APPROVAL SHEET

Title of Dissertation: Analysis of Longitudinal Interval Reported Binary Recurrent Event Data & Statistical Model for Subgroup Identification in Enrichment Design

Name of Candidate: Wenxin Lu
Doctor of Philosophy, 2019

Dissertation and Abstract Approved: _____

Yi Huang
Associate Professor
Department of Mathematics and Statistics

Date Approved: 8/30/2019

Title of dissertation: Analysis of Longitudinal Interval Reported Binary Recurrent Event Data
&
Statistical Model for Subgroup Identification in Enrichment Design

Wenxin Lu, Doctor of Philosophy, 2019

Dissertation directed by: Dr. Yi Huang
Department of Mathematics and Statistics

## ABSTRACT

This thesis contains two research projects. The first thesis project investigates the analytical methods for longitudinal interval reported binary recurrent event data. This project is motivated by the post hip fracture infection project using Baltimore Hip studies (BHS). The infection related outcomes were collected longitudinally using questionnaire items like "Since the last time we spoke in (Provide Month), have you ever had fever?". Even though a subject could miss a few scheduled visits, such questionnaire design only captured the available longitudinal fever data across available visiting months, where the missing visits were skipped and merged into the reporting interval. Another feature is that the recurrent events of interest was observed dichotomously - only the binary status of occurrence in the reporting interval, without frequency counts information nor when they re-occur in this interval. Even though the literature on longitudinal binary data are quite comprehensive, the longitudinal models accounting for interval reported and binary recurrent event features are quite limited. We proposed two longitudinal models in this project, where discrete survival modeling technique and Poisson process are used to account for interval censored reporting system between longitudinal visits and binary nature of recurrent events outcomes. The intensity function follows Cox regression structure allowing for both subject's baseline characteristics and time-varying covariates, which leads to varying intensities over longitudinal visits but fixed intensity within each reporting interval. Simulation studies are used to compare the proposed models vs. standard longitudinal models with logit link to see how well they will capture the significant cross-sectional and longitudinal effects, especially with or without considering interval reporting nature, with or without time-varying covariates, and some other sensitivity analyses to model mis-specifications. Various simulation studies confirm the great performance of the proposed GLMM model with comp(log-log) link. Then, I implemented both the proposed and standard methods on the infection project using BHS data. Out of all 4 models, only the proposed GLMM model with comp(log-log) link detected the statistically significant monthly increasing trend of hazard of infection re-occurrence during the first year hip fracture post-surgery re-

covery time. And, all models confirmed no sex difference in various measures of infection re-occurrence risks on average over time during the first year follow ups.

The second research project is on the statistical model for subgroup identification in enriched clinical trial design. Enrichment designs have been widely used in randomized clinical trials (RCT) for years across pharmaceutical industry and academia, because such designs are often more efficient, such as smaller sample size, shortened development time, and reduced cost. The enrichment design strategies can be summarized into three categories, which are well documented in the FDA guidance as the prospective use of any patient's characteristics to select a target study sub-population (called "subgroup") smartly, so that the drug effects (if one is in fact present) are easier and clearer to be detected than the unselected population. Based on the information from the phase II RCTs and prior scientific knowledge from historical literature and other studies, the current popular practice of enrichment design strategy is to use individual indicator variable as the criteria for subgroup identification. However, this strategy becomes infeasible when the number of associated variables or criteria increases. Thus, in this thesis project, we build a subgroup selection model using many patients' characteristics, which could be estimated by outcome regression model, inverse probability weighted estimator (IPWE), and doubly robust inverse probability weighted estimator (DRIPWE). The purpose is to facilitate the subgroup identification and find the ideal target sub-population for phase III participants, making phase III RCT more efficient. Simulation studies are used to compare the three proposed methods on building the optimal subgroup selection model and demonstrate the importance to include as many covariates as possible into the model.

# Analysis of Longitudinal Interval Reported Binary Recurrent Event Data & Statistical Model for Subgroup Identification in Enrichment Design

by

Wenxin Lu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Dr. Yi Huang, Advisor
Dr. Yehenew Kifle (Alphabetical Order)
Dr. Aiyi Liu
Dr. Thomas Mathew
Dr. DoHwan Park

# Dedication

To mom and dad.

# Acknowledgments

First of all, I would like to express my deepest thanks to my advisor, Dr. Yi Huang, for giving me the continuous support, caring, guidance, and mentoring on my doctoral study, Baltimore Hip Studies' projects, and research over the past five years. I really appreciate your patience, guidance, and encouragement on teaching me how to be a statistician. Thank you for giving me invaluable advice not only on my research but also on my career.

I would like to thank Dr. Thomas Mathew and Dr. DoHwan Park, for serving as my thesis reader and giving me insightful comments. Thank you for your valuable time and help on my thesis.

I would like to thank the rest of my committee members, Dr. Yehenew Kifle and Dr. Aiyi Liu, for serving as my committee members and giving me helpful suggestions. Thank you for your support on my thesis.

I would like to thank the gerontology group at medical school on Baltimore Hip Studies for providing the opportunity collaborating with you, which motivates my thesis paper one. Thank you for allowing me to use the data on my thesis.

I would like to thank Dr. Atul Bhattaram and Dr. Anindya Roy for giving me the opportunity working with you in FDA. Your guidance on the theoretical and coding part have taught me a lot, which are helpful in the future career. Thank you for your advice on our project.

I would like to thank professors in our department for those wonderful courses you provide. It is those courses that have made my thesis possible.

I would like to thank my family. It is impossible for me to pursue my degrees in the US without the support from my parents. Thank you for giving me this opportunity. A special thanks to my beloved husband, Jin Wang, who has been there for me always. Thank you for your support and encouragement throughout my doctoral life.

Finally, I would like to thank God, the dearest Father. Thank you for leading me, being with me through those difficulties, and letting me obtain the degree. Thank you for giving me peace and joy in my heart no matter how hard the life is. Thank you for everything you provide.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Review Longitudinal Data Analysis and Survival Analysis

### 1.1.1 Longitudinal Data Analysis

Longitudinal data is defined as the repeated measurements over time on each study subject. On the contrary, it is cross-sectional data where each subject has only a single outcome. Even though both longitudinal and cross-sectional studies could address the same scientific questions, the major advantage of the former is its capacity to draw inferences on both cohort effects and longitudinal effects. Longitudinal effect is often called "age effect" quantifying the outcome changes over time within the individual, while cohort effect is often called "cross-sectional effect" quantifying the outcome differences average over time across different groups of subjects (Diggle et al., 2002). Longitudinal studies can be conducted either prospectively, following subjects forward in time, or retrospectively, recalling historical information multiple times per subject in surveys or extracting multiple measurements from a subject's historical records. Some important nationwide longitudinal studies, such as National Health and Nutrition Examination Survey (NHANES) and Framingham Heart Study, exemplify how important longitudinal studies play the role in public

health study. With the wide popularity of longitudinal studies in the public health and medical research, their study design can be complicated given its complexity of dealing with human subjects and feasibility in medical practice. Resulting from this, the longitudinal data can have complicated structures and missing data patterns, which will be further discussed in Chapter 2.

Longitudinal data analysis is built upon the generalized linear models (GLMs) framework, extending the inference and estimation framework from a single outcome to the clusters of correlated outcomes per subject. For inference framework of a single outcome, GLMs provide a unified class of regression models of independent but diverse types of univariate responses on mixed types of covariates. The natural exponential family of outcome distributions can accommodate various common types of outcome variables, such as continuous, binary, ordinal, and count (Fitzmaurice et al., 2011). Based on the standard linear regression inference framework, the GLM framework includes two components: (1) systematic component, $g[E(Y|X)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$; and (2) random component, probability distribution of $f_{Y|X}$ (Agresti, 2002). The systematic component associates the the mean responses of subjects to subjects' characteristics captured by various covariates through $g$ function, the link function. The link function makes a transformation to the mean responses and then links the covariates.

Depend on the types of outcomes, the link function can be very different. For continuous outcomes, in particular the normally distributed outcomes, identify link

is often used as the canonical link with nice large sample properties, leading to linear regression. For count data, in particular the Poisson type of outcomes, log link is the most popular one as the canonical link, leading to Poisson log-linear model. Depend on the complicity of count data, especially for the over dispersion problems, zero-inflated Poisson model and negative binomial GLM are also popular models to use. For binary outcomes, which is the focus in our Chapter 2, the most popular link is logit as the canonical link. For medical data, binary variable are often generated based on the dichotomization of an underlying continuous variable by a cutoff point. Depending on the nature of how binary outcomes are generated, the legitimate non-canonical links which are often used in practice are probit link, inverse CDF link functions, complementary log-log link, and identity link, as explained in Agresti's book (Agresti, 2002). Our chapter 2 will focus on the complementary log-log link as a gateway to model the complicated longitudinal interval reported binary recurrent event data. For ordinal and nominal categorical data, various types of models using logit link have been proposed according to whether accounting for the natural ordering structure of outcome or not, including multinomial logistic model (or, polytomous logistic regression), proportional odds model (or, cumulative logit model), and conditional odds model (Agresti, 2002).

If there are repeated measurements for each subject across time, generalized linear models need to be extended to account for dependency among repeated measurements obtained from one subject. Let $Y_{ij}$ be the response vairiable for the $i$th subject at the $j$th longitudinal visit, and $\mathbf{X}_{ij}$ be the $p$ covariates associated with

the response $Y_{ij}$, $i = 1, \cdots, N, j = 1, \cdots, n_i$, where $n_i$ indicates the $i$th subject's total number of repeated measurements within the study period. It is assumed that $\mathbf{Y}_i$ are independent of one another, but $Y_{ij}$ is correlated with $Y_{ik}$ on the same subject $i$. Among the $p$ covariates $\mathbf{X}_{ij}$, there may exist time-invariant (between-subject) covariates that do not change within the study period, such as gender, and time-varying (within-subject) covariates that change over time, such as age (Fitzmaurice et al., 2011). To account for the within-subject association, the first method marginal models (generalized linear models for longitudinal data) which has been widely used in the biomedical and health sciences estimated by Generalized Estimating Equations (GEEs) is used, with the only assumption of how the mean response is related with covariates, and there is no assumption on distribution of responses (Liang and Zeger, 1986). Because this method is for marginal model estimation and the inference is for the population, it is also called population-average model estimation. The marginal models have three components, with two of them are the same as GLM based on the standard generalized linear model formulation: (1) the mean of each response is assumed to depend on the covariates through a known link function $g$, $g[E(Y_{ij}|\mathbf{X}_{ij})] = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp}$; (2) given the covariates, the conditional variance of each response is assumed to depend on the mean, $\text{Var}(Y_{ij}|\mathbf{X}_{ij}) = \phi v(E(Y_{ij}|\mathbf{X}_{ij}))$, where $v(\cdot)$ is a known variance function of the mean and $\phi$ is a scale parameter; and (3) given the covariates, the conditional within-subject association among repeated measurements is assumed to depend on the mean and association parameters $\alpha$, which leads to the "working" covariance matrix (Fitzmaurice et al., 2011). Even if the within-subject association have been

incorrectly modeled, the GEE estimator $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$ under the correctly specified model for the mean response with the "robust (empirical)" variance estimator (or "sandwich" estimator), which shows a very appealing robustness property of the GEE estimator and the "sandwich" estimator. The second method is via the introduction of random effects $\mathbf{b}_i$ allowing for some regression coefficients vary randomly from one subject to another and the inference is for the individual, which result in generalized linear mixed effects models (GLMMs) or subject-specific models estimated by maximum likelihood estimation (MLE) (Fitzmaurice et al., 2011). The GLMMs also include three components: (1) the conditional mean of each response is assumed to depend on fixed and random effects through a known link function $g$, $g[E(Y_{ij}|\mathbf{b}_i)] = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp} + b_{0i} + b_{1i} t_{ij}$; (2) given random effects $\mathbf{b}_i$, the response $Y_{ij}$ is assumed to be independent of one another, and the conditional distribution of each response is assumed to belong to the exponential family of distributions with the conditional variance depends on the mean:$\text{Var}(Y_{ij}|\mathbf{b}_i) = \phi v(E(Y_{ij}|\mathbf{b}_i))$, where $v(\cdot)$ is a known variance function of the mean and $\phi$ is a scale parameter; (3) the random effects $\mathbf{b}_i$ are assumed to follow a multivariate normal distribution and be independent of the covariates $\mathbf{X}_{ij}$ (Fitzmaurice et al., 2011). For example, if the response is binary, both marginal model and generalized linear mixed effects model with logit link function are widely used for the relationship between the response and a set of covariates, which can be written as following:

$$\log \frac{\Pr(Y_{ij} = 1|\mathbf{X}_{ij})}{\Pr(Y_{ij} = 0|\mathbf{X}_{ij})} = \alpha_0 + \alpha_1 X_{ij1} + \alpha_2 X_{ij2} + \ldots + \alpha_p X_{ijp}, \tag{1.1}$$

and

$$\log\frac{\Pr(Y_{ij}=1|\mathbf{X}_{ij},b_{0i})}{\Pr(Y_{ij}=0|\mathbf{X}_{ij},b_{0i})}=\beta_0+\beta_1X_{ij1}+\beta_2X_{ij2}+\ldots+\beta_pX_{ijp}+b_{0i}. \tag{1.2}$$

The equation (1.1) is for the marginal model with logit link and a "working" covariance matrix estimated by GEE, while the equation (1.2) is for the generalized linear mixed effects model with logit link estimated by MLE. The interpretations of parameters from the two equations are totally different, with the former have population-average interpretation on changes in the transformed mean response in the study population, and the latter have subject-specific interpretation on changes in the transformed mean response for any individual (Fitzmaurice et al., 2011). Thus, the parameter $\alpha_k$ can be interpreted as the changes in the log odds of response in the study population with one unit change in the corresponding covariate $X_{ijk}$ after controlling for the other covariates. The parameter $\beta_k$ can be interpreted as the changes in an individual's log odds of response with one unit change in the corresponding covariate $X_{ijk}$ after controlling for the other covariates and the random effect.

Even though subjects are required to do measurements at each time of follow-up for most longitudinal studies, it is very common to see missing data which results in different missing patterns for different subjects. Because different subjects have different number of repeated measurements at a common set of occasions, the longitudinal data is unbalanced over time (Fitzmaurice et al., 2011). There are three different types of missing data mechanisms, missing completely at random (MCAR),

missing at random (MAR), and not missing at random (NMAR) (Fitzmaurice et al., 2011). Missing completely at random (MCAR) is defined as the missingness in response is completely at random and is unrelated with the observed or unobserved responses. Missing at random (MAR) is defined as the missingness in response is only related with the observed response but not the unobserved responses. Not missing at random (NMAR) is defined as the missingness in response is related with the unobserved responses, which is also called nonignorable missingness or informative missingness. When data are missing completely at random (MCAR) or missing at random (MAR), GLMMs can yield valid estimates of regression parameters; while the GEE methods can only provide valid estimates when data are MCAR but not MAR, but the methods can be adapted for MAR data to provide valid estimates, which is known as the inverse probability weighted (IPW) GEE method. However, when data are not missing at random (NMAR), both GEE methods and GLMMs provide biased estimates without considering the missing data mechanism. In Chapter 2, we will assume that the missing data mechanisms are MCAR.

### 1.1.2 Survival Analysis

Survival analysis is defined as a class of statistical methods for studying occurrence and timing of events (Allison, 2010). The time to a specific event is the primary outcome, and the event can be death, occurrence of a disease, marriage, divorce, and so on. If the time to an event is known in fine detail, it is called continuous time. If the time to an event is known within an interval, it is called grouped time or

interval-censored time. Grouped survival data is a special case of interval-censored data, where all subjects have all information available at the same visiting time and have the same disjoint intervals (Giolo et al., 2009). If the time to an event is not known except for the discrete number of time points, it is called discrete time. However, the time to an event may not be observed for every subject, which is called censoring (Allison, 2010). Right censoring is the most commonly one and we only know that the time to an event is greater than some value. For example, some subjects drop out of a study before the event occurred, or some subjects did not experience the event by the end of the study. Compared to ordinary regression models, survival analysis can handle censored data to have valid estimates. The survivor and hazard functions are the two important functions in survival analysis (Allison, 2010). The survivor function gives the probability of the variable survival time $T$ being greater than a specified time $t$, which can be written as, $S(t) = P(T > t)$. The hazard function gives the instantaneous rate of change of the event probability at a specified time $t$, conditional on the subject survived to that specified time $t$, which can be written as, $h(t) = \frac{d(-\log S(t))}{dt} = \lim_{\Delta t \to 0} \left[ \frac{Pr\{t \leq T < t+\Delta t\}}{\Delta t} \right]$.

If the survival times follow certain distributions, such as exponential distribution (with constant hazard over the time), Weibull distribution, lognormal distribution, and so on, fully parametric methods are used. If the distributions are unknown, nonparametric methods are used. The method developed by Kaplan and Meier (1958) is widely used to estimate and graph survival probabilities as a function of time. To better describe the relationship between survival times and covariates,

Cox regression model was proposed by Cox (1972). It is usually referred to as the proportional hazards model. But it can be generalized to allow for non-proportional hazards. Because the Cox regression model makes no assumption about the baseline hazard function, it is also called semiparametric model. The model focuses on the relationship between the hazard function and the covariates, and it assumes that the log hazards is linearly associated with covariates. If the hazard ratio comparing any two observations are constant over time, the proportional hazards assumption is made.

#### 1.1.2.1 Cox Proportional Hazards Model

The Cox proportional hazards model (Cox, 1972) describes the relationship between the hazard function and the covariates $\mathbf{x}$, which can be written as

$$\log h_i(t, \mathbf{x}) = \log \lambda_0(t) + \beta_1 x_{i1} + \cdots + \beta_k x_{ik},$$

or,

$$h_i(t, \mathbf{x}) = \lambda_0(t) \exp\left(\beta_1 x_{i1} + \cdots + \beta_k x_{ik}\right),$$

where $\lambda_0(t)$ is the baseline hazard function.

Regression coefficients from the above model are log-hazard ratios. By exponentiating a regression coefficient, it means the relative change in the hazard of occurrence of event of interest which is associated with one unit increase in the corresponding predictor (Austin, 2017).

#### 1.1.2.2 Discrete Time Models

Suppose $T$ is a discrete random variable indicating the time of occurrence of an event, and let $t_{ij}$ be the $j$th discrete time point for the $i$th subject, where $0 < t_{i1} < t_{i2} < \cdots < t_{iJ_i}$, with $J_i$ indicating the total number of time points ($i = 1, \cdots, n; j = 1, \cdots, J_i$) (Vermunt, 2009).

The probability of experiencing an event at $T = t_{ij}$ for the $i$th subject is given as $f(t_{ij}) = P(T = t_{ij})$.

The survival function is $S(t_{ij}) = P(T > t_{ij}) = 1 - P(T \leq t_{ij}) = 1 - \sum_{k=1}^{j} f(t_{ik})$.

The conditional probability, which cannot be called hazard rate, that the event occurs at $T = t_{ij}$, given that the event did not occur prior to $T = t_{ij}$ is defined as

$$\lambda(t_{ij}) = P(T = t_{ij} \mid T \geq t_{ij}) = \frac{P(T = t_{ij})}{P(T \geq t_{ij})} = \frac{S(t_{i(j-1)}) - S(t_{ij})}{S(t_{i(j-1)})} = 1 - \frac{S(t_{ij})}{S(t_{i(j-1)})}.$$

If time $T$ is a continuous variable which is measured discretely, the conditional probability of experiencing an event at $t_{ij}$ can be expressed as

$$\lambda(t_{ij}) = 1 - \exp\left(-\int_{t_{i(j-1)}}^{t_{ij}} h(u)du\right).$$

If the hazard rate is constant ($h(u) = h$) in time interval $(t_{i(j-1)}, t_{ij}]$, this expression can be simplified to $\lambda(t_{ij}) = 1 - \exp\left(-h(t_{ij} - t_{i(j-1)})\right)$.

## 1.2 Review Clinical Trials and Advanced Clinical Trial Designs

### 1.2.1 Clinical Trials

A clinical trial is the evaluation of intervention (treatment) on disease in a controlled experimental setting (Friedman et al., 2010). After the preliminary tests conducted on animals showing promising results for humans, a clinical trial is done on humans, which includes four phases (Friedman et al., 2010). Phase I explores a tolerated dose and the pharmacology of the drug. It is known that the therapeutic effect increases with dose, but the toxic effects increases as well. Doses are increasing on a small number of subjects to determine the maximum dose that can be tolerated by most patients with the disease. And the possible side effects of the drug are also documented during this phase. Phase II conducts initial assessment for therapeutic effects and further assesses toxicities. A small study is conducted before the costly large study to assess whether the efficacy of the drug is sufficient. After showing sufficient efficacy in phase II, it moves to phase III. Phase III compares the intervention to the standard treatment with respect to efficacy and toxicity. The study is large enough to detect any significant difference if there is one and obtain unbiased estimates using reasonable statistical methods during this phase. After a drug is approved by FDA and on the market, a large number of patients can take it. Some adverse events occur after a long time use of the drug. A monitoring system is needed to detect those adverse events. Phase IV is a post marketing study, which studies adverse effects using observational data. The previous description is a general idea about clinical trials, and the way how to select patients into the study

will be discussed in the next paragraph.

Starting from a target population - a general population with certain disease of interest, this population are screened for enrollment, which is called screened population after excluding some subjects from the target population considering specific characteristics. They are excluded according to age, geographical settings, severity of the disease, and so on. The screened population is further screened using inclusion and exclusion criteria, and the informed consent is distributed at the same time. Inclusion criteria are certain standards that participants must meet to stay in the clinical trial. Exclusion criteria are factors that exclude subjects from the clinical trial. For those satisfying eligibility criteria and signing informed consent, they are called study population, which will be randomized into two or more treatments to estimate treatment efficacy and safety.

### 1.2.2 Advanced Clinical Trial Designs

#### 1.2.2.1 Enrichment Design

In randomized controlled trials, patients are randomized into two or more groups. However, the population in clinical trials is not selected randomly from the general population considering the costs, treatment effects and ethical issues. There are some inclusion and exclusion criteria for those entering clinical trials. They are chosen because they have good compliance with treatment, or they do not have placebo

responses, or they have certain disease, or they have certain biomarkers suggesting good responses to the treatment, and so on. The above ways choosing the study population can be also called enrichment design (targeted design).

The enrichment design was originally defined as the additional screening processes with the active treatments evaluated in the study by Temple (1994). The additional screening processes with the active treatments were performed after either the screening period or the placebo run-in (or, placebo lead-in, placebo washout, placebo baseline) period to identify potential patients who are likely to benefit the test drug in the early phase of the trial (Liu, 2003). A classic example of this type can be found in the area of arrhythmia such as the Cardiac Arrhythmia Suppression Trial (CAST), which randomly assigned patients to a long-term study of encainide, flecainide, or morcizine compared with placebo only if they had ejection fraction of at least 30% (Echt et al., 1991; Liu, 2003; Temple, 2005). This type of study design can be also called randomized withdrawal design or randomized discontinuation design. Freidlin and Simon (2005) concluded that the randomized discontinuation designs can be useful in some settings in the early development of targeted agents where a reliable assay to select patients expressing the target is not available. They developed a study design for cytostatic agents, which is similar with CAST; that is, patients who responded to the treatment continued on the drug, while those whose disease progressed were taken off the study, and patients with stable disease were randomly assigned between the continued drug and placebo (Freidlin and Simon, 2005). Temple (2005) mentioned that the randomized withdrawal design is con-

siderably more efficient if there is a responder subpopulation, especially when the responder population is relatively low (30%).

Then, the FDA (2012) issued a draft about the enrichment strategies for clinical trials to support approval of human drugs and biological products. The guidance defined the enrichment as the prospective use of any patient characteristic to select a study population in which detection of a drug effect (if one is in fact present) is more likely than it would be in an unselected population. There are three broad categories for the enrichment strategies - strategies to decrease heterogeneity, prognostic enrichment strategies, and predictive enrichment strategies - in this guidance. Strategies to decrease heterogeneity include selecting patients with decreased inter-patient variability (e.g., include those whose baseline measurements are in a narrow range) and decreased intra-patient variability (e.g., exclude those whose disease or symptoms improve spontaneously); prognostic enrichment strategies choose patients with a greater likelihood of having a disease-related endpoint event or a substantial worsening in condition; predictive enrichment strategies choose patients more likely to respond to the drug treatment than other patients with the condition being treated. The enrichment strategies can give us efficient and powerful results with smaller sample size, shortened development time, and reduced cost, especially for pharmaceutical companies. But how to generalize and how to apply the results are the two challenges. In the following, prognostic enrichment strategies and predictive enrichment strategies are fully explained.

Prognostic enrichment strategies are widely used in cardiovascular outcome trials. This guidance describes that the severity of the illness, a history of recent myocardial infarction or stroke, the presence of concomitant illness such as diabetes, hypertension, or hyperlipidemia, and certain blood markers, such as high LDL (low-density lipoprotein) cholesterol, low HDL (high-density lipoprotein) cholesterol and high C-reactive protein, have been used to identify high risk patients for cardiovascular events. In the enalapril trials (CONSENSUS Trial Study Group, 1987), mortality reduction and decreases in morbid events (such as hospitalization) were first assessed in a very ill CHF (congestive heart failure) population of NYHA (New York Heart Association) Class IV patients. Later trials by Yusuf et al. (1991) in less ill patients were much longer and with much larger sample size because of the lower mortality rate. In the JUPITER study (Ridker et al., 2008), the rosuvastatin was effective reducing the incidence of major CV events in patients with normal LDL cholesterol levels of less than 130 mg per deciliter but with elevated high-sensitivity C-reactive protein levels of greater than or equal to 2.0 mg per liter.

Predictive enrichment strategies are not commonly used in cardiovascular studies but in heart failure and hypertrophic cardiomyopathy studies, where the systolic and diastolic subtypes would be expected to respond differently to different treatments for heart failure, and where patients with obstructive and nonobstructive physiologies could respond differently for hypertrophic cardiomyopathy (Blaus et al., 2015). For example, the combination of isosorbide dinitrate and hydralazine showed significant reduction for heart failure among black patients (Taylor et al., 2004).

Predictive enrichment strategies are also popular in oncology studies because of genetic characteristics of tumors. In 2001, trastuzumab showed the clinical benefit of first-line chemotherapy in metastatic breast cancer that overexpressed HER2 in a targeted randomized phase III trial (Slamon et al., 2001). In Iressa Pan-Asia Study (IPASS), gefitinib had better outcome for patients with EGFR (epidermal growth factor receptor) mutations on tumors (Mok et al., 2009). For the drug erlotinib, there was a highly significant survival difference for EGFR-positive patients, while only little effects seen among the EGFR-negative patients (Temple, 2005). In 2013, tarceva (erlotinib) was approved by FDA for the first-line treatment of patients with metastatic non-small cell lung cancer (NSCLC) whose tumors have EGFR exon 19 deletions or exon 21 (L858R) substitution mutations.

In the guidance issued by the FDA, it is important to include a reasonable sample of marker-negative patients if there exists uncertainty about the marker cut-off and responsiveness of marker-negative patients. The marker-negative patients are for those not satisfying the enrichment criteria, while marker-positive patients are for those selected subpopulation in the study. Yang et al. (2015) raised a novel design strategy by augmenting biomarker-negative patients into biomarker-positive patients. To assess the overall treatment effect, the biomarker-negative patients were enrolled after the biomarker-positive subpopulation was sufficiently powered. They combined the two estimates from biomarker-positive patients and biomarker-negative patients using weighted statistic which was determined from the screening information.

#### 1.2.2.2 Sequential Parallel Comparison Design

The sequential parallel comparison design (SPCD), also called sequential parallel design (SPD), is a clinical trial methodology developed by Fava et al. (2003) to reduce both the overall placebo response rate and the sample size for double-blind, placebo-controlled trials in psychiatric disorders, while the widely used design strategy - implement a placebo lead-in phase prior to randomization - only reduces placebo rate. The placebo response represents an apparent improvement for those randomly assigned to the placebo group in a clinical trial (e.g., a pre-posttreatment change within the placebo group) (Schatzberg and Kraemer, 2000; Fava et al., 2003). With the novel study design, it can reduce the cost and time remarkably for the evaluation of new drugs. The sequential parallel comparison design has two different formats. Both Formats include two double-blind treatment phases of equal duration. For Format 1, the first phase is an unbalanced randomization with more patients randomized to placebo compared with active treatment. For those placebo non-responders, they are randomized to either placebo or active treatment in the second phase to reduce placebo rate. Data from the two phases are pooled to maximize power and reduce required sample size. In Format 2, Fava et al. (2003) combined the two phases into three treatment groups: drug alone (DP), placebo then drug (PD), and placebo then placebo (PP). Then eligible subjects are randomized to one of the above three groups in an unbalanced ratio by $1-2a$, $a$, and $a$. Only placebo non-responders from the first phase are continued on placebo or drug in the second phase and drug non-responders are switched to placebo, while responders from the

first phase will enter open continuation therapy, or discontinue the study. Except for the data from phase 2 in the DP group, all of the data from the two phases are used for estimating weighted average treatment effect based on response rate which is only for binary endpoint - response or not. This novel study design was applied in smoking cessation trial and major depressive disorder (MDD) trial with the permission of Massachusetts General Hospital.

Although the novel design developed by Fava et al. (2003) could obtain considerable efficiency, it did not account for dropouts among placebo non-responders and did not mention any continuous endpoints that could use this design. Tamura and Huang (2007) looked more closely on this novel design and examined the efficiency of the design with the comparison of conventional two arm clinical trial considering both binary and continuous endpoints. Before any analysis, Tamura and Huang (2007) made two modifications of the design proposed by Fava et al. (2003). Instead of entering open continuation therapy for those responders from the first phase, Tamura and Huang (2007) suggested all patients should remain on blinded throughout both phases. The second modification is that all patients in the drug group in the first phase should remain on drug during the second phase to see the efficacy and safety of the drug over a longer period of time. To assess the efficacy of the drug, all data from the first phase and only placebo non-responders data from the second phase are used. With the consideration of drop out rates among placebo non-responders from phase 1 to phase 2, Tamura and Huang (2007) confirmed the results from Fava et al. (2003) for binary endpoints. For continuous endpoints, Tamura and Huang

(2007) used seemingly unrelated regression (SUR) (Zellner, 1962) to account for the within-subject correlation between phase 1 and phase 2 for placebo non-responders, and showed the efficiency of SPD for continuous endpoints without missing data.

Based on sequential parallel design, Chen et al. (2011) proposed sequential parallel design with re-randomization (SPD_ReR) to measure continuous endpoints with the presence of missing data. The SPD_ReR is the same design as Format 1 from SPD with randomization to either placebo or drug in phase 2 for those placebo non-responders. However, those drug non-responders in phase 1 will still be remained in drug group in phase 2, and responders in phase 1 will be blinded throughout the two phases. Chen et al. (2011) only focus on continuous endpoints, which is a complementary of Fava et al. (2003)'s estimates on binary endpoints. With the SPD_ReR design, the simple weighted ordinary least square (OLS) test statistic $Z_{OLS}$ was used instead of the weighted test statistic based on seemingly unrelated regression (SUR) $Z_{SUR}$ proposed by Tamura and Huang (2007). Both weighted test statistics are presented as following.

