11-10-2016

Efficiency of an Unbalanced Design in Collecting Time to Event Data with Interval Censoring

Peiyao Cheng
University of South Florida, peiyaoc@gmail.com

Efficiency of an Unbalanced Design in Collecting Time to Event Data with Interval Censoring

by

Peiyao Cheng

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Epidemiology and Biostatistics
College of Public Health
University of South Florida

Major Professor: Yougui Wu, Ph.D.
Getachew A. Dagne, Ph.D.
Yangxin Huang, Ph.D.
Paul Leaverton, Ph.D.

Date of Approval:
November 7, 2016

Keywords: unbalanced design, balanced design, interval censor, time to event

Dedication

This work is dedicated to my family, who have been supportive of my research all these years.

Acknowledgments

I owe my gratitude to all the people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

I would like to especially thank my major advisor, Dr. Yougui Wu, for his guidance and support during the completion of this dissertation. His invaluable guidance, constructive criticism and encouragement were crucial in the completion of this dissertation.

I am also very grateful to my dissertation committee members, Drs. Getachew A. Dagne, Yangxin Huang and Paul Leaverton for their insightful suggestions and comments. I would also like to thank Jaeb Center for Health Research, Inc. for providing the data used in my research.

Finally, I would like to thank all of the faculty members and staff in the Department of Epidemiology and Biostatistics at College of Public Health for their help during my graduate studies in these years.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CDF	Cumulative Density Function
CVD	Cardiovascular Disease
DCCT	Diabetes Control and Complications Trial
DirecNet	The Diabetes Research in Children Network
DME	Diabetic Macular Edema
DRCR.net	Diabetic Retinopathy Clinical Research Network
EM	Expectation-Maximization
ICM	Iterative Convex Minorant
MMTT	Mixed-Meal Tolerance Test
NPMLE	Nonparametric Maximum Likelihood Estimation
PDF	Probability Density Function
PDR	Diabetic Retinopathy
PH	Proportional Hazard
PRP	Panretinal Photocoagulation
RCTs	Randomized Controlled Trials
SD	Standard Deviation
SE	Standard Error
T1D	Type 1 Diabetes

Abstract

In longitudinal studies, the exact timing of an event often cannot be observed, and is usually detected at a subsequent visit, which is called interval censoring. Spacing of the visits is important when designing study with interval censored data. In a typical longitudinal study, the spacing of visits is usually the same across all subjects (balanced design). In this dissertation, I propose an unbalanced design: subjects at baseline are divided into a high risk group and a low risk group based on a risk factor, and the subjects in the high risk group are followed more frequently than those in the low risk group. Using a simple setting of a single binary exposure of interest (covariate) and exponentially distributed survival times, I derive the explicit formula for the asymptotic sampling variance of the estimate for the covariate effect. It shows that the asymptotic sampling variance can be simply reduced by increasing the number of examinations in the high risk group. The relative reduction tends to be greater when the baseline hazard rate in the high risk group is much higher than that in the low risk group and tends to be larger when the frequency of assessments in the low risk group is relatively sparse. Numeric simulations are also used to verify the asymptotic results in small samples and evaluate the efficiency of the unbalanced design in more complicated settings. Beyond comparing the asymptotic sampling variances, I further evaluate the power and empirical Type I error from unbalanced design and compare against the traditional balanced design. Data from a randomized clinical trial for type 1 diabetes are further used to test the performance of the proposed unbalanced design, and the

parametric analyses of these data confirmed the findings from the theoretical and numerical studies.

Chapter 1: Introduction

Longitudinal study can be defined as a study in which each individual is observed on more than one occasion. In most biomedical researches or epidemiology studies, when designing a longitudinal study, the same visit or assessment schedule is usually planned for all the participants in the study. Although due to some practical reasons, the assessment times will not be exactly the same across all the participants, for example, missed visits, early dropouts, or simply visit times out of the pre-define windows, however, by study design, all the participants should follow the same assessment schedule. In this dissertation, I simply name this type of design as "balanced design" in order to differentiate it from the next design I will introduce here.

In rare occasions, we might also see some longitudinal studies with visit or assessment schedules adaptive to different participants based on their data collected during the studies. In this dissertation, I name this type of design as "unbalanced design". My study in this dissertation will focus on a new type of "unbalanced design" with increased frequency of assessment among those participants who have higher risk for the study outcome, as defined by one or more risk factors measured in the study, when collecting interval-censored time to event data. I hypothesize that this type of design can improve the efficiency of parameter estimation and power of the study when collecting and analyzing interval-censored time to event data.

In this chapter, first, I will review some examples of longitudinal studies with either balanced or unbalanced design, then, I will explain the rationale on why I intend to study the use of unbalanced design when collecting interval-censored time to event data. To help with understanding the issues associated with interval-censored time to event data and the theoretical derivation process in Chapter 2, as well as data analyses in Chapter 3, I will also review some basic theories and methods

associated with survival analysis for interval-censored data. This chapter, by no means, provides an extensive review for all the analysis methods developed interval-censored time to event data, instead, I mainly focus on the basic theory for those methods used in the following two chapters.

1.1 Balanced Design

As previously defined, in balanced design, all study participants have the same visit/assessment schedule. The visits/assessments in the longitudinal studies can be either evenly spaced or unevenly spaced throughout the study follow-up period. Examples for longitudinal studies with evenly spaced visit/assessment schedule include the well-known long-running Framingham Heart Study, which identifies common factors that contribute to cardiovascular disease (CVD). This study started in 1948, and since then, the participants return to the study every two years for a detailed medical history, physical examination, and laboratory test (Framingham Heart Study, n.d.). Another example is the HIV Vaccine Trial in Thai Adults, which is a phase III placebo-controlled HIV prevention trial conducted at Thailand. This study enrolled more than 16,000 HIV negative participants, and after the treatments were given, HIV infection was assessed every 6 months for 3 years (ClinicalTrials.gov, 2012).

