontinuous risk intervals when both (M2) and (M3) were satisfied. Following (4.3), the model of interest for binary outcomes in the presence of a latent variable representing visit propensity was:

$$\Pr[Y_i(t) = 1 \mid X_i(t)] = \text{expit}\{\mu(t) + \beta_1 X_{i1} + \eta_{i1}\}, \tag{4.9}$$

for which $\mu(t) = -0.5 - 0.1t^{-1/2}$, $\epsilon_i(t) \sim$ Normal(0,1), and β_1 was the parameter of interest. The time-independent exposure variable $X_{i1} \sim$ Bernoulli(0.5).

Following the simulation procedure of Bůžková and Lumley (2009) based on a probit link approximation, we generated binary outcomes under both (M2) and (M3) with the following equation:

$$Y_i(t) = I\Big[f^*(t) + \beta_1^* X_{i1}(t) + \beta_2 X_{i2} + \eta_{i1}^* + \phi_i + \epsilon_i(t) > 0\Big], \tag{4.10}$$

for which $f^*(t) = \mu(t)M - \beta_2 \mathrm{E}[X_{i2} \mid X_{i1}]$, $\beta_1^* = \beta_1 M$. We included an additional covariate X_{i2} drawn from a mixture distribution, for which $X_{i2} \sim$ Normal(0.5,1) if $X_{i1} = 1$ and $X_{i2} \sim$ Normal(1,0.5) if $X_{i1} = 0$. We defined $\eta_{i1}^* = \eta_{i1} M$ and $M = \sqrt{\sigma_\epsilon^2 + \sigma_\phi^2 + \beta_2^2 \mathrm{var}[X_{i2} \mid X_{i1}]}/1.7$. The parameter ϕ_i was a subject-specific latent variable that induced an exchangeable correlation structure on the outcomes from the same subject. We assumed ϕ_i was normally distributed with mean 0 and variance $\sigma_\phi^2 = 0.25$.

Model (4.10) described the case when X_{i2} affects the covariate-outcome association between X_{i1} and $Y_i(t)$. Proper marginalization over the additional covariate X_{i2}, the random effect, and the error term in (4.10) resulted in the marginal semi-parametric outcome model (4.9).

We generated observation times T_{ij} from a non-homogeneous Poisson process with intensity function:

$$\lambda_i(t) = \eta_{i2} \lambda_0(t) \exp\{\gamma_1 X_{i1} + \gamma_2 X_{i2}\}, \tag{4.11}$$

in which $\lambda_0(t) = \frac{\sqrt{t}}{40}$. Note that X_{i2} induced additional correlation between the outcome and observation-time processes, and X_{i2} was specified in the observation-time model but not in the marginal outcome model (4.9).

We generated the covariate-dependent latent variable η_{i2} in the observation-time model from a mixture distribution, following Uniform[0.5,1.5] if $X_{i1} = 1$ and Gamma distribution with mean 1 and variance 0.7 if $X_{i1} = 0$. The covariate-dependent latent variable induces (M3). The latent variable η_{i1} was defined as $\mathrm{E}[\eta_{i1} \mid \eta_{i2}] = \theta(\eta_{i2} - 1)$. The independent censoring time C_i was $\tau = 113$ as was the end of study time in the malaria example.

For subject i, given the $(j-1)^{\text{th}}$ observation-gap end-time $L_{i,j-1}$ with $V_{i,0} = 0$, we generated observation times and corresponding binary outcomes in the presence of observation gaps as follows:

Step 1. Given X_{i1}, X_{i2} and η_{i2}, generate the subject's time to j^{th} observation time since $V_{i,j-1}$, denoted by T_{ij}, according to the intensity function (4.11).

Step 2. Generate the outcome at T_{ij}, denoted by Y_{ij}, using the outcome model (4.10).

Step 3. If $Y_{ij} = 1$, set U_{ij} as the observation gap duration, otherwise set $U_{ij} = 0$.

Step 4. Define the j^{th} observation-gap end-time as $L_{ij} = T_{ij} + U_{ij}$.

The coefficients were defined as $(\beta_1, \beta_2, \theta) = \log(2, 0.75, 0.5)$, $(\gamma_1, \gamma_2) = (0.2, 0.4)$. In Step 3 of the procedure above we consider both a fixed duration $\{U = (0, 7, 14, 21, 28)\}$ or a variable duration $U \sim$ Uniform[7,28].

Results

Table 4.1 provides the estimated bias, empirical standard error estimates, and mean squared error estimates for the estimation of β_1 in model (4.9). When there were no discontinuous risk intervals, i.e., $U = 0$, then the weighted-IEE (with and without adjusted risk set) and the proposed method (with and without adjusted risk set) performed well. Biases in the estimate of β_1 were negligible. The bias under the IEE was larger because it was not able to address (M2) and (M3). With non-zero observation gap durations, we observed that the bias under the proposed method with adjusted risk set, i.e., proper use of $\Delta_i(t)$ instead of $\xi_i(t)$, was smaller than the proposed method with unadjusted risk set. The bias was much smaller under all cases in which $U > 0$. The performance of the weighted-IEE was comparable to the proposed method, in that the weighted-IEE with adjusted risk set yielded smaller bias than the weighted-IEE with unadjusted risk set; both methods provided comparable bias and mean squared error of the covariate effect.

We observed similar conclusions when the duration of the observation gaps varied according to Uniform[7,28]. The proposed method with adjusted risk set yielded smaller bias than the proposed method with unadjusted risk set.

Table 4.1: Simulation results for $\beta_1 = \log(2)$ when fixed observation gaps occur only after $Y_i(t) = 1$: Bias, $\hat{\beta}_1 - \beta_1$; ESE, empirical sample error; MSE, mean squared error

				Weighted-IEE		Proposed method	
N	$U^{\dagger}$		IEE	Unadjusted RS*	Adjusted RS**	Unadjusted RS*	Adjusted RS**
200	0	Bias	0.374	−0.001	−0.001	−0.001	−0.001
		ESE	0.173	0.157	0.157	0.157	0.157
		MSE	0.170	0.025	0.025	0.025	0.025
	7	Bias	0.371	−0.039	−0.016	−0.041	−0.017
		ESE	0.194	0.184	0.171	0.185	0.171
		MSE	0.175	0.035	0.029	0.036	0.029
	14	Bias	0.384	0.021	−0.003	0.014	−0.004
		ESE	0.199	0.194	0.178	0.196	0.178
		MSE	0.187	0.038	0.032	0.038	0.032
	21	Bias	0.380	0.052	−0.001	0.043	−0.002
		ESE	0.208	0.212	0.186	0.215	0.186
		MSE	0.188	0.048	0.034	0.048	0.034
	28	Bias	0.402	0.092	0.026	0.083	0.025
		ESE	0.217	0.219	0.198	0.223	0.198
		MSE	0.209	0.056	0.040	0.057	0.040
	[7, 28]	Bias	0.383	0.043	−0.021	0.035	−0.023
		ESE	0.195	0.201	0.180	0.206	0.180
		MSE	0.184	0.042	0.033	0.044	0.033

All outcome models were fitted with B-splines with 3 degrees of freedom.
* With unadjusted risk set, i.e., does not account for observation gaps
** With adjusted risk set via $\Delta_i(t)$
† Duration of observation gaps, either fixed $(U_{ij} = 0, 7, 14, 21, 28)$ or variable ($U_{ij} \sim$ Uniform[7,28]).

4.5. Application

4.5.1. Background

In this section, we apply our proposed model to data from a randomized controlled phase-II trial among healthy children 2–3 years old living in or near the village of Bancoumana, Mali. The study was designed to assess the safety, immunogenicity, and biologic impact of a malaria vaccine candidate AMA1-C1 (Sagara et al., 2009). The study randomized 289 subjects into two arms: the vaccine candidate AMA1-C1 $(n = 139)$ and the active control $(n = 140)$. Both vaccines were administered on Days 0 and 28. Parasitologic follow up began 14 days after the second vaccination on Study Day 42 and censoring time C_i was defined as the end of the parasitologic follow-up period

(Day 154).

The study protocol specified monthly follow-up visits, during which blood samples were obtained for malaria smears and hemoglobin count. Physicians also scheduled as-needed visits in between protocol-required visits based on the patient's clinical presentation. If the patient was determined to have a malaria episode during the visit, the patient would be given malaria treatment. Malaria vaccines that target the blood stage of infection, such as those used in the trial, are not intended to prevent infection by parasites such as *P. falciparum*. Instead, they are expected to reduce parasite density, thereby reducing morbidity and mortality due to severe malaria. The *P. falciparum* parasite density was calculated using individual white blood cell counts from blood smears. The primary outcome $Y_i^1(t)$ was binary: 1 if *P. falciparum* parasite $> 3000/\mu$L at t, and 0 otherwise. We considered the primary outcome model: $\Pr[Y_i^1(t) = 1 \mid X_i, B_i]$ = expit$\{\mu(t) + \beta' X_i + \theta B_i\}$. The primary exposure was vaccine arm. Because children were enrolled in two cohorts (May-June 2006 and July-August 2006), we included the cohort indicator as a covariate in the outcome model. Recall that B_i included the latent variable from the observation-time model and represented the subject-level propensity for visit. The outcome model assumed that the effect of the latent variable was not modified by any of the outcome-model covariates.

We also considered the impact of vaccination on hemoglobin (Hb) level. The secondary outcome $Y_i^2(t)$ is the continuous measurement of hemoglobin. Hemoglobin < 8.5 g/dL often suggests anemia. We considered the outcome model: $\mathrm{E}[Y_i^2(t) \mid X_i, B_i]$ = $\{\mu(t) + \beta' X_i + \theta B_i\}$.

The observation-time model was defined as: $\mathrm{E}[d\Lambda_i(t) \mid Z_i(t)] = \eta_{i2} \exp\{\gamma' Z_i(t)\} d\lambda_0(t)$ for both the primary and secondary outcomes. The set of covariates $Z_i(t)$ included whether the parasite count was $> 3000/\mu$L at the previous visit. Thus, the observation-time model mirrored the clinical management of patients with history of malaria symptoms or episodes and required additional follow-up. Other baseline characteristics were also screened for inclusion in the observation-time model. Univariable recurrent event models were used to assess unadjusted covariate associations with the observation times. The time at risk was the period of histologic follow-up (Day 42–154) minus 28 days after each treatment for malaria, constituting discontinuous risk intervals. Hence, the at-risk indicator $\Delta_i(t)$ was 1 only when the subject was not in the 28-day observation gap and was not censored at t.

Table 4.2: Parameter estimates and 95% CI of γ from the observation-time model

	Unadjusted RS*	Adjusted RS**
	$\hat{\gamma}$ (95% CI)	$\hat{\gamma}$ (95% CI)
Vaccine Arm		
Active control	–	–
Vaccine candidate	$-0.02\,(-0.08, 0.03)$	$0.01\,(-0.05, 0.06)$
Cohort		
A	–	–
B	$-0.06\,(-0.12, 0.00)$	$-0.03\,(-0.10, 0.05)$
Parasite $> 3000/\mu$L at previous visit	$0.19\,(\ 0.12, 0.26)$	$0.13\,(\ 0.05, 0.22)$

* With unadjusted risk set, i.e., does not account for observation gaps
** With adjusted risk set via $\Delta_i(t)$

To estimate 95% confidence intervals (CI), we performed a cluster bootstrap in which subjects were sampled with replacement. The sampling procedure was repeated 1000 times and the 95% bootstrap CI was obtained from the 2.5^{th} and 97.5^{th} percentile of the empirical distribution produced from these 1000 estimates of β. For the primary outcome of parasite $> 3000/\mu$L, the β estimates and 95% CI from the outcome models were exponentiated to obtain the odds ratios and corresponding 95% CI. Odds ratios less than 1 indicated decreased odds of *P. falciparum* $> 3000/\mu$L. For comparison, we fit a GEE with a working independence correlation structure (IEE) and two weighted-IEEs that incorporated observation-level weights $\rho_i(t; \hat{\gamma}, \hat{\delta})$ and $\hat{B}_i(t)$ as a covariate calculated from an observation-time model either with or without proper specification of the at risk intervals. All outcome models used B-splines with 4 degrees of freedom to approximate $\mu(t)$.

4.5.2. Results

The estimates of γ from the observation-time model indicated that neither the vaccine arm or cohort was significantly associated with the observation times (Table 4.2). However, patients were significantly more likely to have a visit if the parasite level was $> 3000/\mu$L at the previous visit: the estimate of γ under the observation-time model with unadjusted risk set was higher than the estimate with adjusted risk set [$\hat{\gamma}$, 0.19; 95% CI:(0.12, 0.26) versus $\hat{\gamma}$, 0.13; 95% CI:(0.05, 0.22)]. The observation-level weights calculated from the observation-time model with adjusted risk set had a median of 1.00 and ranged from 0.87 to 1.00.