Let $\theta^{(1)}$ be the treatment effect for the endpoint at phase 1, $\theta^{(2)}$ be the treatment effect for the endpoint at phase 2. To test the null hypothesis for continuous endpoints, $H_0 : \theta^{(1)} = \theta^{(2)} = 0$, the weighted test statistic based on SUR can be written as (Tamura and Huang, 2007):

$$Z_{SUR} = \frac{w\hat{\theta}^{(1)} + (1-w)\hat{\theta}^{(2)}}{\sqrt{w^2 Var\left(\hat{\theta}^{(1)}\right) + 2w(1-w)Cov\left(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}\right) + (1-w)^2 Var\left(\hat{\theta}^{(2)}\right)}},$$

where $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ are the estimates of treatment effects $\theta^{(1)}$ and $\theta^{(2)}$ respectively, and $w$ is the prespecified weight on $\theta^{(1)}$ for weighted average treatment effect.

Based on the ordinary least square (OLS) estimates $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ for $\theta^{(1)}$ and $\theta^{(2)}$ respectively, the weighted test statistic can be written as (Chen et al., 2011):

$$Z_{OLS} = \frac{w\hat{\theta}^{(1)} + (1-w)\hat{\theta}^{(2)}}{\sqrt{w^2 Var\left(\hat{\theta}^{(1)}\right) + (1-w)^2 Var\left(\hat{\theta}^{(2)}\right)}},$$

which can be used an alternative method for testing the above null hypothesis.

When data are missing at random (MAR) and the dropout rate is moderate, both SUR and OLS weighted test statistic can be used with missing data imputed by last observation carried forward (LOCF) or multiple imputation (MI), and weighted test statistic based on mixed-effect model for repeated measures (MMRM) estimates on observed data. MMRM was shown to have the most robust test statistic for SPD_ReR design with large power and accurate estimation while controlling for the type I error rate.

#### 1.2.2.3 Adaptive Enrichment Design

The standard designs always enroll a broad range of subjects to decide a subset of subjects that may benefit from the treatment because the evidence is not strong enough when subjects enter into the phase III trial, which can expose many subjects to unnecessary side effects and decrease the treatment efficiency especially

when there is only a small subset of subjects benefit. To exclude those cannot benefit from the treatment, Simon and Simon (2013) introduced a class of adaptive enrichment designs that allow the enrollment criteria to change adaptively during the trial. With the inadequate information about the candidate predictive biomarkers from early phase clinical trials, the adaptive enrichment designs proposed by Simon and Simon (2013) for phase III trial start without any restrictions based on those candidate predictive biomarkers and restrict the enrollment sequentially in an adaptive manner using biomarkers. Such designs do not need to choose a subset of patients at the beginning of phase III trial but they will end to a subset of subjects who will benefit from the treatment after changing the enrollment criteria during the study, and eventually improve the treatment efficiency. Compared to the standard designs which enroll a large number of subjects and use post-hoc subgroup analysis to identify those who may benefit from the treatment, the adaptive enrichment designs expose fewer subjects to experience side effects and increase the efficiency. This class of designs not only preserve the type I error but also increase the power without increasing sample size especially when a small subset of subjects benefit from the treatment.

Simon and Simon (2013) applied the adaptive enrichment design to a setting where there is only a single candidate predictive biomarker $x$ but the cutpoint is unknown. To find the true cutpoint $x^*$, there is a set of discrete candidate cutpoints which are denoted by $\xi_1, \ldots, \xi_K$. Let $p_T(x)$ and $p_C(x)$ be the probabilities of response for a patient with the single biomarker $x$ under treatment and control by some mod-

els. Assume that the treatment effect $p_T(x) - p_C(x)$ is either 0 or $\delta$; that is, it is monotone non-decreasing with a jump only at one of the candidate cutpoints. With two unknown constants, $p_0 \leq p_1$, $p_C(x) = p_0$ for all $x$, while $p_T(x) = p_0$ when $x \leq \xi_k$ and $p_T(x) = p_1$ when $x > \xi_k$ with the maximized log-likelihood $l(\xi_k)$. Thus, the cutpoint $\xi_k$ is chosen as the estimate of the true cutpoint $x^*$, and the enrollment criterion is updated to enroll those whose biomarker value is greater than $\xi_k$. A general procedure for the adaptive enrichment design is described as following.

Let $z_i$ be the treatment assignment for subject $i$ where $z_i = 1$ for treatment group and $z_i = 0$ for control group, $\mathbf{x}_i$ be a vector of covariates, and $y_i$ be the outcome where $y_i = 1$ for response and $y_i = 0$ for non-response.

$$f(\mathbf{x}_i) = I\{p_T(\mathbf{x}_i) > p_C(\mathbf{x}_i)\},$$

where $f(\mathbf{x}_i)$ is an indicator function indicates whether the $i$th subject with covariates $\mathbf{x}_i$ will perform better on treatment than control. Let $\hat{f}_m(\mathbf{x}_i)$ be the estimate of $f_m(\mathbf{x}_i)$ after enrolling $m$ subjects, where $\hat{f}_m(\mathbf{x}_i)$ are based on $\mathbf{x}_1, \ldots, \mathbf{x}_{m-1}, z_1, \ldots, z_{m-1}$, and $y_1, \ldots, y_{m-1}$.

Firstly, $m_0$ subjects are randomly assigned to the treatment or placebo group without any restrictions on the enrollment to get the initial estimate $\hat{f}_{m_0}(\mathbf{x}_i)$ after modeling the response as a function of covariates in both treatment and control groups. Based on the estimated response models, a new subject's response can be estimated given his/her covariates, and lead to the decision whether or not enroll this sub-

ject. Secondly, $\hat{f}_m(\mathbf{x}_i)$ is estimated with updating models from the previous $(m-1)$ subjects after enrolling $m$ subjects, where $m > m_0$. With the updated models, the enrolling criteria are restricted to enroll those with $\hat{f}_m(\mathbf{x}_i) = 1$. Thirdly, the previous step is repeated until the total number of subjects $n$ have been enrolled.

However, Simon and Simon (2013) did not discuss how to best estimate $f(\mathbf{x})$ but only focus on how to preserve the type I error. To preserve the type I error, they introduced two methods for binary responses with a single interim analysis time at which the enrollment criteria can be modified. The first test statistic is presented as following:

$$S_1 = \sum_{i=1}^{n} [z_i y_i + (1 - z_i)(1 - y_i)],$$

where $S_1$ is the number of responses in the treatment group and the number of non-responses in the control group. Then, under the null hypothesis that both treatment and control group have the same probability of responses, $H_0 : p_T(\mathbf{x}) = p_C(\mathbf{x})$ for all $\mathbf{x}$,

$$S_1 \sim \text{binomial}(n, 0.5).$$

If subjects are randomized in pairs, one goes to the treatment group and the other one goes to the control group. Let $y_{i,T}$ and $y_{i,C}$ be the outcome under treatment and control group respectively for the pair $i$. Then, the second test statistic is presented as following:

$$S_2 = \sum_{i=1}^{n} \left[ I\{y_{i,T} > y_{i,C}\} - I\{y_{i,T} < y_{i,C}\} \right],$$

where $S_2$ is the number of pairs favoring treatment minus the number of pairs favoring control. Then, under the null hypothesis that each pair has the same favoring for treatment and control,

$$\frac{S_2 + u}{2} \sim \text{binomial}(n, 0.5),$$

where $u$ is the pre-specified number of pairs that we need in total. The test is also called McNemar's test.

The above two test statistics can protect the type I error for binary outcomes no matter what methods are used for changing the enrollment criteria adaptively.

When interim analysis is more than one time, or the adaptiveness is in a group sequential manner, Simon and Simon (2013) proposed other methods to preserve the type I err.

For continuous outcomes, Simon and Simon (2013) proposed the following statistic:

$$\frac{1}{\sqrt{n}} \sum_{k \leq K} \sqrt{n_k} \left( \frac{\bar{y}_{(T,k)} - \bar{y}_{(C,k)}}{\sqrt{\hat{\sigma}^2_{(T,k)}/(n_{T,k} - 1) + \hat{\sigma}^2_{(C,k)}/(n_{C,k} - 1)}} \right),$$

where $n_{T,k}, n_{C,k}, \bar{y}_{(T,k)}, \bar{y}_{(C,k)}, \hat{\sigma}^2_{(T,k)}$, and $\hat{\sigma}^2_{(C,k)}$ are sample sizes, sample means and variances for treatment and control group respectively in the $k$th block, $n_k$ is the total sample size in the $k$th block, and $n$ is the total sample size across all blocks.

For binary outcomes, Simon and Simon (2013) proposed the following statistic:

$$\frac{1}{\sqrt{n/2}} \sum_{k \leq K} \sqrt{n_k/2} \left( \frac{\hat{p}_{(T,k)} - \hat{p}_{(C,k)}}{2\sqrt{\hat{p}_{(pool,k)}(1 - \hat{p}_{(pool,k)})/n_k}} \right),$$

where $\hat{p}_{(T,k)}$ and $\hat{p}_{(C,k)}$ are sample success proportions in treatment and control group respectively in the $k$th block, and $\hat{p}_{(pool,k)} = (\hat{p}_{(T,k)} + \hat{p}_{(C,k)})/2$.

Simon (2015) discussed three scenarios for the adaptive enrichment design that are more effective than standard designs. The intent for standard designs is to develop a treatment that can treat the entire population with a specific disease. However, the standard designs have been unsuccessful for many cases because causal mechanisms for the same disease may be different among subjects and there is only a small subset of the population that benefit from the treatment. Therefore, targeted treatments that only treat a subset of the diseased population are created, and approaches for the three scenarios showing how to select the subset are presented as following. The first scenario is "single categorical biomarker", which defines strata for patients. It runs a group sequential trial and drops strata at interim analyses; that is, if the treatment shows ineffective in some strata, patients from these strata will not be enrolled again in the trial. The second scenario is "single continuous biomarker with unknown cut point". If this biomarker is broken into several predetermined discrete categories, the method used in the first scenario can be adopted here. Otherwise, the Simon block-sequential approach (Simon design, or model-based adaptive enrichment design) can be applied, which builds two models separately under treatment and control groups using response as the dependent variable and covariates as the

independent variables, and updates the models sequentially by blocks with the first block enrolled without restriction. If the number of subjects in the first block is too small, instead of enrolling those might benefit, Simon (2015) suggested to enroll all subjects except those do not benefit with strong evidence. The third scenario is "multidimensional biomarkers / combining multiple candidate biomarkers". The Simon designs can be also used under this scenario by modeling the response as a function of multiple biomarkers in both treatment and control groups.

However, because the null hypothesis for the Simon design is "there exists no subgroup for which treatment is more effective than control", it is impossible to know which subgroup shows significance if the null hypothesis is rejected, so Simon (2015) recommended using the characteristics of population from the final stage as the enrollment criteria. Also, there exist difficulties estimating the treatment effect in the target population because of the selection bias. With updating enrollment criteria sequentially, the whole process can become very complicated and time consuming. Therefore, an improvement of the Simon design is needed to address the above concerns.

#### 1.2.2.4 SMART Design

Sequential Multiple Assignment Randomized Trials (SMARTs) involve multiple intervention stages or multi-stage randomized trial, and each participant moves through the multiple stages; each stage corresponds to one of the critical decisions

involved in the adaptive intervention; each participant is randomly (re)assigned to one of intervention options at each stage (Lei et al., 2012). This design was developed for building optimal adaptive interventions (Lavori and Dawson, 2000, 2004; Murphy, 2005). Adaptive interventions can be also called dynamic treatment regimes or adaptive/multi-stage treatment strategies. An adaptive intervention is a sequence of individually tailored decision rules which are based on patients' characteristics or clinical presentation to alter the type or the dosage of the intervention offered to patients at critical decision points in the course of care, and then repeatedly adjusted over time in response to their ongoing performance (Almirall et al., 2011; Lei et al., 2012). Lei et al. (2012) also provided four elements for an adaptive intervention: (I) a sequence of critical decisions in a patient's care; (II) a set of possible intervention options at each critical decision point; (III) a set of tailoring variables for indicating when the intervention should be altered and identifying which intervention option is the best; (IV) a sequence of decision rules, one rule per critical decision. By making sequential decisions according to patients' characteristics and intermediate factors during the intervention such as a patient's response and adherence, this approach can be helpful to improve clinical practice because it solves the problem that patients' different response to an intervention, or the changing effectiveness of an intervention to a patient. Although SMART is an innovative name developed in recent years, a lot of SMART designs have been conducted. Four selected examples of SMART studies which are completed were discussed by Lei et al. (2012): (I) the Adaptive "Characterizing Cognition in Nonverbal Individuals with Autism" (CCNIA) Developmental and Augmented Intervention (Kasari, 2009) for school age,

nonverbal, children with autism-spectrum disorders; (II) the Adaptive Pharmacological and Behavioral Treatments for children with attention-deficit hyperactivity disorder (ADHD) (Nahum-Shani et al., 2012); (III) the Adaptive Reinforcement-Based Treatment for Pregnant Drug Abusers (RBT) (Jone, 2010); (IV) the Extending Treatment Effectivceness of Naltrexone (ExTENd) study for alcohol dependent individuals (Oslin, 2005).

### 1.2.3 Subgroup Identification

Because subjects with the same disease may have different causal mechanisms, there is only a subset of the population that benefit from the treatment. This subset can be determined based on subjects' gender, age, geographical settings, genomics, and other covariates which can be also called biomarkers. Therefore, subgroup identification is defined as identify subsets of a population by rules based on subjects' biomarkers (Lipkovich et al., 2017). The subgroup can be also defined as any subset of the recruited patient population that falls into the same category with regard to one or more biomarkers (Alemayehu et al., 2017). Pocock et al. (2002) mentioned that there are about 70 % of 50 trial reports in four major journals containing some results of subgroup analyses in late-stage clinical trials. However, identifying a subgroup that benefits from the treatment based on a selection model before the phase III clinical trial according to the information from the phase II and prior scientific knowledge is our main interest. How to identify a subgroup that maximize benefits from a treatment and how to select biomarkers becomes an important step for

personalized/precision medicine which is based on subjects' biomarkers in clinical trials. Therefore, the selection model and corresponding biomarkers used for selecting patients into the phase III trial should be pre-specified in phase III protocol, which is called confirmatory subgroup analysis. Confirmatory subgroup analyses and exploratory subgroup analyses are the two classifications commonly used for subgroup identification. For confirmatory subgroup analyses which are guidance-driven, subgroups are pre-specified in the protocol based on biomarkers and type I error rate needs to be controlled for multiple hypothesis tests problem (Lipkovich et al., 2017). In the following, we will only focus on exploratory subgroup analyses which are data-driven without pre-specifying the subgroups and are commonly used in phase III clinical trials to fully understand statistical methods on subgroup identification for building selection model in Chapter 3.

#### 1.2.3.1 Univariate Regression Model

Let $\mathbf{X}, Y, Z$ be random variables on covariates, outcome, and treatment assignment for an arbitrary patient. For the $i$th ($i = 1, \ldots, n$) patient in a clinical trial comparing a treatment and a control, let $z_i$ be the treatment actually received, where $z_i = 1$ indicates receiving treatment while $z_i = 0$ indicates receiving control; $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ be a vector of $p$ observed baseline biomarkers; $y_i$ be the observed continuous response variable; $s(\mathbf{x}_i)$ be selection option, where $s(\mathbf{x}_i) = 0$ or 1 based on $\mathbf{x}_i$. For example, $s(\mathbf{X}) = I\{X_1 > c_1\}$ indicates that only patients with biomarker $X_1$ greater than $c_1$ are selected. Then, the estimated subgroup can be obtained by

$\hat{s}(\mathbf{X}) = I\{x_{i1} > c_1, i = 1, \ldots, n\}$ (Lipkovich et al., 2017).

Let $f(\mathbf{x}, z) = E(Y|\mathbf{X} = \mathbf{x}, Z = z)$ be the expected outcome for a patient with information $\mathbf{x}$ that were to receive treatment $z$. The expected outcome can be written as (Lipkovich et al., 2017),

$$f(\mathbf{x}, z) = f(\mathbf{x}, 0) + (f(\mathbf{x}, 1) - f(\mathbf{x}, 0))\, z.$$

The above equation includes two parts, the main effect plus an interaction term of treatment effect and treatment assignment.

In general, the outcome function can be written as,

$$f(\mathbf{x}, z) = l\{g(\mathbf{x}) + m[h(\mathbf{x})z]\},$$

where $g(\cdot)$ is a prognostic effect function, $h(\cdot)$ is a predictive effect function, and $m(\cdot)$ is a monotone function. Prognostic effect evaluates a patient's outcome no matter what the treatment is, while predictive effect evaluates treatment effects (treatment-modifying covariates, or treatment-moderators) of biomarkers (Ondra et al., 2016).

The univariate regression model using $t$-test is the simplest selection method which is only based on a single biomarker in each model. First, a series of univariate regression models with terms of a single biomarker, treatment, and biomarker-by-treatment interaction are fitted. For example, a univariate regression model on biomarker $X_1$ for the $i$th patient can be written as, $y_i = \beta_0 + \beta_1 x_{i1} + I\{x_{i1} > c\}z_i$, where $I\{x_{i1} > c\} = E(Y|\mathbf{X} = x_{i1}, Z = 1) - E(Y|\mathbf{X} = x_{i1}, Z = 0)$ is treatment effect.

Second, the interaction term is tested based on the significance level of 0.1. Third, biomarkers with significant interaction term are chosen for defining subgroups. For binary biomarkers, subgroups can be defined according to the model. For continuous or ordinal biomarkers, dichotomization should be considered based on appropriately defined 'optimal' cutoff values or clinically relevant cutoffs before subgroups are defined (Lipkovich et al., 2017). However, if there exist interactions among biomarkers, the performance of this method is poor.

#### 1.2.3.2 Tree Based Regression Model

Tree based regression models are one of the best and mostly used supervised learning methods which can be completely nonparametric (Analytics Vidhya Content Team, 2016). Popular used methods are decision trees, random forest, and gradient boosting. The decision trees, which can be applied to both classification and regression problems, is also called Classification and Regression Tree (CART). It is one of the simplest methods introduced by Breiman et al. (1984). If the outcome is a categorical variable, it is classification tree; otherwise, it is regression tree. With a set of candidate biomarkers and binary treatment indicator, the tree based regression models have high-order interaction effects and subgroups can be defined by multiple biomarkers (Lipkovich et al., 2017). Also, cutoffs do not need to be pre-specified for continuous or ordinal biomarkers, which can be estimated in the process of partitioning the data (Lipkovich et al., 2017). By partitioning the data recursively, tree based models can obtain increasingly homogenous groups by minimizing resid-

ual sums of squares (RSS) for regression tree and Gini index or cross-entropy for classification tree. Finally, this process partitions biomarkers into non-overlapping regions with no region contains more than five patients, which are known as terminal nodes. With the 'overgrown' tree, cross-validation is used to prune the tree to get the optimal subtree. Any patient can fall into only one region according to his/her biomarkers, and the outcome can be predicted by the mean of the outcome values within this region (Lipkovich et al., 2017). Decision trees are easy to explain by displaying the tree graphically, but the prediction accuracy may be poor. Although the prediction accuracy can be improved dramatically with complicated methods, the results become hard to interpret and the cost on computation is high. Most importantly, the generalizability of tree based models is poor with small datasets from phase II clinical trials, which results in choosing patients who may not benefit from treatment into the phase III clinical trial.

#### 1.2.3.3 Optimal Treatment Regimes

A patient's treatment option is made by the patient's characteristics. Decisions based on synthesizing all information of a patient can lead to the best outcome for the patient. For example, Gail and Simon (1985) used the data from a trial conducted by the National Surgical Adjuvant Breast and Bowel Project comparing L-Phenylalanine mustard, 5-Fluorouracil, Tamoxifen (PFT) and L-Phenylalanine mustard, 5-Fluorouracil (PF) in patients with primary operable breast cancer and positive nodes (Fisher et al., 1983). Investigators from this project found "evidence

for a heterogeneity in response to PFT therapy that is both age and progesterone receptor dependent". Gail and Simon (1985) developed a likelihood ratio test for qualitative interaction by partitioning the data into four subgroups defined by age and progesterone receptor levels. The results showed that patients less than 50 years old with progesterone receptor levels less than 10 fmole have better results on the treatment PF while the others have better results on PFT. In other words, if age $< 50$ years and progesterone $< 10$ fmole for a patient, give this patient treatment $z_1$ (PF); otherwise, give $z_0$ (PFT). This example is the case of single decision point. We will focus on single decision point in the following (Tstiatis and Davidian, 2016).

Let $\mathbf{x}$ be all available information on a patient; $z_0$ and $z_1$ be the two treatment options, $\{z_0, z_1\} = \{0, 1\}$; $d(\mathbf{x})$ be the treatment regime, which can be 0 or 1 based on $\mathbf{x}$.

For example, if $\mathbf{x} = \{\text{Age}, \text{WBC}, \text{Gender}\}$, then $d(\mathbf{x}) = I\{\text{Age} < 50, \text{WBC} < 10, \text{and Gender} = \text{Female}\}$, which involves cut-offs, or $d(\mathbf{x}) = I\{\text{Age} + 8\text{WBC} + 0.5\text{Gender} - 60 > 0\}$, which involves a linear combination.

Let $\mathcal{D}$ be the class of all possible treatment regimes. The optimal regime is denoted as $d^{opt} \in \mathcal{D}$. If a patient received the optimal treatment, the patient's expected outcome would be as large as possible given his/her available information. If all patients received the optimal treatment, the expected outcome for the population

would be as large as possible. It can be written as

$$E\{Y^*(d)|\mathbf{X}=\mathbf{x}\} \leq E\{Y^*(d^{opt})|\mathbf{X}=\mathbf{x}\}; E\{Y^*(d)\} \leq E\{Y^*(d^{opt})\},$$

where $Y^*(d)$ is the potential outcome for a patient with baseline information $\mathbf{X}$ that were to receive the treatment $d$ developed by Rubin (1974). Splawa-Neyman (1923) started the Neyman-Rubin framework that each subject has two potential outcomes, but the observed outcome can be only under treatment or control group on randomized studies. Rubin (1974) developed the model into a general framework for observational studies. Thus, $Y^*(1)$ is the outcome that would be achieved if a patient were to receive the treatment 1; $Y^*(0)$ is the outcome that would be achieved if a patient were to receive the treatment 0. With the combination of $Y^*(1)$ and $Y^*(0)$, $Y^*(d)$ can be written as $Y^*(d) = Y^*(1)d(\mathbf{X}) + Y^*(0)(1-d(\mathbf{X}))$. $E\{Y^*(1)\}$ is the expected outcome if all patients in the population were to receive the treatment 1; $E\{Y^*(0)\}$ is the expected outcome if all patients in the population were to receive the treatment 0. $E\{Y^*(d)|\mathbf{X}=\mathbf{x}\}$ is the expected outcome for a patient with information $\mathbf{x}$ that were to receive the treatment $d$. $E\{Y^*(d)\} = E\left[E\{Y^*(d)|\mathbf{X}\}\right]$ is the expected outcome for the population if all patients were to receive the treatment $d$. Therefore, $d^{opt}$ can make $E\{Y^*(d)\}$ have the largest value among $d \in \mathcal{D}$.

$E\{Y^*(d)\}$ can be also called as the value of treatment $d$, which is written as

$V(d) = E\{Y^*(d)\}$. Thus,

$$\begin{aligned} V(d) &= E\{Y^*(d)\} = E\left[E\{Y^*(d)|\mathbf{X}\}\right] \\ &= E\left[E\{Y^*(1)d(\mathbf{X}) + Y^*(0)(1-d(\mathbf{X}))|\mathbf{X}\}\right] \\ &= E\left[E\{Y^*(1)|\mathbf{X}\}d(\mathbf{X}) + E\{Y^*(0)|\mathbf{X}\}(1-d(\mathbf{X}))\right]. \end{aligned}$$

Then, the optimal regime can be written as,

$$d^{opt}(\mathbf{x}) = I\left[E\{Y^*(1)|\mathbf{X}=\mathbf{x}\} > E\{Y^*(0)|\mathbf{X}=\mathbf{x}\}\right];$$

that is, the optimal regime assigns the treatment to a patient that his/her expected outcome would be larger conditional on $\mathbf{x}$.

For observational studies, let $\mathbf{X}$ be the baseline covariates, $Z$ be the treatment actually received (0 or 1), $Y$ be the observed outcome. Since the mechanism of treatment assignment is unknown, two assumptions are considered in the following (Rubin, 1986; Rosenbaum and Rubin, 1983):

a. Consistency assumption: $Y = Y^*(1)Z + Y^*(0)(1-Z)$;

b. No unmeasured confounders assumption: $Y^*(0), Y^*(1) \perp Z|\mathbf{X}$.

The first assumption shows that the potential outcomes for a subject will be the same irrespective of the mechanism used to assign the treatment to that subject and irrespective of which treatments the other subjects receive. The second assumption shows that the potential outcomes are conditionally independent of treatment assignments given measured covariates.

With the two assumptions, we have:

$$\begin{aligned} E\{Y^*(1)\} &= E[E\{Y^*(1)|\mathbf{X}\}] = E[E\{Y^*(1)|\mathbf{X}, Z=1\}] \\ &= E[E\{Y^*(1)Z + Y^*(0)(1-Z)|\mathbf{X}, Z=1\}] = E[E\{Y|\mathbf{X}, Z=1\}], \end{aligned}$$

$$\begin{aligned} E\{Y^*(0)\} &= E[E\{Y^*(0)|\mathbf{X}\}] = E[E\{Y^*(0)|\mathbf{X}, Z=0\}] \\ &= E[E\{Y^*(1)Z + Y^*(0)(1-Z)|\mathbf{X}, Z=0\}] = E[E\{Y|\mathbf{X}, Z=0\}], \end{aligned}$$

$$\begin{aligned} V(d) &= E\{Y^*(d)\} = E\left[E\{Y^*(d)|\mathbf{X}\}\right] \\ &= E[E\{Y^*(1)d(\mathbf{X}) + Y^*(0)(1-d(\mathbf{X}))|\mathbf{X}\}] \\ &= E[E\{Y^*(1)|\mathbf{X}\}d(\mathbf{X}) + E\{Y^*(0)|\mathbf{X}\}(1-d(\mathbf{X}))] \\ &= E[E\{Y|\mathbf{X}, Z=1\}d(\mathbf{X}) + E\{Y|\mathbf{X}, Z=0\}(1-d(\mathbf{X}))], \end{aligned}$$

$$d^{opt}(\mathbf{x}) = I\left[E\{Y|\mathbf{X}=\mathbf{x}, Z=1\} > E\{Y|\mathbf{X}=\mathbf{x}, Z=0\}\right].$$

Let $Q(\mathbf{x}, z) = E\{Y|\mathbf{X} = \mathbf{x}, Z = z\}$, then $Q(\mathbf{x}, 1) = E\{Y|\mathbf{X} = \mathbf{x}, Z = 1\}$ and $Q(\mathbf{x}, 0) = E\{Y|\mathbf{X} = \mathbf{x}, Z = 0\}$; so $d^{opt}(\mathbf{x})$ and $V(d)$ can be also written as $d^{opt}(\mathbf{x}) = I\left[Q(\mathbf{x}, 1) > Q(\mathbf{x}, 0)\right]$, and $V(d) = E[Q(\mathbf{x}, 1)d(\mathbf{x}) + Q(\mathbf{x}, 0)(1 - d(\mathbf{x}))]$.

$Q(\mathbf{x}, z)$ is not known, but it can be modeled as a linear or logistic regression $Q(\mathbf{x}, z; \boldsymbol{\beta})$, and $\boldsymbol{\beta}$ can be estimated by least squares, maximum likelihood, or other appropriate methods. For example, $Q(\mathbf{x}, z; \boldsymbol{\beta}) = \beta_0 + \boldsymbol{\beta}_1\mathbf{x} + \beta_2 z + \boldsymbol{\beta}_3 z\mathbf{x}$.

With the known model $Q(\mathbf{x}, z; \hat{\boldsymbol{\beta}}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + z(\hat{\beta}_4 + \hat{\beta}_5 x_1 + \hat{\beta}_6 x_2 + \hat{\beta}_7 x_3)$, $V(d)$ can be written as

$$\hat{V}(d) = n^{-1} \sum_{i=1}^{n} \left[ Q(\mathbf{X}_i, 1; \hat{\boldsymbol{\beta}}) d(\mathbf{X}_i) + Q(\mathbf{X}_i, 0; \hat{\boldsymbol{\beta}})(1 - d(\mathbf{X}_i)) \right].$$

Then, $d^{opt}(\mathbf{x})$ and $V(d^{opt})$ can be estimated as $\hat{d}_Q^{opt}(\mathbf{x}) = I\left[ Q(\mathbf{x}, 1; \hat{\boldsymbol{\beta}}) > Q(\mathbf{x}, 0; \hat{\boldsymbol{\beta}}) \right]$ and

$$\hat{V}(\hat{d}^{opt}) = n^{-1} \sum_{i=1}^{n} \left[ Q(\mathbf{X}_i, 1; \hat{\boldsymbol{\beta}}) \hat{d}^{opt}(\mathbf{X}_i) + Q(\mathbf{X}_i, 0; \hat{\boldsymbol{\beta}})(1 - \hat{d}^{opt}(\mathbf{X}_i)) \right].$$

Also, the estimated optimal regime can be simplified as $\hat{d}_\beta^{opt}(\mathbf{x}) = I[\hat{\beta}_4 + \hat{\beta}_5 x_1 + \hat{\beta}_6 x_2 + \hat{\beta}_7 x_3 > 0]$ with a subset of elements of $\mathbf{x}$. It can be rewritten as $I[x_1 > \hat{\eta}_0 + \hat{\eta}_1 x_2 + \hat{\eta}_2 x_3]$ if $\hat{\beta}_5$ is positive, or $I[x_1 < \hat{\eta}_0 + \hat{\eta}_1 x_2 + \hat{\eta}_3 x_3]$ if $\hat{\beta}_5$ is negative, with $\hat{\eta}_0 = -\hat{\beta}_4/\hat{\beta}_5, \hat{\eta}_1 = -\hat{\beta}_6/\hat{\beta}_5$ and $\hat{\eta}_2 = -\hat{\beta}_7/\hat{\beta}_5$ (Zhang et al., 2012). In general, $\boldsymbol{\eta}$ is a function of $\boldsymbol{\beta}$, which can be written as, $\boldsymbol{\eta} = \eta(\boldsymbol{\beta})$. Therefore, with the correct fitted outcome regression model $Q(\mathbf{x}, z; \boldsymbol{\beta})$, $d_\eta^{opt}(\mathbf{x}) = d(\mathbf{x}, \boldsymbol{\eta}^{opt})$ can be estimated for $d^{opt}$ to make $E\{Y^*(d_\eta)\}$ have the largest value among $d_\eta \in \mathcal{D}_\eta$. If the outcome regression model is incorrectly fitted, $d^{opt}$ may not be in $\mathcal{D}_\eta$ and $\hat{d}_\eta^{opt}(\mathbf{x})$ may be far away from $d^{opt}$. It also happens when outcome regression models are too complex, so Zhang et al. (2012) proposed an alternative method using only a key subset of elements of $\mathbf{X}$ based on interpretability, cost, and feasibility in practice to define a class of regimes by $\boldsymbol{\eta}$. For example, $d_\eta(\mathbf{x}) = d(\mathbf{x}, \boldsymbol{\eta}) = I\{x_1 < \eta_0, x_2 < \eta_1, x_3 < \eta_2\}$ without using a regression model. Therefore, the optimal regime $d_\eta^{opt}$ based on the mis-specified outcome regression model with parameter estimators $\hat{\boldsymbol{\beta}}$ can lead to poor performance on $E\{Y^*(d_\eta^{opt})\}$.