Examples for longitudinal studies with unevenly spaced visit/assessment schedule include the Dunedin Multidisciplinary Health and Development Study, an ongoing, longitudinal study of the health, development and well-being of a general sample of New Zealanders born between April 1, 1972 and March 31, 1973. They were studies at birth, followed up and assessed at age three, and then at ages 5, 7, 9, 11, 13, 15, 18, 21, 26, 32 and 38 (the most recent assessment) (Dunedin Multidisciplinary Health & Development Research Unit, n.d.). Another example is the COLON Study, a currently ongoing longitudinal, observations study on nutrition and lifestyle factors that may influence colorectal tumour recurrence, survival and quality of life. In this study, at least 1000 incident colorectal cancer patients will be recruited from 11 hospitals in the Netherlands, and data will be collected at recruitment, after 6 months, 2 years, and 5 years (Winkels et al., 2014).

Besides above examples from observational longitudinal studies, unevenly spaced visit/assessment schedule is also very common in randomized controlled trials (RCTs). For example, in an ongoing phase III randomized, placebo-controlled clinical trial for assessing hormone therapy with or without everolimus in treating patients with breast cancer, participants are treated by either endocrine therapy with placebo or endocrine therapy with everolimus for one year. After completion of study treatment, participants are followed up every 6 months for 2 years and then yearly thereafter for 10 years (ClinicalTrials.gov, 2016).

Among those studies with unevenly spaced visit/assessment schedule, although assessment times in some studies do not have obvious pattern, a lot of them tend to have more frequent observations at the beginning of the study than the later phase of the study, because it is expected that changes will happen faster during the early phase of the study, such as those studies which enrolled newborns (Dunedin Multidisciplinary Health & Development Research Unit, n.d.), newly diagnosed patients (Winkels et al., 2014), or a new treatment (ClinicalTrials.gov, 2016).

No matter evenly spaced or unevenly spaced assessment times, in each of above studies, the follow-up schedule is the same among all subjects in the study when ignoring early dropout and missed visits. Therefore, they belong to the concept of "balanced design" as I defined previously. This design is widely used because it is convenience for study management and statistical analysis is more straightforward (e.g. direct group comparisons can be done at every data collection time point).

1.2 Unbalanced Design

As previously defined, in unbalanced design, visit/assessment schedule varies across different subjects based on some observed data or simply the assigned treatment groups in RCTs. For example, in a randomized clinical trial conducted by the Diabetic Retinopathy Clinical Research Network (DRCR.net) for comparing panretinal photocoagulation (PRP) vs. intravitreous ranibizumab for the treatment of proliferative diabetic retinopathy (PDR), the study eyes were

randomly assigned to two treatment arms. One treatment arm received PRP with ranibizumab as needed for diabetic macular edema (DME) treatment, the other treatment arm received 0.5 mg ranibizumab by injection with PRP allowed for cases of treatment failure. In the PRP treatment group, assessment visits occurred every 16 weeks; the ranibizumab group had more frequent visits than the PRP treatment group – besides the assessment visits at every 16 weeks, the participants in this group also had additional treatment visits every 4 weeks during the first year and every 4 to 16 weeks during the second year depending on treatment (Diabetic Retinopathy Clinical Research Network [DRCR.net], 2015).

An another example is also a RCT conducted by the same research group for comparing the efficacy of ranibizumab plus prompt or deferred laser with triamcinolone plus prompt laser for DME treatment. In this study, the study eyes were randomly assigned to 1 of 4 treatment groups: group 1 received sham injection plus prompt laser treatment, group 2 received 0.5 mg intravitreal ranibizumab plus prompt laser treatment, group 3 received 0.5 mg intravitreal ranibizumab with deferred laser treatment, and group 4 received 4 mg intravitreal triamcinolone plus prompt laser treatment. During the first year, the follow-up visits occurred every 4 weeks for all 4 groups, and after the first year, the follow-up visits occurred every 4 to 16 weeks depending on the treatment group, disease course, and treatment administered (DRCR.net, 2010).

In above two examples, the follow-up visits schedule was determined by the particular treatment received during the trials. Based on some private conversations with key personnel in the DRCR.net study group, the reason for such kind of unbalanced design with one treatment group having more frequent visits than another was mainly due to medical necessity.

Another example of unbalanced design is the TrialNet TN-01 Pathway to Prevention Study (originally called Natural History Study of the Development of Type 1 Diabetes), which is an ongoing observational longitudinal study to evaluate the development of type 1 diabetes (T1D). This study includes three phases: screening, baseline risk assessment, and follow-up risk assessments. In a more recent revision of the study protocol, the study follow-up schedule changed from every 6

months for all the subjects in the follow-up risk assessment phase (Type 1 Diabetes TrialNet, 2009) to either annual monitoring or semi-annual monitoring depending on the risk factor measured from the study (Type 1 Diabetes TrialNet, 2011). In the revised protocol, at screening, participants who are positive for at least two autoantibodies on the same sample directly enter into semi-annual monitoring; participants with a single autoantibody need to undergo a baseline monitoring visit. The results of lab tests conducted at baseline monitoring visit determine whether a participant enters into annual monitoring or semi-annual monitoring. Participants with ≥ 2 autoantibody, abnormal glucose tolerance, an HbA1c $\geq 6.0\%$, or a DPT-1 Risk Score ≥ 6.5 enter into semi-annual monitoring, and the others enter into annual monitoring. In subsequent visits, the same lab tests are conducted among those participants in the annual monitoring group, those who develop ≥ 2 positive autoantibodies, an HbA1c level $\geq 6.0\%$, or an increase in the HbA1c level $\geq 0.5\%$ compared with the previous HbA1c level enter into the semi-annual monitoring stage (Type 1 Diabetes TrialNet, 2011). This study is an example that visit schedule depends on certain risk factors measured during the study, however, it is unknown why this study adopted this type of unbalanced design.