Based on the empirical distribution of η_{i2} by vaccine arm and cohort (Figure 4.2), there was little evidence that the distribution of η_{i2} from the observation-time model differed by vaccine arm or cohort. We retained θB_i in the outcome model to investigate the relationship between the propensity

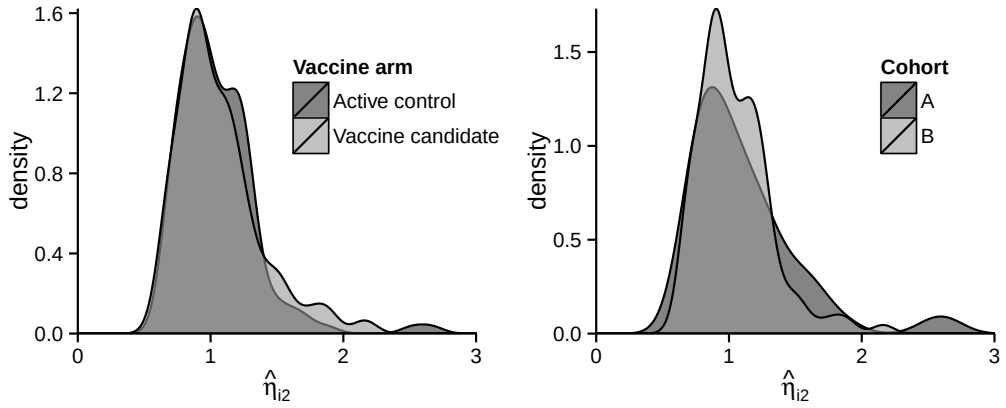


Figure 4.2: The empirical distribution of η_{i2} by arm and cohort.

of visit and the outcome.

The odds ratios (ORs) and 95% CIs from the outcome models for the primary outcome of parasite $> 3000/\mu$L are presented in Table 4.3. We first focus on the results under naïve IEE and the proposed method with unadjusted risk sets. The estimate of the difference between the log odds of parasite level $> 3000/\mu$L in the two arms was smaller under IEE than the proposed method that incorporated both observation-level weights and a latent variable representing patient-level visit propensity. Under the proposed method, the β for each covariate represents the difference in log odds of the parasite level $> 3000/\mu$L between populations of individuals given 'vaccine candidate' and individuals given 'active control' with the same visit propensity.

Next, we compare the results under the proposed method with unadjusted and adjusted risk sets. In the proposed method with adjusted risk set, the accommodation of the 28-day observation gaps through the modified at-risk indicators drew the estimates further away from the null, even though the confidence intervals were wide [OR, 0.93; 95% CI: (0.73, 1.20)]. The estimate of θ was positive, implying that the outcome and observation-time processes were positively associated, such that patients with greater odds of parasite $> 3000/\mu$L had more frequent visits. By addressing the discontinuous risk intervals, we prevented underestimation of the effect of the vaccine candidate.

In the original analysis, Sagara et al. (2009) compared the incidence rates between the two vaccine arms. The incidence rate was defined as the rate of parasite $> 3000/\mu$L/day at risk, in which time at risk was the period of parasitologic follow up minus 28 days after each treatment for malaria.

They found that the median rate of parasite $> 3000/\mu$L/day at risk was 0.016 in the vaccine candidate group and 0.014 in the active control group (Hodges-Lehmann rate ratio (vaccine arm in the denominator)=1.02, p=0.67). Our results agreed with the conclusion of the original analysis in that there were no significant differences between the vaccine arms, the direction of the effect was reversed in our analysis.

The β estimates and 95% CIs for the secondary outcome of continuous hemoglobin levels are presented in Table 4.4. The results under the proposed method with adjusted risk set indicated that the population of individuals on the vaccine candidate had lower hemoglobin levels than the population on the active control, although this difference was not significant. As was the case for the primary outcome, the effect size for the vaccine arm for the secondary outcome under IEE was slightly greater than the proposed method with adjusted risk set. The observation-level weights applied to both the primary and secondary outcomes were close to 1 and there was little evidence of covariate-dependent propensity of visit; therefore there were only minor differences between IEE and the weighted-IEE or proposed methods.

4.6. Discussion

In this chapter, we presented an extension of the comprehensive model from Chapter 3 to analyze longitudinal outcomes in the presence of both outcome-dependent observation times and discontinuous risk intervals. There is a tradeoff between longer blackout periods and total time at risk. Naïve IEE analysis methods may miscategorize a patient who experienced multiple lengthy observation gaps with low risk or odds of a negative outcome if they do not account for the blackout period. We introduced our proposed method with a modified at-risk indicator to address observation gaps during the study when patients are not at risk for a visit or measured on any outcomes. We showed that our proposed method is applicable under any combination of outcome-observation dependence mechanisms in the presence of discontinuous risk intervals. In simulations, our proposed method performed as well as the weighted-IEE that incorporated observation-level weights and latent variables calculated using an observation-time model with adjusted risk sets. Both methods performed better than the naïve IEE and weighted-IEE or proposed method with unadjusted risk sets that assumed all subjects were at risk at all times before time of censoring.

Several key features of our approach are worth noting. First, we assume that the duration of obser-

Table 4.3: Odds ratios (OR) and 95% confidence intervals (CI) for parasite $> 3000\mu$/L

		Weighted-IEE		Proposed method	
	IEE	Unadjusted RS*	Adjusted RS**	Unadjusted RS*	Adjusted RS**
	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
Vaccine arm					
Active control	–	–	–	–	–
Vaccine candidate	$0.99\,(\,0.77, 1.25)$	$0.97\,(\,0.74, 1.26)$	$0.95\,(\,0.76, 1.23)$	$0.96\,(\,0.75, 1.25)$	$0.93\,(\,0.73, 1.20)$
Cohort					
A	–	–	–	–	–
B	$0.72\,(\,0.53, 1.25)$	$0.74\,(\,0.54, 1.14)$	$0.77\,(\,0.55, 1.14)$	$0.77\,(\,0.55, 1.07)$	$0.80\,(\,0.58, 1.13)$
θ	–	$0.06\,(\,0.03, 0.11)$	$3.57\,(\,2.53, 5.52)$	$0.07\,(\,0.05, 0.11)$	$4.02\,(\,2.85, 6.30)$

* With unadjusted risk set, i.e., does not account for observation gaps

** With adjusted risk set via $\Delta_i(t)$

Table 4.4: Estimated β and 95% confidence intervals (CI) for continuous hemoglobin level

		Weighted-IEE		Proposed method	
	IEE	Unjusted RS *	Adjusted RS **	Undjusted RS *	Adjusted RS **
	$\hat{\beta}\,(95\%\text{CI})$	$\hat{\beta}\,(95\%\text{CI})$	$\hat{\beta}\,(95\%\text{CI})$	$\hat{\beta}\,(95\%\text{CI})$	$\hat{\beta}\,(95\%\text{CI})$
Vaccine arm					
Active control	–	–	–	–	–
Vaccine candidate	$-0.18\,(-0.42, 0.06)$	$-0.18\,(-0.42, 0.06)$	$-0.17\,(-0.42, 0.06)$	$-0.18\,(-0.43, 0.06)$	$-0.20\,(-0.44, 0.05)$
Cohort					
A	–	–	–	–	–
B	$0.08\,(-0.24, 0.39)$	$0.05\,(-0.27, 0.36)$	$0.06\,(-0.27, 0.38)$	$0.07\,(-0.26, 0.39)$	$0.08\,(-0.24, 0.41)$
θ	–	$0.68\,(\,0.15, 1.18)$	$-0.25\,(-0.65, 0.19)$	$0.78\,(\,0.19, 1.27)$	$-0.21\,(-0.60, 0.18)$

* With unadjusted risk set, i.e., does not account for observation gaps

** With adjusted risk set via $\Delta_i(t)$

vation gaps is predetermined. However, one can envision cases when the duration of observation gaps may be dependent on covariate values or subject-level characteristics, e.g., certain treatment cycles may take longer to complete or two subjects may respond differently to the same treatment. In addition, the censoring may also be related to the longitudinal outcomes, such as a terminal event of death. Joint model approaches to accommodate informative or incomplete observation gaps and informative censoring warrant future research. We can also consider modifications to the proposed method to jointly model multiple outcomes such as the parasite and hemoglobin levels in the malaria vaccine example.

Next, we recognize that there may be situations in which an observation gap follows after every visit, such as the hospitalization cost data introduced earlier (Zhu et al., 2013). The duration of the observation gap may also depend on the covariate values or the outcomes. We can consider alternative specification of the observation-level weights, such as $\rho_i(t) = \exp\{\gamma X_i(t)\}\eta_{i2}$. Consideration of more general form of non-standard data-collection schedules and the target of inference on the population warrants future research.

Lastly, the observation-time process can be modeled on two time scales: total time scale (i.e., time-to-events model) and gap time scale (i.e., time-between-events model). The proposed method in this paper adopts the total time scale, but it may be of interest to consider alternative parameterizations. Gap time analysis, sometimes referred to as renewal process, is commonly adopted for recurrent episodic illness in which treatment is related to the duration of observation gaps or time between events. The gap time scale is attractive as the individual is considered renewed after each event. However, the gap time analysis approach may suffer from dependent censoring because longer gaps are more likely to be censored (Yan and Fine, 2008). Nevertheless, it may be of interest to investigate the utility of the gap time scale in the context of joint models.

CHAPTER 5

DISCUSSION

5.1. Summary

This dissertation examined statistical models for the analysis of longitudinal outcomes in the presence of outcome-dependent observation times. This topic has gained great interest in recent years, but there is a lack of a clear framework of the relationships between the outcome and observation-time processes. We proposed a framework of three potential outcome-observation dependence mechanisms and provided various model-checking procedures to guide the selection of appropriate analysis. We proposed a set of semi-parametric joint models based on estimating equations for continuous outcomes that can accommodate any combination of the outcome-observation dependence mechanisms.

For binary outcomes, we proposed a new semi-parametric method to estimate covariate-outcome associations under any outcome-observation dependence mechanism. In simulations, we showed that our method performs better than the naïve marginal longitudinal data analysis approach. We provided additional clarification of the interpretation of the estimated parameters from the outcome model. The comprehensive semi-parametric estimator can be applied to other types of longitudinal outcomes with the appropriate link functions.

Lastly, we extended the general semi-parametric estimator to accommodate the presence of discontinuous risk intervals. We showed that with an adjusted risk set, our method can properly account for observation gaps and other non-standard data-collection schedules, thereby greatly maximizing the utility of our proposed method.

5.2. Future directions

5.2.1. Prediction modeling

Beyond the standard analysis of longitudinal outcomes to derive outcome-covariate associations, the adoption of longitudinal data for predicting subsequent outcomes is regularly the focus of diagnostic studies. There is great value in developing dynamic models to be repeatedly applied to

a subject's longitudinal outcomes and visit profile to predict a subsequent outcome. Albert (2012) recently considered a linear mixed model for predicting a poor pregnancy outcome based on a series of ultrasound measurements collected over time. In our case, the outcomes are available at all times instead of at the end of the study period (i.e., not a marked point process). As future research, it would be useful to exploit the joint models in the previous chapters to estimate prediction rules for future outcomes and visit times. From the patient's perspective, a dynamic prediction model can provide insight into the risk of future poor outcomes and inform time to next follow-up visit. A dynamic model could be an asset at the institutional level to optimize resources, infrastructure and staffing.

5.2.2. Multi-level model

In the previous chapters, we assumed that patients were independent. However, patients may be nested within physicians if the visit schedules are determined by the physician. Different physicians have different attitudes and practices toward the frequency and intensity of follow-up visits. In such cases, we need to recognize the physician-level variability in the propensity of scheduling a visit for a particular patient. Furthermore, visit schedules may depend on hospital resources. In a multi-center study, we may need to consider hospital-level variability in the observation-time process. Therefore, in addition to unmeasured patient characteristics, we may need to account for physician-level and hospital-level characteristics. Incorporating multiple sources of correlation, such as in a multi-level model, warrants future research.

Therefore, in addition to unmeasured patient characteristics, we may need to account for physician-level characteristics. Incorporating multiple sources of correlation, such as in a multi-level model, warrants future research.

5.2.3. Software

Implementation and broader application of the methods introduced in the previous chapters are likely hampered by the lack of available general-purpose statistical software. The computation of the proposed estimators has been implemented using R and are provided in the appendices. We hope to create an R package to analyze longitudinal data with outcome-dependent observation times, along with help files and additional model-checking procedures.