#### 1.2.3.4 Optimal Treatment Regimes Based on IPWE

From the previous section, $\boldsymbol{\eta}^{opt}$ should be estimated to obtain the maximum value of $E\{Y^*(d_\eta)\}$. With the estimator $\hat{\boldsymbol{\eta}}^{opt}$, the optimal regime $d_\eta^{opt}$ is estimated by $\hat{d}_\eta^{opt}(\mathbf{X}) = d(\mathbf{X}, \hat{\boldsymbol{\eta}}^{opt})$. With fixed $\boldsymbol{\eta}$, let $C_\eta = Zd(\mathbf{X}, \boldsymbol{\eta}) + (1-Z)(1-d(\mathbf{X}, \boldsymbol{\eta}))$, $C_\eta$ can be 1 or 0. For subjects with $C_\eta = 1$, they receive treatment 0 or 1 following the regime $d_\eta$, which means their outcomes are observed with $Y^*(d_\eta) = Y$. For the others with $C_\eta = 0$, their outcomes $Y^*(d_\eta)$ following the regime $d_\eta$ are unknown, which means they are missing. Only observed subjects are used for estimating $E\{Y^*(d_\eta)\}$. Since $C_\eta$ is a function of $\{Z, \mathbf{X}\}$, and $Y^*(d_\eta) = Y^*(1)d(\mathbf{X}, \boldsymbol{\eta}) + Y^*(0)(1-d(\mathbf{X}, \boldsymbol{\eta}))$ is a function of $\{Y^*(0), Y^*(1), \mathbf{X}\}$ with the fixed $\boldsymbol{\eta}$, $C_\eta$ is independent of $Y^*(d_\eta)$ given $\mathbf{X}$ under the second assumption from the last section, which means that the missing mechanism on $Y^*(d_\eta)$ is missing at random (MAR) (Cao et al., 2009; Zhang et al., 2012). Therefore, the probability of being observed given $\mathbf{X}$ can be written as

$$\begin{aligned} P(C_\eta = 1|\mathbf{X}) &= P(Zd(\mathbf{X}, \boldsymbol{\eta}) + (1-Z)(1-d(\mathbf{X}, \boldsymbol{\eta})) = 1|\mathbf{X}) \\ &= d(\mathbf{X}, \boldsymbol{\eta})P(Z=1|\mathbf{X}) + (1-d(\mathbf{X}, \boldsymbol{\eta}))P(Z=0|\mathbf{X}) \\ &= d(\mathbf{X}, \boldsymbol{\eta})e(\mathbf{X}) + (1-d(\mathbf{X}, \boldsymbol{\eta}))(1-e(\mathbf{X})), \end{aligned}$$

where $e(\mathbf{X}) = P(Z=1|\mathbf{X})$ is the propensity score for treatment 1 group. Therefore, $P(C_\eta = 1|\mathbf{X}) = e_c(\mathbf{X}; \boldsymbol{\eta}) = d(\mathbf{X}, \boldsymbol{\eta})e(\mathbf{X}) + (1-d(\mathbf{X}, \boldsymbol{\eta}))(1-e(\mathbf{X}))$. For clinical trials, the propensity score $e(\mathbf{X})$ is known as a constant. For observational studies, the propensity score $e(\mathbf{X})$ is not known and need to be estimated through a parametric model $e(\mathbf{X}; \boldsymbol{\gamma})$. With the estimated propensity score $e(\mathbf{X}; \hat{\boldsymbol{\gamma}})$ and the fixed $\boldsymbol{\eta}$, the

inverse probability weighted estimator (IPWE) can be written as (Lunceford and Davidian, 2004; Cao et al., 2009; Zhang et al., 2012)

$$\begin{aligned}\hat{V}_{IPW}(d_\eta) &= n^{-1}\sum_{i=1}^{n}\frac{C_{\eta,i}Y_i}{e_c(\mathbf{X}_i;\boldsymbol{\eta},\hat{\boldsymbol{\gamma}})} \\ &= n^{-1}\sum_{i=1}^{n}\frac{C_{\eta,i}Y_i}{d(\mathbf{X}_i,\boldsymbol{\eta})e(\mathbf{X}_i;\hat{\boldsymbol{\gamma}})+(1-d(\mathbf{X}_i,\boldsymbol{\eta}))(1-e(\mathbf{X}_i;\hat{\boldsymbol{\gamma}}))} \\ &= n^{-1}\sum_{i=1}^{n}\frac{C_{\eta,i}Y_i}{e(\mathbf{X}_i;\hat{\boldsymbol{\gamma}})^{Z_i}[1-e(\mathbf{X}_i;\hat{\boldsymbol{\gamma}})]^{1-Z_i}}.\end{aligned}$$

By maximizing the above IPW estimator $\hat{V}_{IPW}(d_\eta)$, the optimal regime $d_\eta^{opt}$ can be estimated. If the propensity score model $e(\mathbf{X};\hat{\boldsymbol{\gamma}})$ is correctly specified with $e(\mathbf{X};\hat{\boldsymbol{\gamma}}) = e_0(\mathbf{X})$, the IPW estimator is consistent with $E[\hat{V}_{IPW}(d_\eta)] = E(Y^*(d_\eta))$. The proof is given as following (Lunceford and Davidian, 2004).

$$\begin{aligned}E[\hat{V}_{IPW}(d_\eta)] &= E\left[\frac{C_\eta Y}{e_{c0}(\mathbf{X};\boldsymbol{\eta})}\right] = E\left\{E\left[\frac{C_\eta Y}{e_{c0}(\mathbf{X};\boldsymbol{\eta})}|Y^*(d_\eta),\mathbf{X}\right]\right\} \\ &= E\left\{E\left[\frac{C_\eta Y^*(d_\eta)}{e_{c0}(\mathbf{X};\boldsymbol{\eta})}|Y^*(d_\eta),\mathbf{X}\right]\right\} = E\left\{E\left[\frac{I\{C_\eta=1\}Y^*(d_\eta)}{e_{c0}(\mathbf{X};\boldsymbol{\eta})}|Y^*(d_\eta),\mathbf{X}\right]\right\} \\ &= E\left\{E\left[\frac{I\{Zd(\mathbf{X},\boldsymbol{\eta})+(1-Z)(1-d(\mathbf{X},\boldsymbol{\eta}))=1\}Y^*(d_\eta)}{e_{c0}(\mathbf{X};\boldsymbol{\eta})}|Y^*(d_\eta),\mathbf{X}\right]\right\} \\ &= E\left\{\frac{Y^*(d_\eta)}{e_{c0}(\mathbf{X};\boldsymbol{\eta})}E\left[I\{Zd(\mathbf{X},\boldsymbol{\eta})+(1-Z)(1-d(\mathbf{X},\boldsymbol{\eta}))=1\}|Y^*(d_\eta),\mathbf{X}\right]\right\} \\ &= E\left\{\frac{Y^*(d_\eta)}{e_{c0}(\mathbf{X};\boldsymbol{\eta})}\left[d(\mathbf{X},\boldsymbol{\eta})E(Z|Y^*(d_\eta),\mathbf{X})+(1-d(\mathbf{X},\boldsymbol{\eta}))(1-E(Z|Y^*(d_\eta),\mathbf{X}))\right]\right\} \\ &= E\left\{\frac{Y^*(d_\eta)}{e_{c0}(\mathbf{X};\boldsymbol{\eta})}\left[d(\mathbf{X},\boldsymbol{\eta})e_0(\mathbf{X})+(1-d(\mathbf{X},\boldsymbol{\eta}))(1-e_0(\mathbf{X}))\right]\right\} = E(Y^*(d_\eta)),\end{aligned}$$

where $e_{c0}(\mathbf{X};\boldsymbol{\eta}) = d(\mathbf{X},\boldsymbol{\eta})e_0(\mathbf{X}) + (1-d(\mathbf{X},\boldsymbol{\eta}))(1-e_0(\mathbf{X}))$.

The IPW estimator is always consistent for randomized clinical trials with constant true propensity scores. However, for observational studies, the IPW estimator

is not consistent if the propensity score model is misspecified. With the misspecified propensity score model, an augmentation term including outcome regression model is added to the IPW estimator to improve efficiency and provide another protection, which will be presented in the following section.

#### 1.2.3.5 Optimal Treatment Regimes Based on Doubly Robust IPWE

To provide protection against the misspecification of propensity score model leading to the inconsistent IPW estimator from the last section and improve efficiency, an augmentation term including outcome regression model is added. If the outcome regression model is correctly specified, the new estimator is consistent no matter the propensity score model is right or wrong, and vice versa. Therefore, either the propensity score model or the outcome regression model is right, the new estimator is consistent, which is called doubly robust inverse probability weighted estimator (DRIPWE) with double protections. Based on the IPWE, the DRIPWE is written as following (Robins et al., 1994; Cao et al., 2009; Zhang et al., 2012),

$$\hat{V}_{DRIPW}(d_\eta) = n^{-1}\sum_{i=1}^{n}\left\{\frac{C_{\eta,i}Y_i}{e_c(\mathbf{X}_i;\boldsymbol{\eta},\hat{\boldsymbol{\gamma}})} - \frac{C_{\eta,i}-e_c(\mathbf{X}_i;\boldsymbol{\eta},\hat{\boldsymbol{\gamma}})}{e_c(\mathbf{X}_i;\boldsymbol{\eta},\hat{\boldsymbol{\gamma}})}m(\mathbf{X}_i;\boldsymbol{\eta},\hat{\boldsymbol{\beta}})\right\},$$

where $e_c(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\gamma}}) = d(\mathbf{X},\boldsymbol{\eta})e(\mathbf{X},\hat{\boldsymbol{\gamma}}) + (1-d(\mathbf{X},\boldsymbol{\eta}))(1-e(\mathbf{X},\hat{\boldsymbol{\gamma}}))$, $m(\mathbf{X}_i;\boldsymbol{\eta},\hat{\boldsymbol{\beta}}) = E[Y^*(\hat{d}_\eta)|\mathbf{X}] = Q(\mathbf{X},1;\hat{\boldsymbol{\beta}})d(\mathbf{X},\boldsymbol{\eta}) + Q(\mathbf{X},0;\hat{\boldsymbol{\beta}})(1-d(\mathbf{X},\boldsymbol{\eta}))$.

By maximizing the above DRIPW estimator $\hat{V}_{DRIPW}(d_\eta)$, the optimal regime $d_\eta^{opt}$ can be estimated. If the propensity score model $e(\mathbf{X};\hat{\boldsymbol{\gamma}})$ is correctly specified and the outcome regression model may be misspecified, the DRIPW estimator is consistent

with $E[\hat{V}_{DRIPW}(d_\eta)] = E(Y^*(d_\eta))$. The proof is given as following.

$$E[\hat{V}_{DRIPW}(d_\eta)] = E\left[\frac{C_\eta Y}{e_{c0}(\mathbf{X};\boldsymbol{\eta})}\right] - E\left[\frac{C_\eta - e_{c0}(\mathbf{X};\boldsymbol{\eta})}{e_{c0}(\mathbf{X};\boldsymbol{\eta})} m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}})\right] = E_1 - E_2;$$

$E_1 = E(Y^*(d_\eta))$, which has been proved in the last section.

$$\begin{aligned}
E_2 &= E\left[\frac{C_\eta - e_{c0}(\mathbf{X};\boldsymbol{\eta})}{e_{c0}(\mathbf{X};\boldsymbol{\eta})} m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}})\right] \\
&= E\left\{E\left[\frac{C_\eta - e_{c0}(\mathbf{X};\boldsymbol{\eta})}{e_{c0}(\mathbf{X};\boldsymbol{\eta})} m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}})|\mathbf{X}\right]\right\} \\
&= E\left\{E\left[\frac{I\{Zd(\mathbf{X},\boldsymbol{\eta}) + (1-Z)(1-d(\mathbf{X},\boldsymbol{\eta})) = 1\} - e_{c0}(\mathbf{X};\boldsymbol{\eta})}{e_{c0}(\mathbf{X};\boldsymbol{\eta})} m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}})|\mathbf{X}\right]\right\} \\
&= E\left\{\frac{m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}})}{e_{c0}(\mathbf{X};\boldsymbol{\eta})} E[I\{Zd(\mathbf{X},\boldsymbol{\eta}) + (1-Z)(1-d(\mathbf{X},\boldsymbol{\eta})) = 1\} - e_{c0}(\mathbf{X};\boldsymbol{\eta})|\mathbf{X}]\right\} \\
&= E\left\{\frac{m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}})}{e_{c0}(\mathbf{X};\boldsymbol{\eta})} [d(\mathbf{X},\boldsymbol{\eta})E(Z|\mathbf{X}) + (1-d(\mathbf{X},\boldsymbol{\eta}))(1-E(Z|\mathbf{X})) - e_{c0}(\mathbf{X};\boldsymbol{\eta})]\right\} \\
&= E\left\{\frac{m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}})}{e_{c0}(\mathbf{X};\boldsymbol{\eta})} [d(\mathbf{X},\boldsymbol{\eta})e_0(\mathbf{X}) + (1-d(\mathbf{X},\boldsymbol{\eta}))(1-e_0(\mathbf{X})) - e_{c0}(\mathbf{X};\boldsymbol{\eta})]\right\} \\
&= E\left\{\frac{m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}})}{e_{c0}(\mathbf{X};\boldsymbol{\eta})} [e_{c0}(\mathbf{X};\boldsymbol{\eta}) - e_{c0}(\mathbf{X};\boldsymbol{\eta})]\right\} \\
&= 0;
\end{aligned}$$

thus,

$$E[\hat{V}_{DRIPW}(d_\eta)] = E(Y^*(d_\eta)) - 0 = E(Y^*(d_\eta)),$$

where $e_{c0}(\mathbf{X};\boldsymbol{\eta}) = d(\mathbf{X},\boldsymbol{\eta})e_0(\mathbf{X}) + (1-d(\mathbf{X},\boldsymbol{\eta}))(1-e_0(\mathbf{X}))$.

If the outcome regression model $m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}})$ is correctly specified with $m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}}) = m_0(\mathbf{X};\boldsymbol{\eta})$ and the propensity score model may be misspecified, the DRIPW estima-

tor is consistent with $E[\hat{V}_{DRIPW}(d_\eta)] = E(Y^*(d_\eta))$. The proof is given as following.

$$\begin{aligned}
E[\hat{V}_{DRIPW}(d_\eta)] &= E\{E[\hat{V}_{DRIPW}(d_\eta)|Y^*(d_\eta), \mathbf{X}]\} \\
&= E\left\{E\left[\frac{C_\eta Y}{e_c(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\gamma}})}|Y^*(d_\eta), \mathbf{X}\right]\right\} - E\left\{E\left[\frac{C_\eta m_0(\mathbf{X};\boldsymbol{\eta})}{e_c(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\gamma}})}|Y^*(d_\eta), \mathbf{X}\right]\right\} \\
&+ E\left\{E[m_0(\mathbf{X};\boldsymbol{\eta})|Y^*(d_\eta), \mathbf{X}]\right\} \\
&= E_1 - E_2 + E_3;
\end{aligned}$$

$$\begin{aligned}
E_1 &= E\left\{E\left[\frac{C_\eta Y}{e_c(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\gamma}})}|Y^*(d_\eta), \mathbf{X}\right]\right\} \\
&= E\left\{E\left[\frac{C_\eta Y^*(d_\eta)}{e_c(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\gamma}})}|Y^*(d_\eta), \mathbf{X}\right]\right\},
\end{aligned}$$

$$\begin{aligned}
E_2 &= E\left\{E\left[\frac{C_\eta m_0(\mathbf{X};\boldsymbol{\eta})}{e_c(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\gamma}})}|Y^*(d_\eta), \mathbf{X}\right]\right\} \\
&= E\left\{E\left[\frac{C_\eta E(Y^*(d_\eta)|\mathbf{X})}{e_c(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\gamma}})}|Y^*(d_\eta), \mathbf{X}\right]\right\} \\
&= E\left\{E\left[\frac{C_\eta Y^*(d_\eta)}{e_c(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\gamma}})}|Y^*(d_\eta), \mathbf{X}\right]\right\},
\end{aligned}$$

$E_3 = E\left\{E[m_0(\mathbf{X};\boldsymbol{\eta})|Y^*(d_\eta), \mathbf{X}]\right\} = E\{E[E(Y^*(d_\eta)|\mathbf{X})|Y^*(d_\eta), \mathbf{X}]\} = E(Y^*(d_\eta));$

thus,

$$\begin{aligned}
E[\hat{V}_{DRIPW}(d_\eta)] &= E\left\{E\left[\frac{C_\eta Y^*(d_\eta)}{e_c(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\gamma}})}|Y^*(d_\eta), \mathbf{X}\right]\right\} - E\left\{E\left[\frac{C_\eta Y^*(d_\eta)}{e_c(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\gamma}})}|Y^*(d_\eta), \mathbf{X}\right]\right\} \\
&+ E(Y^*(d_\eta)) \\
&= E(Y^*(d_\eta)).
\end{aligned}$$

# Chapter 2

# Analysis of Longitudinal Interval Reported Binary Recurrent Event Data

## 2.1 Introduction

Longitudinal interval reported binary recurrent events data are fairly common in medical studies, including clinical trials or observational studies where subjects/ patients are assessed periodically based on fixed follow-up schedule. Even though patients are asked to visit or call the medical center at fixed follow-up time, many patients may miss a few visits throughout the study period, and report whether there is any onset of disease symptoms since last visit rather than in this specific scheduled interval (Thall and Lachin, 1988). Those symptoms are recurrent and can happen anytime, but the exact occurrence time and total count of recurrence during this reporting interval between two visits are not available. For example, females more than 20 years of age are interviewed every two years for ten years, and the event of interest is whether they have ever been pregnant (including miscarriages) since last interview. If someone miss some interviews, their reporting intervals would be longer than two years (Allison, 1982). When a recurrent event occurs on subjects who are observed only at discrete time points in clinical trials or longitudinal

observational studies, it is called panel-count data (Zhu et al., 2017). For this type of interval reported data, even though the time to recurrence is unknown due to interval censoring, the recurrent event counts in each reported interval are observed (Sun and Zhao, 2013; Zhu et al., 2017). However, for this paper, we only observe the binary status of whether or not the event of interest happens in the reporting interval, without information on the frequency counts nor on when they re-occur in this interval. And, we call this type of data, longitudinal interval reported binary recurrent event data.

The outcomes we observed, in post hip fracture infection project using Baltimore Hip studies (BHS), are longitudinal interval reported binary recurrent event data. During in-person and phone interviews, subjects were asked about status change since last interview and they might miss a few interviews. For example, "Since the last time we spoke in (Provide Month) have you ever had fever" is one of the typical questionnaire items measuring the outcome infection longitudinally. Even though infection is recurrent, we only know that at least one infection occurred at some reporting interval, with the exact time and the number of recurrent infections in the reporting interval are unknown. It is common to see this type of questionnaire in clinical trials or observational studies. The aim of this project is to investigate how the post-surgery infection developed over the first year follow-ups in hip fracture patients, and whether the infection trend differed between men and women. Considering interval reporting issue and accounting for missing interviews by the varying interval lengths between two visits across subjects are very important for

the validity of model inference.

Even though there are many statistical methods dealing with longitudinal binary data or interval censored discrete survival data, there are not many methods dealing with longitudinal interval reported binary recurrent event data. For those interval-censored discrete survival data without considering the nature of recurrent events in a longitudinal study, several statistical models are available. A discrete model developed from the Cox proportional hazards model has been used when non-recurrent failure time data (data are either exact or right-censored observations) are broadly grouped, with failure times grouped into fixed intervals $A_i$ for all patients (Kalbfleisch and Prentice, 1973; Prentice and Gloeckler, 1978). In 1986, Finkelstein developed a methodology for applying a proportional hazards model on nonrecurrent interval-censored failure time data (i.e., the exact response time is unknown, but is known within some time interval) which may be censored into overlapping and non-disjoint intervals (Finkelstein, 1986). In 2017, Austin illustrated two models on multilevel survival analysis with non-recurrent events, a "piecewise exponential survival model" assuming constant hazard within each interval and taking the length of interval into account after dividing the follow-up time into mutually exclusive intervals (Allison, 2010); and a "complementary log-log generalized linear model" (Rodriguez, 2008) for discrete survival time focusing on whether or not an event occurred within an interval without considering the length of interval (Austin, 2017).

Without considering interval reported binary recurrent event nature of the data,

only focusing on the longitudinal binary outcomes, the marginal model under GLM framework estimated by Generalized Estimating Equations (GEEs) can be used (Liang and Zeger, 1986), in particular the logistic regression model. Also, generalized linear mixed effect models are very popular for taking the population heterogeneity into account. With considering recurrent events and longitudinal nature, the hierarchical linear model (or multilevel model) developed by Snijders (1996) considered varying number of repeated measures across subjects (e.g., income, physiological measurements, et al.) due to design or missing and varying reporting intervals due to different measurement time points or missing data across subjects of longitudinal data (Snijders, 1996). Sutradhar et al. (2011) constructed a multistate model for interval-censored longitudinal data showing changes in a patient's health condition over time. With the intermittent assessment on cancer patients, the transition, which is the progression of performance status over time representing a patient's health condition, is only known within an interval, so information on transition time (the time when changes occur) is incomplete. Thus, different subjects may have different number and timing of assessments, but the outcomes are not binary reported nor recurrent. Also, the multistate methods developed under interval censoring do not consider covariates in the model which may change the transition rates (Sutradhar et al., 2011). However, for our study, we should notice that the recurrent events reported in each visit are dichotomized due to questionnaire design, also the reporting intervals between visits within subject change over time, and visiting/ interviewing schedules change across subjects too due to various missing patterns. The literature on methods dealing with this type of data are very limited.

In this paper, we developed a relatively simple and flexible longitudinal model framework to deal with this type of data, where discrete survival modeling technique and Poisson process are used to account for interval censored reporting system between longitudinal visits/interviews and binary nature of recurrent events reporting within each interval. Specifically, the probability of observing an event in an interval is based on the Poisson process of the events with the intensity follows a Cox proportional hazards model. The hazard function follows the Cox proportional hazards model allowing both baseline covariates and time-varying covariates. In another words, our hazard function is flexible enough to incorporate a subject's characteristics and time-varying nature across the longitudinal visits, and the subject's hazard function within each reporting interval stays fixed over time. This model setting simplified the joint likelihood into a generalized linear mixed effects model framework with binary responses and complementary log-log link, leading to widely available software for estimation. This method was applied to our post hip fracture infection project using Baltimore Hip studies (BHS). Section 2.2 describes the data from Baltimore Hip studies (BHS) and the infection project which motivated our work. Section 2.3 shows the theoretical development of our longitudinal model using discrete survival technique and Poisson process to deal with interval censoring problem and our longitudinal binary recurrent data. Section 2.4 shows the results of a simulation study. Section 2.5 presents an application of the proposed method to the post hip fracture infection data.

## 2.2 Data

The project motivated our study came from the seventh cohort of Baltimore Hip Studies (BHS-7), a longitudinal study investigating consequences of hip fracture with metabolic, physiologic, neuromuscular, functional, and clinical outcomes and differences between men and women in the first year of post hip fracture recovery period (Resnick et al., 2011; Magaziner, 2012; Orwig et al., 2018). Our study aimed to investigate how the post-surgery infection developed over the first-year follow-ups in hip fracture patients, and whether the infection trend differed between men and women. The BHS-7 study collected baseline health information and comprehensive assessment of psychosocial, physical, and physiological outcomes of community-dwelling patients with surgical repair of a non-pathological hip fracture from eight hospitals during 2006 and 2011. Patients were eligible if they were at age 65 or older and admitted to one of the eight study hospitals during the study period with a diagnosis of hip fracture (ICD9 code 820) (Magaziner, 2012; Orwig et al., 2018). There were 339 hip fracture patients enrolled within 15 days of admission, and women enrollment was frequency-matched to men in each hospital (Resnick et al., 2011; Orwig et al., 2018).

The outcome considered in our analysis was infection status post hip fracture surgery. Infection was defined by having at least one of the following symptoms: fever, antibiotic use, cough with green or bloody phlegm or sputum, burning with urination, cloudy urine, foul-smelling urine, bloody urine, chills, and discharge or swelling at

the surgical site. Comprehensive assessments of psychosocial, physical, and physiological outcomes, including infection outcomes, are available at baseline and at 2, 6, and 12 months post-fracture through in-person interviews. Additionally, phone interviews were conducted monthly between in-person interviews asking questions about status change since the last time spoken. This longitudinal design leads to 11 scheduled infection outcomes per patient to be measured in this study. Since monthly interviews asked questions about status change since the last time spoken and subjects might miss some interviews, this interval censored reporting system and design lead to varying lengths of intervals between repeated measures. Another nature of this data is that infection events are reported dichotomizely, i.e., we only know whether at least one infection occurred during some reporting intervals, but the exact time and number of infections are unknown.

The primary predictors considered in the model were gender, time, age, education, race (white vs. non-white), Charlson comorbidity index, body mass index, elevated white blood cell, and combined urinary tract infection. The sample size for this study was 288 with available measurements on outcome and covariates, excluding those subjects who did not have any post-discharge follow-up outcome data among 11 scheduled interviews or with missing covariates' values.

## 2.3 Methods

There are multiple effect measures for quantifying the longitudinal infection trends and cross-sectional gender differences in our longitudinal interval reported binary recurrent event data. Since the longitudinal binary outcome results from dichotomizing the process of recurrent infection events with interval censoring problem, we focus on the two popular measures log odds ratio and log hazard ratio for comparison in this paper. Following typical longitudinal model notation, let $t_{ij}$ be the discrete time for the $j$th longitudinal visit of the $i$th patient, and $0 < t_{i1} < t_{i2} < \cdots < t_{iJ_i}$, $i = 1, \cdots, n, j = 1, \cdots, J_i$, where $J_i$ indicates the $i$th patient's total number of visits within the study period. $Y^*_{ij}$ is the unobserved total count of recurrent infection events during the reporting interval $(t_{i(j-1)}, t_{ij}]$, and $Y_{ij}$ is the observed reported binary infection status at $j$th visit of $i$th patient.

$$\begin{cases} Y_{ij} = 1, & \text{if } Y^*_{ij} \geq 1 \\ Y_{ij} = 0, & \text{if } Y^*_{ij} = 0 \end{cases}, \tag{2.1}$$

$Y^*_{ij} \sim \text{Poisson}\left(\lambda(\mathbf{X}_{ij})(t_{ij} - t_{i(j-1)})\right)$ or,

$$P(Y^*_{ij} = y^*_{ij}) = e^{-\lambda(\mathbf{X}_{ij})(t_{ij} - t_{i(j-1)})} \left[\lambda(\mathbf{X}_{ij})(t_{ij} - t_{i(j-1)})\right]^{y^*_{ij}} / y^*_{ij}!. \tag{2.2}$$

$Y_{ij} = 1$ means infection occurred at least once during the interval $(t_{i(j-1)}, t_{ij}]$, even though we don't know when and how many times it happened; while, $Y_{ij} = 0$ means no infection occurrence during the reporting interval $(t_{i(j-1)}, t_{ij}]$. Assuming the occurrences of infection follows a Poisson process with intensity $\lambda(\mathbf{X}_{ij})$ in $(t_{i(j-1)}, t_{ij}]$ is a fairly common and flexible assumption, allowing the intensity vary across indi-

vidual reporting intervals defined by their available visits within patient and vary across patients. Since our recurrent infection events are binary and interval censored, $\lambda(\mathbf{X}_{ij})$ is not a function of continuous time $t$, but a function of $(\mathbf{X}_i, \mathbf{X}_{ij})$ only. As a result, the average counts of recurrent infections in the interval $(t_{i(j-1)}, t_{ij}]$ is $\lambda(\mathbf{X}_{ij})(t_{ij} - t_{i(j-1)})$. Based on (2.1) and (2.2), the probability density function of $Y_{ij}$ in interval $(t_{i(j-1)}, t_{ij}]$ is:

$$P(Y_{ij} = 0) = P(Y_{ij}^* = 0) = \exp\{-\lambda(\mathbf{X}_{ij})(t_{ij} - t_{i(j-1)})\}, \text{and}$$

$$P(Y_{ij} = 1) = P(Y_{ij}^* \geq 1) = 1 - P(Y_{ij} = 0) = 1 - \exp\{-\lambda(\mathbf{X}_{ij})(t_{ij} - t_{i(j-1)})\}. \quad (2.3)$$

To capture how the intensity $\lambda(\mathbf{X}_{ij})$ vary with $(\mathbf{X}_i, \mathbf{X}_{ij})$, we choose the widely popular Cox proportional hazards model to model $\lambda(\mathbf{X}_{ij})$, i.e.,

$$\lambda(\mathbf{X}_{ij}) = \lambda_{0i}\exp\{\boldsymbol{\beta}_1\mathbf{X}_i + \boldsymbol{\beta}_2\mathbf{X}_{ij}\} = \exp\{\beta_{0i} + \boldsymbol{\beta}_1\mathbf{X}_i + \boldsymbol{\beta}_2\mathbf{X}_{ij}\} \quad (2.4)$$

in each interval $(t_{i(j-1)}, t_{ij}], i = 1, \cdots, n, j = 1, \cdots, J_i$. And, $\lambda_{0i} = \exp\{\beta_{0i}\}$ is the individual baseline hazard function, $\mathbf{X}_i$ are baseline covariates, while $\mathbf{X}_{ij}$ are time-varying covariates across visits, $j = 1, \cdots, J_i$. The coefficients, $\boldsymbol{\beta}$s, can be interpreted as individual log hazard ratios and/or individual log intensity ratios, after conditioning on individual baseline hazard and other confounders. After plugging the model (2.4) into the probability density function (2.3), it turns into:

$$E(Y_{ij}|\mathbf{X}, \beta_{0i}) = P(Y_{ij} = 1|\mathbf{X}, \beta_{0i}) = 1 - \exp\left\{-\exp\{\beta_{0i} + \boldsymbol{\beta}_1\mathbf{X}_i + \boldsymbol{\beta}_2\mathbf{X}_{ij}\}(t_{ij} - t_{i(j-1)})\right\},$$

i.e.,

$$\log\left\{-\log[1 - E(Y_{ij}|\mathbf{X}, \beta_{0i})]\right\} = \beta_{0i} + \boldsymbol{\beta}_1\mathbf{X}_i + \boldsymbol{\beta}_2\mathbf{X}_{ij} + \log(t_{ij} - t_{i(j-1)}), \quad (2.5)$$

where $\beta_{0i} \sim N(\beta_0, \sigma^2)$.

Given the random effects $\beta_{0i}$, it is assumed that $Y_{ij}$ are independent of one another. Then, the probability density function of $i$th patient's longitudinal interval reported binary recurrent event $\mathbf{Y}_i$ can be written as

$$f(\mathbf{Y}_i|\mathbf{X}, \beta_{0i}) = \int \prod_{j=1}^{J_i} \{P(Y_{ij} = 1|\mathbf{X}, \beta_{0i})\}^{Y_{ij}} \{1 - P(Y_{ij} = 1|\mathbf{X}, \beta_{0i})\}^{1-Y_{ij}} f(\beta_{0i}) d\beta_{0i}.$$

Thus, the likelihood function is

$$L(\beta) = \prod_{i=1}^{n} \int \prod_{j=1}^{J_i} \{P(Y_{ij} = 1|\mathbf{X}, \beta_{0i})\}^{Y_{ij}} \{1 - P(Y_{ij} = 1|\mathbf{X}, \beta_{0i})\}^{1-Y_{ij}} f(\beta_{0i}) d\beta_{0i}.$$

This likelihood function turns into a typical random effect longitudinal likelihood with complementary log-log link. The parameter estimates can be obtained from standard statistical packages of generalized linear mixed effects model of binary outcome with a complementary log-log link. And, the varying reporting interval can be taken into account by the offset $\log(t_{ij} - t_{i(j-1)})$. Remark that the interpretations of regression parameters are individual log hazard ratios of the infection occurrence across covariates' values to quantify the longitudinal infection trends and cross-sectional gender effect.