Overall, the longitudinal studies with unbalanced design are much less than the longitudinal studies with balanced design. Among those limited examples, the main factor determining the unbalanced visit schedule was based on medical consideration, not on statistical consideration. And currently, there is no publication systematically evaluating the balanced design and unbalanced design in statistical perspective (i.e. bias and precision).

1.3 Interval-Censored Time to Event Data and Analysis Methods

1.3.1 Survival Analysis

Survival analysis is a branch of statistics which concerns about failure time, or event time, i.e., the time elapsed from a specified starting point until a failure or event occurs. In biomedical or epidemiology research, a failure time could be, for example, the age when a subject develops

certain disease, the time from a treatment until disease progression, or the time of death after developing certain disease, etc.

In survival analysis, event time can only be non-negative ($T \geq 0$). The basic quantity to describe time to an event is the survival function, i.e., the probability of an individual surviving beyond time t. It is defined as $S(t) = Pr(X > t)$. When X is a continuous random variable, we have $S(t) = 1 - F(t) = \int_t^\infty f(u)du$, where $F(t)$ is the cumulative density function (CDF) and $f(t)$ is the corresponding probability density function (PDF).

Another basic quantity in survival analysis is the hazard function. The hazard rate is defined by $h(t) = \lim_{\Delta t \to 0} P[t \leq X < t + \Delta t]/\Delta t$. When X is a continuous random variable, the hazard rate $h(t) = \frac{f(t)}{S(t)} = -d\ln[S(t)]/dt$, and the cumulative hazard function is defined by $H(t) = \int_0^t h(u)du = -\ln[S(t)]$.

1.3.2 Censored Data

A common feature of the data sets with failure time or time to an event outcome is that they often contain censored observations. In statistics, the term censoring refer to a condition in which observed data contain incomplete information. Censored data arises when the event is known to occur only in a certain period of time.

The most common encountered type of censoring in biomedical or epidemiology research is right censoring. This type of censoring occurs when a subject leaves the study before an event occurs, or the study ends before the event has happened. For example, consider a 5-year clinical trial to study the effect of a treatment on stoke occurrence, for those patients who have had no strokes by the end of 5 years, then their event time will be estimated by $(5, \infty)$.

Another common type of censoring is called interval censoring, where the only information available is that the event occurs within certain interval. This type of censoring occurs when patients in a clinical trial or longitudinal study have periodic follow-up and the event time is only

known to fall in an interval $(L_i, R_i]$, where L_i is the left endpoint and R_i is the right endpoint of the censoring interval for the ith subject.

Another type of censoring is left censoring, where the individual already experienced the event of interest before the first visit in the study. Left-censoring can be treated as a special case of interval censoring with $L_i = 0$, so the true event time T_i falls in the interval $(0, R_i]$ and R_i is the period of time from the beginning of the study until the first visit. Right censoring can also be treated as a special case of interval censoring where $R_i = \infty$, so T_i falls in the interval (L_i, ∞) and L_i is the last observation time.

A general assumption in analyzing censored data is the censoring times and event times are independent (also called noninformative censoring) (Klein and Moeschberger, 2003). The methods described in below sections and used in this dissertation all imply this assumption.

1.3.3 Parametric Methods

The parametric methods assume that event times follow a specific parametric distribution, then parameters in the function can be easily estimated by using the on maximum likelihood theory. For interval censored data, the likelihood function $L \propto \prod_{i \in I}[S(L_i) - S(R_i)]$, where L_i is the left endpoint and R_i is the right endpoint of the censoring interval for the ith subject. For right-censored data, we have $S(R_i) = S(\infty) = 0$, and for left-censored data, we have $S(L_i) = S(0) = 1$.

Exponential Distribution

Since the event time is always positive and the distribution is usually right-skewed, in the parametric models used for estimation of survival functions, exponential distribution is the simplest and convenient choice. It only has one parameter to estimate and assumes constant hazard rate. It can be represented in following way (Klein and Moeschberger, 2003):

$$T \sim \exp(\lambda), \lambda > 0 \qquad f(t) = \lambda \exp(-\lambda t)$$

$$h(t) = \lambda \qquad S(t) = \exp(-\lambda t)$$

Therefore, the likelihood function for a data set which only contains interval-censored data (or in combination with left- or right-censored data) with exponential distribution of event time can be written as

$$L = \prod_{i=1}^{n} [\exp(-\lambda L_i) - \exp(-\lambda R_i)], \tag{1.1}$$

where $L_i = 0$ for left-censored observations and $R_i = \infty$ for right-censored observations.

Weibull Distribution

Weibull distribution is another important and commonly used parametric model in estimation of survival functions. It has two parameters and can be represented in following way (Klein and Moeschberger, 2003):

$$T \sim Wb(\alpha, \lambda), \alpha, \lambda > 0 \qquad f(t) = \alpha\lambda t^{\alpha-1} \exp(-\lambda t^{\alpha})$$

$$h(t) = \alpha\lambda t^{\alpha-1} \qquad S(t) = \exp(-\lambda t^{\alpha})$$

where λ is a scale parameter and α is a shape parameter. Therefore, the likelihood function for interval-censored data with Weibull distribution of event time can be written as

$$L = \prod_{i=1}^{n} [\exp(-\lambda L_i^{\alpha}) - \exp(-\lambda R_i^{\alpha})], \tag{1.2}$$

where $L_i = 0$ for left-censored observations and $R_i = \infty$ for right-censored observations.