APPENDIX A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

In this supplement, we provide theoretical results for the Weighted-Liang and Weighted-Sun methods. We include R code to reproduce estimates and standard errors from Chapter 2 Case Study Table 4 using publicly available data. The standard errors are calculated from 1000 cluster-bootstrap samples in which the subjects are the sampling units.

A.1. Theoretical results for extensions to the Liang method: Weighted-Liang

In this section, we provide the theoretical results for our extension to the Liang method to allow for time-dependent covariates in the observation-time model as well as to accommodate M2. We assume that the latent variable η_{i2} follows a Gamma distribution with mean 1, variance σ^2 and $\mathrm{E}(\eta_{i1} \mid \eta_{i2}) = \theta(\eta_{i2} - 1)$. Following the notation from Liang, Lu, and Ying (2009), we first define:

$$
\begin{aligned}
S_Z^{(k)}(t,\gamma) &= n^{-1}\sum_{i=1}^{n} \xi_i(t) Z_i^k(t) \exp\{\gamma' Z_i(t)\} \\
P_X^{(k)}(t,\delta,\Lambda) &= n^{-1}\sum_{i=1}^{n} \xi_i(t) \exp\{\delta' X_i(t)\} X_i^k \frac{m_i}{\pi(C_i;Z_i)} \\
P_B^{(1)}(t,\delta,\Lambda) &= n^{-1}\sum_{i=1}^{n} \xi_i(t) \exp\{\delta' X_i(t)\} B_i \frac{m_i}{\pi(C_i;Z_i)},
\end{aligned}
$$

in which $k = 0, 1$, and $k = 1$ indexes the presence of the covariate $X_i(t)$ or $Z_i(t)$. We let $s_Z^{(k)}(t)$, $p_X^{(k)}(t)$, $p_B^{(1)}(t)$, $\mu_Z(t)$, $\tilde{\mu}_X(t)$, and $\tilde{\mu}_B(t)$ be the asymptotic limit of $S_Z^{(k)}(t,\gamma)$, $P_X^{(k)}(t,\Lambda_0)$, $P_B^{(1)}(t,\Lambda)$, $\frac{S_Z^{(1)}(t,\gamma)}{S_Z^{(0)}(t,\gamma)}$, $\frac{P_X^{(1)}(t,\Lambda)}{P_X^{(0)}(t,\Lambda)}$, and $\frac{P_B^{(1)}(t,\Lambda)}{P_X^{(0)}(t,\Lambda)}$. Let $\phi = (\beta', \alpha')'$.
Furthermore, define the mean-zero processes as:

$$
M_i(t) = M_i(t,\beta,\theta,\mathcal{A},\Lambda,B_i) = \int_0^t \frac{1}{\rho_i(s;\gamma,\delta)} \left[\{Y_i(s) - \beta' X_i(s) - \theta' B_i(s)\} dN_i(s) - \xi_i(s) m_i \frac{d\mathcal{A}(s)}{\Lambda(C_i)} \right],
$$

and:

$$
M_i^*(t) = N_i(t) - \int_0^t \xi_i(u) \exp\{\gamma' X_i(t)\} d\Lambda(u)
$$

and the positive definite matrix:

$$A = E\left[\int_0^\tau \{Z_i - \mu_Z(t)\}^{\otimes 2}\xi_i(t)\exp\{\gamma' Z_i(t)\}d\Lambda(t)\right]$$

A.1.1. Asymptotic results for $\hat{\gamma}$

Lin et al. (2000) showed the consistency of $\hat{\gamma}$, which can be written as:

$$\sqrt{n}(\hat{\gamma} - \gamma) = A^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \{Z_i(t) - \mu_Z(t)\}dM_i^*(t) + o_p(1). \tag{A.1}$$

A.1.2. Asymptotic results for $\hat{\Lambda}(t)$

Additionally, $\sqrt{n}(\hat{\Lambda}(t) - \Lambda(t))$ is asymptotically equivalent to

$$\begin{aligned}\sqrt{n}\{\hat{\Lambda}(t) - \Lambda(t)\} = &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t \frac{dM_i^*(u)}{S_Z^{(0)}(u)} \\ &- \int_0^t \mu_Z'(u)d\Lambda(u)A^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \{Z_i(t) - \mu_Z(t)\}dM_i^*(t) + o_p(1).\end{aligned} \tag{A.2}$$

A.1.3. Asymptotic results for $\hat{\sigma}^2$

According to Liang, Lu, and Ying (2009),

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) = \frac{1}{\sqrt{n}}\frac{1}{T}\sum_{i=1}^{n}\{m_i^2 - m_i - T(\sigma^2 + 1)\} + o_p(1),$$

in which $T = \lim_{n\to\infty}(1/n)\sum_{i=1}^{n}\exp\{\gamma' Z_i(t)\}\Lambda(C_i)$.

A.1.4. Asymptotic properties of $(1/\sqrt{n})U_1(\phi, \hat{\Lambda}, \hat{B})$

Let

$$\begin{aligned}U_{11}(\phi, \hat{\Lambda}, \hat{B}) &= \sum_{i=1}^{n}\int_0^\tau \frac{1}{\rho_i(t;\hat{\gamma},\delta)}\{X_i(t) - \bar{X}(t)\}\{Y_i(t) - \beta' X_i(t) - \theta'\hat{B}_i(t)\}dN_i(t), \\ U_{12}(\phi, \hat{\Lambda}, \hat{B}) &= \sum_{i=1}^{n}\int_0^\tau \frac{1}{\rho_i(t;\hat{\gamma},\delta)}\{\hat{B}_i(t) - \bar{\hat{B}}(t)\}\{Y_i(t) - \beta' X_i(t) - \theta'\hat{B}_i(t)\}dN_i(t),\end{aligned}$$

in which $\bar{X}(t)$ and $\bar{\bar{B}}(t)$ are as previously defined. Expressions to derive the asymptotic properties of $(1/\sqrt{n})U_1(\phi, \hat{\Lambda}, \hat{B})$ largely follows that outlined in Liang, Lu, and Ying (2009), with the inclusion of the observation-level inverse weights $\frac{1}{\rho_i(t;\gamma,\delta)}$, changing V_i to $Z_i(t)$, and replacing the sub-functions (i.e., $S_Z^{(k)}(t,\gamma)$, $M_i(t,\beta,\theta,\mathcal{A},\Lambda,B_i)$) as we have described above.

A.2. Theoretical results for extensions to the Sun method: Weighted-Sun

In this section, we provide the theoretical results for our extension to the Sun method to accommodate M2 under the assumption that the observation-time process is conditionally independent of the censoring times, following the notation from Sun, Song, and Zhou (2011).
We first define:

$$
\begin{aligned}
S_z^{(k)}(t;\gamma) &= n^{-1}\sum_{i=1}^{n}\xi_i(t)\exp\{\gamma' Z_i(t)\}Z_i^k(t)\\
S^{(0)}(t;\delta) &= n^{-1}\sum_{i=1}^{n}\xi_i(t)\exp\{\delta' X_i(t)\}\ m_i/\pi(C_i;Z_i)\\
S_x^{(k)}(t;\delta) &= n^{-1}\sum_{i=1}^{n}\xi_i(t)\exp\{\delta' X_i(t)\}X_i^k(t)\ m_i/\pi(C_i;Z_i)\\
S_\eta^{(1)}(t;\delta) &= n^{-1}\sum_{i=1}^{n}\xi_i(t)\exp\{\delta' X_i(t)\}\hat{\eta}_i(t)\ m_i/\pi(C_i;Z_i)\\
S_{\eta x}^{(2)}(t;\delta) &= n^{-1}\sum_{i=1}^{n}\xi_i(t)\exp\{\delta' X_i(t)\}\hat{\eta}_i X_i(t)\ m_i/\pi(C_i;Z_i)
\end{aligned}
$$

Let $s_z^{(0)}(t)$, $s_z^{(1)}(t)$, $s^{(0)}(t)$, $s_x^{(1)}(t)$, $s_x^{(2)}(t)$, $s_\eta^{(1)}(t)$, $s_{\eta x}^{(2)}(t)$ and $\mu_Z(t)$ denote the limiting values of $S_z^{(0)}(t;\gamma_0)$, $S_z^{(1)}(t;\gamma_0)$, $S^{(0)}(t;\delta)$, $S_x^{(1)}(t;\delta)$, $S_x^{(2)}(t;\delta)$, $S_\eta^{(1)}(t;\delta)$, $S_{\eta x}^{(2)}(t;\delta)$ and $s_z^{(1)}(t)/s_z^{(0)}(t)$. Furthermore, let $\bar{x}(t) = s_x^{(1)}(t)/s^{(0)}(t)$ and $\bar{\eta}(t) = s_\eta^{(1)}(t)/s^{(0)}(t)$. Let A be as defined in the previous section.

If the observation-time process is conditionally independent of the censoring times, the asymptotic results of γ and Λ follows (A.1) and (A.2).

Let

$$U_1(\beta,\alpha;\hat{\gamma}) = \sum_{i=1}^{n}\int_0^{\tau}\frac{W(t)}{\rho_i(t;\hat{\gamma},\delta)}[\{X_i(t)-\bar{X}(t)\}\{Y_i(t)-\beta' X_i(t)-\alpha\hat{\eta}_i\}]dN_i(t)$$

$$U_2(\beta,\alpha;\hat{\gamma}) = \sum_{i=1}^{n}\int_0^{\tau}\frac{W(t)}{\rho_i(t;\hat{\gamma},\delta)}[\{\hat{\eta}_i-\bar{\eta}(t)\}\{Y_i(t)-\beta' X_i(t)\}-\alpha\{\hat{\Omega}_i-\hat{\eta}_i\bar{\eta}(t)\}]dN_i(t),$$

Define $U(\beta,\alpha;\hat{\gamma}) = (U_1(\beta,\alpha;\hat{\gamma})', U_1(\beta,\alpha;\hat{\gamma})')$. The asymptotic properties of $(1/\sqrt{n})U(\beta,\alpha;\hat{\gamma})$ follows from the expressions of Sun, Song, and Zhou (2011), except with the assumption of non-informative censoring, the inclusion of the observation-level inverse weights $\frac{1}{\rho_i(t;\gamma,\delta)}$, and replacing the sub-functions (i.e., $S_z^{(k)}(t;\gamma)$) as we have described above.

We develop the asymptotic results for A.1 and A.2 under the assumption that δ is a fixed value. In practice, δ is estimated by $\hat{\delta}$ using the data. Following Liang and Zeger (1986), the variability from estimating δ does not affect the asymptotic behavior of $\hat{\beta}$ using stabilized weights.

A.3. Relative efficiency

We examine the relative efficiency of weighted versus unweighted methods under the simulation Setting 1 in Chapter 2 to determine the potential loss of efficiency when methods include an additional covariate when it is not necessary.