Using similar idea but ignoring the heterogeneity of individual baseline hazard, we could specify a marginal longitudinal model to estimate the longitudinal and cross-sectional effects with a complementary log-log link, offset $\log(t_{ij} - t_{i(j-1)})$ as below, and a working correlation matrix, estimated by GEE with "robust" variance

estimator:

$$\log\{-\log[1-E(Y_{ij}|\mathbf{X})]\} = \alpha_0 + \boldsymbol{\alpha}_1\mathbf{X}_i + \boldsymbol{\alpha}_2\mathbf{X}_{ij} + \log(t_{ij} - t_{i(j-1)}) \tag{2.6}$$

Remark that the interpretations of $\boldsymbol{\alpha}$ are log hazard ratios of the infection occurrence across two stratums of patients with different covariates' values, after adjusting for all the other confounders. The estimates of regression coefficients in the marginal models ($\boldsymbol{\alpha}$) are not expected to be similar as random effect models ($\boldsymbol{\beta}$), due to different modeling frameworks and the non-linearity of these effect measures. One exception is that when individual baseline hazard heterogeneity $\sigma^2$, $\text{var}(\beta_{0i})$, is small, and those estimates could be close. However, a large heterogeneity $\sigma^2$ can make individual effects and marginal effects far apart. For complementary log-log link and logit link, the individual effects are often larger in effect sizes than the marginal effects with increasing $\sigma^2$.

Dealing with such longitudinal study, it is quite intuitive to ignore the interval reported binary recurrent event data nature of this infection outcome, just consider it as a standard binary longitudinal outcome. Then, a marginal model with logit link and a working correlation matrix estimated by GEE with "robust" variance estimator as well as a generalized linear mixed effects model with logit link are common methods to use, as below:

$$\begin{cases} \text{cov}(\mathbf{Y}_i|\mathbf{X}) \sim \text{Working correlation matrix} \\ \text{logit}[E(Y_{ij}|\mathbf{X})] = \alpha_0^* + \boldsymbol{\alpha}_1^*\mathbf{X}_i + \boldsymbol{\alpha}_2^*\mathbf{X}_{ij} \end{cases} \tag{2.7}$$

$$\begin{cases} \text{logit}[E(Y_{ij}|\mathbf{X}), \beta_{0i}^*] = \beta_{0i}^* + \boldsymbol{\beta}_1^*\mathbf{X}_i + \boldsymbol{\beta}_2^*\mathbf{X}_{ij} \\ \beta_{0i}^* \sim N(\beta_0^*, \sigma^2) \end{cases} \tag{2.8}$$

Remark that the log odds ratios $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ are used to quantify the longitudinal and cross-sectional effects. Interpretations of $\boldsymbol{\alpha}^*$ are log odds ratios of the infection occurrence across two stratums of patients with different covariates' values, after adjusting for all the confounders, while interpretations of $\boldsymbol{\beta}^*$ are individual log odds ratios of the infection occurrence.

In summary, for analyzing longitudinal interval reported binary recurrent event data, we developed a generalized linear mixed effects model (2.5) which allowing population heterogeneity in baseline hazards, as well as a marginal model estimated by GEE (2.6) without considering such heterogeneity, to quantify their longitudinal and cross-sectional effects. Specifically, the probability of observing an event in an interval is based on the Poisson process of the events with the intensity follows a Cox proportional hazards model. Based on the Cox proportional hazards model, our hazard function is flexible enough to capture the variability caused by a subject's baseline characteristics and time-varying covariates across the longitudinal visits, while the subject's hazard function within each reporting interval stays fixed over time. As a result, complementary log-log link and offset $\log(t_{ij} - t_{i(j-1)})$ in generalized linear model framework were used to account for the interval censored reporting system between longitudinal visits/ interviews and binary nature of recurrent events reporting in intervals, leading to widely available software for estimation, such as the

geepack and glmmML package in R and corresponding models in SAS. To evaluate the numerical performances of the proposed models, simulation studies were carried out to compare them with the standard generalized linear models with logit link (2.7) (2.8). Even though those models use different effect measures to quantify the longitudinal and cross-sectional effects, we want to see which model performs better in capturing significant effects when dealing with longitudinal interval reported binary recurrent event data, whether standard longitudinal models ignoring these features of outcomes could reach good results. Simulation studies were done using R 3.4.2, while the real data application was analyzed using SAS 9.4.

## 2.4 Simulation Study

We generate data to mimic the real infection data structure listed in section 2.2, which has 11 months in total from month 2 to month 12 with missing values existing in some of the months. The simulation process was provided as following:

1. The total number of longitudinal visits per subject was generated randomly with the range from 1 to 11. Then, we randomly selected months of assessment for each subject consistent with the chosen total visits of this subject, called "month", ranging from month 2 to month 12. For example, subject $i$ can be observed at month 3, 5, 6, 9, and 12 if his/her total number of visits is 5. Based on $\text{month}_{ij}$(or $t_{ij}$) $= \{3, 5, 6, 9, 12\}$, the $i$th subject's reporting intervals are $\{3, 2, 1, 3, 3\}$ months.

2. We generated two baseline variables - age and gender $(x_1, x_2)$, $x_1 \sim N(80, 8^2)$, and $x_2 \sim Bin(n, 0.5)$, as a simple representation of the mixture of continuous and categorical covariates in typical longitudinal studies.

3. We generated seven rationales of intensity based on Cox proportional hazards model for the $i$th subject in the following:

$$\begin{cases} \lambda_{k1}(\mathbf{X}_{ij}) = \exp\{\alpha_{0k} + b_{0i} + \alpha_{1k}x_{1i} + \alpha_{2k}x_{2i}\} \\ \lambda_{k2}(\mathbf{X}_{ij}) = \exp\{\alpha_{0k} + b_{0i} + \alpha_{1k}x_{1i} + \alpha_{2k}x_{2i} + \alpha_{3k}t_{ij}\} \\ \lambda_{k3}(\mathbf{X}_{ij}) = \exp\{\alpha_{0k} + b_{0i} + \alpha_{1k}x_{1i} + \alpha_{2k}x_{2i} + \alpha_{3k}t_{ij} + \alpha_{4k}x_{2i}t_{ij}\} \\ \lambda_{3}(\mathbf{X}_{ij}) = \exp\{\alpha_{01} + b_{0i} + \alpha_{11}x_{1i}\} \end{cases}, \quad (2.9)$$

where $i = 1, \ldots, n, j = 1, \cdots, J_i, k = 1, 2,$ and $b_{0i} \sim N(0, \sigma^2)$ accounting for heterogeneity of the baseline hazard in the study population. Sensitivity studies to $\sigma = \{0.5, 1, 2\}$ are carried out, addressing how well the model performance will be under small, moderate and large heterogeneity respectively.

$$\begin{cases} \lambda_{11}(\mathbf{X}_{ij}) = \exp\{-1 + b_{0i} + 0.005x_{1i} - 1.5x_{2i}\} \\ \lambda_{12}(\mathbf{X}_{ij}) = \exp\{-1 + b_{0i} + 0.005x_{1i} - 1.5x_{2i} - 0.1t_{ij}\} \\ \lambda_{13}(\mathbf{X}_{ij}) = \exp\{-1 + b_{0i} + 0.005x_{1i} - 1.5x_{2i} - 0.1t_{ij} - 0.2x_{2i}t_{ij}\} \end{cases} \quad (2.10)$$

$$\begin{cases} \lambda_{21}(\mathbf{X}_{ij}) = \exp\{-0.3 + b_{0i} - 0.02x_{1i} - 0.5x_{2i}\} \\ \lambda_{22}(\mathbf{X}_{ij}) = \exp\{-0.3 + b_{0i} - 0.02x_{1i} - 0.5x_{2i} + 0.05t_{ij}\} \\ \lambda_{23}(\mathbf{X}_{ij}) = \exp\{-0.3 + b_{0i} - 0.02x_{1i} - 0.5x_{2i} + 0.05t_{ij} - 0.03x_{2i}t_{ij}\} \end{cases} \quad (2.11)$$

$$\lambda_3(\mathbf{X}_{ij}) = \exp\{-1 + b_{0i} + 0.005x_{1i}\} \tag{2.12}$$

Under large covariates' effect of gender and month ($k = 1$) on the frequency of recurrent event outcomes, $\{\lambda_{11}(\mathbf{X}_{ij}), \lambda_{12}(\mathbf{X}_{ij}), \lambda_{13}(\mathbf{X}_{ij})\}$ (2.10) are generated, representing three rationales of time-fixed intensity $\lambda_{11}(X_{ij})$ and two time-varying intensities $\{\lambda_{12}(\mathbf{X}_{ij}), \lambda_{13}(\mathbf{X}_{ij})\}$ during the first year of follow-up. Under moderate covariates' effect of gender and month ($k = 2$) on outcomes, $\{\lambda_{21}(X_{ij}), \lambda_{22}(X_{ij}), \lambda_{23}(X_{ij})\}$ (2.11) are generated respectively. In addition, $\lambda_3(\mathbf{X}_{ij})$ (2.12) is generated only based on age.

4. We generated the underlying unobservable counts ($Y^*_{ij}$) of recurrent events in the interval $(t_{i(j-1)}, t_{ij}]$ using Poisson models, then dichotomized them to get the observed interval reported binary recurrent event outcome:

$$\begin{cases} Y_{ij} = I(Y^*_{ij} > 0) \\ P(Y^*_{ij} = y^*_{ij}) = e^{-\lambda_{km}(\mathbf{X}_{ij})(t_{ij} - t_{i(j-1)})} \left[\lambda_{km}(\mathbf{X}_{ij})(t_{ij} - t_{i(j-1)})\right]^{y^*_{ij}} / y^*_{ij}! \end{cases}, \tag{2.13}$$

where $m = 1, 2, 3$, accounting for one time-fixed intensity and two time-varying intensities, and $km = 3$, accounting for the time-fixed intensity only based on age.

The simulation was done in statistical software R 3.4.2, and the geepack and glmmML library were utilized for parameter estimates. 500 simulations per scenario was carried, and sample size was usually set at 200. Sensitivity of model performance to small sample size is carried out too, where sample size is set at 50. After obtaining four types of outcomes - longitudinal binary recurrent event data - under the four

different Poisson processes (time-fixed vs. time-varying ones), we did the analysis using models in (2.5 - 2.8) with logistic or complementary-log-log link. With or without time covariate in the model, there were 3 models for each approach. Therefore, there were 12 models in total, and 12 different parameter estimations for the time-fixed covariate "gender" and time-varying covariate "month" were presented correspondingly. To compare the performances of the 12 models, eight scenarios were provided, with the first three without model mis-specifications and the last five with model mis-specifications: (1) All models were "correctly" specified with sample size 200 using three types of outcomes based on large covariates' effect; (2) All models were "correctly" specified with sample size 200 using three types of outcomes based on moderate covariates' effect; (3) All models were "correctly" specified with small sample size 50 using three types of outcomes based on large covariates' effect; (4) All models were fitted on the same outcomes which were generated through the Poisson distribution with time-varying intensity $\lambda_{13}(\mathbf{X}_{ij})$ with sample size 200; (5) All models were fitted on the same outcomes which were generated through the Poisson distribution with time-varying intensity $\lambda_{23}(\mathbf{X}_{ij})$ with sample size 200; (6) All models were fitted on the same outcomes which were generated through the Poisson distribution with time-varying intensity $\lambda_{12}(\mathbf{X}_{ij})$ with sample size 200; (7) All models were fitted on the same outcomes which were generated through the Poisson distribution with time-varying intensity $\lambda_{22}(\mathbf{X}_{ij})$ with sample size 200; (8) All models were fitted on the same outcomes which were generated through the Poisson distribution with time-fixed intensity $\lambda_3(\mathbf{X}_{ij})$ with sample size 200. To measure the performance of the 12 models among the seven scenarios, estimates on "gender" and

"month", bias, standard errors, percentages on significant effects (power), coverage probability of confidence intervals, AIC, and BIC were calculated.

### 2.4.1 Simulation Results - Part I

In Part I, we want to see how different types of models behave under small to large population heterogeneity in baseline hazard, large or moderate covariates' effect, small or large sample sizes, without model mis-specification problems under various outcome generating mechanisms (Table 2.1 - Table 2.3).

First scenario is when all marginal models and generalized linear mixed effects models were "correctly" specified with large covariates' effect and sample size 200 on three types of outcomes, generated based on $\{\lambda_{11}(\mathbf{X}_{ij}), \lambda_{12}(\mathbf{X}_{ij}), \lambda_{13}(\mathbf{X}_{ij})\}$ (2.10), and results are shown in Table 2.1. Among all of the models, the proposed GLMMs with complementary log-log link have the best performance under various population heterogeneity of the baseline hazards. Under small population heterogeneity, GEEs with both logit link and complementary log-log link also have similar effect estimates as the proposed GLMMs, with "sex" effects close to $-1.5$ and "time" effects close to $-0.1$, small standard errors, and high percentages on power. However, as population heterogeneity increases, all models except for the proposed GLMMs with complementary log-log link become unstable, where both "sex" and "time" effects increase for the GLMMs with logit link, while "sex" and "time" effects decrease for GEEs. GLMMs with logit link have the largest standard errors compared with the

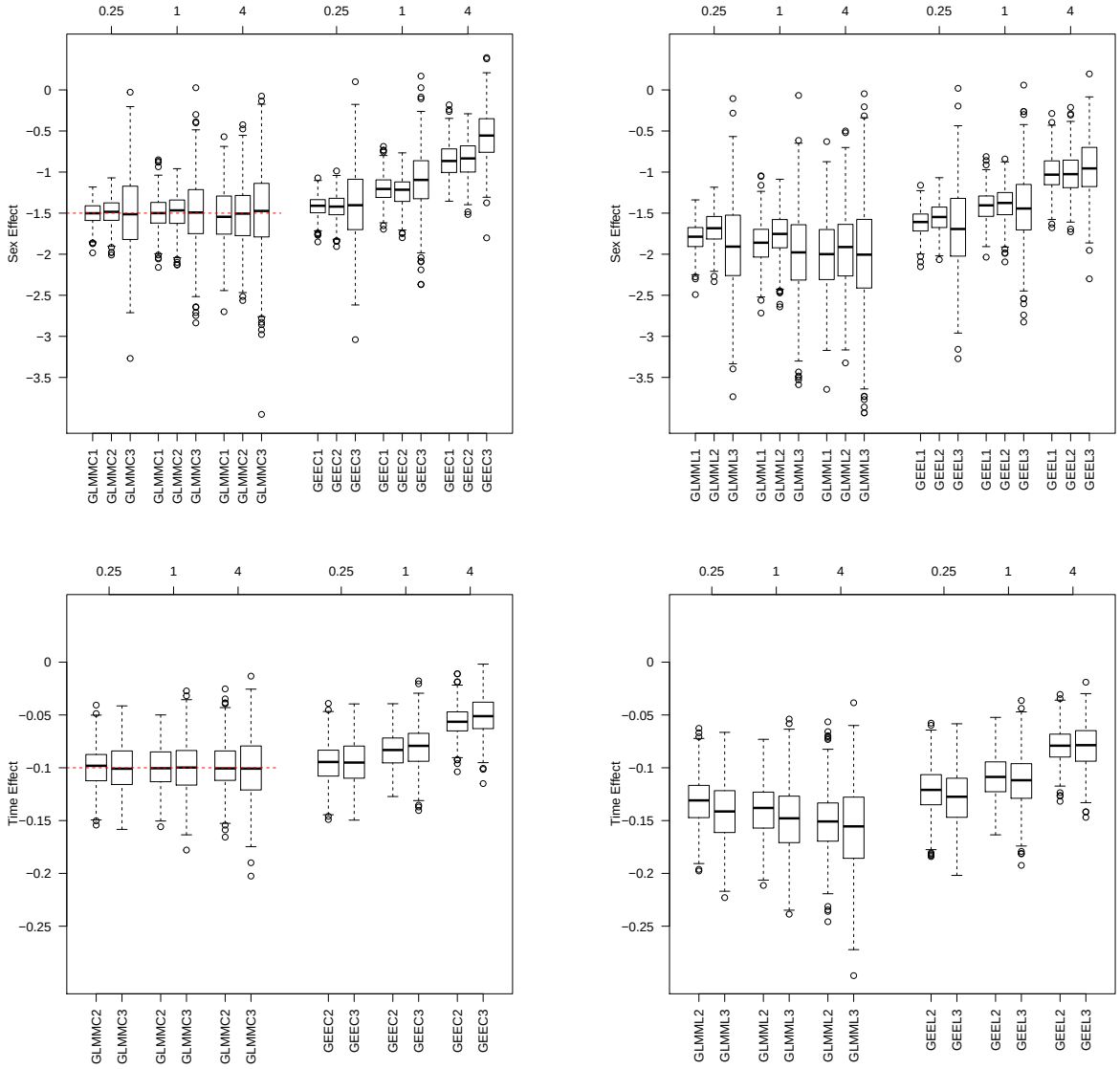


Figure 2.1: Boxplots on sex and time effect estimates using "correctly" specified models under large covariates' effect and various study population heterogeneity ($n = 200$)

Table 2.1: Covariates' effect estimates using "correctly" specified models under large covariates' effect and various study population heterogeneity ($n = 200$)

| | Random effects model estimated by MLE, with complementary log-log link (True Sex effect = -1.5, True Time effect = -0.1) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | GLMM[1] | | | GLMM[2] | | | GLMM[3] | | |
| $\sigma^2$ (heterogeneity) | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -1.5028 | -1.4916 | -1.5216 | -1.4883 | -1.4857 | -1.5296 | -1.4957 | -1.4878 | -1.4864 |
| SE | 0.1301 | 0.1918 | 0.3361 | 0.1576 | 0.2038 | 0.3529 | 0.4566 | 0.4255 | 0.5273 |
| Power (%) | 1 | 1 | 0.994 | 1 | 1 | 0.992 | 0.926 | 0.946 | 0.844 |
| Time | - | - | - | -0.0994 | -0.0998 | -0.0994 | -0.1000 | -0.1001 | -0.1004 |
| SE | - | - | - | 0.0192 | 0.0200 | 0.0217 | 0.0224 | 0.0246 | 0.0298 |
| Power (%) | - | - | - | 1 | 1 | 0.988 | 0.998 | 0.986 | 0.938 |
| AIC | 1303.45 | 1290.69 | 1139.41 | 1109.32 | 1137.88 | 1074.98 | 901.88 | 921.10 | 913.74 |
| BIC | 1323.81 | 1311.06 | 1159.78 | 1134.76 | 1163.34 | 1100.44 | 932.42 | 951.66 | 944.29 |
| | Marginal model estimated by GEE, with complementary log-log link | | | | | | | | |
| | GEE[1] | | | GEE[2] | | | GEE[3] | | |
| $\sigma^2$ (heterogeneity) | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -1.4185 | -1.2049 | -0.8549 | -1.4257 | -1.2322 | -0.8404 | -1.3970 | -1.1074 | -0.5567 |
| SE | 0.1238 | 0.1546 | 0.2202 | 0.1486 | 0.1729 | 0.2174 | 0.4491 | 0.3746 | 0.3192 |
| Power (%) | 1 | 1 | 0.990 | 1 | 1 | 0.986 | 0.896 | 0.848 | 0.428 |
| Time | - | - | - | -0.0951 | -0.0834 | -0.0562 | -0.0945 | -0.0802 | -0.0509 |
| SE | - | - | - | 0.0182 | 0.0169 | 0.0136 | 0.0209 | 0.0201 | 0.0180 |
| Power (%) | - | - | - | 1 | 1 | 0.974 | 0.998 | 0.980 | 0.812 |
| | Random effects model estimated by MLE, with logit link | | | | | | | | |
| | GLMM[1] | | | GLMM[2] | | | GLMM[3] | | |
| $\sigma^2$ (heterogeneity) | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -1.7894 | -1.8636 | -1.9946 | -1.6887 | -1.7686 | -1.9459 | -1.8920 | -1.9906 | -2.0174 |
| SE | 0.1746 | 0.2582 | 0.4524 | 0.1956 | 0.2571 | 0.4607 | 0.5325 | 0.5316 | 0.6967 |
| Power (%) | 1 | 1 | 0.994 | 1 | 1 | 0.988 | 0.954 | 0.970 | 0.876 |
| Time | - | - | - | -0.1316 | -0.1403 | -0.1521 | -0.1412 | -0.1493 | -0.1570 |
| SE | - | - | - | 0.0237 | 0.0257 | 0.0299 | 0.0287 | 0.0327 | 0.0415 |
| Power (%) | - | - | - | 1 | 1 | 1 | 1 | 0.996 | 0.982 |
| AIC | 1411.39 | 1375.73 | 1195.06 | 1192.25 | 1208.05 | 1125.62 | 967.74 | 976.08 | 954.73 |
| BIC | 1431.75 | 1396.09 | 1215.42 | 1217.70 | 1233.52 | 1151.08 | 998.27 | 1006.64 | 985.28 |
| | Marginal model estimated by GEE, with logit link | | | | | | | | |
| | GEE[1] | | | GEE[2] | | | GEE[3] | | |
| $\sigma^2$ (heterogeneity) | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -1.6179 | -1.4112 | -1.0096 | -1.5534 | -1.3900 | -1.0209 | -1.6800 | -1.4381 | -0.9494 |
| SE | 0.1538 | 0.1890 | 0.2224 | 0.1719 | 0.2016 | 0.2437 | 0.5025 | 0.4248 | 0.3594 |
| Power (%) | 1 | 1 | 0.994 | 1 | 1 | 0.986 | 0.944 | 0.932 | 0.766 |
| Time | - | - | - | -0.1211 | -0.1093 | -0.0792 | -0.1281 | -0.1127 | -0.0795 |
| SE | - | - | - | 0.0217 | 0.0202 | 0.0158 | 0.0264 | 0.0253 | 0.0212 |
| Power (%) | - | - | - | 1 | 1 | 0.998 | 1 | 0.994 | 0.986 |

[1] Outcomes generated by $\lambda_{11}(\mathbf{X}_{ij})$ without time-dependent covariates in the model; [2] Outcomes generated by $\lambda_{12}(\mathbf{X}_{ij})$ with time-dependent covariates in the model; [3] Outcomes generated by $\lambda_{13}(\mathbf{X}_{ij})$ with time-dependent covariates and time interaction term in the model

other models. The results can be seen clearly in Figure 2.1 displaying the median for each boxplot with the dashed red line shows the true "sex" and "time" effects. Power in Table 2.1 is defined as the percentage of significant effect with p-value less than 0.05 among the 500 simulations. Our proposed GLMMs provide stable and unbiased "sex" and "time" effect estimates and inference regardless the population baseline hazard heterogeneity is large or small, with effects close to the true effects $-1.5$ and $-0.1$ respectively. When there are true "time" and/ or "sex" effect in the data, this model is very likely to capture those signals, based on high power. Also, those GLMMs have smaller AIC and BIC values, comparing to GLMMs with logit link.

To see the influence of covariates' effect size on the proposed models and other models, the second scenario, all models with moderate covariates' effect on three types of outcomes, generated based on $\{\lambda_{21}(\mathbf{X}_{ij}), \lambda_{22}(\mathbf{X}_{ij}), \lambda_{23}(\mathbf{X}_{ij})\}$ (2.11), results are shown in Table 2.2 with similar settings as the first scenario, and boxplots are shown in Figure 2.2. When the size of covariates' effect were changed from large to moderate, summaries of Table 2.2 are very similar as Table 2.1, and an additional conclusion from Table 2.2 is that GLMMs with complementary log-log link can mostly capture the significance of "sex" and "time" effect with the highest percentage on power, especially for "time" effect. For "sex" effect, the power are similar between the proposed GLMMs and the other models; while for "time" effect, the power are higher for models using complementary log-log link compared with those using logit link.

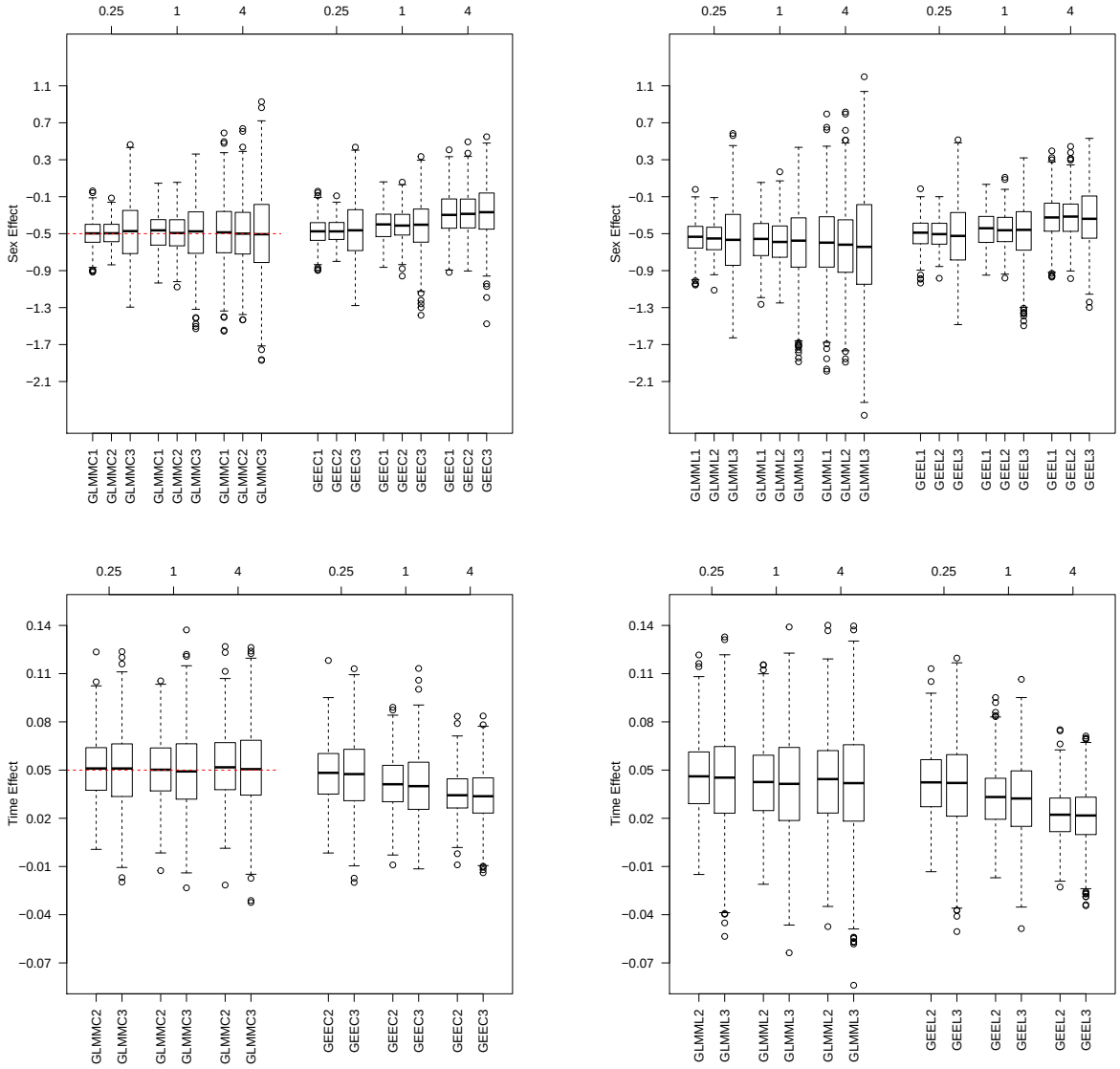


Figure 2.2: Boxplots on sex and time effect estimates using "correctly" specified models under moderate covariates' effect and various study population heterogeneity ($n = 200$)

Table 2.2: Sensitivity of model performance ("correct" model specification) under moderate covariates' effect and various study population heterogeneity ($n = 200$)

| | Random effects model estimated by MLE, with complementary log-log link (True Sex effect = -0.5, True Time effect = 0.05) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | GLMM[1] | | | GLMM[2] | | | GLMM[3] | | |
| $\sigma^2$ (heterogeneity) | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -0.4931 | -0.4825 | -0.4902 | -0.4920 | -0.4935 | -0.4985 | -0.4836 | -0.5001 | -0.4943 |
| SE | 0.1504 | 0.1998 | 0.3554 | 0.1346 | 0.1896 | 0.3500 | 0.3269 | 0.3285 | 0.4704 |
| Power (%) | 0.910 | 0.660 | 0.294 | 0.944 | 0.718 | 0.296 | 0.356 | 0.298 | 0.194 |
| Time | - | - | - | 0.0503 | 0.0503 | 0.0521 | 0.0502 | 0.0496 | 0.0507 |
| SE | - | - | - | 0.0195 | 0.0195 | 0.0210 | 0.0249 | 0.0247 | 0.0271 |
| Power (%) | - | - | - | 0.770 | 0.766 | 0.718 | 0.548 | 0.500 | 0.442 |
| AIC | 1095.97 | 1146.92 | 1093.62 | 1249.36 | 1269.26 | 1139.14 | 1202.53 | 1230.45 | 1124.37 |
| BIC | 1116.33 | 1167.28 | 1113.99 | 1274.80 | 1294.70 | 1164.59 | 1233.06 | 1260.98 | 1154.91 |
| | Marginal model estimated by GEE, with complementary log-log link | | | | | | | | |
| | GEE[1] | | | GEE[2] | | | GEE[3] | | |
| $\sigma^2$ (heterogeneity) | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -0.4759 | -0.4078 | -0.2904 | -0.4712 | -0.4111 | -0.2778 | -0.4683 | -0.4169 | -0.2661 |
| SE | 0.1460 | 0.1732 | 0.2332 | 0.1295 | 0.1629 | 0.2267 | 0.3162 | 0.2820 | 0.2948 |
| Power (%) | 0.904 | 0.636 | 0.294 | 0.940 | 0.686 | 0.312 | 0.344 | 0.254 | 0.142 |
| Time | - | - | - | 0.0475 | 0.0416 | 0.0350 | 0.0470 | 0.0406 | 0.0338 |
| SE | - | - | - | 0.0187 | 0.0167 | 0.0136 | 0.0239 | 0.0207 | 0.0164 |
| Power (%) | - | - | - | 0.748 | 0.710 | 0.702 | 0.540 | 0.488 | 0.402 |
| | Random effects model estimated by MLE, with logit link | | | | | | | | |
| | GLMM[1] | | | GLMM[2] | | | GLMM[3] | | |
| $\sigma^2$ (heterogeneity) | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -0.5360 | -0.5591 | -0.6083 | -0.5482 | -0.5849 | -0.6298 | -0.5585 | -0.6049 | -0.6185 |
| SE | 0.1780 | 0.2429 | 0.4481 | 0.1683 | 0.2356 | 0.4492 | 0.3918 | 0.4110 | 0.6182 |
| Power (%) | 0.856 | 0.632 | 0.268 | 0.888 | 0.670 | 0.286 | 0.340 | 0.286 | 0.186 |
| Time | - | - | - | 0.0451 | 0.0428 | 0.0435 | 0.0439 | 0.0414 | 0.0415 |
| SE | - | - | - | 0.0232 | 0.0248 | 0.0287 | 0.0311 | 0.0326 | 0.0369 |
| Power (%) | - | - | - | 0.542 | 0.452 | 0.356 | 0.326 | 0.258 | 0.172 |
| AIC | 1173.00 | 1214.93 | 1144.35 | 1345.88 | 1346.94 | 1193.12 | 1292.28 | 1305.22 | 1176.64 |
| BIC | 1193.35 | 1235.28 | 1164.72 | 1371.33 | 1372.39 | 1218.58 | 1322.81 | 1335.75 | 1207.18 |
| | Marginal model estimated by GEE, with logit link | | | | | | | | |
| | GEE[1] | | | GEE[2] | | | GEE[3] | | |
| $\sigma^2$ (heterogeneity) | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -0.4964 | -0.4447 | -0.3234 | -0.5018 | -0.4548 | -0.3224 | -0.5123 | -0.4742 | -0.3233 |
| SE | 0.1641 | 0.1931 | 0.2397 | 0.1536 | 0.1833 | 0.2333 | 0.3620 | 0.3190 | 0.3241 |
| Power (%) | 0.856 | 0.610 | 0.268 | 0.886 | 0.666 | 0.292 | 0.356 | 0.284 | 0.208 |
| Time | - | - | - | 0.0417 | 0.0334 | 0.0224 | 0.0407 | 0.0322 | 0.0213 |
| SE | - | - | - | 0.0213 | 0.0194 | 0.0149 | 0.0284 | 0.0249 | 0.0188 |
| Power (%) | - | - | - | 0.546 | 0.470 | 0.374 | 0.346 | 0.280 | 0.190 |