Weibull distribution has the advantage of being adaptable to many different shapes: when $\alpha = 1$, it reduces to the exponential distribution with constant hazard rates; when $\alpha < 1$, it represents decreasing hazard rates; and when $\alpha > 1$, it represents increasing hazard rates. This property, coupled with the relatively simple hazard, survival and probability density functions, have made it a very popular parametric model.

Although there are also other distributions used in parametric survival analysis, such as Gamma distribution, log normal distribution, log logistic distribution, etc., in this dissertation, I only use exponential distribution and Weibull distribution to explore the gained efficiency by using unbalanced design.

Regression Models

When there are covariates present in the data, we need to use regression models to specify how the covariates affect the failure time of interest. There are multiple choices of regression models available for survival data with covariates, herein, we only introduce two commonly used models, with one of them be used in theoretical derivation process in Chapter 2.

One commonly used model is the proportional hazard (PH) model. Let Z be a vector of covariates, the PH model assumes that the hazard function of T has the form

$$h(t; \boldsymbol{Z}) = h_0(t) \exp(\boldsymbol{Z'\beta}) \tag{1.3}$$

given covariates $\boldsymbol{Z}$ (Cox, 1972). In the above formula, $\lambda_0(t)$ is an arbitrary baseline hazard function, and β is the vector of regression parameters. In the parametric case when T follows exponential distribution, we have $h(t) = \lambda$, and thus $\lambda = \lambda_0 \exp(\boldsymbol{Z'\beta})$. Therefore, equation (1.1) can be expressed as

$$L = \prod_{i=1}^{n} [\exp(-\lambda_0 e^{\boldsymbol{Z'\beta}} L_i) - \exp(-\lambda_0 e^{\boldsymbol{Z'\beta}} R_i)]. \tag{1.4}$$

This model is used in Chapter 2.

Another commonly used model is the accelerated failure time (AFT) model. In the AFT model, the effect of covariates directly works on the failure time T instead of hazard function as above. This model assumes $T = T_0 \exp(\boldsymbol{Z'\beta})$, where T_0 is the failure time for the individual with

covariate value 0. When taking natural logarithms, the AFT model can be expressed as

$$\log(T) = \log(T_0) + \boldsymbol{Z'\beta}.$$

If we assume that $\log(T_0) = \mu + \sigma W$, where W is a random variable, then above model can be written in a linear form:

$$\log(T) = \mu + \boldsymbol{Z'\beta} + \sigma W. \tag{1.5}$$

This model is used by PROC LIFEREG module in SAS.

1.3.4 Imputation Based Methods

The purpose of imputation when analyzing interval-censored data is to generate one or multiple sets of right-censored data, then apply the standard methods for the right-censored failure time data on the imputed data in order to make inference. The imputation approaches include single point imputation and multiple imputation.

Single Point Imputation Approach

Single point imputation is the simplest imputation approach which is commonly used in practice. For subject i, if the true failure time T_i is within an interval $(L_i, R_i], i = 1, \ldots, n$, a conventional approach adopted in the industry is to impute the right-point (R_i) of the time interval as the true failure time. Another commonly adopted approach is to impute the mid-point of the time interval as the true failure time, and a less frequently adopted approach is to impute the left-point of the time interval as the true failure time. For intervals with $R_i = \infty$ or right-censored observations, the original observations are kept and no imputation is needed. Then we have a set of right-censored failure time data, and can apply standard survival analysis methods for right-censored data to the imputed data. To analyze right-censored data, there is a widely accepted standard by the pharmaceutical industry: the Kaplan-Meier estimator (Kaplan and Meier, 1958) is used in estimation of survival curve; the log-rank test is used for hypothesis testing of treatment

effect; and the Cox's proportional hazards model is used to estimate treatment effect given other covariates (Cox, 1972).

When all the finite observation time intervals are narrow, above three imputation methods will give similar results. In addition, when all the assessment intervals are equal for all subjects (i.e. evenly spaced visits under balanced design), all three imputation methods will yield the same result because the ordering of event times remains intact no matter which imputation method is used. The biggest advantage of the single-point imputation approach lies in its simplicity, since inference can be easily performed using existing software packages. If all the intervals are narrow or there are only minimal overlapping among the intervals, this approach can provide a reasonable approximation to the inference based on observed data. However, in general, this approach may not be reliable, and it can create serious bias especially when the assessment schedule is different among treatment groups (e.g. unbalanced design) (Sun and Chen, 2010; Tang, Holland, and Sridhara, 2008).

Multiple Imputation Approach

For multiple imputations on the true failure times, we need to use some data augmentation algorithms (Tanner, 1991; Tanner and Wong, 1987) to impute values for T_i several times and get estimates iteratively. When our interest is to make inference about some unknown parameter θ, the general steps for multiple imputation are as follows:

Step 0. Given an initial value $\hat{\theta}^{(0)}$ and set $\hat{S}^{(0)}(t) = S(t; \hat{\theta}^{(0)})$.

step 1. At the lth iteration and kth imputation: let $T_i^{(k,l)} = L_i$ and $\delta_i^{(k,l)} = 0$ if $R_i = \infty$, otherwise, sample $T_i^{(k,l)}$ from $\hat{S}^{(l-1)}$ conditional on $T_i^{(k,l)} \in (L_i, R_i]$ and $\delta_i^{(k,l)} = 1$. This gives M sets of right-censored data

$$\{T_i^{(k,l)}, \delta_i^{(k,l)}, \boldsymbol{Z_i}; i = 1, ..., n\}, \tag{1.6}$$

$k = 1, ..., M$.