A.4. R code

We provide R code to reproduce Table 4 from the case study in Chapter 2 and implement model-checking procedures.

```
##############################################
## Case Study Table 4: Bladder Data
## Additional Covariate:
## cumulative # of tumors since baseline
##############################################

## Load extension packages
library(Hmisc)
library(plyr)
library(zoo)
library(nleqslv)

## Load dataset
## The data is available from the delisted r package spef,
## available at http://cran.r-project.org/src/contrib/Archive/spef/
## The dataset: blaTum
library(spef); data(blaTum);
```

Table A.1: Simulation results for β_1 under (M2): Bias, $\hat{\beta}_1 - \beta_1$, $\beta_1 = 1$; ESE, empirical sample error; ERE, estimated relative efficiency

n	β_2	γ_2	**Lin**		**Bůžková**			**Liang** (extension)		**Weighted-Liang** (extension)			**Sun**		**Weighted-Sun** (extension)		
			Bias	ESE	Bias	ESE	ERE[a]	Bias	ESE	Bias	ESE	ERE[b]	Bias	ESE	Bias	ESE	ERE[c]
100	0	0	0.003	0.284	0.003	0.284	1.000	0.004	0.280	−0.006	0.393	1.970	0.004	0.282	−0.006	0.393	1.942
		−0.2	−0.001	0.292	−0.002	0.289	0.980	0.003	0.287	−0.012	0.428	2.224	0.003	0.288	−0.013	0.428	2.209
		0.5	−0.020	0.342	−0.009	0.289	0.714	−0.018	0.324	−0.007	0.379	1.368	−0.018	0.329	−0.008	0.379	1.327
	0.3	0	0.003	0.318	0.002	0.313	0.969	0.004	0.313	−0.008	0.411	1.724	0.004	0.315	−0.007	0.411	1.702
	1	0	0.003	0.543	−0.001	0.521	0.921	0.004	0.536	−0.010	0.577	1.159	0.004	0.539	−0.010	0.577	1.146
200	0	0	−0.004	0.203	−0.003	0.203	1.000	−0.002	0.202	−0.010	0.284	1.977	−0.003	0.202	−0.010	0.285	1.991
		−0.2	−0.004	0.213	−0.004	0.209	0.963	−0.001	0.210	−0.009	0.304	2.096	−0.001	0.211	−0.009	0.304	2.076
		0.5	0.001	0.258	0.004	0.202	0.613	−0.003	0.237	0.002	0.269	1.288	−0.001	0.243	0.001	0.269	1.225
	0.3	0	−0.007	0.227	−0.007	0.224	0.974	−0.006	0.225	−0.014	0.299	1.766	−0.006	0.226	−0.014	0.299	1.750
	1	0	−0.014	0.389	−0.015	0.374	0.924	−0.015	0.386	−0.023	0.421	1.190	−0.014	0.388	−0.023	0.421	1.177