[1] Outcomes generated by $\lambda_{21}(\mathbf{X}_{ij})$ without time-dependent covariates in the model; [2] Outcomes generated by $\lambda_{22}(\mathbf{X}_{ij})$ with time-dependent covariates in the model; [3] Outcomes generated by $\lambda_{23}(\mathbf{X}_{ij})$ with time-dependent covariates and time interaction term in the model

Table 2.3: Sensitivity of model performance under small sample size ($n = 50$, with "correct" model specification) under various study population heterogeneity

| | Random effects model estimated by MLE, with complementary log-log link (True Sex effect = -1.5, True Time effect = -0.1) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | GLMM[1] | | | GLMM[2] | | | GLMM[3] | | |
| $\sigma^2$ (heterogeneity) | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -1.4940 | -1.5592 | -1.5997 | -1.5476 | -1.5453 | -1.6169 | -1.4383 | -1.4112 | -1.6134 |
| SE | 0.2907 | 0.3832 | 0.6990 | 0.3460 | 0.4512 | 0.7315 | 1.0540 | 1.0430 | 1.1451 |
| Power (%) | 1 | 0.986 | 0.558 | 1 | 0.968 | 0.582 | 0.357 | 0.330 | 0.292 |
| Time | - | - | - | -0.1030 | -0.1030 | -0.1035 | -0.1004 | -0.0986 | -0.1047 |
| SE | - | - | - | 0.0398 | 0.0420 | 0.0466 | 0.0484 | 0.0471 | 0.0595 |
| Power (%) | - | - | - | 0.746 | 0.724 | 0.656 | 0.575 | 0.514 | 0.416 |
| | Marginal model estimated by GEE, with complementary log-log link | | | | | | | | |
| | GEE[1] | | | GEE[2] | | | GEE[3] | | |
| $\sigma^2$ (heterogeneity) | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -1.4125 | -1.2817* | -0.6333* | -1.4806 | -1.2885 | -0.9145* | -1.3454 | -1.0569 | -0.6529* |
| SE | 0.2677 | 0.3270* | 6.2330* | 0.3248 | 0.3846 | 0.4942* | 1.0282 | 0.9150 | 0.7225* |
| Power (%) | 1 | 0.984 | 0.548 | 1 | 0.948 | 0.500 | 0.378 | 0.290 | 0.148 |
| Time | - | - | - | -0.0983 | -0.0857 | -0.0597* | -0.0949 | -0.0785 | -0.0542* |
| SE | - | - | - | 0.0374 | 0.0360 | 0.0298* | 0.0455 | 0.0389 | 0.0361* |
| Power (%) | - | - | - | 0.748 | 0.726 | 0.530 | 0.602 | 0.486 | 0.288 |

[1] Outcomes generated by $\lambda_{11}(\mathbf{X}_{ij})$ without time-dependent covariates in the model; [2] Outcomes generated by $\lambda_{12}(\mathbf{X}_{ij})$ with time-dependent covariates in the model; [3] Outcomes generated by $\lambda_{13}(\mathbf{X}_{ij})$ with time-dependent covariates and time interaction term in the model; * Trim the unstable values which are too big ($> 1000$) or too small ($< -1000$)

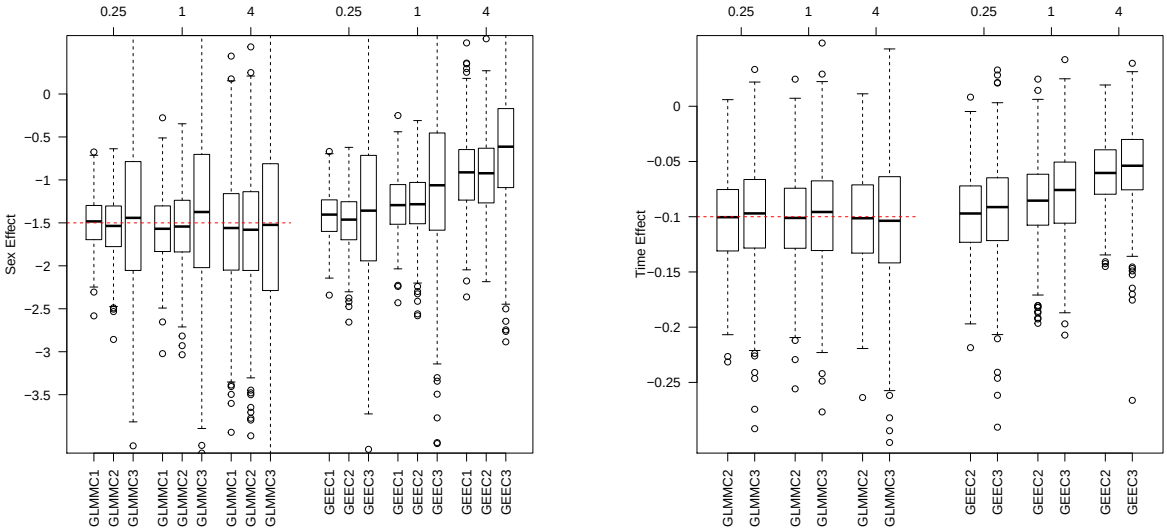


Figure 2.3: Boxplots on sex and time effect estimates using "correctly" specified models under small sample size and various study population heterogeneity ($n = 50$)

To investigate whether the inference and estimation of the proposed models sensitive to small sample size, the third scenario, GLMMs and GEEs with complementary log-log link under small sample size 50 on three types of outcomes, generated based on $\{\lambda_{11}(\mathbf{X}_{ij}), \lambda_{12}(\mathbf{X}_{ij}), \lambda_{13}(\mathbf{X}_{ij})\}$ (2.10), results are shown in Table 2.3 with similar settings as the first scenario, and boxplots are shown in Figure 2.3. Comparing Table 2.1 and Table 2.3, with small sample size, the proposed GLMMs and GEEs with complementary log-log link provide unbiased and stable "sex" and "time" effect estimates under various population heterogeneity, just with increased standard error estimates. Regarding outcomes generated based on $\{\lambda_{11}(\mathbf{X}_{ij}), \lambda_{12}(\mathbf{X}_{ij})\}$, as long as the population heterogeneity is not too large, i.e., $\sigma^2 = (0.5^2, 1.0^2)$, the proposed GLMMs and GEEs perform reasonable well. Even though the SE estimates increased dramatically at small sample size, the proposed GLMMs still capture the significant cross-sectional and longitudinal effects well. However, when heterogeneity is very large, $\sigma^2 = 2^2$, SE estimates increase dramatically, leading to decreased power. Regarding outcomes generated based on $\lambda_{13}(\mathbf{X}_{ij})$, both GLMM and GEE still produce good effect estimates under various population heterogeneity, but with large SE estimates. As a result, the power of GLMM$^3$ and GEE$^3$ dramatically reduce from $> 90\%$ in Table 2.1 with sample size 200 to $< 38\%$ for "sex" effects and $< 61\%$ for "time" effects.

### 2.4.2 Simulation Results - Part II

In Part II, we want to see how sensitive are the proposed models and standard models to various model mis-specifications, under small to large population heterogeneity, large or moderate covariates' effect. Table 2.4 and Table 2.5 are the two scenarios with missing a large "sex" and "time" interaction term, under large covariates' effect on outcomes $Y_{13}$ and moderate covariates' effect on outcomes $Y_{23}$ respectively; while Table 2.6 and Table 2.7 are the two scenarios with missing a large "time" effect term, or over-fitting model with correct model nested, under large covariates' effect on outcomes $Y_{12}$ and moderate covariates' effect on outcomes $Y_{22}$ respectively. Table 2.8 is the eighth scenario with missing a large "sex" and "time" effect terms.

Table 2.4 shows results from the fourth scenario, where all marginal models and generalized linear mixed effects models were mis-specified with large covariates' effect and sample size 200 on outcomes generated through the Poisson distribution with time-varying intensity $\lambda_{13}(\mathbf{X}_{ij})$, and Figure 2.4 lists boxplots displaying clearer results on the distribution of "sex" and "time" effect estimates. All models' inferences are sensitive to model mis-specification with bias close to $-1$ for "sex" effect and $-0.02$ for "time" effect under various population heterogeneity. Also, the small coverages show poor coverage probability of 95% confidence intervals of covariates' effect throughout all models. For "time" effect, the inference is better behaved in the proposed GLMMs and GEEs with complementary log-log link than the standard longitudinal models with logit link. As model selection criteria, AIC and BIC values

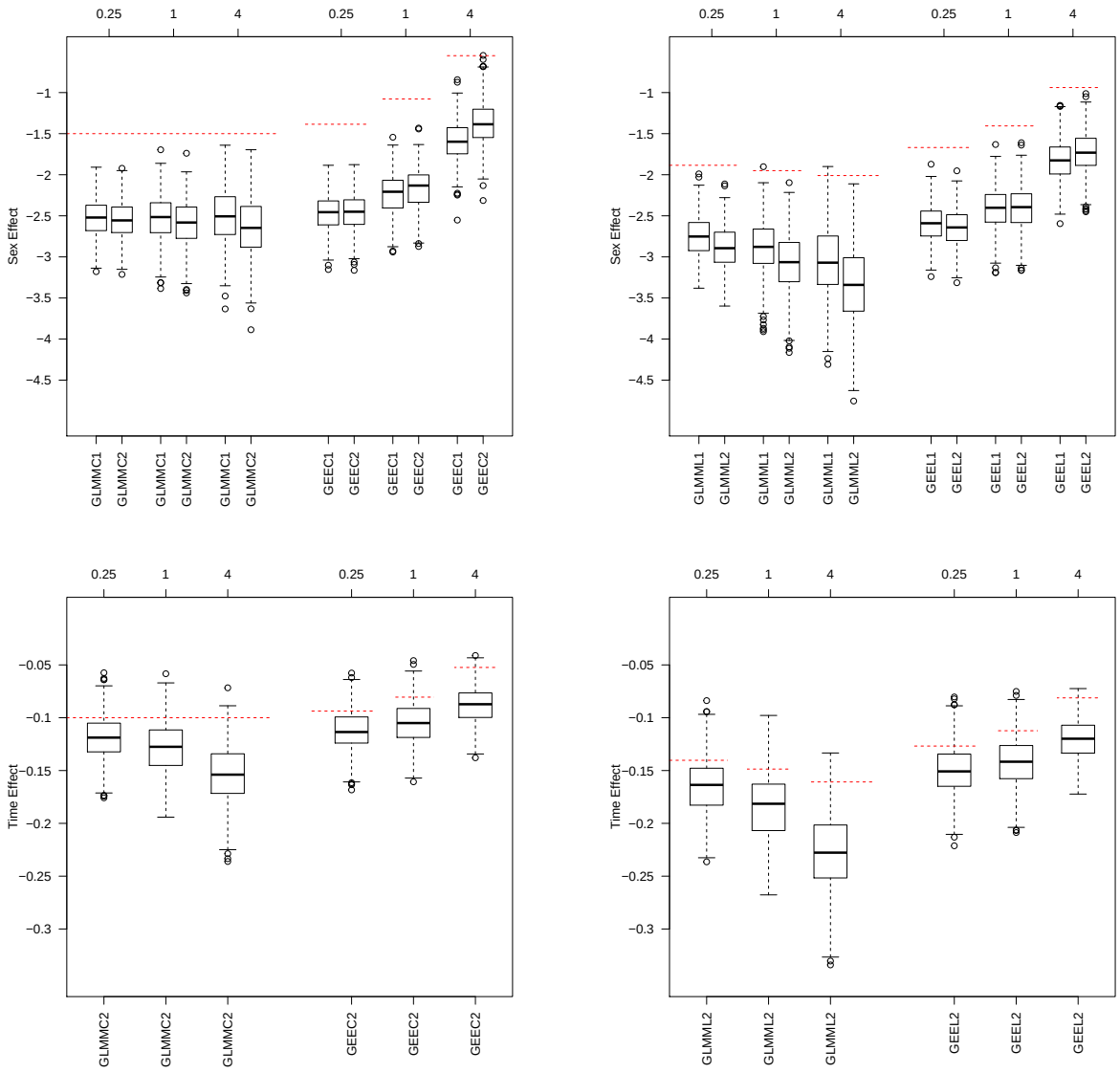


Figure 2.4: Boxplots on sex and time effect estimates using incorrectly specified models missing a large sex and time interaction term, under various study population heterogeneity and large covariates' effect on outcomes based on $\lambda_{13}(\mathbf{X}_{ij})$ ($n = 200$)

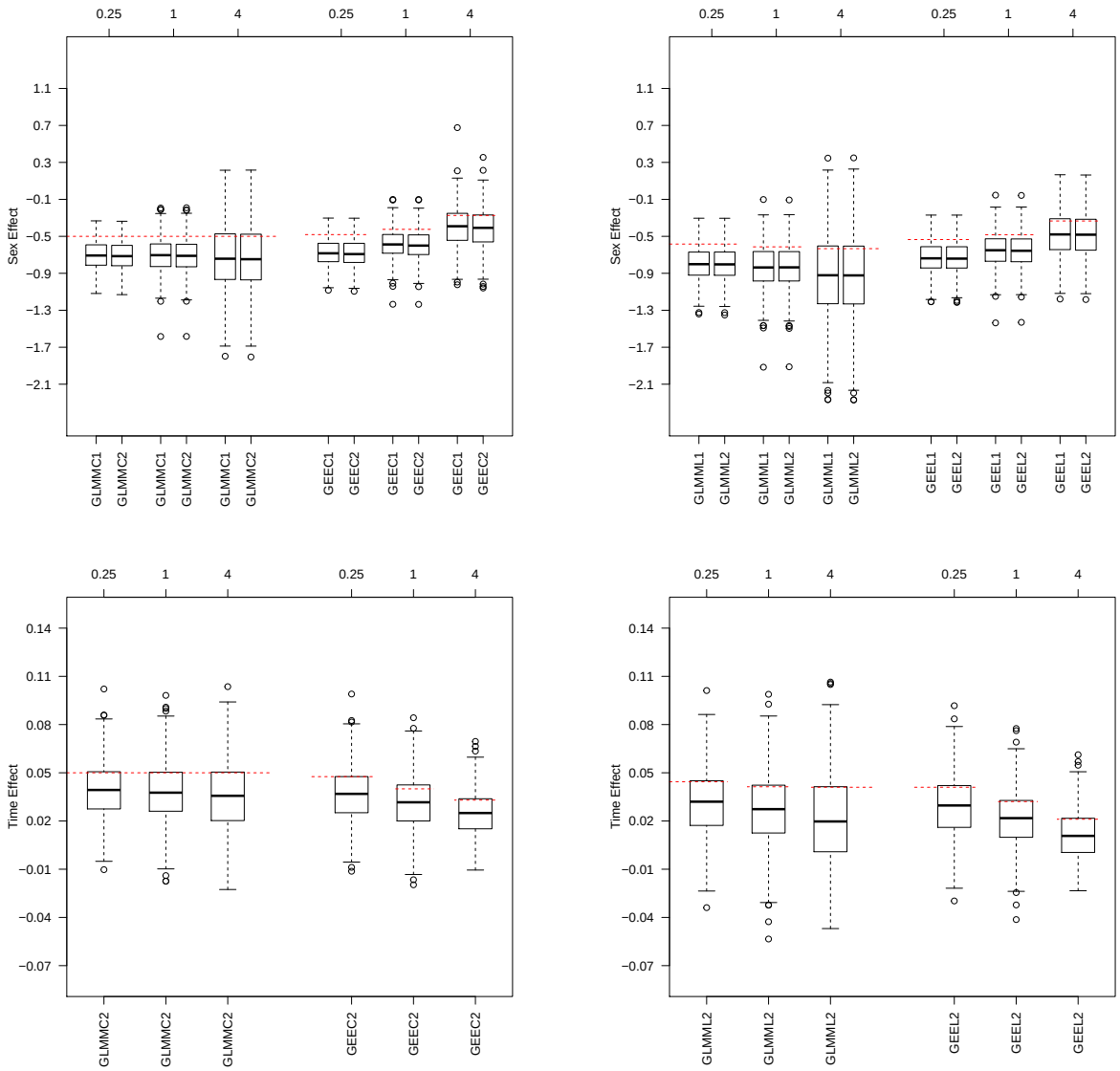


Figure 2.5: Boxplots on sex and time effect estimates using incorrectly specified models missing a large sex and time interaction term, under various study population heterogeneity and moderate covariates' effect on outcomes based on $\lambda_{23}(\mathbf{X}_{ij})$ $(n = 200)$

Table 2.4: Sensitivity of model inferences to missing a large sex and time interaction term, under various study population heterogeneity and large covariates' effect on outcomes based on $\lambda_{13}(\mathbf{X}_{ij})$ $(n = 200)$

| | Random effects model, with complementary log-log link (True Sex effect = -1.5, True Time effect = -0.1) | | | | | | Marginal model, with complementary log-log link (Sex effect = (-1.3854, -1.0773, -0.5509)*, Time effect = (-0.0937, -0.0805, -0.0524)*) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GLMM[1] | | | GLMM[2] | | | GEE[1] | | | GEE[2] | | |
| $\sigma^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -2.5304 | -2.5271 | -2.5013 | -2.5607 | -2.5935 | -2.6489 | -2.4664 | -2.2391 | -1.5858 | -2.4566 | -2.1676 | -1.3670 |
| Bias | -1.0304 | -1.0271 | -1.0013 | -1.0607 | -1.0935 | -1.1489 | -1.0810 | -1.1618 | -1.0349 | -1.0712 | -1.0903 | -0.8161 |
| SE | 0.2284 | 0.2708 | 0.3465 | 0.2286 | 0.2764 | 0.3690 | 0.2227 | 0.2407 | 0.2426 | 0.2229 | 0.2462 | 0.2813 |
| Cov. | 0 | 0.006 | 0.166 | 0 | 0.002 | 0.126 | 0 | 0 | 0.01 | 0 | 0.004 | 0.09 |
| Time | - | - | - | -0.1190 | -0.1286 | -0.1544 | - | - | - | -0.1126 | -0.1051 | -0.0882 |
| Bias | - | - | - | -0.0190 | -0.0286 | -0.0544 | - | - | - | -0.0189 | -0.0246 | -0.0358 |
| SE | - | - | - | 0.0216 | 0.0246 | 0.0276 | - | - | - | 0.0201 | 0.0200 | 0.0167 |
| Cov. | - | - | - | 0.868 | 0.752 | 0.422 | - | - | - | 0.872 | 0.746 | 0.488 |
| AIC | 939.88 | 960.28 | 962.40 | 909.29 | 927.08 | 921.83 | - | - | - | - | - | - |
| BIC | 960.24 | 980.64 | 982.76 | 934.74 | 952.53 | 947.28 | - | - | - | - | - | - |
| | Random effects model, with logit link (Sex effect = (-1.8849, -1.9500, -2.0099)*, Time effect = (-0.1403, -0.1486, -0.1608)*) | | | | | | Marginal model, with logit link (Sex effect = (-1.6685, -1.4058, -0.9387)*, Time effect = (-0.1268, -0.1123, -0.0812)*) | | | | | |
| | GLMM[1] | | | GLMM[2] | | | GEE[1] | | | GEE[2] | | |
| $\sigma^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -2.7493 | -2.8884 | -3.0490 | -2.8901 | -3.0810 | -3.3487 | -2.5954 | -2.4128 | -1.8211 | -2.6471 | -2.4068 | -1.7244 |
| Bias | -0.8644 | -0.9384 | -1.0391 | -1.0052 | -1.1310 | -1.3388 | -0.9269 | -1.0070 | -0.8824 | -0.9786 | -1.0010 | -0.7857 |
| SE | 0.2535 | 0.3256 | 0.4418 | 0.2631 | 0.3446 | 0.4861 | 0.2326 | 0.2544 | 0.2525 | 0.2344 | 0.2570 | 0.2575 |
| Time | - | - | - | -0.1653 | -0.1847 | -0.2280 | - | - | - | -0.1502 | -0.1424 | -0.1203 |
| Bias | - | - | - | -0.0250 | -0.0361 | -0.0672 | - | - | - | -0.0234 | -0.0301 | -0.0391 |
| SE | - | - | - | 0.0277 | 0.0313 | 0.0369 | - | - | - | 0.0245 | 0.0235 | 0.0186 |
| AIC | 1011.48 | 1020.05 | 1010.93 | 974.25 | 978.77 | 961.01 | - | - | - | - | - | - |
| BIC | 1031.84 | 1040.41 | 1031.29 | 999.70 | 1004.22 | 986.46 | - | - | - | - | - | - |

[1] Without time-dependent covariates in the model; [2] With time-dependent covariates in the model; * True sex and time effects come from Table 2.1

Table 2.5: Sensitivity of model inferences to missing a large sex and time interaction term, under various study population heterogeneity and moderate covariates' effect on outcomes based on $\lambda_{23}(\mathbf{X}_{ij})$ ($n = 200$)

| | Random effects model, with complementary log-log link (True Sex effect = -0.5, True Time effect = 0.05) | | | | | | Marginal model, with complementary log-log link (Sex effect = (-0.4811, -0.4244, -0.2740)*, Time effect = (0.0476, 0.0400, 0.0330)*) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GLMM[1] | | | GLMM[2] | | | GEE[1] | | | GEE[2] | | |
| $\sigma^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -0.7072 | -0.7022 | -0.7263 | -0.7143 | -0.7081 | -0.7307 | -0.6786 | -0.5840 | -0.3976 | -0.6863 | -0.5933 | -0.4137 |
| Bias | -0.2072 | -0.2022 | -0.2263 | -0.2143 | -0.2081 | -0.2307 | -0.1975 | -0.1596 | -0.1236 | -0.2052 | -0.1689 | -0.1397 |
| SE | 0.1469 | 0.1867 | 0.3468 | 0.1484 | 0.1876 | 0.3479 | 0.1437 | 0.1577 | 0.2222 | 0.1454 | 0.1589 | 0.2227 |
| Cov. | 0.674 | 0.858 | 0.888 | 0.660 | 0.850 | 0.888 | 0.696 | 0.872 | 0.888 | 0.682 | 0.866 | 0.874 |
| Time | - | - | - | 0.0393 | 0.0376 | 0.0352 | - | - | - | 0.0369 | 0.0307 | 0.0247 |
| Bias | - | - | - | -0.0107 | -0.0124 | -0.0148 | - | - | - | -0.0107 | -0.0093 | -0.0083 |
| SE | - | - | - | 0.0179 | 0.0190 | 0.0223 | - | - | - | 0.0173 | 0.0166 | 0.0142 |
| Cov. | - | - | - | 0.930 | 0.892 | 0.876 | - | - | - | 0.924 | 0.910 | 0.908 |
| AIC | 1206.27 | 1230.42 | 1128.42 | 1203.11 | 1227.45 | 1126.25 | - | - | - | - | - | - |
| BIC | 1226.63 | 1250.78 | 1148.78 | 1228.56 | 1252.90 | 1151.70 | - | - | - | - | - | - |
| | Random effects model, with logit link (Sex effect = (-0.5836, -0.6137, -0.6333)*, Time effect = (0.0444, 0.0413, 0.0410)*) | | | | | | Marginal model, with logit link (Sex effect = (-0.5347, -0.4823, -0.3338)*, Time effect = (0.0410, 0.0321, 0.0211)*) | | | | | |
| | GLMM[1] | | | GLMM[2] | | | GEE[1] | | | GEE[2] | | |
| $\sigma^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -0.7983 | -0.8357 | -0.9196 | -0.8003 | -0.8375 | -0.9213 | -0.7321 | -0.6533 | -0.4813 | -0.7352 | -0.6564 | -0.4833 |
| Bias | -0.2147 | -0.2220 | -0.2863 | -0.2167 | -0.2238 | -0.2880 | -0.1974 | -0.1710 | -0.1475 | -0.2005 | -0.1741 | -0.1495 |
| SE | 0.1762 | 0.2345 | 0.4523 | 0.1764 | 0.2349 | 0.4532 | 0.1600 | 0.1824 | 0.2402 | 0.1604 | 0.1826 | 0.2402 |
| Time | - | - | - | 0.0315 | 0.0272 | 0.0214 | - | - | - | 0.0293 | 0.0214 | 0.0113 |
| Bias | - | - | - | -0.0129 | -0.0141 | -0.0196 | - | - | - | -0.0117 | -0.0107 | -0.0098 |
| SE | - | - | - | 0.0213 | 0.0231 | 0.0294 | - | - | - | 0.0197 | 0.0180 | 0.0155 |
| AIC | 1294.93 | 1303.51 | 1178.89 | 1294.12 | 1303.26 | 1179.16 | - | - | - | - | - | - |
| BIC | 1315.29 | 1323.86 | 1199.25 | 1319.57 | 1328.71 | 1204.61 | - | - | - | - | - | - |

[1] Without time-dependent covariates in the model; [2] With time-dependent covariates in the model; * True sex and time effects come from Table 2.2

indicate that the proposed GLMMs are better than the standard GLMMs with logit link. To see the influence of covariates' effect size on different mis-specified models, the fifth scenario, all models with moderate covariates' effect on outcomes which were based on the time-varying intensity $\lambda_{23}(\mathbf{X}_{ij})$ are shown in Table 2.5 with similar settings as the fourth scenario in Table 2.4, and boxplots are shown in Figure 2.5. With the size of covariates' effect changed from large to moderate, the performances of all models are better with smaller bias and higher coverage for "sex" and "time" effect under various population heterogeneity.