Step 2. For each of the M sets of right-censored data generated in step 1, obtain an estimate $\hat{\theta}^{(k,l)}$

and the variance parameter $\hat{\Sigma}^{(k,l)}$.

Step 3. Determine the updated estimator θ by

$$\hat{\theta}^{(l)} = \frac{1}{M}\sum_{k=1}^{M}\hat{\theta}^{(k,l)}, \tag{1.7}$$

and the variance function can be estimated by

$$\hat{\Sigma}^{(l)} = \frac{1}{M}\sum_{k=1}^{M}\hat{\Sigma}^{(k,l)} + \left(1+\frac{1}{M}\right)\frac{\sum_{k=1}^{M}[\hat{\theta}_t^{(k,l)} - \hat{\theta}_t^{(l)}][\hat{\theta}_t^{(k,l)} - \hat{\theta}_t^{(l)}]'}{M-1}. \tag{1.8}$$

Step 4. Repeat steps 1-3 until converge.

In Step 3, the first term of (1.8) represents the within-imputation estimation and the second term represents the between-imputation estimation for variance of θ. For Step 0, the initial value can be simply obtained by applying the single point imputation approach (e.g. mid-point imputation) to the observed data and use the resulting estimate.

1.3.5 Nonparametric Methods

Nonparametric Maximum Likelihood Estimation (NPMLE)

In the case of right-censored failure time data, the Kaplan-Meier estimator (Kaplan and Meier, 1958) provides the NPMLE of a survival function, which is very simple and has been extensively studied in the literature. However, for the interval-censored failure time data, the nonparametric inference can be quite complicated and the NPMLE of a survival function usually does not have a closed form, thus can only be estimated using iterative algorithms.

The first person who proposes a nonparametric method for estimating the survival function of interval-censored data is Peto (1973), who uses a Newton-Raphson method to estimate the NPMLE, then Turnbull (1976) formulates an self-consistency algorithm to estimate the NPMLE

for interval-censored data. This algorithm can be regarded as an application of the expectation-maximization (EM) algorithm. Consider a failure time study which consists n independent subjects with survival function $S(t)$, let T_i denote the true failure time of subject $i, i = 1, ..., n$, and T_i is censored by interval $[L_i, R_i]$ (both Peto (1973) and Turnbull (1976) assume a closed censoring interval). Then the likelihood function is

$$L = \prod_{i=1}^{n} [S(L_i) - S(R_i^+)]. \tag{1.9}$$

where $S(R_i^+)$ means $\lim_{\Delta \to 0^+} S(R_i + \Delta)$.

Since the observed event times only occur within potentially overlapping intervals, the survival curve can only jump within so-called equivalence sets $[q_j, p_j], j = 1, ..., m$, where $q_j \leq p_j < q_{j+1} \leq ....$ Between p_j and q_{j+1} the curve is flat. The estimate of $S(t)$ is unique only up to these equivalence classes, therefore, any function that jumps the appropriate amount within the equivalence class will yield the same likelihood. This estimator has no closed form, and Turnbull's algorithm is presented below.

To construct the estimator, define $s_j = S(q_j) - S(p_j), 1 \leq j \leq m$, then the vector $\boldsymbol{s} = (s_1, ..., s_m)'$ where $\sum s_j = 1$ and $s_j \geq 0$, defines equivalence classes on the space of the survival functions which are flat outside these equivalence classes. For $1 \leq i \leq n, 1 \leq j \leq m$, let $\alpha_{ij} = 1$ if $T_i \in [q_i, p_i]$ and 0 otherwise. Then the expected number of events occur within time interval $[q_j, p_j]$ is given by

$$d_j = \sum_{i=1}^{n} \frac{\alpha_{ij} s_j}{\sum_{k=1}^{m} \alpha_{ik} s_k}, j = 1, ...m. \tag{1.10}$$

The steps for obtaining Turnbull's estimator are as follow:

Step 1. Make an initial guess on $s_j^0 (1 \leq j \leq m)$. This can be any set of positive numbers summing to unity, e.g. $s_j = 1/m$ for all j.

Step 2. Compute expected value d_j from equation (1.10) and obtain an updated estimator $s_j^1 =$

d_j/n for $1 \leq j \leq m$.

Step 3. Return to step 1 with $\boldsymbol{s^1}$ replacing $\boldsymbol{s^0}$, etc.

Step 4. Stop when the required accuracy has been achieved.

Besides the self-consistency algorithm developed by Turnbull (1976), an iterative convex minorant (ICM) algorithm is proposed by Groeneboom and Wellner (1992) to estimate the survival function, and later, Wellner and Zhan (1997) develop a hybrid algorithm which combines the self-consistency and ICM algorithms together, named as EM-ICM algorithm. All three algorithms are implemented in SAS PROC ICLIFETEST module.

Comparison of Survival Functions

Comparison of treatment is usually one of the primary objectives in most biomedical studies such as clinical trials. In the case of right-censored data, log-rank test is usually applied as a standard test. The two sample log-rank test statistic can be written as

$$U = \sum_{j=1}^{m} \left(d_{1j} - \frac{d_j n_{1j}}{n_j} \right),$$

where m is the number of failure times, d_{1j} is the number of failures in group one and d_j is the number of failures in both groups at time j. Similarly, n_{1j} is the number at risk for a failure in the first group, and n_j is the number at risk in both groups at time j. Then, the statistic U is divided by its standard error and compared to a standard normal distribution.

For interval-censored data, Finkelstein (1986) derives a score test under continuous proportional hazards model. Based on the discrete logistic model, Sun (1996) proposes a test statistic which is considered as generalization of the original log-rank test. Fay (1996) obtains another test under the proportional odds model, he categorizes all three tests as a family of weighted log-rank tests and attempts to construct a unified framework of comparing survival functions for interval-censored data (Fay, 1999).