Estimated relative efficiency was calculated for unbiased estimators with:
[a] the variance of the Lin parameter estimate in the denominator;
[b] the variance of the Liang parameter estimate in the denominator;
[c] the variance of the Sun parameter estimate in the denominator.

```
m <- NULL    # m = number of observations per subject
for (ii in unique(blaTum$id)){
  m[blaTum$id==ii] <- nrow(blaTum[blaTum$id==ii,])
}

#########################################################################
## X3 <- additional covariate in observation-time model: # of new tumors since baseline
prev_cumtumors <- ave(blaTum$count, blaTum$id, FUN = function(x) Lag(cumsum(x), shift=1))
prev_cumtumors[is.na(prev_cumtumors)]<-0

## outcome Y = cumulative number of new tumors since baseline
Ycumtumors <- ave(blaTum$count, blaTum$id, FUN = cumsum)
#########################################################################

#########################################################################
# Analysis dataset
# Q=Q(t) from Liang approach: Q(t) may be an outcome model covariate
# Here: Q=blaTum$treatment
#########################################################################
sim.data <- data.frame(ID=blaTum$id, t=blaTum$time, m=m, Y = log(Ycumtumors+1), X1=blaTum$treatment,
                       X2=log(blaTum$num+1), X3=log(prev_cumtumors+1), Q=blaTum$treatment, C=53) # Q=1

#/****** Select first record of each patient **/
baseData <- ddply(sim.data, .(ID), function(x) x[1, ])
meannumvisits <- mean(baseData$m)
udt <- unique(sort(sim.data$t[sim.data$t>0]))
N <- length(baseData$ID)

## Start function
DepObsTimes <- function(sim.data, baseData, udt){
  ###########################################################################
  #### expand time-varying covariates to full set of t=53 rows              ####
  #### Analyst may consider other methods such as imputation or closest neighbor
  ###########################################################################
  testdata  <- data.frame(ID=sim.data$ID, t=sim.data$t, X3=sim.data$X3)
  testdata2 <- reshape(testdata, timevar="t", idvar="ID", direction="wide", v.names="X3")
  testdata3 <- reshape(testdata2, idvar="ID",direction="long")
  testdata3 <- testdata3[order(testdata3$ID,testdata3$t),]
  testdata4 <- NULL
  for (i in unique(testdata3$ID)){testdata4 <- rbind(testdata4, na.locf(testdata3[testdata3$ID==i,]))}
  testdata4[is.na(testdata4)] <- 0

  ###################################################################
  # GAMMA.HAT (observation-time model covariates)
  ###################################################################
  f <- function(gamma){
    exp_gamma <- function(tt){
                 exp(gamma[1]*baseData$X1+gamma[2]*baseData$X2+gamma[3]*testdata4$X3[testdata4$t==tt])}
    numer1 <- sapply(sim.data$t, function(u){
                           sum( (baseData$X1*exp_gamma(u))[u<=baseData$C], na.rm=T) } )
    numer2 <- sapply(sim.data$t, function(u){
                           sum( (baseData$X2*exp_gamma(u))[u<=baseData$C], na.rm=T) } )
    numer3 <- sapply(sim.data$t, function(u){
                           sum( (testdata4$X3[testdata4$t==u]*exp_gamma(u))[u<=baseData$C], na.rm=T) } )
    denom  <- sapply(sim.data$t, function(u){
                           sum( (exp_gamma(u))[u<=baseData$C], na.rm=T) } )
    Vbar <- cbind(numer1/denom, numer2/denom, numer3/denom)

    bigV <-cbind(sim.data$X1, sim.data$X2, sim.data$X3)
    temp <- colSums((bigV-Vbar)/N, na.rm=T)
    temp
  }
  gamma <- c(0.5, 0.5, 0.5)
  gamma.hat <- nleqslv(gamma, f)$x

  ###################################################################
  # DELTA.HAT (outcome model covariates)
  ###################################################################
  f <- function(gamma){
    exp_delta <- exp(gamma[1]*baseData$X1+gamma[2]*baseData$X2)
    numer1 <- sapply(sim.data$t, function(u){sum( (baseData$X1*exp_delta)[u<=baseData$C], na.rm=T) } )
    numer2 <- sapply(sim.data$t, function(u){sum( (baseData$X2*exp_delta)[u<=baseData$C], na.rm=T) } )
    denom  <- sapply(sim.data$t, function(u){sum( (exp_delta)[u<=baseData$C], na.rm=T) } )
    Vbar   <- cbind(numer1/denom, numer2/denom)
    Vbar.long  <- t(sapply(sim.data$t, function(tt) Vbar[tt,]))

    bigV <-cbind(sim.data$X1, sim.data$X2)
```

```
  temp <- colSums((bigV-Vbar.long)/N, na.rm=T)
  temp
}
gamma <- c(0, 0)
delta.hat <- nleqslv(gamma, f)$x

#####################################################################
# Set-Up
#####################################################################
exp_gamma <- function(u){exp(gamma.hat[1]*baseData$X1+
                             gamma.hat[2]*baseData$X2+gamma.hat[3]*testdata4$X3[testdata4$t==u])}
denom_gamma <- sapply(udt, function(u){sum( (exp_gamma(u))[u<=baseData$C], na.rm=T) } )
exp_delta <- exp(delta.hat[1]*baseData$X1+delta.hat[2]*baseData$X2)
denom_delta <- sapply(udt, function(u){sum( (exp_delta)[u<=baseData$C], na.rm=T) } )

#/**** : estimated dLam(t) under X1+X2 in observation-time model ****/
estlam.t.delta <- sapply(1:length(udt), function(u) sum( ((sim.data$t==udt[u])/denom_delta[u])) )

#/**** : estimated dLam(t) under X1+X2+X3 in observation-time model ****/
estlam.t.gamma <- sapply(1:length(udt), function(u) sum( ((sim.data$t==udt[u])/denom_gamma[u])) )

#/**** : estimated Ybar_star (closest neighbor): Only used in LY and Buzkova ****/
Y_star <- function(t){
  sapply(baseData$ID, function(n){
    tail(sim.data$Y[sim.data$ID==n]
    [(abs(sim.data$t[sim.data$ID==n]-t)==min(abs(sim.data$t[sim.data$ID==n]-t)))],1) })
}

numer <- sapply(udt, function(u) sum((Y_star(u)*exp_delta)[ baseData$C >= u ]) )
Ybar_star <- (numer/denom_delta)
Ybar_starX <- (sapply(sim.data$t, function(tt) Ybar_star[tt]))

########################################
########################################
# LIN & YING METHOD
########################################
########################################
#/**** : estimated Xbar1 & Xbar2 ****/
denom  <- sapply(sim.data$t, function(u) sum(exp_delta[ baseData$C >= u  ]) )
numer1 <- sapply(sim.data$t, function(u) sum((baseData$X1*exp_delta)[ baseData$C >= u ]) )
numer2 <- sapply(sim.data$t, function(u) sum((baseData$X2*exp_delta)[ baseData$C >= u ]) )
Xbar <- cbind(numer1/denom, numer2/denom)
bigX <- as.matrix(cbind(sim.data$X1, sim.data$X2))

f <- function(beta){
  temp <- rep(0,length=ncol(bigX))
  temp[1] <- sum(((bigX[,1]-Xbar[,1])*(sim.data$Y-Ybar_starX-beta[1]*(bigX[,1]-Xbar[,1])
                             -beta[2]*(bigX[,2]-Xbar[,2]) )), na.rm=T)
  temp[2] <- sum(((bigX[,2]-Xbar[,2])*(sim.data$Y-Ybar_starX-beta[1]*(bigX[,1]-Xbar[,1])
                             -beta[2]*(bigX[,2]-Xbar[,2]) )), na.rm=T)
  temp
}
beta <- c(0,0)
LY.beta <- nleqslv(beta, f)$x

########################################
########################################
# BUZKOVA METHOD
########################################
########################################
##################/**** calculate weights (iirr2 = rho) ****/#############
Z <- cbind(sim.data$X1, sim.data$X2, sim.data$X3)
X <- cbind(sim.data$X1, sim.data$X2)
iirr2 <- exp(Z %*% as.matrix(gamma.hat))/exp(X %*% as.matrix(delta.hat))

bigX <- as.matrix(cbind(sim.data$X1, sim.data$X2))
f <- function(beta){
  temp <- rep(0,length=ncol(bigX))
  temp[1] <- sum( 1/iirr2*((bigX[,1]-Xbar[,1])*(sim.data$Y-Ybar_starX-beta[1]*
                                                  (bigX[,1]-Xbar[,1])-beta[2]*(bigX[,2]-Xbar[,2]) )), na.rm=T)
  temp[2] <- sum( 1/iirr2*((bigX[,2]-Xbar[,2])*(sim.data$Y-Ybar_starX-beta[1]*
                                                  (bigX[,1]-Xbar[,1])-beta[2]*(bigX[,2]-Xbar[,2]) )), na.rm=T)
  temp
}
beta <- c(0,0)
Buzkova.stable.beta <- nleqslv(beta, f)$x
```

```
########################################
#####################################
# LIANG METHOD
#####################################
########################################
gammaV_b <- sapply(baseData$ID, function(nn){sum((exp(delta.hat[1]*baseData$X1[baseData$ID==nn]+
               delta.hat[2]*baseData$X2[baseData$ID==nn])*estlam.t.delta)[udt<=baseData$C[baseData$ID==nn]])})

#/*** estsigma ***/
estsigma2 <- max(sum((baseData$m^2-baseData$m-(gammaV_b)^2)/sum((gammaV_b)^2)),0)

#/**** Bhat ****/
Bhat_i <- ( (1+baseData$m*estsigma2)/(1+gammaV_b*estsigma2)-1)
Bhat_long <- sapply(sim.data$ID, function(i) Bhat_i[baseData$ID==i])
Bhat <- sim.data$Q*Bhat_long

#/**** vector of "observed": ****/
bigX <- as.matrix(cbind(sim.data$X1, sim.data$X2, Bhat))
bigX_base <- as.matrix(cbind(baseData$X1, baseData$X2, Bhat_i*baseData$Q))

#/**** : estimated Xbar1 & Xbar2 & Bhat****/
denom  <- sapply(sim.data$t, function(u){
                  sum((exp_delta*(baseData$m/gammaV_b))[ baseData$C >= u ])} )
numer1 <- sapply(sim.data$t, function(u){
                  sum((bigX_base[,1]*exp_delta*(baseData$m/gammaV_b))[ baseData$C >= u ])} )
numer2 <- sapply(sim.data$t, function(u){
                  sum((bigX_base[,2]*exp_delta*(baseData$m/gammaV_b))[ baseData$C >= u ])} )
numer3 <- sapply(sim.data$t, function(u){
                  sum((bigX_base[,3]*exp_delta*(baseData$m/gammaV_b))[ baseData$C >= u ])} )

if (estsigma2 != 0 ){
  Xbar <- cbind(numer1/denom,numer2/denom,numer3/denom)
  f <- function(beta){
    temp <- rep(0,length=ncol(bigX))
    temp[1] <- sum(( (bigX[,1]-Xbar[,1])*
                  (sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2]-beta[3]*bigX[,3])), na.rm=T)
    temp[2] <- sum(( (bigX[,2]-Xbar[,2])*
                  (sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2]-beta[3]*bigX[,3])), na.rm=T)
    temp[3] <- sum(( (bigX[,3]-Xbar[,3])*
                  (sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2]-beta[3]*bigX[,3])), na.rm=T)
    temp
  }
  beta <- c(0,0,0)
  Liang.beta <- nleqslv(beta, f)$x
}

if (estsigma2 == 0 ){
  Xbar <- cbind(numer1/denom,numer2/denom)
  f <- function(beta){
    temp <- rep(0,length=ncol(Xbar))
    temp[1] <- sum(( (bigX[,1]-Xbar[,1])*(sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2])), na.rm=T)
    temp[2] <- sum(( (bigX[,2]-Xbar[,2])*(sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2])), na.rm=T)
    temp
  }
  beta <- c(0,0)
  Liang.beta <- c(nleqslv(beta, f)$x, NA)
}

########################################
#####################################
# WEIGHTED-LIANG METHOD
#####################################
########################################
gammaV_b <- sapply(baseData$ID, function(nn){
  sum((exp(gamma.hat[1]*baseData$X1[baseData$ID==nn]+gamma.hat[2]*baseData$X2[baseData$ID==nn]+
             gamma.hat[3]*testdata4$X3[testdata4$ID==nn])*estlam.t.gamma)[udt<=baseData$C[baseData$ID==nn]])})

#/*** estsigma ***/
estsigma2 <- max(sum((baseData$m^2-baseData$m-(gammaV_b)^2)/sum((gammaV_b)^2)),0)

#/**** Bhat ****/
Bhat_i <- ( (1+baseData$m*estsigma2)/(1+gammaV_b*estsigma2)-1)
Bhat_long <- sapply(sim.data$ID, function(i) Bhat_i[baseData$ID==i])
Bhat <- sim.data$Q*Bhat_long

#/**** vector of "observed": ****/
bigX <- as.matrix(cbind(sim.data$X1, sim.data$X2, Bhat))
```

```
bigX_base <- as.matrix(cbind(baseData$X1, baseData$X2, Bhat_i*baseData$Q))

#/**** : estimated Xbar1 & Xbar2 & Bhat****/
denom  <- sapply(sim.data$t, function(u) sum((exp_delta*(baseData$m/gammaV_b))[ baseData$C >= u ]) )
numer1 <- sapply(sim.data$t, function(u) sum((bigX_base[,1]*exp_delta*(baseData$m/gammaV_b))[baseData$C>=u]))
numer2 <- sapply(sim.data$t, function(u) sum((bigX_base[,2]*exp_delta*(baseData$m/gammaV_b))[baseData$C>=u]))
numer3 <- sapply(sim.data$t, function(u) sum((bigX_base[,3]*exp_delta*(baseData$m/gammaV_b))[baseData$C>=u]))

if (estsigma2 != 0 ){
  Xbar <- cbind(numer1/denom,numer2/denom,numer3/denom)
  f <- function(beta){
    temp <- rep(0,length=ncol(bigX))
    temp[1] <- sum(1/iirr2*((bigX[,1]-Xbar[,1])*
                            (sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2]-beta[3]*bigX[,3])), na.rm=T)
    temp[2] <- sum(1/iirr2*((bigX[,2]-Xbar[,2])*
                            (sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2]-beta[3]*bigX[,3])), na.rm=T)
    temp[3] <- sum(1/iirr2*((bigX[,3]-Xbar[,3])*
                            (sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2]-beta[3]*bigX[,3])), na.rm=T)
    temp
  }
  beta <- c(0,0,0)
  Weighted.Liang.beta <- nleqslv(beta, f)$x
}

if (estsigma2 == 0 ){
  Xbar <- cbind(numer1/denom,numer2/denom)
  f <- function(beta){
    temp <- rep(0,length=ncol(Xbar))
    temp[1] <- sum(1/iirr2*((bigX[,1]-Xbar[,1])*(sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2])), na.rm=T)
    temp[2] <- sum(1/iirr2*((bigX[,2]-Xbar[,2])*(sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2])), na.rm=T)
    temp
  }
  beta <- c(0,0)
  Weighted.Liang.beta <- c(nleqslv(beta, f)$x, NA)
}

##########################################
########################################
# SUN METHOD
########################################
##########################################
piCi <- sapply(baseData$ID, function(n, t){sum( (exp(delta.hat[1]*baseData$X1[baseData$ID==n]+
                      delta.hat[2]*baseData$X2[baseData$ID==n])*estlam.t.delta)
                      [t<=baseData$C[baseData$ID==n]], na.rm=T) }, t=udt  )

### Zhat & Ohat ###
Zhat <- (baseData$m-1)/piCi
Ohat <- (baseData$m-1)*(baseData$m-2)/piCi^2

### Xbar ###
S0  <-  sapply(sim.data$t, function(u){sum((exp_delta*(baseData$m/piCi))[u<=baseData$C], na.rm=T)})
Sx1 <-  sapply(sim.data$t, function(u){sum((exp_delta*baseData$X1*(baseData$m/piCi))[u<=baseData$C], na.rm=T)})
Sx2 <-  sapply(sim.data$t, function(u){sum((exp_delta*baseData$X2*(baseData$m/piCi))[u<=baseData$C], na.rm=T)})
Sz  <-  sapply(sim.data$t, function(u){sum((exp_delta*Zhat*(baseData$m/piCi))[u<=baseData$C], na.rm=T)})

Xbar1 <- Sx1/S0
Xbar2 <- Sx2/S0
Zbar <- Sz/S0

bigX <- cbind(sim.data$X1, sim.data$X2)
Zhat.long <- sapply(sim.data$ID, function(i) Zhat[baseData$ID==i])
Ohat.long <- sapply(sim.data$ID, function(i) Ohat[baseData$ID==i])

f <- function(beta){
  temp <- rep(0,length=ncol(bigX)+1)