Table 2.6 shows results from the sixth scenario, where all marginal models and generalized linear mixed effects models were mis-specified with large covariates' effect and sample size 200 on outcomes generated through the Poisson distribution with time-varying intensity $\lambda_{12}(\mathbf{X}_{ij})$, and boxplots are shown in Figure 2.6. The performances of all models are much better compared with Table 2.4 with smaller bias and higher coverages under various population heterogeneity. For "sex" effect, the inference is better behaved in the over-fitting GLMM$^3$ with complementary log-log link than the GLMM$^1$. As model selection criteria, AIC and BIC values indicate that the proposed GLMMs are better than the standard GLMMs with logit link. To see the influence of covariates' effect size on different mis-specified models, the seventh scenario, all models with moderate covariates' effect on outcomes which were based on the time-varying intensity $\lambda_{22}(\mathbf{X}_{ij})$ are shown in Table 2.7 with similar settings as the sixth scenario in Table 2.6, and boxplots are shown in Figure 2.7. Comparing the results in Table 2.7 and Table 2.6, all models' inferences are similar

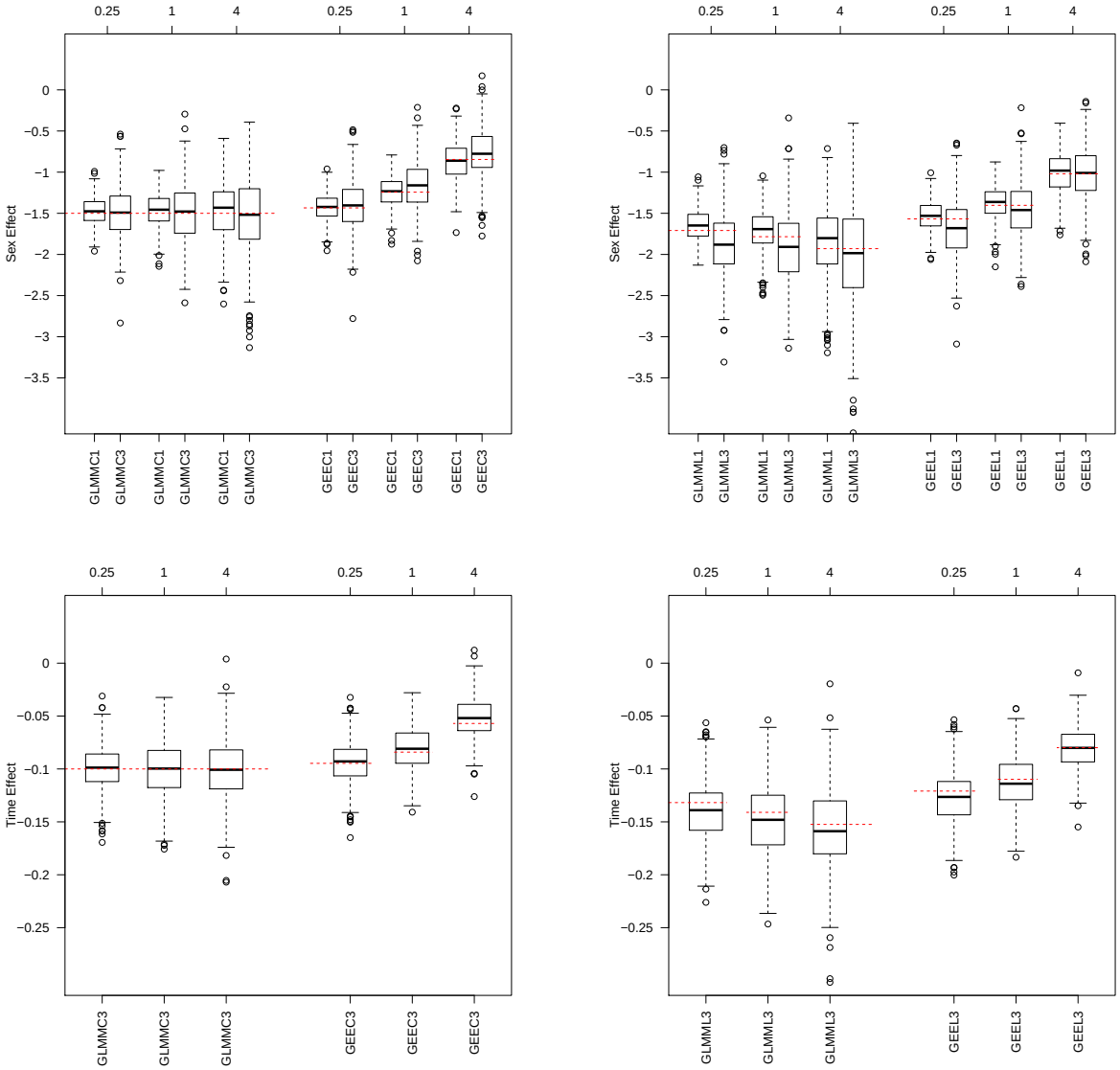


Figure 2.6: Boxplots on sex and time effect estimates using incorrectly specified models missing a large time effect term, or over-fitting model with correct model nested, under various study population heterogeneity and large covariates' effect on outcomes based on $\lambda_{12}(\mathbf{X}_{ij})$ ($n = 200$)

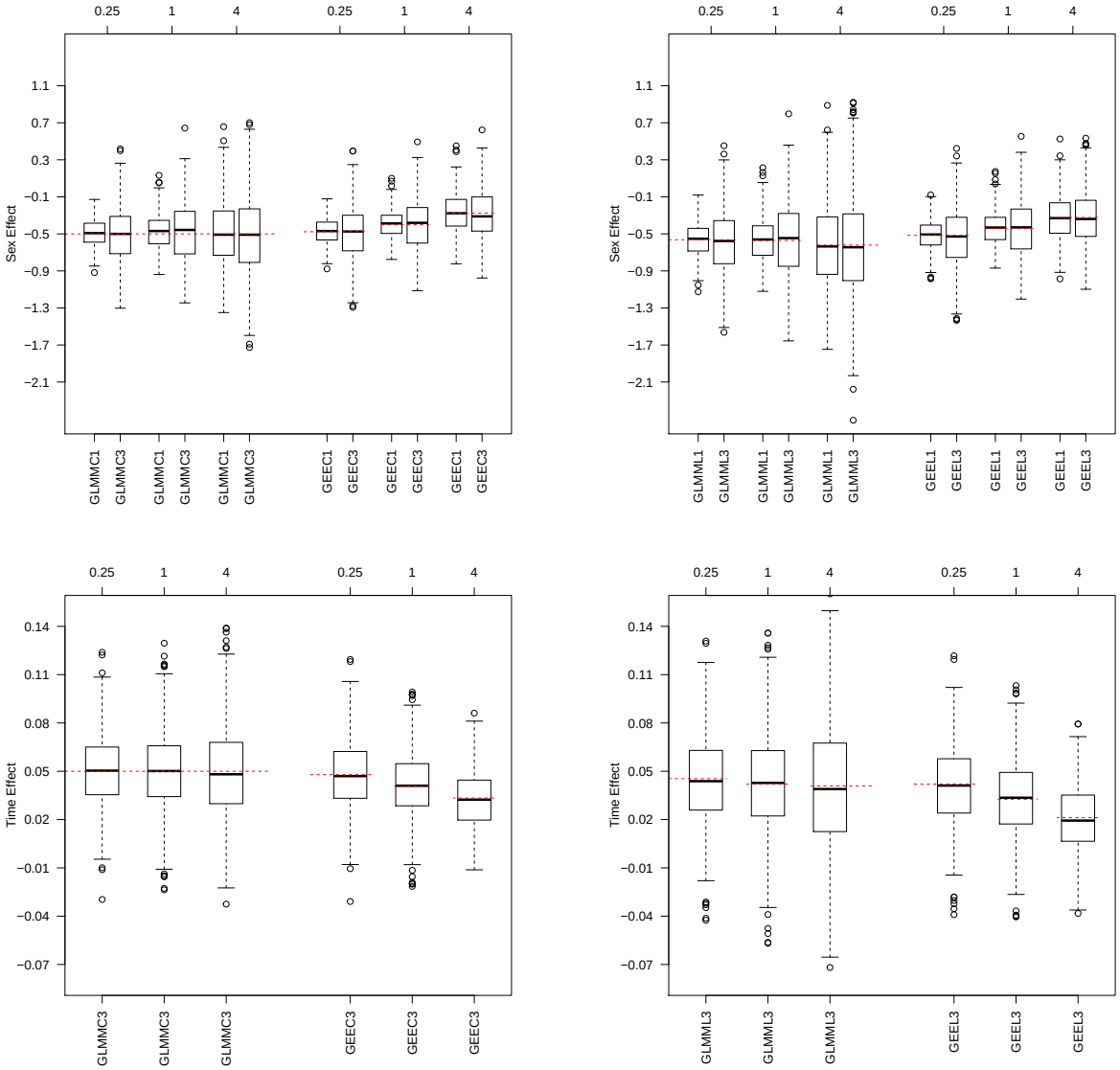


Figure 2.7: Boxplots on sex and time effect estimates using incorrectly specified models missing a large time effect term, or over-fitting model with correct model nested, under various study population heterogeneity and moderate covariates' effect on outcomes based on $\lambda_{22}(\mathbf{X}_{ij})$ ($n = 200$)

Table 2.6: Sensitivity of model inferences to missing a large time effect term, or over-fitting model with correct model nested, under various study population heterogeneity and large covariates' effect on outcomes based on $\lambda_{12}(\mathbf{X}_{ij})$ $(n = 200)$

| | Random effects model, with complementary log-log link (True Sex effect = -1.5, True Time effect = -0.1) | | | | | | Marginal model, with complementary log-log link (Sex effect = (-1.4342, -1.2428, -0.8455)*, Time effect = (-0.0947, -0.0841, -0.0570)*) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GLMM[1] | | | GLMM[3] | | | GEE[1] | | | GEE[3] | | |
| $\sigma^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -1.4760 | -1.4619 | -1.4692 | -1.4953 | -1.4883 | -1.5305 | -1.4243 | -1.2428 | -0.8666 | -1.4071 | -1.1632 | -0.7634 |
| Bias | 0.0240 | 0.0381 | 0.0308 | 0.0047 | 0.0117 | -0.0305 | 0.0099 | 0 | -0.0211 | 0.0271 | 0.0796 | 0.0821 |
| SE | 0.1641 | 0.2039 | 0.3480 | 0.3184 | 0.3483 | 0.4752 | 0.1603 | 0.1758 | 0.2312 | 0.3080 | 0.2950 | 0.2889 |
| Cov. | 0.930 | 0.946 | 0.944 | 0.950 | 0.954 | 0.948 | 0.946 | 0.960 | 0.930 | 0.946 | 0.934 | 0.918 |
| Time | - | - | - | -0.0991 | -0.1004 | -0.1004 | - | - | - | -0.0937 | -0.0804 | -0.0516 |
| Bias | - | - | - | 0.0009 | -0.0004 | -0.0004 | - | - | - | 0.0010 | 0.0037 | 0.0054 |
| SE | - | - | - | 0.0216 | 0.0256 | 0.0286 | - | - | - | 0.0202 | 0.0207 | 0.0174 |
| Cov. | - | - | - | 0.952 | 0.934 | 0.954 | - | - | - | 0.952 | 0.936 | 0.956 |
| AIC | 1135.94 | 1160.46 | 1096.18 | 1111.57 | 1136.24 | 1076.45 | - | - | - | - | - | - |
| BIC | 1156.31 | 1180.82 | 1116.54 | 1142.11 | 1166.78 | 1106.99 | - | - | - | - | - | - |
| | Random effects model, with logit link (Sex effect = (-1.7090, -1.7843, -1.9296)*, Time effect = (-0.1318, -0.1410, -0.1524)*) | | | | | | Marginal model, with logit link (Sex effect = (-1.5682, -1.4030, -1.0196)*, Time effect = (-0.1209, -0.1098, -0.0799)*) | | | | | |
| | GLMM[1] | | | GLMM[3] | | | GEE[1] | | | GEE[3] | | |
| $\sigma^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -1.6433 | -1.7095 | -1.8401 | -1.8690 | -1.9073 | -2.0003 | -1.5282 | -1.3721 | -1.0079 | -1.6864 | -1.4478 | -1.0156 |
| Bias | 0.0657 | 0.0748 | 0.0895 | -0.1600 | -0.1230 | -0.0707 | 0.0400 | 0.0309 | 0.0117 | -0.1182 | -0.0448 | 0.0040 |
| SE | 0.1921 | 0.2502 | 0.4508 | 0.3821 | 0.4406 | 0.6333 | 0.1761 | 0.1968 | 0.2470 | 0.3505 | 0.3388 | 0.3257 |
| Time | - | - | - | -0.1399 | -0.1481 | -0.1570 | - | - | - | -0.1268 | -0.1124 | -0.0797 |
| Bias | - | - | - | -0.0081 | -0.0071 | -0.0046 | - | - | - | -0.0059 | -0.0026 | 0.0002 |
| SE | - | - | - | 0.0281 | 0.0331 | 0.0388 | - | - | - | 0.0253 | 0.0248 | 0.0199 |
| AIC | 1223.46 | 1234.99 | 1151.87 | 1194.71 | 1205.35 | 1126.08 | - | - | - | - | - | - |
| BIC | 1243.83 | 1255.35 | 1172.23 | 1225.25 | 1235.89 | 1156.62 | - | - | - | - | - | - |

[1] Without time-dependent covariates in the model; [3] With time-dependent covariates and time interaction term in the model; * True sex and time effects come from Table 2.1

Table 2.7: Sensitivity of model inferences to missing a large time effect term, or over-fitting model with correct model nested, under various study population heterogeneity and moderate covariates' effect on outcomes based on $\lambda_{22}(\mathbf{X}_{ij})$ $(n = 200)$

| | Random effects model, with complementary log-log link (True Sex effect = -0.5, True Time effect = 0.05) | | | | | | Marginal model, with complementary log-log link (Sex effect = (-0.4770, -0.4014, -0.2762)*, Time effect = (0.0479, 0.0407, 0.0334)*) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GLMM[1] | | | GLMM[3] | | | GEE[1] | | | GEE[3] | | |
| $\sigma^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -0.4930 | -0.4788 | -0.4871 | -0.5050 | -0.4660 | -0.5050 | -0.4723 | -0.3967 | -0.2737 | -0.4852 | -0.3893 | -0.2921 |
| Bias | 0.0070 | 0.0212 | 0.0129 | -0.0050 | 0.0340 | -0.0050 | 0.0047 | 0.0047 | 0.0025 | -0.0082 | 0.0121 | -0.0159 |
| SE | 0.1410 | 0.1773 | 0.3409 | 0.2986 | 0.3200 | 0.4415 | 0.1373 | 0.1486 | 0.2139 | 0.2886 | 0.2749 | 0.2762 |
| Cov. | 0.956 | 0.972 | 0.946 | 0.972 | 0.966 | 0.940 | 0.940 | 0.964 | 0.928 | 0.968 | 0.966 | 0.962 |
| Time | - | - | - | 0.0507 | 0.0505 | 0.0488 | - | - | - | 0.0476 | 0.0415 | 0.0324 |
| Bias | - | - | - | 0.0007 | 0.0005 | -0.0012 | - | - | - | -0.0003 | 0.0008 | -0.0010 |
| SE | - | - | - | 0.0233 | 0.0252 | 0.0293 | - | - | - | 0.0223 | 0.0209 | 0.0180 |
| Cov. | - | - | - | 0.960 | 0.942 | 0.940 | - | - | - | 0.934 | 0.948 | 0.968 |
| AIC | 1258.36 | 1278.35 | 1152.99 | 1252.85 | 1273.06 | 1148.94 | - | - | - | - | - | - |
| BIC | 1278.72 | 1298.71 | 1173.36 | 1283.39 | 1303.60 | 1179.49 | - | - | - | - | - | - |
| | Random effects model, with logit link (Sex effect = (-0.5633, -0.5716, -0.6210)*, Time effect = (0.0454, 0.0420, 0.0409)*) | | | | | | Marginal model, with logit link (Sex effect = (-0.5160, -0.4432, -0.3226)*, Time effect = (0.0420, 0.0326, 0.0212)*) | | | | | |
| | GLMM[1] | | | GLMM[3] | | | GEE[1] | | | GEE[3] | | |
| $\sigma^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -0.5604 | -0.5691 | -0.6177 | -0.5852 | -0.5622 | -0.6300 | -0.5132 | -0.4414 | -0.3216 | -0.5379 | -0.4392 | -0.3285 |
| Bias | 0.0029 | 0.0025 | 0.0033 | -0.0219 | 0.0094 | -0.0090 | 0.0028 | 0.0018 | 0.0010 | -0.0219 | 0.0040 | -0.0059 |
| SE | 0.1701 | 0.2275 | 0.4457 | 0.3555 | 0.4025 | 0.5821 | 0.1552 | 0.1755 | 0.2336 | 0.3266 | 0.3115 | 0.3031 |
| Time | - | - | - | 0.0444 | 0.0426 | 0.0405 | - | - | - | 0.0409 | 0.0329 | 0.0208 |
| Bias | - | - | - | -0.0010 | 0.0006 | -0.0004 | - | - | - | -0.0011 | 0.0003 | -0.0004 |
| SE | - | - | - | 0.0289 | 0.0326 | 0.0396 | - | - | - | 0.0264 | 0.0248 | 0.0201 |
| AIC | 1351.44 | 1352.33 | 1202.40 | 1349.32 | 1351.06 | 1202.26 | - | - | - | - | - | - |
| BIC | 1371.80 | 1372.69 | 1222.77 | 1379.87 | 1381.59 | 1232.81 | - | - | - | - | - | - |

[1] Without time-dependent covariates in the model; [3] With time-dependent covariates and time interaction term in the model; * True sex and time effects come from Table 2.2

under various population heterogeneity.

Table 2.8: Sensitivity of model inferences to missing a large sex and time effect terms, under various study population heterogeneity on outcomes based on $\lambda_3(\mathbf{X}_{ij})$ $(n = 200)$

| | Random effects model estimated by MLE, with complementary log-log link (No Sex and Time effect in the true model) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | GLMM[1] | | | GLMM[2] | | | GLMM[3] | | |
| $\sigma^2$ (heterogeneity) | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -0.0050 | 0.0028 | -0.0054 | -0.0048 | 0.0029 | -0.0054 | 0.0096 | 0.0004 | -0.0131 |
| SE | 0.1212 | 0.1735 | 0.3272 | 0.1214 | 0.1737 | 0.3280 | 0.2188 | 0.2747 | 0.3968 |
| Power (%) | 0.050 | 0.036 | 0.042 | 0.050 | 0.036 | 0.042 | 0.052 | 0.040 | 0.042 |
| Time | - | - | - | -0.0003 | 0.0008 | -0.0013 | 0.0007 | 0.0007 | -0.0020 |
| SE | - | - | - | 0.0134 | 0.0146 | 0.0183 | 0.0188 | 0.0217 | 0.0266 |
| Power (%) | - | - | - | 0.046 | 0.044 | 0.040 | 0.056 | 0.058 | 0.042 |
| | Marginal model estimated by GEE, with complementary log-log link | | | | | | | | |
| | GEE[1] | | | GEE[2] | | | GEE[3] | | |
| $\sigma^2$ (heterogeneity) | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ | $0.5^2$ | $1.0^2$ | $2^2$ |
| Sex | -0.0045 | -0.0011 | -0.0057* | -0.0043 | -0.0011 | -0.0018* | 0.0101 | -0.0024 | -0.0060* |
| SE | 0.1132 | 0.1442 | 0.2183* | 0.1133 | 0.1445 | 0.2369* | 0.2030 | 0.2140 | 0.2436* |
| Power (%) | 0.046 | 0.058 | 0.100 | 0.046 | 0.062 | 0.104 | 0.058 | 0.064 | 0.062 |
| Time | - | - | - | 0.0008 | 0.0041 | 0.0063* | 0.0019 | 0.0041 | 0.0061* |
| SE | - | - | - | 0.0124 | 0.0113 | 0.0102* | 0.0173 | 0.0169 | 0.0148* |
| Power (%) | - | - | - | 0.042 | 0.046 | 0.086 | 0.058 | 0.050 | 0.060 |

[1] Without time-dependent covariates in the model; [2] With time-dependent covariates in the model; [3] With time-dependent covariates and time interaction term in the model; * Trim the unstable values which are too big ($> 1000$) or too small ($< -1000$)

Table 2.8 shows results from the eighth scenario, where all marginal models and generalized linear mixed effects models were mis-specified with sample size 200 on outcomes generated through the Poisson distribution with time-fixed intensity $\lambda_3(\mathbf{X}_{ij})$. When there are no true "sex" and/ or "time" effect in the data, the proposed GLMM and marginal models with complementary log-log link cannot capture those significant signals, based on low power.

## 2.5 Case Study

After excluding subjects with missing values at baseline, the total sample size of BHS-7 study was 296 with 144 men and 152 women. Table 2.9 shows demographic and clinical characteristics of hip fracture patients by gender at baseline. Regarding the three continuous characteristics of patients, men patients have a significantly higher Charlson comorbidity index than women with p-value 0.0002. And, men and women have similar age and body mass index in this study sample. Regarding the categorical characteristics of patients, women have a marginally significant higher percentage on education (completed high school), and elevated white blood cell count on admission (greater than or equal to 13.6 K/mcL) than men.

Table 2.9: Demographic and clinical characteristics of hip fracture patients at baseline, by gender

| Characteristics | All ($n = 296$) | Men ($n = 144$) | Women ($n = 152$) | |
|---|---|---|---|---|
| | Mean (SD) | Mean (SD) | Mean (SD) | P-value |
| Age, years | 80.70 (7.86) | 80.26 (7.77) | 81.12 (7.95) | 0.3509 |
| Body mass index | 25.29 (5.14) | 25.51 (4.51) | 25.09 (5.67) | 0.4753 |
| Charlson comorbidity index | 2.01 (1.77) | 2.40 (1.88) | 1.64 (1.59) | **0.0002** |
| | % | % | % | P-value |
| Completed high school | 79.4 | 75.0 | 83.4 | 0.0753 |
| White race | 91.1 | 89.4 | 92.7 | 0.3327 |
| Elevated white blood cell count on admission ($\geq 13.6$ vs. $< 13.6$) | 21.0 | 16.7 | 25.2 | 0.0733 |
| Urinary tract infection during hospital stay | 19.3 | 16.0 | 22.4 | 0.1631 |

Figure 2.8 explores longitudinal trajectories of monthly infection by gender during the first-year post-hip fracture recovery period. Prevalence of infection between men and women do not have big differences across the time except for the month 2, 6, and 12 after hip fracture. For month 2, 6, and 12, the prevalence of infection is much higher for men compared with women. Because the prevalence of infection

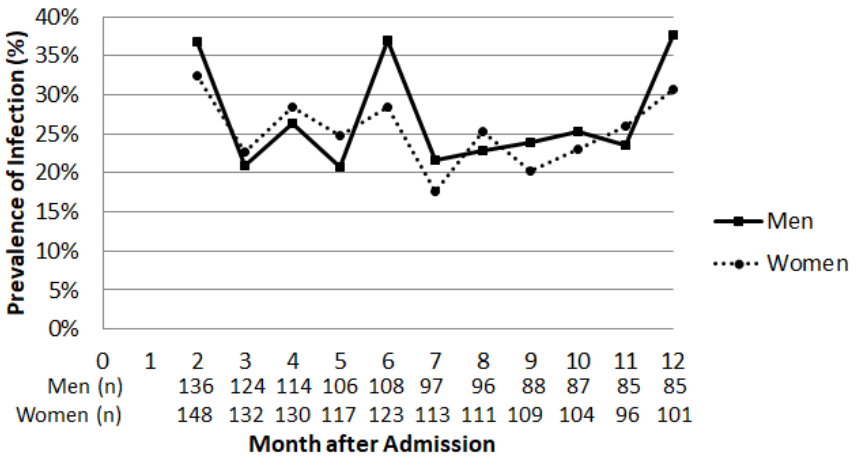


Figure 2.8: Longitudinal trajectories of monthly infection by gender during the first-year post-hip fracture recovery period

for men is higher at some time points but lower at the other time points, it is very difficult to tell the difference between men and women from the figure. Also, the trajectories between men and women are very similar across time, showing no difference of infectious pattern between men and women across time.

To quantify the "sex" effect and capture the "time" trajectory trend of post surgery infection over the first year of recovery period, we fit all 12 marginal models and generalized linear mixed effects models with both complementary log-log link and logit link (2.5 - 2.8), with controlling for all confounding factors. Results are summarized in Table 2.10. With moderate and significant population heterogeneity ($\hat{\sigma}^2 = 1$, P $< 0.0001$), moderate covariates' effect, and insignificant time interaction term, Table 2.10 shows that GLMM$^2$ estimated by MLE with complementary log-log link should be the chosen model, which has the smallest AIC and BIC. All of the models show, on average over time, insignificant "sex" differences on infection. Only GLMMs with

Table 2.10: Comparisons of model performances quantifying risk factors' effects on infectious outcomes, after controlling for confounders, including education, race, CCI, BMI, admission WBC count, and in-hospital UTI

| | Random effects model estimated by MLE, with complementary log-log link | | | Marginal model estimated by GEE, with complementary log-log link | | |
|---|---|---|---|---|---|---|
| | GLMM[1] | **GLMM[2]** | GLMM[3] | GEE[1] | GEE[2] | GEE[3] |
| $\hat{\sigma}^2$ (heterogeneity) | 0.99 | **1.00** | 1.00 | - | - | - |
| Sex | -0.0035 | **-0.0115** | 0.1142 | 0.0050 | -0.0018 | 0.0806 |
| SE | 0.1591 | **0.1598** | 0.2337 | 0.1325 | 0.1332 | 0.2134 |
| P-value | 0.9827 | **0.9426** | 0.6252 | 0.9701 | 0.9892 | 0.7057 |
| Age | -0.0212 | **-0.0214** | -0.0217 | -0.0171 | -0.0168 | -0.0167 |
| SE | 0.0105 | **0.0105** | 0.0105 | 0.0088 | 0.0088 | 0.0088 |
| P-value | **0.0432** | **0.0425** | **0.0392** | 0.0514 | 0.0571 | 0.0574 |
| Time | - | **0.0298** | 0.0399 | - | 0.0264 | 0.0331 |
| SE | - | **0.0136** | 0.0194 | - | 0.0136 | 0.0198 |
| P-value | - | **0.0284** | **0.0396** | - | 0.0524 | 0.0937 |
| Time×Sex | - | - | -0.0205 | - | - | -0.0131 |
| SE | - | - | 0.0271 | - | - | 0.0271 |
| P-value | - | - | 0.4501 | - | - | 0.6276 |
| AIC | 2372.23 | **2369.54** | 2370.89 | - | - | - |
| BIC | 2408.86 | **2409.83** | 2414.85 | - | - | - |
| | Random effects model estimated by MLE, with logit link | | | Marginal model estimated by GEE, with logit link | | |
| | GLMM[1] | GLMM[2] | GLMM[3] | GEE[1] | GEE[2] | GEE[3] |
| $\hat{\sigma}^2$ (heterogeneity) | 1.62 | 1.62 | 1.63 | - | - | - |
| Sex | -0.0438 | -0.0333 | 0.0991 | -0.0241 | -0.0240 | 0.0763 |
| SE | 0.1998 | 0.1998 | 0.2983 | 0.1562 | 0.1562 | 0.2649 |
| P-value | 0.8265 | 0.8678 | 0.7399 | 0.8775 | 0.8780 | 0.7732 |
| Age | -0.0283 | -0.0265 | -0.0269 | -0.0221 | -0.0221 | -0.0221 |
| SE | 0.0132 | 0.0132 | 0.0132 | 0.0108 | 0.0108 | 0.0108 |
| P-value | **0.0319** | **0.0442** | **0.0419** | **0.0408** | **0.0407** | **0.0402** |
| Time | - | 0.0007 | 0.0120 | - | -0.0005 | 0.0077 |
| SE | - | 0.0174 | 0.0251 | - | 0.0168 | 0.0237 |
| P-value | - | 0.9665 | 0.6342 | - | 0.9752 | 0.7441 |
| Time×Sex | - | - | -0.0217 | - | - | -0.0159 |
| SE | - | - | 0.0348 | - | - | 0.0335 |
| P-value | - | - | 0.5337 | - | - | 0.6353 |
| AIC | 2449.64 | 2451.67 | 2453.31 | - | - | - |
| BIC | 2486.27 | 2491.96 | 2497.26 | - | - | - |

[1] Without time-dependent covariates in the model; [2] With time-dependent covariates in the model; [3] With time-dependent covariates and time interaction term in the model

complementary log-log link show significant "time" effect, which is consistent with the simulation results that the proposed GLMM with complementary log-log link can mostly capture the significance of longitudinal effect. The significant "time" effect from the final model (P = 0.0284) can be explained as the hazard of infection re-occurrence increases by 1.0302 ($e^{0.0298}$) time with one month increase on time for an individual after controlling for the other covariates and the random effect. All GLMM and GEE models with logit link and complementary log-log link show consistent and significant "age" effect on infection re-occurrence. In Aging study, age is the natural biological risk factor contributing to outcome, which is a commonly expected result. The significant "age" effect (P = 0.0425) on infection re-occurrence from the final model shows that one year increase on age for an individual is associated with a 0.0212 ($1 - e^{-0.0214}$) decrease in the hazard of infection re-occurrence after controlling for the other covariates and the random effect, while the decrease is 0.0219 ($1 - e^{-0.0221}$) in the odds of infection re-occurrence for the population according to the GEE[2] with logit link. With the consistent results as the simulation, GLMMs with logit link also have the largest standard errors compared with the other models, and higher AIC, BIC, showing poor performance of the GLMMs with logit link. Therefore, even though the GEE with logit link can have similar effects on time-fixed covariates as the GLMM with complementary log-log link when the random effect is moderate, the explanations are totally different from the two models.

After comparing different models in Table 2.10, Table 2.11 shows the results for

all risk factors using the final model GLMM$^2$. In addition to the significant "time" and "age" effect discussed in the last paragraph, only the risk factor "Charlson comorbidity index" shows significant effect, which means that the hazard of infection re-occurrence increases by 1.2269 ($e^{0.2045}$) time with one unit increase on Charlson comorbidity index for an individual after controlling for the other covariates and the random effect.

Table 2.11: Longitudinal analysis quantifying risk factors' effect on infectious outcomes using the final model GLMM$^2$ with com(log-log) link

| Risk Factors | Estimate | 95% CI | P-value |
|---|---|---|---|
| Sex | -0.0115 | (-0.3249, 0.3019) | 0.9426 |
| Time | 0.0298 | (0.0032, 0.0564) | **0.0284** |
| Age | -0.0214 | (-0.0420, -0.0007) | **0.0425** |
| Education | -0.0140 | (-0.0605, 0.0326) | 0.5564 |
| Race | 0.1890 | (-0.3990, 0.7771) | 0.5285 |
| Charlson comorbidity index | 0.2045 | (0.1169, 0.2921) | **<.0001** |
| Body mass index | -0.0080 | (-0.0380, 0.0219) | 0.5984 |
| Elevated white blood cell count on admission | 0.0336 | (-0.3438, 0.4111) | 0.8613 |
| Urinary tract infection during hospital stay | 0.2080 | (-0.1860, 0.6020) | 0.3006 |

## 2.6 Discussion

Even though interval reported binary recurrent event data are commonly seen in longitudinal studies, marginal models or generalized linear mixed effects models with logit link without considering interval reporting are usually used. To consider interval reporting, we developed a relatively simple and flexible longitudinal model framework, using discrete survival modeling technique accounting for interval censored reporting system between longitudinal visits and Poisson process accounting for binary nature of recurrent events reporting within each interval. The intensity in the Poisson process follows a Cox proportional hazards model allowing for both

time-fixed and time-varying covariates, which leads to varying intensities across the longitudinal visits but the intensity stays fixed within each reporting interval. This model setting simplified the joint likelihood into a generalized linear mixed effects model with binary responses and complementary log-log link, which can be estimated by widely available software.

Without considering interval reporting, the standard marginal models or generalized linear mixed effects models with logit link are usually used, measuring log odds ratios of the event of interest. With considering interval reporting, the proposed generalized linear mixed effects models with complementary log-log link allowing for population heterogeneity in baseline hazards, as well as marginal models with complementary log-log link are used for measuring log hazard ratios of the event of interest. To evaluate the numerical performances of the proposed models, simulation studies were carried out to compare them with the standard longitudinal models with logit link, which have the following conclusions dealing with longitudinal interval reported binary recurrent event data. First, the proposed GLMMs with complimentary log-log link have the best performance with stable and unbiased "sex" and "time" effect estimates, high power, and small AIC & BIC values, regardless of the population heterogeneity, covariates' effect size, and sample size; while the standard GLMMs with logit link have poor performance all the time. Second, with the size of covariates' effect changed from large to moderate, the proposed GLMMs with complementary log-log can mostly capture the significance of "sex" and "time" effect with the highest percentage on power compared with the

other models, especially for longitudinal effect "time". Third, GEEs with both logit link and complementary log-log link can also have similar effect estimates as the proposed GLMMs under small population heterogeneity and moderate covariates' effect size regardless of sample size, especially for time-fixed covariates. Fourth, when a true model includes time-dependent covariates but without time interaction terms, all models are not sensitive to model mis-specification with small bias and high coverages with or without time-dependent covariates, especially for models with moderate covariates' effect. For "sex" effect, the inference is better behaved in the over-fitting GLMM with complementary log-log link than the GLMM without considering the time effect term. Fifth, when a true model includes time-dependent covariates and time interaction terms, all models' inferences are sensitive to model mis-specification with large bias and small coverages, especially for models with large covariates' effect. Therefore, when random effects are small, both GLMMs with complimentary log-log link and GEEs with logit link or complementary log-log can be used for longitudinal interval reported binary recurrent event data, but GEEs are not as good as GLMMs capturing the significant effects on time-varying covariates. Overall, the proposed GLMMs with complimentary log-log link have the best performance.

Regarding the BHS-7 study, whether considering interval reporting or not, the results from all models show that there is no significant difference between men and women hip fracture patients on infection re-occurrence during the 12-month post-discharge follow-up interval. Only GLMMs with complementary log-log link show

significant "time" effect, which is consistent with the simulation results. Because random effects are moderate and significant in the real data analysis, the marginal model with logit link and generalized linear mixed effects model with complementary log-log link show very similar effects on time-fixed covariates but with different explanations, which is also consistent with the simulation results.

In summary, the proposed generalized linear mixed effects model with complementary log-log can deal with interval reported binary recurrent event data but it has its limitations. First, our simulation scenarios are not comprehensive, which needs to be broadened including more sensitivity analysis, e.g., scenarios with small covariates' effect of gender and time, to see their performances on capturing significant signals; adding $t_{ij} - t_{i(j-1)}$ or the number of intervals as a covariate to those standard longitudinal models with logit link, to see if they can take care of the interval reporting issue. Second, we only consider Cox proportional hazards model for the intensity, and more intensity structures will be discussed in the future, e.g., adding random slopes to the intensity in addition to the random intercept. Third, the performance of the proposed method can be poor if one of the assumptions is violated, e.g., the fixed intensity within each interval. Further investigations are needed to explore the proposed method to make it more efficient.