For interval-censored data $T_i \in (L_i, R_i]$, we assume $L_i, R_i \in \{s_0, ..., s_{m+1}\}$, where $s_{m+1} = \infty$ for right-censored data. Let $\alpha_{ij} = 1$ if $L_i < s_j \leq R_i$, and 0 otherwise. We wish to test for any difference between k treatments. For the ith subject, let $\boldsymbol{z_i}$ be a $k \times 1$ vector of zeros except for the lth row which is one. Under the null hypothesis, the statistic for the lth treatment group, where $1 \leq l \leq k$, takes the form

$$U_l = \sum_{j=1}^{m} w_j \left[d'_{jl} - \frac{n'_{jl} d'_j}{n'_j} \right], \tag{1.11}$$

where w_j is the weight, depending on the particular model specified (Fay, 1999), d'_{jl} is the expected number of failures in time interval $(s_{j-1}, s_j]$ for the lth treatment group, d'_j represents the expected total number of failures in all treatment groups, and similarly, n'_{jl} and n'_j represent the expected numbers at risk. Under the null hypothesis, we have

$$d'_{jl} = \sum_{i=1}^{n} z_{il} \frac{\alpha_{ij}(P_{j-1} - P_j)}{\sum_{j=1}^{m+1} \alpha_{ij}(P_{j-1} - P_j)} \tag{1.12}$$

and

$$n'_{jl} = \sum_{i=1}^{n} z_{il} \frac{\sum_{k=j}^{m+1} \alpha_{ik}(P_{k-1} - P_k)}{\sum_{j=1}^{m+1} \alpha_{ij}(P_{j-1} - P_j)}, \tag{1.13}$$

where $P_j = Pr(T > s_j | z_i)$.

Regression Analysis

As I mentioned previously, Cox's PH model is perhaps the most commonly used method in regression analysis of right-censored failure time data, because of the availability of the partial likelihood function derived under the model (Cox, 1972). A key advantage of the partial likelihood method is that one does not have to deal with the underlying baseline hazard function. In the case of interval-censored data, several procedures are also developed for regression analysis based on the proportional hazards model. Different from the PH model for right-censored data, incorporating of interval censoring into the PH model does not enable cancelling the baseline hazard function, therefore, estimation of the regression coefficients and derivation of its asymptotic properties are more challenging.

One of the pioneering papers is given by Finkelstein (1986), which uses the full likelihood approach to fit the proportional hazards model to interval-censored data by partitioning the time axis based on the endpoints of the event time intervals. Among others, Betensky et al. (2002) use a local likelihood to jointly estimate the regression coefficient and the baseline hazard function. In order to avoid estimation of baseline hazard function, Glenn (2011) proposes an estimating equation based method to select event time pairs when the ordering is unambiguous under PH model; Sun, Feng, and Zhao (2015) propose two simple estimation approaches, motivated by the imputation approach, that do not need estimation of the baseline cumulative hazard function.

Other procedures have also been proposed, such as those based on the proportional odds model, the accelerated failure time model, and the logistic model, overall, all the procedures for regression analysis of interval-censored failure time data are quite complicated and none of them are currently implemented by standard statistical software packages.

1.4 Influence of Study Design on Parameter Estimation

When designing a biomedical or epidemiology follow-up study with interval-censored outcome, the assessment schedule is usually decided based on some convenient selection (e.g. semi-annually, annually or bi-annually) and available budget. There is no theoretical guideline on how often the assessments should be scheduled. In addition, there are very few studies which evaluate whether and how the scheduling of assessments can influence the bias and precision of parameter estimation in interval-censored failure time data.

Alexander (2008) evaluates the influence of number of assessments in fixed study duration on the precision of event rate estimation. By assuming a constant event rate and uniform interval-censoring throughout the study, he derives the explicit formula for the Fisher information of the estimated event rate as below:

$$I = \frac{nT^2(1-e^{-\lambda T})}{m^2(e^{\lambda T/m}-1)(1-e^{-\lambda T/m})}, \tag{1.14}$$

where λ is the event rate, T is the total study follow-up time, n is number of subjects, and m is number of evenly spaced assessments within time T. Then the sampling variance of $\hat{\lambda}$ can be obtained by taking the reciprocal of the Fisher information. He studies the relationship between number of observations and ratio of sampling variance from interval-censored data vs. continuous surveillance (exact event time from exponential distribution). His study shows that the sampling variance of $\hat{\lambda}$ is always greater than that from the continuous surveillance, and this difference begin to decrease as number of assessments increase, with ratio close to 1 when $m \to \infty$. At fixed number of assessments, the ratio is bigger when event rate per study duration is higher.

In Glenn (2011), simulation studies results are presented which evaluate the bias and variability of a covariate coefficient estimate $(\hat{\beta})$ in the PH regression model using the method he proposes and compare with the results from right-point imputation approach. In his simulation, two types of monitoring schedule are used: every 12 months and every 24 months. His simulation results show that bias is minimal in either monitoring schedule, and comparing with the every 12 months monitoring schedule, the variability of the coefficient estimate is larger from the every 24 months monitoring schedule. When using the right-point imputation approach followed by the standard Cox's PH model, significant positive bias has been observed, which is even higher when subjects are followed every 24 months than every 12 months. Since it is imputation based approach, the precision of the estimate is not influenced by the interval width or the monitoring schedule.