  temp[1] <- sum(((bigX[,1]-Xbar1)*(sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2]
                                    -beta[3]*Zhat.long )), na.rm=T)
  temp[2] <- sum(((bigX[,2]-Xbar2)*(sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2]
                                    -beta[3]*Zhat.long )), na.rm=T)
  temp[3] <- sum(((Zhat.long - Zbar)*(sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2])
                 -beta[3]*(Ohat.long - Zhat.long*Zbar) ), na.rm=T)
  temp
}
beta <- c(1,-1,0)
Sun.beta <- nleqslv(beta, f)$x

##########################################
```

```
  ######################################
  # WEIGHTED-SUN METHOD
  ######################################
  ########################################
  piCi <- sapply(baseData$ID, function(n, t){sum( (exp(
                gamma.hat[1]*baseData$X1[baseData$ID==n]+
                 gamma.hat[2]*baseData$X2[baseData$ID==n]+
                 gamma.hat[3]*testdata4$X3[testdata4$t==t & testdata4$ID==n])*
                 estlam.t.gamma)[t<=baseData$C[baseData$ID==n]], na.rm=T) }, t=udt  )

  ### Zhat & Ohat ###
  Zhat <- (baseData$m-1)/piCi
  Ohat <- (baseData$m-1)*(baseData$m-2)/piCi^2

  ### Xbar ###
  S0  <-  sapply(sim.data$t, function(u){sum((exp_delta*(baseData$m/piCi))[u<=baseData$C], na.rm=T)})
  Sx1 <-  sapply(sim.data$t, function(u){sum((exp_delta*baseData$X1*(baseData$m/piCi))[u<=baseData$C], na.rm=T)})
  Sx2 <-  sapply(sim.data$t, function(u){sum((exp_delta*baseData$X2*(baseData$m/piCi))[u<=baseData$C], na.rm=T)})
  Sz  <-  sapply(sim.data$t, function(u){sum((exp_delta*Zhat*(baseData$m/piCi))[u<=baseData$C], na.rm=T)})

  Xbar1 <- Sx1/S0
  Xbar2 <- Sx2/S0
  Zbar <- Sz/S0

  bigX <- cbind(sim.data$X1, sim.data$X2)
  Zhat.long <- sapply(sim.data$ID, function(i) Zhat[baseData$ID==i])
  Ohat.long <- sapply(sim.data$ID, function(i) Ohat[baseData$ID==i])

  f <- function(beta){
    temp <- rep(0,length=ncol(bigX)+1)
    temp[1] <- sum( ( (1/iirr2)*(bigX[,1]-Xbar1)*(sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2]
                                                  -beta[3]*Zhat.long )), na.rm=T)
    temp[2] <- sum( ( (1/iirr2)*(bigX[,2]-Xbar2)*(sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2]
                                                  -beta[3]*Zhat.long )), na.rm=T)
    temp[3] <- sum( ( (1/iirr2)*(Zhat.long - Zbar)*(sim.data$Y-beta[1]*bigX[,1]-beta[2]*bigX[,2])
                      -beta[3]*(Ohat.long - Zhat.long*Zbar) ), na.rm=T)
    temp
  }
  beta <- c(1,-1,1)
  Weighted.Sun.beta <- nleqslv(beta, f)$x

  ## Obtain estimates
  est_out <- data.frame(t(gamma.hat),t(delta.hat), t(LY.beta), t(Buzkova.stable.beta),
                        t(Liang.beta),t(Weighted.Liang.beta), t(Sun.beta), t(Weighted.Sun.beta))
  est_out
}

###### Run function and save original estimates
sim.data.orig  <- sim.data
Orig_Est <- round(DepObsTimes(sim.data=sim.data, baseData=baseData, udt=udt),3)

#########################################################
###### CLUSTER-BOOTSTRAP for SD of current dataset ######
######## Subjects are sampled with replacement ##########
simout.sd <- NULL
Bcluster <- 1000
for (bbc in 1:Bcluster){
  if (bbc %% 2 ==0) cat("inner=", bbc, "\n")

  set.seed(bbc)
  ###########################################################

  sim.data.sd <- NULL
  for(i in 1:N){
    select <- sample(sim.data.orig$ID,1)
    m <- nrow(sim.data.orig[sim.data.orig$ID==select,])
    sim.data.sd <- rbind(sim.data.sd, data.frame(ID=rep(i, nrow=m), sim.data.orig[sim.data.orig$ID==select,]))
    i <- i+1
  }
  baseData.sd <- ddply(sim.data.sd, .(ID), function(x) x[1, ])
  udt.sd <- unique(sort(sim.data.sd$t[sim.data.sd$t>0]))

  #### Run function
  Est_SD <- DepObsTimes(sim.data=sim.data.sd, baseData=baseData.sd, udt=udt.sd)


  #### Cluster-ootstrap coefficients
```

```
  simout.sd <- rbind(simout.sd, Est_SD)
  bbc <- bbc+1
}

sd <- round(apply(simout.sd,2, sd, na.rm=T),3)

cat("\n\n ================================================\n\n",
    "\n Bladder Cancer Case Study: Table 4\n",
    "\n (SE: Cluster-bootstrap based on", Bcluster, "repetitions)\n",
    "\n\n Estimation of gamma's (Observation-time model):",
    "\n gamma  ",paste("g1=",Orig_Est[1],"(",sd[1],"),
                      g2=",Orig_Est[2],"(",sd[2],"),g3=",Orig_Est[3],"(",sd[3],")"),
    "\n\n Estimation of betas using different methods:",
    "\n LY     ",paste("b1=",Orig_Est[6], "(",sd[6],"), b2=",Orig_Est[7],"(", sd[7], ")"),
    "\n Buzkova",paste("b1=",Orig_Est[8], "(",sd[8],"), b2=",Orig_Est[9],"(", sd[9], ")"),
    "\n Liang  ",paste("b1=",Orig_Est[10],"(",sd[10],"),
                      b2=",Orig_Est[11],"(",sd[11],"), theta=",Orig_Est[12],"(",sd[12],")"),
    "\n W-Liang",paste("b1=",Orig_Est[13],"(",sd[13],"),
                      b2=",Orig_Est[14],"(",sd[14],"), theta=",Orig_Est[15],"(",sd[15],")"),
    "\n Sun    ",paste("b1=",Orig_Est[16],"(",sd[16],"),
                      b2=",Orig_Est[17],"(",sd[17],"), alpha=",Orig_Est[18],"(",sd[18],")"),
    "\n W-Sun  ",paste("b1=",Orig_Est[19],"(",sd[19],"),
                      b2=",Orig_Est[20],"(",sd[20],"), alpha=",Orig_Est[21],"(",sd[21],")"),
    "\n\n ================================================\n\n" )


#########################################################
#########################################################
# CHECK if latent variables are covariate-dependent
# Using density curves of estimated latent variables
#########################################################
#########################################################
#install.packages("ggplot2")
library(ggplot2)
library(grid)
require(gridExtra)


## First run original sim.data (from blaTum) within DepObsTimes() function to obtain individual estimates.

#/******** Density Curve of \eta_i under WEIGHTED-SUN METHOD: Gamma distributed eta-i2
gammaV_b <- sapply(baseData$ID, function(nn){sum((exp(gamma.hat[1]*baseData$X1[baseData$ID==nn]+
             gamma.hat[2]*baseData$X2[baseData$ID==nn]+gamma.hat[3]*testdata4$X3[testdata4$ID==nn])*
             estlam.t.gamma)[udt<=baseData$C[baseData$ID==nn]])})

#/*estimated variance and \eta_{i2}
estsigma2 <- max(sum((baseData$m^2-baseData$m-(gammaV_b)^2)/sum((gammaV_b)^2)),0)
eta_i2 <- (1+baseData$m*estsigma2)/(1+gammaV_b*estsigma2)

## combined
plot1 <- ggplot(baseData)+geom_density(alpha=.6, aes(x=eta_i2))+
  theme_bw()+theme(axis.line = element_line(color = 'black'))+
  scale_x_continuous(expand = c(0, 0)) + scale_y_continuous(expand = c(0, 0))+
  theme( plot.background = element_blank()
         ,panel.grid.major = element_blank()
         ,panel.grid.minor = element_blank()
         ,panel.border = element_blank()
         ,panel.background = element_blank()) +xlab(expression(hat(eta)[i2]))
## by treatment group
plot2 <- ggplot(baseData)+geom_density(alpha=.6, aes(x=eta_i2, fill=as.factor(baseData$X1)))+
  theme_bw()+scale_fill_manual(values=c("grey20", "grey60"), name="Group", breaks=c("0", "1"),
                               labels=c("Placebo", "Treatment"))+theme(axis.line = element_line(color = 'black'))+
  scale_x_continuous(expand = c(0, 0)) + scale_y_continuous(expand = c(0, 0))+
  theme( plot.background = element_blank()
         ,panel.grid.major = element_blank()
         ,panel.grid.minor = element_blank()
         ,panel.border = element_blank()
         ,panel.background = element_blank()) + theme(legend.position = c(.8, .7))+xlab(expression(hat(eta)[i2]))

grid.arrange(plot1, plot2, ncol=2)


#/******** Density Curve of \eta_i under WEIGHTED-SUN METHOD
piCi <- sapply(baseData$ID, function(n, t){sum( (exp(gamma.hat[1]*baseData$X1[baseData$ID==n]+
         gamma.hat[2]*baseData$X2[baseData$ID==n]+gamma.hat[3]*testdata4$X3[testdata4$t==t & testdata4$ID==n])*
         estlam.t.gamma)[t<=baseData$C[baseData$ID==n]], na.rm=T) }, t=udt  )
```

```
piCi <- sapply(baseData$ID, function(n, t){sum( (exp(delta.hat[1]*baseData$X1[baseData$ID==n]+
        delta.hat[2]*baseData$X2[baseData$ID==n])*
        estlam.t.delta)[t<=baseData$C[baseData$ID==n]], na.rm=T) }, t=udt  )
#/*estimated \eta_i
Zhat <- (baseData$m-1)/piCi

## combined
plot3 <- ggplot(baseData)+geom_density(alpha=.6, aes(x=Zhat))+
  theme_bw()+theme(axis.line = element_line(color = 'black'))+
  scale_x_continuous(expand = c(0, 0)) + scale_y_continuous(expand = c(0, 0))+
  theme( plot.background = element_blank()
         ,panel.grid.major = element_blank()
         ,panel.grid.minor = element_blank()
         ,panel.border = element_blank()
         ,panel.background = element_blank()) +xlab(expression(hat(eta)[i]))
## by treatment group
plot4 <- ggplot(baseData)+geom_density(alpha=.6, aes(x=Zhat, fill=as.factor(baseData$X1)))+
  theme_bw()+scale_fill_manual(values=c("grey20", "grey60"), name="Group", breaks=c("0", "1"),
                               labels=c("Placebo", "Treatment"))+theme(axis.line = element_line(color = 'black'))+
  scale_x_continuous(expand = c(0, 0)) + scale_y_continuous(expand = c(0, 0))+
  theme( plot.background = element_blank()
         ,panel.grid.major = element_blank()
         ,panel.grid.minor = element_blank()
         ,panel.border = element_blank()
         ,panel.background = element_blank()) + theme(legend.position = c(.8, .7))+xlab(expression(hat(eta)[i]))

grid.arrange(plot3, plot4, ncol=2)




##############################################
##############################################
# Check overall model-fit using residuals
# Only focus on models with weights (case 2)
##############################################
##############################################
##/* setup */
exp_delta <- exp(delta.hat[1]*baseData$X1+delta.hat[2]*baseData$X2)
denom_delta <- sapply(udt, function(u){sum( (exp_delta)[u<=baseData$C], na.rm=T) } )
estlam.t.delta <- sapply(1:length(udt), function(u) sum( ((sim.data$t==udt[u])/denom_delta[u])) )

exp_gamma <- function(u){exp(gamma.hat[1]*baseData$X1+
                               gamma.hat[2]*baseData$X2+gamma.hat[3]*testdata4$X3[testdata4$t==u])}
denom_gamma <- sapply(udt, function(u){sum( (exp_gamma(u))[u<=baseData$C], na.rm=T) } )
estlam.t.gamma <- sapply(1:length(udt), function(u) sum( ((sim.data$t==udt[u])/denom_gamma[u])) )

plot_resid <- function(est_resid, addtitle){
  plot(est_resid[sim.data$X1==0]~sim.data$t[sim.data$X1==0], pch=1, ylim=c(-3, 3), xlim=c(0,54),
       xlab="Observation times", ylab="Residuals", las=1, cex=.7)
  par(new=T)
  plot(est_resid[sim.data$X1==1]~sim.data$t[sim.data$X1==1], pch=2, ylim=c(-2, 3), xlim=c(0,54),
       xlab=" ", ylab=" ", xaxt="n", yaxt='n', cex=.5)
  abline(h=0)
  lines(smooth.spline(est_resid~sim.data$t, df = 10), lty = 2, col = "red", lwd=2)
  mtext(addtitle, 3, line=-1.2, cex=.7)
}

##/************** Buzkova method **************/##
#/*** d\mathcal{A}(t) by id ***/
numer <- (1/iirr2)*(sim.data$Y-Buzkova.stable.beta[1]*sim.data$X1-Buzkova.stable.beta[2]*sim.data$X2)
mathcal_A.delta <- sapply(1:length(udt), function(uu) sum(numer[sim.data$t==udt[uu]]/denom_delta[uu], na.rm=T) )

alpha_hat_step <- mathcal_A.delta/estlam.t.gamma
alpha_hat_sim  <- sapply(sim.data$t, function(uu) alpha_hat_step[udt==uu] )

#/*** residuals ***/#
pred_Y_Buzkova <- Buzkova.stable.beta[1]*sim.data$X1+Buzkova.stable.beta[2]*sim.data$X2
est_resid_Buzkova <- sim.data$Y - alpha_hat_sim -
                  Buzkova.stable.beta[1]*sim.data$X1-Buzkova.stable.beta[2]*sim.data$X2
plot_resid(est_resid_Buzkova, "Buzkova method")


##/************** Weighted-Liang method: Q=X1 **************/##
#/*** Bhat ****/
Bhat2_i    <- ( (1+baseData$m*estsigma2)/(1+gammaV_b*estsigma2)-1)
Bhat_long <- sapply(sim.data$ID, function(i) Bhat2_i[baseData$ID==i])*sim.data$Q
```

```
#/*** d\mathcal{A}(t) by id ***/
numer <- (1/iirr2)*(sim.data$Y-Weighted.Liang.beta[1]*sim.data$X1-Weighted.Liang.beta[2]*sim.data$X2-
                    Weighted.Liang.beta[3]*Bhat_long)
mathcal_A.delta <- sapply(1:length(udt), function(uu) sum(numer[sim.data$t==udt[uu]]/denom_delta[uu], na.rm=T) )

alpha_hat_step <- mathcal_A.delta/estlam.t.gamma
alpha_hat_sim  <- sapply(sim.data$t, function(uu) alpha_hat_step[udt==uu] )

#/*** residuals ***/#
pred_Y_WLiang_X1 <- Weighted.Liang.beta[1]*sim.data$X1+
                    Weighted.Liang.beta[2]*sim.data$X2+Weighted.Liang.beta[3]*Bhat_long
est_resid_WLiang_X1 <- sim.data$Y - alpha_hat_sim - Weighted.Liang.beta[1]*sim.data$X1-
  Weighted.Liang.beta[2]*sim.data$X2-Weighted.Liang.beta[3]*Bhat_long
plot_resid(est_resid_WLiang_X1, "Weighted-Liang method")


##/************** Weighted-Sun method **************/##

#/**** function for observation-level weights ****/
Z <- function(uu){cbind(baseData$X1, baseData$X2, testdata4$X3[testdata4$t==udt[uu]])}
X <- function(uu){cbind(baseData$X1, baseData$X2)}
iirr2b <- function(uu){exp(Z(uu) %*% as.matrix(gamma.hat))/exp(X(uu) %*% as.matrix(delta.hat))}

#/****/
piCi <- sapply(baseData$ID, function(n, t){sum( (exp(gamma.hat[1]*baseData$X1[baseData$ID==n]+
         gamma.hat[2]*baseData$X2[baseData$ID==n]+gamma.hat[3]*testdata4$X3[testdata4$t==t & testdata4$ID==n])*
         estlam.t.gamma)[t<=baseData$C[baseData$ID==n]], na.rm=T) }, t=udt  )

### Zhat & Ohat ###
Zhat <- (baseData$m-1)/piCi
Zhat.long <- sapply(sim.data$ID, function(i) Zhat[baseData$ID==i])
numer <- (1/iirr2)*(sim.data$Y-Weighted.Sun.beta[1]*sim.data$X1-
                              Weighted.Sun.beta[2]*sim.data$X2-Weighted.Sun.beta[3]*Zhat.long)
mathcal_A.delta <- sapply(1:length(udt), function(uu) sum(numer[sim.data$t==udt[uu]]/denom_delta[uu], na.rm=T) )

alpha_hat_step <- mathcal_A.delta/estlam.t.gamma
alpha_hat_sim  <- sapply(sim.data$t, function(uu) alpha_hat_step[udt==uu] )

#/*** residuals ***/#
pred_Y_WSun <- alpha_hat_sim + Weighted.Sun.beta[1]*sim.data$X1+
                              Weighted.Sun.beta[2]*sim.data$X2+Weighted.Sun.beta[3]*Zhat.long
est_resid_WSun <- sim.data$Y - alpha_hat_sim - Weighted.Sun.beta[1]*sim.data$X1-
  Weighted.Sun.beta[2]*sim.data$X2-Weighted.Sun.beta[3]*Zhat.long
plot_resid(est_resid_WSun, "Weighted-Sun method")
```