# Chapter 3

# Statistical Model for Subgroup Identification in Enrichment Design

## 3.1 Introduction

The enrichment design was defined as the prospective use of any patient characteristic to select a study population in which detection of a drug effect (if one is in fact present) is more likely than it would be in an unselected population by FDA (2012). According to the enrichment strategies used, there are three broad categories, strategies to decrease heterogeneity, prognostic enrichment strategies, and predictive enrichment strategies. Strategies to decrease heterogeneity include selecting patients with decreased inter-patient variability and decreased intra-patient variability; prognostic enrichment strategies choose patients with a greater likelihood of having a disease-related endpoint event or a substantial worsening in condition; predictive enrichment strategies choose patients more likely to respond to the drug treatment than other patients with the condition being treated. For example, in the JUPITER study (Ridker et al., 2008), the rosuvastatin was effective reducing the incidence of major cardiovascular events in patients with LDL cholesterol levels of less than 130 mg per deciliter but with elevated high-sensitivity C-reactive protein levels of greater than or equal to 2.0 mg per liter, where LDL cholesterol and CRP are two prognostic biomarkers with high values indicating high risk of cardiovascu-

lar events. For the drug erlotinib, there was a highly significant survival difference for EGFR-positive patients, while only little effects seen among the EGFR-negative patients (Temple, 2005), where EGFR is a predictive biomarker. Therefore, enrichment strategies can give us efficient and powerful results with smaller sample size, shortened development time, and reduced cost. But how to select patients that maximize benefits from a treatment based on the information from the phase II and prior scientific knowledge becomes an important step for enrichment design.

The enrichment design was originally defined as the additional screening processes with the active treatments evaluated in the study by Temple (1994). The additional screening processes with the active treatments were performed after either the screening period or the placebo run-in period to identify potential patients who are likely to benefit the test drug in the early phase of the trial (Liu, 2003). It can be also called randomized withdrawal design or randomized discontinuation design, which belongs to predictive enrichment strategies. Temple (2005) mentioned that the randomized withdrawal design is considerably more efficient if there is a responder subpopulation, especially when the responder population is relatively low (30%). During the past decades, many advanced clinical trial designs emerged. Compared to the widely used design strategy implementing a placebo lead-in phase prior to randomization that only reduces placebo rate, the sequential parallel comparison design (SPCD), or sequential parallel design (SPD), is a clinical trial methodology developed by Fava et al. (2003) to reduce both the overall placebo response rate and the sample size for double-blind, placebo-controlled trials in psychiatric disorders.

This novel study design can reduce the cost and time remarkably for the evaluation of new drugs. Reducing placebo response rate belongs to strategies to decrease heterogeneity. Unlike the standard designs that enroll a broad range of subjects to decide a subset of subjects that may benefit from the treatment, Simon and Simon (2013) introduced a class of adaptive enrichment designs that allow the enrollment criteria to change adaptively during the trial. However, the whole process is very complicated and time consuming. With adaptive interventions, which are different from the previous adaptive enrichment designs, Sequential Multiple Assignment Randomized Trials (SMARTs) involve multiple intervention stages, and each participant moves through the multiple stages to build optimal adaptive interventions (Lavori and Dawson, 2000, 2004; Murphy, 2005; Lei et al., 2012). The SMART designs have been conducted in several clinical trials in recent years. Even though there are lot of advanced clinical trial designs, their aim is to improve treatment efficiency and reduce risks with choosing the right subset of population who are benefit from the treatment before the study or during the study.

If the overall treatment effect is not significant from the phase III clinical trial of standard designs, it is known that subgroup analysis is popular to identify a subgroup that benefits from the treatment. However, a lot of patients are under risk during the phase III clinical trial if they do not benefit from the treatment. In this paper, we focus on choosing a subset of population who will benefit from the treatment before phase III clinical trial. The subset is identified based on a selection model built from the information from the phase II randomized clinical trial.

## 3.2 Methods

For a phase II randomized clinical trial for the patient $i(i = 1, \ldots, n)$, let $\mathbf{x}_i$ be all available information on the patient; $z_i$ be the treatment actually received by the patient in phase II, which can be 0 or 1 with 0 means control group and 1 means treatment group; $Y_i$ be the observed outcome. Then, the observed data for the $n$ independent and identically distributed patients from phase II can be written as $(\mathbf{x}_i, z_i, Y_i)$. With the observed data from phase II, the goal is to build a selection model recruiting patients into phase III that can maximize benefits from the treatment.

Let $s(\mathbf{X})$ be the selection model, it can be 0 or 1 based on values of $\mathbf{X}$. For a new patient with covariates $\mathbf{x}$, the patient would be selected if $s(\mathbf{x}) = 1$ or would not be selected if $s(\mathbf{x}) = 0$. For example, if $x = \{\text{Age}\}$, then $s(x) = I\{\text{Age} < 50\}$, which means that patients less than 50 years of age would be chosen for phase III clinical trial.

To find the optimal selection model, let $Y^*(0)$ and $Y^*(1)$ be potential outcomes for a subject that would receive control or treatment (Rubin, 1974). Then, under the consistency assumption that the observed outcome is the potential outcome that would be seen under the group actually been assigned (Rubin, 1986; Zhang et al., 2012), the observed outcome can be written as $Y = Y^*(1)Z + Y^*(0)(1 - Z)$. Since it is a randomized clinical trial, observed and unobserved covariates' distributions

are balanced between control and treatment groups, which shows that the observed outcome $Y$ is independent of $Z$, or the potential outcomes $\{Y^*(0), Y^*(1)\}$ are independent of $Z$. Thus, $E\{Y^*(z)\}$ is the expected outcome for the population if all patients were to receive treatment $z$ ($z = 0$ or 1). It can be also written as,

$$\begin{aligned} E\{Y^*(z)\} &= E[E\{Y^*(z)|\mathbf{X}\}] = E[E\{Y^*(z)|\mathbf{X}, Z = z\}] \\ &= E[E\{Y^*(1)z + Y^*(0)(1-z)|\mathbf{X}, Z = z\}] = E[E\{Y|\mathbf{X}, Z = z\}]. \end{aligned}$$

Under a selection option $s$, the potential outcome $Y^*(s)$ can be written as $Y^*(s) = Y^*(1)s(\mathbf{X}) + Y^*(0)(1 - s(\mathbf{X}))$, which is for a patient with baseline information $\mathbf{X}$ that were to be selected or not according to $s$. Let $\mathcal{S}$ be the class of all possible selection options. The optimal selection option is denoted as $s^{opt} \in \mathcal{S}$. For a patient being selected under the optimal selection option, the patient's expected outcome would be as large as possible given his/her available information. For all patients being selected under the optimal selection option, the expected outcome for the population would be as large as possible. It can be written as

$$E\{Y^*(s)I\{s(\mathbf{X}) = 1\}|\mathbf{X} = \mathbf{x}\} \leq E\{Y^*(s^{opt})I\{s^{opt}(\mathbf{X}) = 1\}|\mathbf{X} = \mathbf{x}\};$$

$$E\{Y^*(s)I\{s(\mathbf{X}) = 1\}\} \leq E\{Y^*(s^{opt})I\{s^{opt}(\mathbf{X}) = 1\}\}.$$

Therefore, $s^{opt}$ can make $E\{Y^*(s)\}$ have the largest value for the selected population among $s \in \mathcal{S}$.

With the consistency assumption, we have:

$$\begin{aligned} E\{Y^*(s)\} &= E\left[E\{Y^*(s)|\mathbf{X}\}\right] \\ &= E[E\{Y^*(1)s(\mathbf{X}) + Y^*(0)(1-s(\mathbf{X}))|\mathbf{X}\}] \\ &= E[E\{Y|\mathbf{X}, Z=1\}s(\mathbf{X}) + E\{Y|\mathbf{X}, Z=0\}(1-s(\mathbf{X}))]. \end{aligned}$$

Then, the optimal selection option can be written as,

$$s^{opt}(\mathbf{x}) = I\{E[Y|\mathbf{X}=\mathbf{x}, Z=1] > E[Y|\mathbf{X}=\mathbf{x}, Z=0]\};$$

that is, under the optimal selection option, a patient would be selected if his/her expected outcome was larger in the treatment group than the control group conditional on $\mathbf{x}$, or would not be selected if his/her expected outcome was larger or equal in the control group than the treatment group.

Let $\mu(\mathbf{x}, z) = E[Y|\mathbf{X}=\mathbf{x}, Z=z]$ be a patient's expected outcome given his/her available information and treatment $z$, and $V(s) = E[Y^*(s)I\{s(\mathbf{X})=1\}]$ be an expected outcome for the selected population, then $s^{opt}(\mathbf{x})$ and $E[Y^*(s)I\{s(\mathbf{X})=1\}]$ can be also written as $s^{opt}(\mathbf{x}) = I\{\mu(\mathbf{x},1) > \mu(\mathbf{x},0)\}$, and $V(s) = E[Y^*(s)I\{s(\mathbf{X})=1\}] = E[\mu(\mathbf{x},1)s(\mathbf{x})]$.

### 3.2.1 Outcome Regression Model

To build the optimal selection model, $\mu(\mathbf{x}, z)$ can be modeled as a linear or logistic regression $\mu(\mathbf{x}, z; \boldsymbol{\beta})$, and $\boldsymbol{\beta}$ can be estimated by ordinary least squares, maximum

likelihood, weighted least squares, or other appropriate methods. With the estimated outcome regression model $\mu(\mathbf{x}, z; \hat{\boldsymbol{\beta}})$, $s^{opt}(\mathbf{x})$ and $E[Y^*(s^{opt})I\{s^{opt}(\mathbf{X}) = 1\}]$ can be estimated as

$$\hat{s}_{\mu}^{opt}(\mathbf{x}) = I\left\{\mu(\mathbf{x}, 1; \hat{\boldsymbol{\beta}}) > \mu(\mathbf{x}, 0; \hat{\boldsymbol{\beta}})\right\}$$

and

$$\hat{V}(\hat{s}_{\mu}^{opt}) = n_1^{-1} \sum_{i=1}^{n_1} \left[\mu(\mathbf{X}_i, 1; \hat{\boldsymbol{\beta}}) \hat{s}_{\mu}^{opt}(\mathbf{X}_i)\right], \tag{3.1}$$

where $n_1$ is the number of patients being selected. For example, if $\mu(\mathbf{x}, z; \hat{\boldsymbol{\beta}}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + z(\hat{\beta}_4 + \hat{\beta}_5 x_2 + \hat{\beta}_6 x_3)$, the estimated optimal selection option can be simplified as $\hat{s}_{\beta}^{opt}(\mathbf{x}) = I\left\{\hat{\beta}_4 + \hat{\beta}_5 x_2 + \hat{\beta}_6 x_3 > 0\right\}$ with a subset of elements of $\mathbf{x}$. It can be rewritten as $I\{x_2 > \hat{\eta}_0 + \hat{\eta}_1 x_3\}$ if $\hat{\beta}_5$ is positive, or $I\{x_2 < \hat{\eta}_0 + \hat{\eta}_1 x_3\}$ if $\hat{\beta}_5$ is negative, with $\hat{\eta}_0 = -\hat{\beta}_4/\hat{\beta}_5$, and $\hat{\eta}_1 = -\hat{\beta}_6/\hat{\beta}_5$ (Zhang et al., 2012). In general, $\boldsymbol{\eta}$ is a function of $\boldsymbol{\beta}$, which can be written as, $\boldsymbol{\eta} = \eta(\boldsymbol{\beta})$. Therefore, with the correct fitted outcome regression model $\mu(\mathbf{x}, z; \boldsymbol{\beta})$, $s_{\eta}^{opt}(\mathbf{x}) = s(\mathbf{x}, \boldsymbol{\eta}^{opt})$ can be estimated for $s^{opt}$ to make $E\{Y^*(s_\eta)I\{s_\eta = 1\}\}$ have the largest value among $s_\eta \in \mathcal{S}_\eta$. If the outcome regression model is incorrectly fitted, $s^{opt}$ may not be in $\mathcal{S}_\eta$ and $\hat{s}_{\eta}^{opt}(\mathbf{x})$ may be far away from $s^{opt}$. It also happens when outcome regression models are too complex, Zhang et al. (2012) proposed an alternative method using only a key subset of elements of $\mathbf{X}$ based on interpretability, cost, and feasibility in practice to define a class of regimes by $\boldsymbol{\eta}$. For example, $s_\eta(\mathbf{x}) = s(\mathbf{x}, \boldsymbol{\eta}) = I\{x_1 < \eta_0, x_2 < \eta_1, x_3 < \eta_2\}$ without using a regression model. Therefore, the optimal selection option $s_{\eta}^{opt}$ based on the mis-specified outcome regression model with parameter estimators $\hat{\boldsymbol{\beta}}$ can

lead to poor performance on $E\{Y^*(s_\eta^{opt})I\{s_\eta^{opt}=1\}\}$, which may result in recruiting patients who would not benefit from the treatment.

### 3.2.2 Inverse Probability Weighted Estimator

From the previous section, $\boldsymbol{\eta}^{opt}$ should be estimated to obtain the maximum value of $E\{Y^*(s_\eta)\}$ for the selected population; that is, patients selected have the maximum benefits from the treatment. With the estimator $\hat{\boldsymbol{\eta}}^{opt}$, the optimal selection model $s_\eta^{opt}$ is estimated by $\hat{s}_\eta^{opt}(\mathbf{X}) = s(\mathbf{X}, \hat{\boldsymbol{\eta}}^{opt})$. Let $C_\eta = Zs(\mathbf{X}, \boldsymbol{\eta}) + (1-Z)(1-s(\mathbf{X}, \boldsymbol{\eta}))$ with fixed $\boldsymbol{\eta}$, $C_\eta$ can be 0 or 1. For patients with $C_\eta = 1$, they received control or treatment following the selection model $s_\eta$, which means their outcomes are observed with $Y^*(s_\eta) = Y$. For the others with $C_\eta = 0$, their outcomes $Y^*(s_\eta)$ following the selection model $s_\eta$ are unknown, which means they are missing. Only observed patients are used for estimating $E\{Y^*(s_\eta)I\{s_\eta = 1\}\}$. Since $C_\eta$ is a function of $\{Z, \mathbf{X}\}$, and $Y^*(s_\eta) = Y^*(1)s(\mathbf{X}, \boldsymbol{\eta}) + Y^*(0)(1-s(\mathbf{X}, \boldsymbol{\eta}))$ is a function of $\{Y^*(0), Y^*(1), \mathbf{X}\}$ with the fixed $\boldsymbol{\eta}$, $C_\eta$ is independent of $Y^*(s_\eta)$, which means that the missing mechanism on $Y^*(s_\eta)$ is missing at random (MAR) (Cao et al., 2009; Zhang et al., 2012). Therefore, the probability of being observed given $\mathbf{X}$ can be written as

$$\begin{aligned} P(C_\eta = 1|\mathbf{X}) &= P(Zs(\mathbf{X}, \boldsymbol{\eta}) + (1-Z)(1-s(\mathbf{X}, \boldsymbol{\eta})) = 1|\mathbf{X}) \\ &= s(\mathbf{X}, \boldsymbol{\eta})P(Z=1) + (1-s(\mathbf{X}, \boldsymbol{\eta}))P(Z=0) \\ &= s(\mathbf{X}, \boldsymbol{\eta})p + (1-s(\mathbf{X}, \boldsymbol{\eta}))(1-p), \end{aligned}$$

where $p = P(Z = 1)$ is a known proportion of patients being assigned to the treatment group. Therefore, $P(C_\eta = 1|\mathbf{X}) = e_c(\mathbf{X}; \boldsymbol{\eta}) = s(\mathbf{X}, \boldsymbol{\eta})p + (1 - s(\mathbf{X}, \boldsymbol{\eta}))(1 - p)$. With the known proportion $p$ and the fixed $\boldsymbol{\eta}$, the inverse probability weighted estimator (IPWE) for $E\{Y^*(s_\eta)I\{s_\eta = 1\}\}$ can be written as (Lunceford and Davidian, 2004; Cao et al., 2009; Zhang et al., 2012),

$$\begin{aligned}
\hat{V}_{IPW}(s_\eta) &= n^{-1}\sum_{i=1}^{n}\frac{C_{\eta,i}I\{s(\mathbf{X}_i,\boldsymbol{\eta})=1\}Y_i}{P(C_{\eta,i}I\{s(\mathbf{X}_i,\boldsymbol{\eta})=1\}=1|\mathbf{X}_i)} \\
&= n^{-1}\sum_{i=1}^{n}\frac{C_{\eta,i}I\{s(\mathbf{X}_i,\boldsymbol{\eta})=1\}Y_i}{\frac{n_1}{n}p} \\
&= n_1^{-1}\sum_{i=1}^{n_1}\frac{Z_i s(\mathbf{X}_i,\boldsymbol{\eta})Y_i}{p}. \qquad (3.2)
\end{aligned}$$

By maximizing the above IPWE $\hat{V}_{IPW}(s_\eta)$, the optimal selection model $s_\eta^{opt}$ can be estimated. Since the proportion $p$ is known, it can be shown that the IPWE is consistent with $E[\hat{V}_{IPW}(s_\eta)] = E[Y^*(s_\eta)I\{s(\mathbf{X}, \boldsymbol{\eta}) = 1\}]$. The proof is given as following.

$$\begin{aligned}
E[\hat{V}_{IPW}(s_\eta)] &= E\left[\frac{C_\eta I\{s(\mathbf{X},\boldsymbol{\eta})=1\}Y}{p}\right] = E\left\{E\left[\frac{C_\eta I\{s(\mathbf{X},\boldsymbol{\eta})=1\}Y}{p}|Y^*(s_\eta),\mathbf{X}\right]\right\} \\
&= E\left\{E\left[\frac{I\{C_\eta=1\}I\{s(\mathbf{X},\boldsymbol{\eta})=1\}Y^*(s_\eta)}{p}|Y^*(s_\eta),\mathbf{X}\right]\right\} \\
&= E\left\{\frac{Y^*(s_\eta)}{p}E\left[I\{Zs(\mathbf{X},\boldsymbol{\eta})=1\}|Y^*(s_\eta),\mathbf{X}\right]\right\} \\
&= E\left\{\frac{Y^*(s_\eta)}{p}I\{s(\mathbf{X},\boldsymbol{\eta})=1\}E\left[I\{Z=1\}|Y^*(s_\eta),\mathbf{X}\right]\right\} \\
&= E[Y^*(s_\eta)I\{s(\mathbf{X},\boldsymbol{\eta})=1\}].
\end{aligned}$$

Therefore, the IPWE is always consistent for randomized clinical trials with constant true propensity scores. To improve efficiency on the IPWE, an augmentation term

including outcome regression model is added, which will be presented in the following section.

### 3.2.3 Doubly Robust Inverse Probability Weighted Estimator

Even though the IPWE is always consistent with known propensity scores in randomized clinical trials, the doubly robust inverse probability weighted estimator (DRIPWE) with an augmentation term including outcome regression model is introduced to improve efficiency. Since either the propensity score model or the outcome regression model is correct, the DRIPWE is consistent with double protections. With the known propensity score $p$, the DRIPWE can be always consistent. Based on the IPWE of (3.2), the DRIPWE for $E\{Y^*(s_\eta)I\{s_\eta = 1\}\}$ is written as following (Robins et al., 1994; Cao et al., 2009; Zhang et al., 2012),

$$\begin{aligned}\hat{V}_{DRIPW}(s_\eta) &= n_1^{-1}\sum_{i=1}^{n_1}\left\{\frac{C_{\eta,i}I\{s(\mathbf{X}_i,\boldsymbol{\eta})=1\}Y_i}{p} - \frac{C_{\eta,i}-e_c(\mathbf{X}_i;\boldsymbol{\eta})}{p}m(\mathbf{X}_i;\boldsymbol{\eta},\hat{\boldsymbol{\beta}})I\{s(\mathbf{X}_i,\boldsymbol{\eta})=1\}\right\} \\ &= n_1^{-1}\sum_{i=1}^{n_1}\left\{\frac{Z_i s(\mathbf{X}_i,\boldsymbol{\eta})Y_i}{p} - \frac{Z_i s(\mathbf{X}_i,\boldsymbol{\eta})-p}{p}m(\mathbf{X}_i;\boldsymbol{\eta},\hat{\boldsymbol{\beta}})I\{s(\mathbf{X}_i,\boldsymbol{\eta})=1\}\right\}, \qquad (3.3)\end{aligned}$$

where $e_c(\mathbf{X};\boldsymbol{\eta}) = s(\mathbf{X},\boldsymbol{\eta})p + (1-s(\mathbf{X},\boldsymbol{\eta}))(1-p)$, $m(\mathbf{X}_i;\boldsymbol{\eta},\hat{\boldsymbol{\beta}}) = E[Y^*(\hat{s}_\eta)|\mathbf{X}] = \mu(\mathbf{X},1;\hat{\boldsymbol{\beta}})s(\mathbf{X},\boldsymbol{\eta}) + \mu(\mathbf{X},0;\hat{\boldsymbol{\beta}})(1-s(\mathbf{X},\boldsymbol{\eta}))$. By maximizing the above DRIPWE $\hat{V}_{DRIPW}(s_\eta)$, the optimal selection model $s_\eta^{opt}$ can be estimated. With the true propensity score $p$, the DRIPWE is always consistent with $E[\hat{V}_{DRIPW}(s_\eta)] = E[Y^*(s_\eta)I\{s(\mathbf{X},\boldsymbol{\eta}) = 1\}]$ no matter the outcome regression model is correct or wrong. The proof is given

as following.

$$
\begin{aligned}
E[\hat{V}_{DRIPW}(s_\eta)] &= E\left[\frac{C_\eta I\{s(\mathbf{X},\boldsymbol{\eta})=1\}Y}{p}\right] - E\left[\frac{C_\eta - e_c(\mathbf{X};\boldsymbol{\eta})}{p} m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}}) I\{s(\mathbf{X},\boldsymbol{\eta})=1\}\right] \\
&= E_1 - E_2;
\end{aligned}
$$

$E_1 = E[Y^*(s_\eta)I\{s(\mathbf{X},\boldsymbol{\eta})=1\}]$, which has been proved in the last section.

$$
\begin{aligned}
E_2 &= E\left[\frac{C_\eta - e_c(\mathbf{X};\boldsymbol{\eta})}{p} m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}}) I\{s(\mathbf{X},\boldsymbol{\eta})=1\}\right] \\
&= E\left\{E\left[\frac{C_\eta I\{s(\mathbf{X},\boldsymbol{\eta})=1\} - p}{p}\mu(\mathbf{X},1;\hat{\boldsymbol{\beta}})s(\mathbf{X},\boldsymbol{\eta})|\mathbf{X}\right]\right\} \\
&= E\left\{\frac{\mu(\mathbf{X},1;\hat{\boldsymbol{\beta}})s(\mathbf{X},\boldsymbol{\eta})}{p} E[I\{Zs(\mathbf{X},\boldsymbol{\eta})=1\} - p|\mathbf{X}]\right\} \\
&= E\left\{\frac{\mu(\mathbf{X},1;\hat{\boldsymbol{\beta}})s(\mathbf{X},\boldsymbol{\eta})}{p} I\{s(\mathbf{X},\boldsymbol{\eta})=1\} E[I\{Z=1\} - p|\mathbf{X}]\right\} \\
&= E\left\{\frac{\mu(\mathbf{X},1;\hat{\boldsymbol{\beta}})s(\mathbf{X},\boldsymbol{\eta})}{p} I\{s(\mathbf{X},\boldsymbol{\eta})=1\}(p-p)\right\} \\
&= 0;
\end{aligned}
$$

thus,

$$
E[\hat{V}_{DRIPW}(s_\eta)] = E[Y^*(s_\eta)I\{s(\mathbf{X},\boldsymbol{\eta})=1\}] - 0 = E[Y^*(s_\eta)I\{s(\mathbf{X},\boldsymbol{\eta})=1\}],
$$

where $e_c(\mathbf{X};\boldsymbol{\eta}) = s(\mathbf{X},\boldsymbol{\eta})p + (1 - s(\mathbf{X},\boldsymbol{\eta}))(1-p)$.

Also, with the true propensity score $p$, the DRIPWE of (3.3) has the minimum variance if the outcome regression model $m(\mathbf{X};\boldsymbol{\eta},\boldsymbol{\beta})I\{s(\mathbf{X},\boldsymbol{\eta}) = 1\}$ is correctly

specified (Cao et al., 2009). The variance of DRIPWE can be written as following,

$$\begin{aligned} var[\hat{V}_{DRIPW}(s_\eta)] &= var\left\{\frac{C_\eta I\{s(\mathbf{X},\boldsymbol{\eta})=1\}Y}{p} - \frac{C_\eta - e_c(\mathbf{X};\boldsymbol{\eta})}{p} m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}}) I\{s(\mathbf{X},\boldsymbol{\eta})=1\}\right\} \\ &= E\{var(A|\mathbf{X},Y)\} + var\{E(A|\mathbf{X},Y)\}, \end{aligned}$$

where $A = \frac{C_\eta I\{s(\mathbf{X},\boldsymbol{\eta})=1\}Y}{p} - \frac{C_\eta - e_c(\mathbf{X};\boldsymbol{\eta})}{p} m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}}) I\{s(\mathbf{X},\boldsymbol{\eta})=1\}$.

It is known that,

$$E(C_\eta I\{s(\mathbf{X},\boldsymbol{\eta})=1\}|\mathbf{X}) = E(C_\eta^2 I\{s(\mathbf{X},\boldsymbol{\eta})=1\}|\mathbf{X}) = p.$$

Thus,

$$E(A|\mathbf{X},Y) = Y.$$

$$\begin{aligned} var(A|\mathbf{X},Y) &= E(A^2|\mathbf{X},Y) - [E(A|\mathbf{X},Y)]^2 = E(A^2|\mathbf{X},Y) - Y^2 \\ &= \frac{pY^2 + pm_\beta^2 + p^2 m_\beta^2 - 2p^2 m_\beta^2 - 2pYm_\beta + 2p^2 Ym_\beta}{p^2} - Y^2 \\ &= \frac{(1-p)Y^2 + (1-p)m_\beta^2 - 2(1-p)Ym_\beta}{p} \\ &= \frac{1-p}{p}(Y - m_\beta)^2, \end{aligned}$$

where $m_\beta = m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}}) I\{s(\mathbf{X},\boldsymbol{\eta})=1\}$.

Therefore,

$$\begin{aligned} var[\hat{V}_{DRIPW}(s_\eta)] &= E\left[\frac{1-p}{p}(Y - m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}}) I\{s(\mathbf{X},\boldsymbol{\eta})=1\})^2\right] \\ &+ var(Y). \end{aligned} \tag{3.4}$$

To minimize the variance of the DRIPWE in (3.3), the first part of (3.4) should be 0, which means that the correctly specified outcome regression model $m(\mathbf{X};\boldsymbol{\eta},\hat{\boldsymbol{\beta}})I\{s(\mathbf{X},\boldsymbol{\eta}) = 1\} = \mu(\mathbf{X},1;\hat{\boldsymbol{\beta}})s(\mathbf{X},\boldsymbol{\eta})$ can lead to minimum variance in (3.3) with the true propensity score.

## 3.3 Simulation Study

We proposed three methods to build a selection model that can maximize the benefits from the treatment, an outcome regression model estimated by ordinary least squares (3.1), an inverse probability weighted estimator (3.2), and a doubly robust inverse probability weighted estimator (3.3), with parameter estimates from the last two methods obtained by the "rgenoud" library in statistical software R 3.4.2. To compare the performance and effectiveness of the three proposed methods, simulation studies were carried out on correctly and incorrectly specified outcome regression model, inverse probability weighted estimator based on the true propensity score, and doubly robust inverse probability weighted estimator based on the true propensity score but with correctly and incorrectly specified outcome regression model.

### 3.3.1 Simulation Design

Simulation studies were used to compare the three proposed methods on building the optimal selection model. The data generating mechanisms are explained in the

following paragraph. Simulations were done in statistical software R 3.4.2, and the "rgenoud" library were utilized for parameter estimates for IPWE and DRIPWE. We carried out 1000 simulations and the sample size was set at 500 for each simulation. Smaller sample size (50) and larger sample size (5,000) were also provided to see the performance of the three methods.

The simulation process was provided as following. First, we generated $n$ independent and identically distributed subjects from phase II randomized clinical trial, $(\mathbf{x}_i, z_i, Y_i)$, where $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})'$ were the three covariates independent with each other, with $x_{i1}$ was uniformly distributed on $(-1, 1)$, $x_{i2}$ was uniformly distributed on $(-1.5, 1.5)$, and $x_{i3}$ was binomial distributed on $(n, 0.5)$. $z_i$ was the treatment actually received by the $i$th subject with 0 indicating the control group and 1 indicating the treatment group, and it was bernoulli distributed with probability being assigned to the treatment group 0.5. Given $\mathbf{x}_i$ and $z_i$, outcomes $Y_i$ can be generated as $Y_i = \exp\{2 - 2x_{i1}^2 - 0.5x_{i2}^2 + x_{i3} + 2x_{i1}x_{i2}x_{i3} + z_i(-0.5 + x_{i1} - 0.7x_{i2} + 0.5x_{i3})\} + e_i$, with $e_i \sim N(0, 0.5^2)$. Based on the generated data, the expected outcome on the treatment and control group can be calculated with $E\{Y(1)\} = 7.41$ and $E\{Y(0)\} = 7.06$, so the treatment effect in the phase II randomized clinical trial is $E\{Y(1)\} - E\{Y(0)\} = 0.35$. Then, based on the way generating outcomes $Y_i$, the true optimal selection model can be written as, $s^{opt}(\mathbf{x}_i) = I\{-0.5 + x_{i1} - 0.7x_{i2} + 0.5x_{i3} > 0\}$, and we only keep those being selected, that is, keep those with $s^{opt}(\mathbf{x}_i) = 1$. Thus, only those benefiting from the treatment were remained. Since $(-0.5, 1, -0.7, 0.5)'$ is only one of the infi-

nite vectors from the coefficients of $-0.5 + x_{i1} - 0.7x_{i2} + 0.5x_{i3} > 0$, to achieve a unique vector for optimal selection model, we scale the vector into a unit vector, $\boldsymbol{\eta}^{opt} = (-0.35, 0.71, -0.50, 0.35)'$. Then, expected outcomes for the selected population were calculated if all of them were assigned to the treatment or control group, with $E[Y^*(s^{opt})I\{s^{opt}(\mathbf{X}) = 1\}] = E\{Y^*(1)\} = 12.41$ and $E\{Y^*(0)\} = 7.93$, and the treatment effect becomes $E\{Y^*(1)\} - E\{Y^*(0)\} = 4.48$. Therefore, if subjects were selected based on the true optimal selection model, the expected outcome on the treatment group $E[Y^*(s^{opt})I\{s^{opt}(\mathbf{X}) = 1\}]$ can be increased from 7.41 to 12.41, and the treatment effect can be increased from 0.35 to 4.48.