Sun and Chen (2010) present the results from their simulation study which compares the performance of conventional imputation-based methods and the nonparametric method by Finkelstein (1986) when analyzing interval-censored data with two treatment arms. In this study, they assess the mean and standard deviation (SD) of the regression coefficient for the treatment effect under both balanced and unbalanced design. In balanced design, patients in the two treatment arms have equal assessment schedules at every 8 time units or 32 time units; in unbalanced design, patients in the two arms have unequal assessment schedules with every 8 time units in one arm and every

16 time units in the other. The results from this study show that under balanced design, Finkelsten's method creates unbiased estimator, the single-point imputation followed by Cox's PH model with Breslow's tie handling method creates biased result, and Efron's tie handling method works much better than Breslow's method. When using the single-point imputation method, larger bias is observed when the assessment schedule is more sparse (assessment at every 32 time units vs. every 8 time units). Under unbalanced design, Finkelsten's method has reasonability well performance. When using the single-point imputation method, the right-point imputation approach tends to overestimate treatment effect when patients in the control arm are more frequently assessed for events and tends to underestimate otherwise. The mid-point imputation method works better than right-point imputation, but overall, Finkelsten's method gives best performance.

1.5 Outline of this Dissertation

Previous studies show that: 1) when appropriate analysis methods are used, we can obtain unbiased results from interval-censored failure time data regardless the interval length; 2) the variability of the parameter estimation can be reduced by increasing the frequency of assessments (or reducing the interval length). Since increasing the frequency of assessments generally leads to increased cost, herein, I mainly evaluate the performance of unbalanced design with only increasing the frequency of assessments in the high risk group, which is determined by a baseline risk factor for the study endpoint, and then compare to the results from balanced design with the same frequency of assessments for all subjects. The hypothesis is that this proposed unbalanced design can help to improve the precision of the parameter estimation in interval-censored time to event data.

This dissertation is organized as follows. Chapter 2 includes the methods and results. In this chapter, by assuming exponential distribution of the true unobserved survival times, evenly spaced visit schedule in order to create interval censoring, and a baseline risk factor which separates the entire study cohort into two strata, I use the maximum likelihood theory to derive the theoretical formula for sampling variance estimator of a binary covariate under parametric setting. Then

I compare the sampling variance estimator obtained from unbalanced design against that from balanced design under various situations. Next, simulation studies are presented to evaluate the accuracy of the derived sampling variance estimator in small samples. Simulations studies are also conducted to assess the relative gain of precision for covariate estimation, by using unbalanced design vs. balanced design, under more complicated situations than those assumed for theoretical formula derivation. In last section, I apply the derived formula to power estimation for both balanced and unbalanced design then compare with results from numerical simulations, and I also evaluate the empirical Type I error rate under both balanced design and unbalanced design.

In Chapter 3, I evaluate the variance estimation of a few covariates effect using data from a study of metabolic control among patients with T1D, which was conducted by the Diabetes Research in Children Network (DirecNet) and the Type 1 Diabetes TrialNet, then compare the results from balanced vs. unbalanced designs using both parametric and nonparametric methods for analyzing interval-censored time to event data.

In Chapter 4, I give some discussion about current study, and in Chapter 5, I provide the concluding remarks and the potential further research on this topic.

Chapter 2: Methods and Results

In this chapter, I only evaluate the parametric case which assume that the time to event outcome follows the distribution from a parametric family. This chapter is organized as follows. In Section 2.1, I use the maximum likelihood theory to derive the theoretical formula for sampling variance estimator of a covariate effect by assuming exponential distribution of event time. In section 2.2, based on the derived formula, I compare the asymptotic sampling variance estimator obtained from unbalanced design against that from balanced design under various situations. Numerical studies results are presented to evaluate the accuracy of the derived sampling variance estimator in small samples. Using simulation samples, I also compare the efficiency of unbalanced design to balanced design under more complicated situations. In Section 2.3, I apply the derived formula to power estimation for both balanced design and unbalanced design then compare with results from numerical simulations. Empirical Type I error is also evaluated in this section.

2.1 Deriving Sampling Variance Estimator of a Covariate Effect

Consider a longitudinal study with evenly spaced visit schedules and time to an event as primary outcome, and the timing of the event is interval-censored by two consecutive visits. Based on a baseline risk factor, the study cohort can be separated into two strata: a low risk stratum and a high risk stratum. The study purpose is to evaluate the common effect of a binary covariate Z on the time to event outcome in both strata. The study duration is fixed time T. For simplicity, I assume no skipping of visits or early dropout. In a regular balanced design, every subject has the same number of visits m within total study duration T, therefore, each interval has length of T/m. In the unbalanced design, the number of visits depends on the baseline risk factor. Assuming the

number of visits in the low risk stratum is m and we want to increase the number of visits in the high risk stratum to $c * m$, where c is a pre-defined constant and $c > 1$. The common effect of covariate Z is $exp(\beta)$.

To simplify formula derivation process, I also assume the event rate follows exponential distribution in each stratum. In the lower baseline risk stratum, I assume the baseline event rate when $Z = 0$ is λ_1, sample size is $2n_1$ ($N = n_1$ for each category of the covariate Z), and let $i = 1, ..., m$, then the probability of remaining free of the event till the ith visit is $\exp\left[-iT\lambda_1 \exp(\beta Z)/m\right]$, and the probability for having the event between visits $i-1$ and i is $\exp\left[-(i-1)T\lambda_1 \exp(\beta Z)/m\right]\{1-\exp\left[-T\lambda_1 \exp(\beta Z)/m\right]\}$. Suppose d_{0i} and d_{1i} subjects have events between visits $i-1$ and i when $Z = 0$ and $Z = 1$, respectively, then the likelihood function when $Z = 0$ among the low risk stratum is