APPENDIX B

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

We provide R code to reproduce estimated bias, empirical standard error estimates and mean squared error estimates from one set of parameters presented in Chapter 3 simulation. We consider Setting 2, in which:

- $(\beta_1, \beta_2, \beta_3) = \{\log(1.5), \log(1.2), \log(0.5)\}$,
- $(\gamma_1, \gamma_2, \gamma_3) = (0.3, 0.2, 0.3)$,
- $Q_i = 1$ and
- $\eta_{i2}^{(2)} : \eta_{i2} \sim I(X_{i1} \leq 0.5)\text{Uniform}[0.5, 1.5] + I(X_{i1} > 0.5)\text{Gamma}(1, 0.7)$

```
## Load extension packages
library("splines")
library("plyr")
library("nleqslv")
library("survival")

################################################################
# Functions to create observation-times
Lam <- function(t, z, x, b, v, c, w, g){
1/2* z * t^( x*b + 3/2 )/( x*b + 3/2 ) *exp(g*w + c*v) }
invLam <- function(t, z, x, b, v, c, w, g){
( t * 2 * ( x*b + 3/2 ) * exp(- g*w - c*v) /z)^(1/(x*b+3/2))}
################################################################

set.seed(234)
simout<- NULL
for (bb in c(1:1000)){
  if (bb %%2 ==0) print(bb)
  set.seed(1)
  ### Set initial values
  Ntot=200 ## number of subjects
  tau=10 ## max study time
  sigma_err <- 1
  sigma_phi <- 1
  b_01 <- log(1.5)
  b_02 <- log(1.2)
  b_03 <- log(.5) ## variable X3 only in observation-time model

  gamma01 <- 0.3
  gamma02 <- 0.2
  gamma03 <- 0.3

  miu_z_1 <- 2
  miu_z_2 <- 0
  sigma_z_1 <- 1
  sigma_z_2 <- 2

  theta0 <- 1
  ### create a full set of outcomes for each subject
  sim.data <- NULL
  for (i in 1:Ntot){
  ### grid of 100 per time unit (t=0.00 is baseline)
```

```
X1 <- X2 <- Z <- phi_i <- err_i <- M <- C <-NULL
C <- runif(1, min=5, max=tau)
X1 <- runif(1,0,1)
X_ind <- as.numeric(X1>0.5)
X3 <- (1-X_ind)*rnorm(n=1, mean=miu_z_1, sd=sigma_z_1) +
(X_ind)*rnorm(n=1, mean=miu_z_2, sd=sigma_z_2)
X2 <- rbinom(1,1,0.5)
expect_X3_X1 <- (1-X_ind)*miu_z_1 + (X_ind)*miu_z_2
var_X3_X1 <- (1-X_ind)^2*(sigma_z_1)^2 + (X_ind)^2*(sigma_z_2)^2
W <- 1 # Q = 1; otherwise X1
sigma_eta2 <- 0.5

##### eta_i(1)
# Z <- eta_2 <- rgamma(1,shape=2,scale=sigma_eta2)
##### eta_i(2)

Z <- eta_2 <- if(X2==1){runif(n=1, min=0.5, max=1.5)} else {rgamma(1,shape=2,scale=sigma_eta2)}
M <- sqrt(sigma_err^2 + (W*sigma_phi)^2 + b_03^2 * var_X3_X1)/1.7
f_0_star <- function(t){(-1+0.5*t^(-1/2))*M - b_03 * expect_X3_X1}
b_01_star <- b_01*M
b_02_star <- b_02*M
phi_i <- rnorm(n=1, mean=0, sd=sigma_phi)
eta_1 <- (theta0*(eta_2-1))*M + phi_i

##########################################
# Generate observations times and outcomes
##########################################
len <- 0; tmpt <- NULL
while ( len < Lam(C, Z, X1, gamma01, X2, gamma02, X3, gamma03) ){
tmpt <- c(tmpt, rexp(1,1) )
len <- sum(tmpt)
}

m <- length(tmpt) - 1
if( m > 0 ){
tt <- invLam( cumsum(tmpt[1:m ]), Z, X1, gamma01, X2, gamma02, X3, gamma03)
} else tt <- 0

Y =as.numeric(f_0_star(tt) + b_01_star * X1 * log(tt) + b_02_star * X2 + b_03 * X3 +
eta_1 * W + rnorm(n=length(tt), mean=0, sd=sigma_err) > 0 )
tmp <- data.frame(ID=i,t=tt[order(tt)],Y=Y, Z=z, m=m, X1=X1, X2=X2, X3=X3, C=C, W=W)
sim.data <- rbind( sim.data, tmp )

i <- i+1
}

######################
# Set-up
######################
sim.data <- sim.data[sim.data$t>0,]
baseData <- ddply(sim.data, .(ID), function(x) x[1, ])
udt <- sort(unique(sim.data$t[sim.data$t>0]))
N<- length(baseData$ID)

##############################
## Bsplines (duration)
##############################
test <- bs(sim.data$t, df = 4, intercept=T)[1:length(unique(sim.data$t)), 1:4]
test.long <- t(sapply(sim.data$t, function(tt) test[unique(sim.data$t)==tt,]))
sim.data$B1 <- test.long[,1]
sim.data$B2 <- test.long[,2]
sim.data$B3 <- test.long[,3]
sim.data$B4 <- test.long[,4]
##############################

#/**** FOR DELTA.HAT (covariates in outcome model) ****/
f <- function(gamma){
  #/**** vector of baseline V: ****/
  bigV <-cbind(baseData$X1, baseData$X2)
  gamma_test <- function(g,u){exp(g[1]*bigV[,1]*log(u)+g[2]*bigV[,2])}

  #/**** Vbar ****/
  denom <- sapply(udt, function(u) sum(gamma_test(gamma,u)[ baseData$C >= u ]) )
  numer1 <- sapply(udt, function(u) sum((bigV[,1]*log(u)*gamma_test(gamma, u))[ baseData$C >= u ]) )
  numer2 <- sapply(udt, function(u) sum((bigV[,2]*gamma_test(gamma, u))[ baseData$C >= u ]) )
  Vbar <- cbind(numer1/denom, numer2/denom)
  Vbar.long <- t(sapply(sim.data$t, function(tt) Vbar[udt==tt,]))
```

```
  #/***** estimating eq for deltas *****/
  bigV <-cbind(sim.data$X1*log(sim.data$t), sim.data$X2)
  temp <- colSums((bigV-Vbar.long)/N, na.rm=T)
  temp
}
gamma <- c(gamma01, gamma02)
gamma.hat.setup <- nleqslv(gamma, f)
delta.hat <- c(gamma.hat.setup$x)

#/**** FOR GAMMA.HAT (covariates from observation-time model) ****/
f <- function(gamma){
  bigV <-cbind(baseData$X1, baseData$X2, baseData$X3)
  gamma_test <- function(g,u){exp(g[1]*bigV[,1]*log(u)+g[2]*bigV[,2]+g[3]*bigV[,3])}

  #/**** Vbar ****/
  denom <- sapply(udt, function(u) sum(gamma_test(gamma, u)[ baseData$C >= u ]) )
  numer1 <- sapply(udt, function(u) sum((bigV[,1]*log(u)*gamma_test(gamma, u))[ baseData$C >= u ]) )
  numer2 <- sapply(udt, function(u) sum((bigV[,2]*gamma_test(gamma, u))[ baseData$C >= u ]) )
  numer3 <- sapply(udt, function(u) sum((bigV[,3]*gamma_test(gamma, u))[ baseData$C >= u ]) )
  Vbar <- cbind(numer1/denom, numer2/denom, numer3/denom)
  Vbar.long <- t(sapply(sim.data$t, function(tt) Vbar[udt==tt,]))

  #/***** estimating eq for gammas *****/
  bigV <-cbind(sim.data$X1*log(sim.data$t), sim.data$X2, sim.data$X3)
  temp <- colSums((bigV-Vbar.long)/N, na.rm=T)
  temp
}
gamma <- c(gamma01, gamma02, gamma03)
gamma.hat.setup <- nleqslv(gamma, f)
gamma.hat <- c(gamma.hat.setup$x)

#/**** : estimated Lam(t) ****/
denom <- sapply(udt, function(u) sum(exp(gamma.hat[1]*baseData$X1*log(u)
            +gamma.hat[2]*baseData$X2+gamma.hat[3]*baseData$X3)[ baseData$C >= u ]) )
estlam.t <- sapply(1:length(udt), function(u) sum( ((sim.data$t==udt[u])/denom[u])) )

#/**** calculate observation-level weights ****/
Z <- cbind(sim.data$X1*log(sim.data$t), sim.data$X2, sim.data$X3)
X <- cbind(sim.data$X1*log(sim.data$t), sim.data$X2)
iirr2 <- exp(Z %*% as.matrix(gamma.hat))/exp(X %*% as.matrix(delta.hat))

#/**** function for observation-level weights ****/
Z <- function(uu){cbind(baseData$X1*log(udt[uu]), baseData$X2, baseData$X3)}
X <- function(uu){cbind(baseData$X1*log(udt[uu]), baseData$X2)}
iirr2b <- function(uu){exp(Z(uu) %*% as.matrix(gamma.hat))/exp(X(uu) %*% as.matrix(delta.hat))}

##################################################
piCi <- sapply(baseData$ID, function(n, t){
             sum( (exp(gamma.hat[1]*baseData$X1[baseData$ID==n]*log(t)+
                       gamma.hat[2]*baseData$X2[baseData$ID==n]+
                       gamma.hat[3]*baseData$X3[baseData$ID==n])*estlam.t)
                       [t<=baseData$C[baseData$ID==n]], na.rm=T) }, t=udt )

estsigma2 <- max(sum((baseData$m^2-baseData$m-(piCi)^2)/sum((piCi)^2)),0)

##/**** Bhat ****/##
baseData_Zhat <- baseData$m/piCi
Zhat_long <- sapply( sim.data$ID, function(i) (baseData_Zhat[baseData$ID==i]))
baseData_Bhat <- baseData$W*(baseData_Zhat-1)
Bhat_long <- sapply( sim.data$ID, function(i) (baseData_Bhat[baseData$ID==i]))

##################################################
## I. IEE
## (GEE with independence correlation structure)
##################################################
X1t <- sim.data$X1*(log(sim.data$t))
iee <-(glm(Y~X1t+X2+B1+B2+B3+B4-1, data=sim.data, family=binomial(link="logit")))$coef

#########################################################################
## II. Weighted-GEE
## (IEE with observation-level weights and latent variable effects)
#########################################################################
weighted.iee <-(glm(Y~X1t+X2+Bhat_long+B1+B2+B3+B4-1, data=sim.data,
family=quasibinomial (link="logit"), weight=(1/iirr2)))$coef
estsigma2 <- max(sum((baseData$m^2-baseData$m-(piCi)^2)/sum((piCi)^2)),0)
```

```
############################
## III. Proposed method
############################
bulk_long <- sapply(1:length(udt), function(uu){
             ((1/iirr2b(uu))*(baseData$m/piCi)
              *exp(gamma.hat[1]*baseData$X1*log(udt[uu])+
                   gamma.hat[2]*baseData$X2+gamma.hat[3]*baseData$X3)*estlam.t[uu])
})

expitin <- function(b1,b2, theta,t1,t2,t3,t4, uu){1/(1+exp(-b1*baseData$X1*log(udt[uu])
                     -b2*baseData$X2 -theta*baseData_Bhat
                     -t1*sim.data$B1[sim.data$t==udt[uu]][1]
                     -t2*sim.data$B2[sim.data$t==udt[uu]][1]
                     -t3*sim.data$B3[sim.data$t==udt[uu]][1]
                     -t4*sim.data$B4[sim.data$t==udt[uu]][1])) }
Ybar1 <- function(b1,b2, theta,t1,t2,t3,t4, inX){sapply(1:length(udt), function(uu) {
                     sum((inX*expitin(b1,b2,theta,t1,t2,t3,t4, uu)
                     *bulk_long[,uu])[ baseData$C >= udt[uu] ]) })}
Ybar1b <- function(b1,b2, theta,t1,t2,t3,t4, inX){sapply(1:length(udt), function(uu) {
                     sum((inX*log(udt[uu])*expitin(b1,b2, theta,t1,t2,t3,t4, uu)
                     *bulk_long[,uu])[ baseData$C >= udt[uu] ]) })}
Ybar3 <- function(b1,b2, theta, t1,t2,t3,t4, inX){sapply(1:length(udt), function(uu) {
                     sum((inX[sim.data$t==udt[uu]][1]*expitin(b1,b2, theta, t1,t2,t3,t4, uu)
                     *bulk_long[,uu])[ baseData$C >= udt[uu] ]) })}

f <- function(beta){

#/***** Equation 7: estimating eq for betas *****/
temp <- rep(0,7)
temp[1] <- sum((1/iirr2)*sim.data$X1*log(sim.data$t)*sim.data$Y, na.rm=T)-
    sum(Ybar1b(beta[1],beta[2],beta[3],beta[4],beta[5],beta[6],beta[7],baseData$X1), na.rm=T)
temp[2] <- sum((1/iirr2)*sim.data$X2*sim.data$Y, na.rm=T)-
   sum(Ybar1(beta[1],beta[2],beta[3],beta[4],beta[5],beta[6],beta[7],baseData$X2), na.rm=T)
temp[3] <- sum((1/iirr2)*Bhat_long*sim.data$Y, na.rm=T)-
   sum(Ybar1(beta[1],beta[2],beta[3],beta[4],beta[5],beta[6],beta[7],baseData_Bhat), na.rm=T)
## 4df Bsplines
temp[4] <- sum((1/iirr2)*sim.data$B1*sim.data$Y, na.rm=T)-
    sum(Ybar3(beta[1],beta[2],beta[3],beta[4],beta[5],beta[6],beta[7],sim.data$B1), na.rm=T)
temp[5] <- sum((1/iirr2)*sim.data$B2*sim.data$Y, na.rm=T)-
    sum(Ybar3(beta[1],beta[2],beta[3],beta[4],beta[5],beta[6],beta[7],sim.data$B2), na.rm=T)
temp[6] <- sum((1/iirr2)*sim.data$B3*sim.data$Y, na.rm=T)-
    sum(Ybar3(beta[1],beta[2],beta[3],beta[4],beta[5],beta[6],beta[7],sim.data$B3), na.rm=T)
temp[7] <- sum((1/iirr2)*sim.data$B4*sim.data$Y, na.rm=T)-
    sum(Ybar3(beta[1],beta[2],beta[3],beta[4],beta[5],beta[6],beta[7],sim.data$B4), na.rm=T)
temp
}

beta <- c(1,1,1,.5,.5,.5,.5)
beta <- nleqslv(beta, f)
proposed.method <- c(beta$termcd, round(beta$x,7))

#####################
# Coefs of each dataset
#####################
meannumvisits <- mean(baseData$m)
simout <- rbind(simout, data.frame(cbind(t(delta.hat), t(gamma.hat),
t(iee), t(weighted.iee), t(proposed.method), meannumvisits)))
bb <- bb+1
save(simout, file="../AppendixCheck.rda")
}

##### Simulation results for beta1
beta01 <- log(1.5)
tableb1 <- simout[, c(6, 12, 20)]
beta_result <- matrix(0,ncol=3,nrow=3)
for(z in c(1:3)){
   beta_result[1,z] <- round(apply(as.matrix(tableb1[,z]), 2, mean, na.rm = T),3) -beta01
   beta_result[2,z] <- round(apply(as.matrix(tableb1[,z]), 2, sd, na.rm = T),3)
   beta_result[3,z] <- round((apply(as.matrix(tableb1[,z]), 2, mean, na.rm = T)-beta01 )^2
    + apply(as.matrix(tableb1[,z]), 2, var, na.rm = T),3)
}
rownames(beta_result) <- c("Bias", "ESE", "MSE")
colnames(beta_result) <- c( "IEE", "Weighted-IEE", "Proposed")
round(beta_result, 2)

##### Simulation results for beta2
beta02 <- log(1.2)
```

```
tableb1 <- simout[, c(7, 13, 21)]
beta_result <- matrix(0,ncol=3,nrow=3)
for(z in c(1:3)){
  beta_result[1,z] <- round(apply(as.matrix(tableb1[,z]), 2, mean, na.rm = T),3) -beta02
  beta_result[2,z] <- round(apply(as.matrix(tableb1[,z]), 2, sd, na.rm = T),3)
  beta_result[3,z] <- round((apply(as.matrix(tableb1[,z]), 2, mean, na.rm = T)-beta01 )^2
    + apply(as.matrix(tableb1[,z]), 2, var, na.rm = T),3)
}
rownames(beta_result) <- c("Bias", "ESE", "MSE")
colnames(beta_result) <- c( "IEE", "Weighted-IEE", "Proposed")
round(beta_result, 2)
```