After obtaining the data, we did analysis on parameter estimates for the optimal selection model based on the three proposed methods. First, two outcome regression models were considered,

$$\mu_T(\mathbf{x}, z; \hat{\boldsymbol{\beta}}) = \exp\left\{\hat{\beta}_0 + \hat{\beta}_1 x_1^2 + \hat{\beta}_2 x_2^2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_1 x_2 x_3 + z(\hat{\beta}_5 + \hat{\beta}_6 x_1 + \hat{\beta}_7 x_2 + \hat{\beta}_8 x_3)\right\},$$

which was a correctly specified model, and

$$\mu_F(\mathbf{x}, z; \hat{\boldsymbol{\beta}}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + z(\hat{\beta}_4 + \hat{\beta}_5 x_1 + \hat{\beta}_6 x_2 + \hat{\beta}_7 x_3),$$

which was an incorrectly specified model. Parameters $\hat{\boldsymbol{\beta}}$ from the two models were estimated by ordinary least squares and the vector for $\hat{\boldsymbol{\eta}}^{opt}$ was scaled into a unit vector for uniqueness. Then, optimal selection model and expected outcomes for the selected population were estimated respectively, $\hat{s}^{opt}_{\eta_T}(\mathbf{x}) = I\{\hat{\eta}_{T0} + \hat{\eta}_{T1} x_1 + \hat{\eta}_{T2} x_2 + \hat{\eta}_{T3} x_3 > 0\}$ and $\hat{s}^{opt}_{\eta_F}(\mathbf{x}) = I\{\hat{\eta}_{F0} + \hat{\eta}_{F1} x_1 + \hat{\eta}_{F2} x_2 + \hat{\eta}_{F3} x_3 > 0\}$. Second, parameter estimates were obtained using the "rgenoud" package based on the IPWE. Because

the generated data followed a randomized clinical trial with the probability being assigned to the treatment group of 0.5, the true propensity score is known as 0.5. Third, parameter estimates were obtained using the same package "rgenoud" based on the DRIPWE with the known propensity score 0.5. To see the influence of outcome regression model on the DRIPWE, both correctly specified model $\mu_T(\mathbf{x}, z; \hat{\boldsymbol{\beta}})$ and incorrectly specified model $\mu_F(\mathbf{x}, z; \hat{\boldsymbol{\beta}})$ were used to estimate the optimal selection model and DRIPWE (3.3) or expected outcome. With the package "rgenoud", both IPWE (3.2) and DRIPWE (3.3) were estimated from the "genoud" function (Zhang et al., 2012). Also, to achieve a unique vector for $\hat{\boldsymbol{\eta}}^{opt}$, the vector was scaled into a unit vector for each simulation.

It would be nice if we can fit a model with all of the necessary covariates. However, the model can be mis-specified without including all of the covariates. Under this scenario, we want to see the performances of all three methods without considering the covariate $x_3$. Without considering $x_3$, both outcome regression models are incorrectly specified, $\mu_{F_1}(\mathbf{x}, z; \hat{\boldsymbol{\beta}}) = \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1^2 + \hat{\beta}_2 x_2^2 + \hat{\beta}_3 x_1 x_2 + z(\hat{\beta}_4 + \hat{\beta}_5 x_1 + \hat{\beta}_6 x_2)\}$ and $\mu_{F_2}(\mathbf{x}, z; \hat{\boldsymbol{\beta}}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + z(\hat{\beta}_3 + \hat{\beta}_4 x_1 + \hat{\beta}_5 x_2)$, with the estimated optimal selection model $\hat{s}^{opt}_{\eta_F}(\mathbf{x}) = I\{\hat{\eta}_{F0} + \hat{\eta}_{F1} x_1 + \hat{\eta}_{F2} x_2 > 0\}$. For the IPWE, even though the true propensity score is known, the estimates are based on the mis-specified selection model without considering $x_3$. Also, the mis-specified selection model without considering $x_3$ is used for the DRIPWE with the two incorrectly specified outcome regression model $\mu_{F_1}(\mathbf{x}, z; \hat{\boldsymbol{\beta}})$ and $\mu_{F_2}(\mathbf{x}, z; \hat{\boldsymbol{\beta}})$. For all of the three methods, the optimal selection model, expected outcome, and treatment effect are

estimated as the last paragraph.

However, what will happen if our methods include some uncessary covariates? Under this scenario, we consider an additional covairate $x_4$, with $x_4 \sim N(2,1)$ in all three methods. Thus, both outcome regression models are incorrectly specified, $\mu_{F_1}(\mathbf{x}, z; \hat{\boldsymbol{\beta}}) = \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1^2 + \hat{\beta}_2 x_2^2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_1 x_2 x_3 x_4 + z(\hat{\beta}_6 + \hat{\beta}_7 x_1 + \hat{\beta}_8 x_2 + \hat{\beta}_9 x_3 + \hat{\beta}_{10} x_4)\}$ and $\mu_{F_2}(\mathbf{x}, z; \hat{\boldsymbol{\beta}}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + z(\hat{\beta}_5 + \hat{\beta}_6 x_1 + \hat{\beta}_7 x_2 + \hat{\beta}_8 x_3 + \hat{\beta}_9 x_4)$, with the estimated optimal selection model $\hat{s}^{opt}_{\eta_F}(\mathbf{x}) = I\{\hat{\eta}_{F0} + \hat{\eta}_{F1} x_1 + \hat{\eta}_{F2} x_2 + \hat{\eta}_{F3} x_3 + \hat{\eta}_{F4} x_4 > 0\}$. For the IPWE and DRIPWE with the two incorrectly specified outcome regression model $\mu_{F_1}(\mathbf{x}, z; \hat{\boldsymbol{\beta}})$ and $\mu_{F_2}(\mathbf{x}, z; \hat{\boldsymbol{\beta}})$, the estimates are based on the mis-specified selection model with considering an additional covariate $x_4$. The same estimators are estimated as above under different methods.

### 3.3.2 Simulation Results

Table 3.1 shows results on building the selection model under the three methods with sample size 500, outcome regression model (3.1), IPWE (3.2) and DRIPWE (3.3). Under the correctly specified outcome regression model, the estimated selection model $\hat{s}^{opt}_{\mu_T} = I\{-0.35 + 0.71 x_1 - 0.50 x_2 + 0.35 x_3\}$ has the best performance, which has the same parameter estimates $\hat{\boldsymbol{\eta}}$ as the true selection model $s^{opt} = I\{-0.35 + 0.71 x_1 - 0.50 x_2 + 0.35 x_3\}$, with the smallest standard errors among all the models. Under the estimated selection model $\hat{s}^{opt}_{\mu_T}$, the expected outcome for

the selected population is 12.42 if all of them were to receive treatment, which is similar with the true expected outcome 12.41. Then, the treatment effect can be calculated as the difference between the expected outcome if all of them were to receive treatment and the expected outcome if all of them were to receive placebo for the selected population, which is 4.49, similar with the true treatment effect 4.48. However, if the outcome regression model is incorrectly specified, the estimated selection model $\hat{s}^{opt}_{\mu_F} = I\{-0.18 + 0.51x_1 - 0.60x_2 + 0.51x_3\}$ is far away from the true selection model, with smaller expected outcome for the selected population $\hat{V}(\hat{s}^{opt}_{\mu_F})$ and smaller treatment effect compared with the true values. With the true propensity score, the selection model $\hat{s}^{opt}_{\text{DRIPWE}_{\mu_T}}$ estimated by the DRIPWE with the correctly specified outcome regression model has the best performance among the three estimated selection models $\hat{s}^{opt}_{\text{IPWE}}, \hat{s}^{opt}_{\text{DRIPWE}_{\mu_T}}$, and $\hat{s}^{opt}_{\text{DRIPWE}_{\mu_F}}$, which is the same as the true selection model with the smallest standard errors. Based on the estimated selection model $\hat{s}^{opt}_{\text{DRIPWE}_{\mu_T}}$, the expected outcome $\hat{V}(\hat{s}^{opt}_{\text{DRIPWE}_{\mu_T}})$ and treatment effect are close to the true values with smaller standard errors, which is consistent with the variance formula in (3.4) that the correctly specified outcome regression model can minimize the variance with the true propensity score. For the selection model $\hat{s}^{opt}_{\text{IPWE}}$ estimated by the IPWE and $\hat{s}^{opt}_{\text{DRIPWE}_{\mu_F}}$ estimated by the DRIPWE but with the incorrectly specified outcome regression model, their parameter estimates are also close to the true selection model. However, the expected outcome $\hat{V}(\hat{s}^{opt}_{\text{IPWE}})$ and treatment effect based on the IPWE are larger than the true values with larger standard errors compared with the values estimated by the DRIPWE. Even though the expected outcome $\hat{V}(\hat{s}^{opt}_{\text{DRIPWE}_{\mu_F}})$ and treatment effect estimated by

the DRIPWE with incorrectly specified outcome regression model are not as good as the one with correctly specified outcome regression model, their values are close to the true values. Therefore, estimates based on the DRIPWE with incorrectly specified outcome regression model still have good performance with small bias.

Table 3.1: Comparisons of model performances and effectiveness on building the selection model ($n = 500$)

| Method | $\hat{\eta}_0$ (SE) ($-0.35$) | $\hat{\eta}_1$ (SE) (0.71) | $\hat{\eta}_2$ (SE) ($-0.50$) | $\hat{\eta}_3$ (SE) (0.35) | $\hat{V}(\hat{s}^{opt})$ (SE) (12.41) | $E\{Y^*(1)\} - E\{Y^*(0)\}$ (SE) (4.48) |
|---|---|---|---|---|---|---|
| $\mu_T$ | -0.35 (0.01) | 0.71 (0.007) | -0.50 (0.005) | 0.35 (0.01) | 12.42 (0.66) | 4.49 (0.28) |
| $\mu_F$ | -0.18 (0.07) | 0.51 (0.17) | -0.60 (0.13) | 0.51 (0.19) | 10.55 (0.70) | 2.52 (0.61) |
| IPWE | -0.32 (0.17) | 0.66 (0.17) | -0.48 (0.16) | 0.30 (0.23) | 12.93 (2.05) | 6.45 (1.76) |
| DRIPWE$_{\mu_T}$ | -0.35 (0.04) | 0.71 (0.03) | -0.50 (0.02) | 0.35 (0.04) | 12.43 (0.72) | 4.47 (0.54) |
| DRIPWE$_{\mu_F}$ | -0.33 (0.11) | 0.72 (0.09) | -0.49 (0.09) | 0.31 (0.13) | 12.31 (1.23) | 4.90 (1.00) |

To investigate whether the inference and estimation of the three methods sensitive to small sample size, the results under small sample size 50 are shown in Table 3.2 while keep the others same as Table 3.1. With small sample size, the standard errors increase for all models under different situations. The selection model $\hat{s}^{opt}_{\mu_T}$ estimated by the correctly specified outcome regression model still performs good, which is close to the true selection model with expected outcome $\hat{V}(\hat{s}^{opt}_{\mu_T})$ and treatment effect similar with the true values; while the results from the incorrectly specified outcome regression model are far away from the true values. With the true propensity score, only the selection model $\hat{s}^{opt}_{\text{DRIPWE}_{\mu_T}}$ estimated by the DRIPWE with the correctly specified outcome regression model closes to the true selection model, which also has good estimator on expected outcome and treatment effect. For the DRIPWE with the incorrectly specified outcome regression model, the estimated selection model $\hat{s}^{opt}_{\text{DRIPWE}_{\mu_F}}$ is far away from the true selection model while the estimator for

the expected outcome is close to the true value but not the treatment effect. The performance of the IPWE becomes worse under small sample size, with estimated selection model $\hat{s}^{opt}_{\text{IPWE}}$, expected outcome, and treatment effect far away from the true values. Therefore, the method DRIPWE with the correctly specified outcome regression model should be used under the small sample size.

Table 3.2: Sensitivity of model performances and effectiveness on building the selection model under smaller sample size ($n = 50$)

| Method | $\hat{\eta}_0$ (SE) ($-0.35$) | $\hat{\eta}_1$ (SE) (0.71) | $\hat{\eta}_2$ (SE) ($-0.50$) | $\hat{\eta}_3$ (SE) (0.35) | $\hat{V}(\hat{s}^{opt})$ (SE) (12.41) | $E\{Y^*(1)\} - E\{Y^*(0)\}$ (SE) (4.48) |
|---|---|---|---|---|---|---|
| $\mu_T$ | -0.35 (0.04) | 0.71 (0.03) | -0.50 (0.02) | 0.35 (0.04) | 12.41 (2.05) | 4.48 (0.85) |
| $\mu_F$ | -0.14 (0.20) | 0.35 (0.46) | -0.40 (0.33) | 0.31 (0.49) | 10.45 (2.80) | 3.54 (1.79) |
| IPWE | -0.21 (0.34) | 0.43 (0.47) | -0.35 (0.34) | 0.17 (0.41) | 14.60 (4.71) | 11.10 (4.81) |
| DRIPWE$_{\mu_T}$ | -0.33 (0.11) | 0.70 (0.09) | -0.51 (0.09) | 0.33 (0.13) | 12.26 (2.09) | 4.44 (1.54) |
| DRIPWE$_{\mu_F}$ | -0.25 (0.31) | 0.37 (0.49) | -0.37 (0.33) | 0.26 (0.39) | 12.77 (3.53) | 6.92 (3.24) |

In addition to the smaller sample size shown in Table 3.2, the results from a larger sample size (5,000) are provided in Table 3.3 while keep the others same as Table 3.1. With larger sample size, the results from both correctly and incorrectly outcome regression model are similar with the results in Table 3.1, with good performance for $\mu_T$ and poor performance for $\mu_F$. With the true propensity score, estimates from the IPWE and DRIPWE with incorrectly outcome regression model are much more close to the true values compared with the results in Table 3.1, while the estimates from the DRIPWE with correctly outcome regression model are similar with the results in Table 3.1. Therefore, estimates from the IPWE and DRIPWE are close to the true values with good performance if we have enough sample size, regardless of the outcome regression model for DRIPWE.

Table 3.3: Sensitivity of model performances and effectiveness on building the selection model under larger sample size ($n = 5,000$)

| Method | $\hat{\eta}_0$ (SE) ($-0.35$) | $\hat{\eta}_1$ (SE) (0.71) | $\hat{\eta}_2$ (SE) ($-0.50$) | $\hat{\eta}_3$ (SE) (0.35) | $\hat{V}(\hat{s}^{opt})$ (SE) (12.41) | $E\{Y^*(1)\} - E\{Y^*(0)\}$ (SE) (4.48) |
|---|---|---|---|---|---|---|
| $\mu_T$ | -0.35 (0.003) | 0.71 (0.002) | -0.50 (0.002) | 0.35 (0.003) | 12.40 (0.19) | 4.47 (0.08) |
| $\mu_F$ | -0.18 (0.02) | 0.55 (0.05) | -0.62 (0.04) | 0.53 (0.05) | 10.53 (0.19) | 2.46 (0.19) |
| IPWE | -0.35 (0.08) | 0.69 (0.08) | -0.50 (0.07) | 0.34 (0.11) | 12.50 (0.86) | 4.94 (0.64) |
| DRIPWE$_{\mu_T}$ | -0.35 (0.02) | 0.71 (0.01) | -0.50 (0.01) | 0.35 (0.02) | 12.41 (0.24) | 4.48 (0.17) |
| DRIPWE$_{\mu_F}$ | -0.35 (0.05) | 0.71 (0.04) | -0.50 (0.04) | 0.35 (0.06) | 12.40 (0.51) | 4.61 (0.38) |

Table 3.4 shows results for all methods without considering the covariate $x_3$. Since both outcome regression models are incorrectly specified, their estimates are far away from the true values with larger standard errors compared with Table 3.1. With the true propensity score, estimates from the IPWE and DRIPWE without considering the covariate $x_3$ in the selection model also far away from the true values except for the estimate of treatment effect by the DRIPWE. Therefore, the performances of all three methods are poor with large bias without including all of the necessary covariates that are needed for estimation.

Table 3.4: Sensitivity of model performances and effectiveness on building the selection model without considering the covariate $x_3$ ($n = 500$)

| Method | $\hat{\eta}_0$ (SE) ($-0.35$) | $\hat{\eta}_1$ (SE) (0.71) | $\hat{\eta}_2$ (SE) ($-0.50$) | $\hat{\eta}_3$ (SE) (0.35) | $\hat{V}(\hat{s}^{opt})$ (SE) (12.41) | $E\{Y^*(1)\} - E\{Y^*(0)\}$ (SE) (4.48) |
|---|---|---|---|---|---|---|
| $\mu_{F_1}$ | -0.07 (0.07) | 0.81 (0.05) | -0.58 (0.06) | - | 10.35 (0.80) | 3.74 (0.62) |
| $\mu_{F_2}$ | 0.09 (0.18) | 0.60 (0.24) | -0.71 (0.18) | - | 9.40 (0.55) | 2.40 (0.65) |
| IPWE | -0.07 (0.18) | 0.78 (0.15) | -0.55 (0.17) | - | 10.97 (1.07) | 5.21 (1.29) |
| DRIPWE$_{\mu_{F_1}}$ | -0.06 (0.12) | 0.81 (0.09) | -0.56 (0.12) | - | 10.70 (0.80) | 4.41 (0.86) |
| DRIPWE$_{\mu_{F_2}}$ | -0.07 (0.13) | 0.80 (0.10) | -0.55 (0.12) | - | 10.71 (0.89) | 4.49 (1.02) |

Table 3.5 shows results for all methods with considering an additional covariate $x_4$. Even though both outcome regression models are incorrectly specified, the results from $\mu_{F_1}$ are much better than $\mu_{F_2}$ with estimates much more close to the true values, but they are not as good as the results from $\mu_T$ in Table 3.1. With the true

propensity score, estimates from the DRIPWE with considering an additional covariate $x_4$ have good performance with estimates close to the true values, especially the one with the incorrectly specified model $\mu_{F_1}$, which has the smallest bias compared to the other models. However, estimates from the IPWE are far away from the true values compared with those from the DRIPWE. Therefore, the method DRIPWE can be used even with additional covariates that are not included in the true model.

Table 3.5: Sensitivity of model performances and effectiveness on building the selection model with considering an additional covariate $x_4$ ($n = 500$)

| Method | $\hat{\eta}_0$ (SE) ($-0.35$) | $\hat{\eta}_1$ (SE) (0.71) | $\hat{\eta}_2$ (SE) ($-0.50$) | $\hat{\eta}_3$ (SE) (0.35) | $\hat{\eta}_4$ (SE) - | $\hat{V}(\hat{s}^{opt})$ (SE) (12.41) | $E\{Y^*(1)\} - E\{Y^*(0)\}$ (SE) (4.48) |
|---|---|---|---|---|---|---|---|
| $\mu_{F_1}$ | -0.40 (0.09) | 0.62 (0.06) | -0.50 (0.05) | 0.42 (0.05) | 0.02 (0.06) | 12.20 (0.75) | 3.76 (0.43) |
| $\mu_{F_2}$ | -0.17 (0.22) | 0.51 (0.17) | -0.58 (0.12) | 0.49 (0.18) | -0.001 (0.11) | 10.55 (0.67) | 2.59 (0.61) |
| IPWE | -0.26 (0.30) | 0.63 (0.17) | -0.48 (0.15) | 0.28 (0.24) | -0.02 (0.16) | 12.83 (1.96) | 6.59 (1.76) |
| DRIPWE$_{\mu_{F_1}}$ | -0.35 (0.12) | 0.70 (0.06) | -0.49 (0.06) | 0.35 (0.08) | -0.0009 (0.06) | 12.53 (0.89) | 4.84 (0.72) |
| DRIPWE$_{\mu_{F_2}}$ | -0.31 (0.17) | 0.70 (0.11) | -0.50 (0.08) | 0.31 (0.14) | -0.009 (0.08) | 12.31 (1.20) | 5.01 (1.03) |

From the previous outputs comparing the three methods, we have the following conclusions. First, the outcome regression model has the best performance with the estimated selection model similar with the true model, smallest bias on expected outcome and treatment effect, and smallest standard errors, regardless of the sample size if we can fit the model correctly. However, it is difficult to fit an outcome regression model correctly, so the DRIPWE should be used under large sample size with small bias and standard errors, even with incorrectly specified outcome regression model. Second, even though the selection model estimated by the IPWE is close to the true selection model, the bias for expected outcome and treatment effect are larger than the DRIPWE. But as the sample size increases, the bias become

smaller, which also shows good performance of the IPWE with enough sample size. Third, the performances of all methods become poor with large bias if they misspecify the model without considering all of the necessary covariates. However, if the model is incorrectly specified with including unnecessary covariates, the DRIPWE still has good estimates on selection model with small bias on expected outcome and treatment effect.

## 3.4 Discussion

Unlike the traditional method that uses one or two indicator variables for choosing a subset of population into the phase III randomized clinical trial based on the information from the phase II randomized clinical trial, historical literatures, and other studies, we have developed a selection model including more variables to identify the subset. The easiest method to build the selection model is fitting an outcome regression model. However, estimates are far away from the true values if the outcome regression model is incorrectly specified. Another method to build the selection model is based on the inverse probability weighted estimator. Even though the estimated selection model is close to the true selection model with the true propensity score, the expected outcome or estimator can be poor with large bias if the sample size is not enough. To improve efficiency of the inverse probability weighted estimator, the doubly robust inverse probability weighted estimator is proposed adding the benefit of robustness, i.e., the performance of doubly robust inverse probability weighted estimator is always good no matter the outcome regres-

sion model is correct or not with the true propensity score. However, if one of the important covariates is missing on building the selection model, all models' performances are poor with large bias. If the model is incorrectly specified with including unnecessary covariates, the DRIPWE can still be used with good estimates on selection model with small bias on expected outcome and treatment effect. Therefore, to make sure at least one of our methods can have good estimates on selection model and expected outcome, it is important to include as many covariates as possible into our model.

Since our proposed methods on building the selection model are only based on the information from the phase II randomized clinical trial, the true propensity score is known from the randomness nature. However, because of the small sample size in phase II randomized clinical trial and different endpoints used between phase II and phase III randomized clinical trials, the information is not enough to build a selection model to make the phase III randomized clinical trial design more efficient. It is necessary to use other information from some historical observational studies. Therefore, our methods need to be extended by building a propensity score model for inverse probability weighted estimator and doubly robust inverse probability weighted estimator, which will be discussed in the future work. Also, we will apply our methods to a real randomized clinical trial in the future to see their performances.

# Bibliography


Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, Inc., 2 edition.

Alemayehu, D., Chen, Y., and Markatou, M. (2017). A comparative study of subgroup identification methods for differential treatment effect: Performance metrics and recommendations. *Statistical Methods In Medical Research*, 0(0):1–21.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. In S. Leinhardt (Ed). *Sociological Methodology*, pages 61–98.

Allison, P. D. (2010). *Survival Analysis Using SAS: A Practical Guide, Second Edition*. SAS Institute Inc.

Almirall, D., Collins, L. M., and Murphy, S. A. (2011). Introduction to adaptive interventions and SMART study design principles. In *COPTR Workshop on Adaptive Interventions*.

Analytics Vidhya Content Team (2016). A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python).

Austin, P. C. (2017). A tutorial on multilevel survival analysis: methods, models and applications. *International Statistical Review*, 85(2):185–203.

Blaus, A., Madabushi, R., Pacanowski, M., Rose, M., Schuck, R. N., Stockbridge, N., Temple, R., and Unger, E. F. (2015). Personalized cardiovascular medicine today: a food and drug administration/center for drug evaluation and research perspective. *Circulation*, 132(15):1425–1432.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth: Belmont, CA.

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

Chen, Y.-F., Yang, Y., Hung, H. M. J., and Wang, S.-J. (2011). Evaluation of performance of some enrichment designs dealing with high placebo response in psychiatric clinical trials. *Contemporary Clinical Trials*, 32(4):592–604.

CONSENSUS Trial Study Group (1987). Effects of enalapril on mortality in severe congestive heart failure. *New England Journal of Medicine*, 316(23):1429–1435.

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

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford, 2 edition.

Echt, D. S., Liebson, P. R., Mitchell, L. B., Peters, R. W., Obias-Manno, D., Barker, A. H., Arensberg, D., Baker, A., Friedman, L., Greene, H. L., et al. (1991). Mortality and morbidity in patients receiving encainide, flecainide, or placebo - the cardiac arrhythmia suppression trial. *The New England Journal of Medicine*, 324(12):781–788.

Fava, M., Evins, A. E., Dorer, D. J., and Schoenfeld, D. A. (2003). The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. *Psychotherapy and Psychosomatics*, 72(3):115–127.

FDA (2012). Draft Guidance/Guidance for industry. Enrichment strategies for clinical trials to support approval of human drugs and biological products.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometric Society*, 42(4):845–854.

Fisher, B., Redmond, C., Brown, A., Wickerham, D. L., Wolmark, N., Allegra, J., Escher, G., Lippman, M., Savlov, E., Wittliff, J., , and Fisher, E. R. (1983). Influence of tumor estrogen and progesterone receptor levels on the response to tamoxifen and chemotherapy in primary breast cancer. *Journal Of Clinical Oncology*, 1(4):227–241.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011). *Applied Longitudinal Analysis*. John Wiley & Sons.

Freidlin, B. and Simon, R. (2005). Evaluation of randomized discontinuation design. *Journal of Clinical Oncology*, 23(22):5094–5098.

Friedman, L. M., Furberg, C. D., and L, D. D. (2010). *Fundamentals of Clinical Trials*. Springer.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41(2):361–372.

Giolo, S. R., Colosimo, E. A., and Demétrio, C. G. B. (2009). Different approaches for modelling grouped survival data: a mango tree study. *Journal of Agricultural, Biological, and Environmental Statistics*, 14(2):154–169.

Jone, H. (2010). Reinforcement-Based Treatment for Pregnant Drug Abusers (HOME II).

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on cox's regression and life model. *Biometrika*, 60(2):267–278.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kasari, C. (2009). Developmental and Augmented Intervention for Facilitating Expressive Language (CCNIA).

Lavori, P. W. and Dawson, R. (2000). A design for testing clinical strategies: biased adaptive withinsubject randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):29–38.

Lavori, P. W. and Dawson, R. (2004). Dynamic treatment regimes: practical design considerations. *Clinical Trials - London - Arnold Journals*, 1(1):9–20.

Lei, H., Murphy, S. A., Nahum-Shani, I., Lynch, K., and Oslin, D. (2012). A "SMART" design for building individualized treatment sequences. *Annual Review of Clinical Phychology*, 8:21–48.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lipkovich, H., Dmitrienko, A., and DAgostino, R. B. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, 36(1):136–196.

Liu, J. P. (2003). Enrichment design. In Chow, S. C., editor, *Encyclopedia of Biopharmaceutical Statistics*, pages 324–326. Marcel Dekker, Inc., New York.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960.

Magaziner, J. (2012). Rehabilitation: The need for multi-component interventions based on what we know about the consequences of hip fracture. In *Fragility Fracture Network Global Congress*.

Mok, T. S., Wu, Y. L., Thongprasert, S., Yang, C. H., Chu, D. T., Saijo, N., Sunpaweravong, P., Han, B., Margono, B., Ichinose, Y., Nishiwaki, Y., Ohe, Y., Yang, J. J., Chewaskulyong, B., Jiang, H., Duffield, E. L., Watkins, C. L., Armour, A. A., and Fukuoka, M. (2009). Gefitinib or carboplatin paclitaxel in pulmonary adenocarcinoma. *New England Journal of Medicine*, 361(10):947–957.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24(10):1455–1482.

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W. E., Gnagy, B., Fabiano, G. A., Waxmonsky, J. G., Yu, J., and Murphy, S. A. (2012). Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychological Methods*, 17(4):457–477.

Ondra, T., Dmitrienko, A., Friede, T., Graf, A., Miller, F., Stallard, N., and Posch, M. (2016). Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. *Journal of Biopharmaceutical Statistics*, 26(1):99–119.

Orwig, D., Hochberg, M. C., Gruber-Baldini, A. L., Resnick, B., Miller, R. R., Hicks, G. E., Cappola, A. R., Shardell, M., Sterling, R., Hebel, J. R., Johnson, R., and Magaziner, J. (2018). Examining differences in recovery outcomes between male and female hip fracture patients: Design and baseline results of a prospective cohort study from the baltimore hip studies. *Journal of Frailty & Aging*, 7(3):162–169.

Oslin, D. W. (2005). Managing Alcoholism in People Who Do Not Respond to Naltrexone (EXTEND).

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21(19):2917–2930.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with applications to breast cancer data. *Biometrics*, 34(1):57–67.

Resnick, B., Galik, E., Boltz, M., Hawkes, W., Shardell, M., Orwig, D., and Magaziner, J. (2011). Physical activity in the post-hip-fracture period. *Journal of Aging & Physical Activity*, 19(4):373–387.

Ridker, P. M., Danielson, E., Fonseca, F. A. H., Genest, J., Gotto, A. M., Kastelein, J. J. P., Koenig, W., Libby, P., Lorenzatti, A. J., MacFadyen, J. G., Nordestgaard, B. G., Shepherd, J., Willerson, J. T., Glynn, R. J., , and JUPITER Study Group (2008). Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *New England Journal of Medicine*, 359(21):2195–2207.

Robins, J. M., Rotnitzky, A., and Ping, Z. L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rodriguez, G. (2008). Multilevel generalized linear models. In Leeuw, J. d. and Meijer, E., editors, *Handbook of Multilevel Analysis*, chapter 9, pages 335–376. Springer, New York.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1986). Statistics and causal inference: comment: which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962.

Schatzberg, A. F. and Kraemer, H. C. (2000). Use of placebo control groups in evaluating efficacy of treatment of unipolar major depression. *Biological Psychiatry*, 47(8):736–744.

Simon, N. (2015). Adaptive enrichment designs: applications and challenges. *Clinical Investigation*, 5(4):383–391.

Simon, N. and Simon, R. (2013). Adaptive enrichment designs for clinical trials. *Biostatistics*, 14(4):613–625.

Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., Baselga, J., and Norton, L. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *New England Journal of Medicine*, 344(11):783–792.

Snijders, T. (1996). Analysis of longitudinal data using the hierarchical linear model. *Quality and Quantity*, 30(4):405–426.

Splawa-Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on Principles. Section 9. Translated from the Polish and edited by Dabrowska, D. M. and Speed, T. P. *Statistical Science*, 5(4):465–472 (1990).

Sun, J. and Zhao, X. (2013). *Statistical Analysis of Panel Count Data*. New York: Springer.

Sutradhar, R., Barbera, L., Seow, H., Howell, D., Husain, A., and Dudgeon, D. (2011). Multistate analysis of interval-censored longitudinal data: application to a cohort study on performance status among patients diagnosed with cancer. *American Journal of Epidemiology*, 173(4):468–475.

Tamura, R. N. and Huang, X. (2007). An examination of the efficiency of the sequential parallel design in psychiatric clinical trials. *Clinical Trials*, 4(4):309–317.

Taylor, A. L., Ziesche, S., Yancy, C., Carson, Peter; D'Agostino, R. J., Ferdinand, K., Taylor, M., Adams, K., Sabolinski, M., Worcel, M., Cohn, J. N., and African-American Heart Failure Trial Investigators (2004). Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *New England Journal of Medicine*, 351(20):2049–2057.

Temple, R. J. (1994). Special study designs: Early escape, enrichment, studies in non-responders. *Communication in Statistics - Theory and Methods*, 23(2):499–531.

Temple, R. J. (2005). Enrichment designs: efficiency in development of cancer treatments. *Journal of Clinical Oncology*, 23(22):4838–4839.

Thall, P. F. and Lachin, J. M. (1988). Analysis of recurrent events: Nonparametric methods for random-interval count data. *Journal of the American Statistical Association*, 83(402):339–347.

Tstiatis, A. and Davidian, M. (2016). Implementing Precision Medicine: Optimal Treatment Regimes and SMARTs. In *NHLBI Biostatistics Workshop on Recent Advances and Challenges in Statistical Methods*.

Vermunt, J. K. (2009). Event history analysis. In Millsap, R. and Maydeu-Olivares, A., editors, *Handbook of Quantitative Methods in Psychology*, pages 658–674. Sage, Thousand Oaks.

Yang, B., Zhou, Y., Zhang, L., and Cui, L. (2015). Enrichment design with patient population augmentation. *Contemporary Clinical Trials*, 42:60–67.

Yusuf, S., Pitt, B., Davis, C. E., Hood, W. B., Cohn, J. N., and SOLVD Investigators (1991). Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *New England Journal of Medicine*, 325(5):293–302.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348–368.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018.

Zhu, L., Zhang, Y., Li, Y., Sun, J., and Robison, L. L. (2017). A semiparametric likelihood-based method for regression analysis of mixed panel-count data. *Biometrics*.