BIBLIOGRAPHY

Albert, PS (2000). A transitional model for longitudinal binary data subject to nonignorable missing data. *Biometrics* 56.2, 602–608. ISSN: 0006-341X.

Albert, PS (2012). A linear mixed model for predicting a binary event from longitudinal data under random effects misspecification. *Statistics in medicine* 31.2, 143–154. ISSN: 1097-0258.

Andrews, DF and Herzberg, A (1985). *Data: A collection of problems from many fields for the student and research worker*. 1st ed. New York: Springer. ISBN: 0387961259.

Brigden, ML, Kay, C, Le, a, Graydon, C, and McLeod, B (1998). Audit of the frequency and clinical response to excessive oral anticoagulation in an out-patient population. *American Journal of Hematology* 59.1, 22–7. ISSN: 0361-8609.

Brumback, BA and Rice, JA (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* 93, 961–976.

Bůžková, P and Lumley, T (2007). Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables. *Canadian Journal of Statistics* 35.4, 485–500.

Bůžková, P and Lumley, T (2008). Semiparametric log-linear regression for longitudinal measurements subject to outcome-dependent follow-up. *Journal of Statistical Planning and Inference* 138.8, 2450–2461. ISSN: 03783758.

Bůžková, P and Lumley, T (2009). Semiparametric modeling of repeated measurements under outcome-dependent follow-up. *Statistics in Medicine* 28, 987–1003.

Cheng, G, Yu, Z, and Huang, JZ (2013). The cluster bootstrap consistency in generalized estimating equations. *Journal of Multivariate Analysis* 115, 33–47. ISSN: 0047259X.

Chernick, MR (2011). *Bootstrap Methods: A Guide for Practitioners and Researchers (Google eBook)*. John Wiley & Sons, 400. ISBN: 1118211596.

Cook, RJ and Lawless, J (2007). *The Statistical Analysis of Recurrent Events*. 1st ed. New York: Springer, 403. ISBN: 0387698094.

Efron, B and Tibshirani, R (1993). *An introduction to the bootstrap*. 1st ed. Boca Raton: Chapman & Hall/CRC.

Field, CA and Welsh, AH (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.3, 369–390. ISSN: 1369-7412.

Fitzmaurice, GM, Lipsitz, SR, Ibrahim, JG, Gelber, R, and Lipshultz, S (2006). Estimation in regression models for longitudinal binary data with outcome-dependent follow-up. *Biostatistics* 7.3, 469–485. ISSN: 1465-4644.

French, B and Heagerty, PJ (2009). Marginal mark regression analysis of recurrent marked point process data. *Biometrics* 65.2, 415–422. ISSN: 1541-0420.

Guo, Z, Gill, TM, and Allore, HG (2008). Modeling repeated time-to-event health conditions with discontinuous risk intervals: an example of a longitudinal study of functional disability among older persons. *Methods of Information in Medicine* 47.2, 107–116.

Heagerty, PJ and Kurland, BF (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* 88.4, 973–985. ISSN: 0006-3444.

Hu, XJ, Sun, J, and Wei, LJ (2003). Regression parameter estimation from panel counts. *Scandinavian Journal of Statistics* 30.1, 25–43.

Hu, XJ, Lorenzi, M, Spinelli, JJ, Ying, SC, and McBride, ML (2011). Analysis of recurrent events with non-negligible event duration, with application to assessing hospital utilization. *Lifetime data analysis* 17.2, 215–233. DOI: 10.1007/s10985-010-9183-8.

Huang, CY, Qin, J, and Wang, MC (2010). Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. *Biometrics* 66.1, 39–49. ISSN: 1541-0420.

Huang, CY, Wang, MC, and Zhang, Y (2006). Analysing panel count data with informative observation times. *Biometrika* 93.4, 763–775. ISSN: 0006-3444.

Huang, X and Liu, L (2007). A joint frailty model for survival and gap times between recurrent events. *Biometrics* 63.2, 389–97. ISSN: 0006-341X.

Hylek, EM, Skate, SJ, Sheehan, MA, and Singer, DE (1996). An analysis of the lowest effective intensity of prophylactic anticoagulation for patients with nonrheumatic atrial fibrillation. *The New England Journal of Medicine* 335, 540–546.

Kalbfleisch, JD and Prentice, RL (2002). *The statistical analysis of failure time data*. 2nd ed. New York: Wiley. ISBN: 047136357X.

Kim, YJ (2014). Regression analysis of recurrent events data with incomplete observation gaps. *Journal of Applied Statistics* 00.0, 1–8. DOI: 10.1080/02664763.2014.885002.

Kimmel, SE, Chen, Z, Price, M, Parker, CS, Newcomb, CW, Samaha, FF, and Gross, R (2007). The Influence of Patient Adherence on Anticoagulation Control With Warfarin: Results from the International Normalized Ratio Adherence and Genetics (IN-RANGE) Study. *Archives of Internal Medicine* 167.3, 229–235.

Li, Y and Ryan, L (2004). Survival analysis with heterogeneous covariate measurement error. *Journal of the American Statistical Association* 99.467, 724–735. ISSN: 0162-1459.

Liang, KY and Zeger, SL (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73.1, 13–22. ISSN: 0006-3444.

Liang, Y, Lu, W, and Ying, Z (2009). Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics* 65.2, 377–384. ISSN: 1541-0420.

Lin, DY and Ying, Z (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* 96.453, 103–126. ISSN: 0162-1459.

Lin, DY, Wei, LJ, Yang, I, and Ying, Z (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.4, 711–730. ISSN: 1369-7412.

Lin, H, Scharfstein, DO, and Rosenheck, RA (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.3, 791–813.

Lin, X and Carroll, RJ (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association* 96.455, 1045–1056. ISSN: 0162-1459.

Lipsitz, S, Fitzmaurice, G, and Ibrahim, J (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics* 58.3, 621–630.

Little, RJA (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 90.431, 1112–1121.

Liu, D, Kalbfleisch, JD, and Schaubel, DE (2011). A positive stable frailty model for clustered failure time data with covariate-dependent frailty. *Biometrics* 67.1, 8–17. ISSN: 1541-0420.

Liu, L, Huang, X, and O'Quigley, J (2008). Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics* 64.3, 950–958. ISSN: 1541-0420.

McCulloch, CE and Neuhaus, JM (2011). Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter. *Statistical Science* 26.3, 388–402. ISSN: 0883-4237. arXiv: arXiv:1201.1980v1.

Neuhaus, JM and McCulloch, CE (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.5, 859–872. ISSN: 1369-7412.

Pepe, MS and Anderson, GL (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation* 23.4, 939–951. ISSN: 0361-0918. DOI: 10.1080/03610919408813210.

Pepe, MS and Cai, J (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association* 88.423, 811–820.

Pepe, MS and Couper, D (1997). Modeling partly conditional means with longitudinal data. *Journal of the American Statistical Association* 92.439, 991–998.

Rizopoulos, D (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Chapman and Hall/CRC Biostatistics Series. Chapman and Hall/CRC. ISBN: 9781439872864.

Robins, JM, Greenland, S, and Hu, FC (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association* 94, 687–700.

Robins, JM, Hernán, MA, and Brumback, B (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* 11.5, 550–560.

Rotnitzky, A and Robins, JM (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* 82.4, 805–820.

Rubin, DB (1976). Inference and missing data. *Biometrika* 63.3, 581–592.

Ryu, D, Sinha, D, Mallick, B, Lipsitz, SR, and Lipshultz, SE (2007). Longitudinal studies with outcome-dependent follow-up. *Journal of the American Statistical Association* 102.479, 952–961. ISSN: 0162-1459.

Sagara, I, Dicko, A, Ellis, RD, Fay, MP, Diawara, SI, Assadou, MH, Sissoko, MS, Kone, M, Diallo, AI, Saye, R, Guindo, Ma, Kante, O, Niambele, MB, Miura, K, Mullen, GED, Pierce, M, Martin, LB, Dolo, A, Diallo, Da, Doumbo, OK, Miller, LH, and Saul, A (2009). A randomized controlled phase 2 trial of the blood stage AMA1-C1/Alhydrogel malaria vaccine in children in Mali. *Vaccine* 27.23, 3090–3098.

Sun, J, Tong, X, and He, X (2007). Regression analysis of panel count data with dependent observation times. *Biometrics* 63.4, 1053–9. ISSN: 0006-341X.

Sun, J and Wei, L (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.2, 293–302.

Sun, J, Park, DH, Sun, L, and Zhao, X (2005). Semiparametric regression analysis of longitudinal data with informative observation times. *Journal of the American Statistical Association* 100.471, 882–889. ISSN: 0162-1459.

Sun, L, Song, X, and Zhou, J (2011). Regression analysis of longitudinal data with time-dependent covariates in the presence of informative observation and censoring times. *Journal of Statistical Planning and Inference* 141.8, 2902–2919. ISSN: 03783758.

Troxel, AB, Lipsitz, SR, Fitzmaurice, GM, Ibrahim, JG, Sinha, D, and Molenberghs, G (2010). A weighted combination of pseudo-likelihood estimators for longitudinal binary data subject to non-ignorable non-monotone missingness. *Statistics in Medicine* 29.14, 1511–1521. ISSN: 1097-0258.

Williamson, JM, Datta, S, and Satten, Ga (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* 59.1, 36–42. ISSN: 0006-341X.

Yan, J and Fine, JP (2008). Analysis of Episodic Data With Application to Recurrent Pulmonary Exacerbations in Cystic Fibrosis Patients. *Journal of the American Statistical Association* 103.482, 498–510. ISSN: 0162-1459.

Zhao, LP, Rotnitzky, A, and Robins, JM (1995). Analysis of Semiparametric Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association* 90.429, 106–121.

Zhao, Q and Sun, J (2006). Semiparametric and nonparametric analysis of recurrent events with observation gaps. *Computational Statistics & Data Analysis* 51.3, 1924–1933. DOI: 10.1016/j.csda.2005.12.006.

Zhu, L, Zhao, H, Sun, J, Pounds, S, and Zhang, H (2013). Joint analysis of longitudinal data and recurrent episodes data with application to medical cost analysis. *Biometrical journal* 55.1, 5–16.