$$L_{10} = \frac{n_1!}{(n_1 - \sum_{i=1}^{m} d_{0i})!} \exp\left[-\lambda_1 T\left(n_1 - \sum_{i=1}^{m} d_{0i}\right)\right] \prod_{i=1}^{m} \frac{1}{d_{0i}!}\left\{\exp\left[-\frac{(i-1)\lambda_1 T}{m}\right]\left[1 - \exp\left(-\frac{\lambda_1 T}{m}\right)\right]\right\}^{d_{0i}}, \tag{2.1}$$

and the likelihood function when $Z = 1$ among the low risk stratum is

$$L_{11} = \frac{n_1!}{(n_1 - \sum_{i=1}^{m} d_{1i})!} \exp\left[-\lambda_1 T e^{\beta}\left(n_1 - \sum_{i=1}^{m} d_{1i}\right)\right] \prod_{i=1}^{m} \frac{1}{d_{1i}!}\left\{\exp\left[-\frac{(i-1)\lambda_1 T e^{\beta}}{m}\right]\left[1 - \exp\left(-\frac{\lambda_1 T e^{\beta}}{m}\right)\right]\right\}^{d_{1i}}. \tag{2.2}$$

In the high baseline risk stratum, assume the baseline event rate when $Z = 0$ is λ_2, sample size is $2n_2$ ($N = n_2$ for each category of the covariate Z). Let $j = 1, ..., c * m$, and suppose h_{0j} and h_{1j} subjects have events between visits $j - 1$ and j when $Z = 0$ and $Z = 1$, respectively, then

the likelihood functions for $Z = 0$ subgroup and $Z = 1$ subgroup are

$$L_{20} = \frac{n_2!}{(n_2 - \sum_{j=1}^{cm} h_{0j})!} \exp\left[-\lambda_2 T \left(n_2 - \sum_{j=1}^{cm} h_{0j}\right)\right] \prod_{j=1}^{cm} \frac{1}{h_{0j}!} \left\{\exp\left[-\frac{(j-1)\lambda_2 T}{cm}\right]\left[1 - \exp\left(-\frac{\lambda_2 T}{cm}\right)\right]\right\}^{h_{0j}} \tag{2.3}$$

and

$$L_{21} = \frac{n_2!}{(n_2 - \sum_{j=1}^{cm} h_{1j})!} \exp\left[-\lambda_2 T e^{\beta} \left(n_2 - \sum_{j=1}^{cm} h_{1j}\right)\right] \prod_{j=1}^{cm} \frac{1}{h_{1j}!} \left\{\exp\left[-\frac{(j-1)\lambda_2 T e^{\beta}}{cm}\right]\left[1 - \exp\left(-\frac{\lambda_2 T e^{\beta}}{cm}\right)\right]\right\}^{h_{1j}}. \tag{2.4}$$

Based on above likelihood functions (2.1)-(2.4), the log-likelihood for the low risk stratum $Z = 0$ subgroup is (omitting terms not involving λ_1)

$$\log(L_{10}) = -\left(n_1 - \sum_{i=1}^{m} d_{0i}\right)\lambda_1 T + \sum_{i=1}^{m}\left[-\frac{d_{0i}(i-1)\lambda_1 T}{m} + d_{0i}\log\left(1 - \exp\left(-\frac{\lambda_1 T}{m}\right)\right)\right], \tag{2.5}$$

and the log-likelihood for the low risk stratum $Z = 1$ subgroup is (omitting terms not involving λ_1 and β)

$$\log(L_{11}) = -\left(n_1 - \sum_{i=1}^{m} d_{1i}\right)\lambda_1 T e^{\beta} + \sum_{i=1}^{m}\left[-\frac{d_{1i}(i-1)\lambda_1 T e^{\beta}}{m} + d_{1i}\log\left(1 - \exp\left(-\frac{\lambda_1 T e^{\beta})}{m}\right)\right)\right]. \tag{2.6}$$

Similarly, the log-likelihood for the two subgroups in the high risk stratum can be written as below:

$$\log(L_{20}) = -\left(n_2 - \sum_{i=1}^{cm} h_{0i}\right)\lambda_2 T + \sum_{i=1}^{cm}\left[-\frac{h_{0i}(i-1)\lambda_2 T}{cm} + h_{0i}\log\left(1-\exp\left(-\frac{\lambda_2 T}{cm}\right)\right)\right], \tag{2.7}$$

$$\begin{aligned}\log(L_{21}) = &-\left(n_2 - \sum_{i=1}^{cm} h_{1i}\right)\lambda_2 T e^{\beta} \\ &+ \sum_{i=1}^{cm}\left[-\frac{h_{1i}(i-1)\lambda_2 T e^{\beta}}{cm} + h_{1i}\log\left(1-\exp\left(-\frac{\lambda_2 T e^{\beta}}{cm}\right)\right)\right].\end{aligned} \tag{2.8}$$

So total log-likelihood is

$$\log(L) = \log(L_{10}) + \log(L_{11}) + \log(L_{20}) + \log(L_{21}). \tag{2.9}$$

There are three parameters to be estimated from the data: β, λ_1 and λ_2, but the parameter of interest is β. The first order derivative of the log-likelihood with respect to each of the three parameters is

$$\begin{aligned}\frac{\partial \log(L)}{\partial \beta} = &-\left(n_1 - \sum_{i=1}^{m} d_{1i}\right)\lambda_1 T e^{\beta} - \left(n_2 - \sum_{j=1}^{cm} h_{1j}\right)\lambda_2 T e^{\beta} \\ &+ \sum_{i=1}^{m}\left[-\frac{d_{1i}(i-1)\lambda_1 T e^{\beta}}{m} + \frac{\lambda_1 T d_{1i}}{m}\frac{e^{\beta}}{\exp(\lambda_1 T e^{\beta}/m) - 1}\right] \\ &+ \sum_{j=1}^{cm}\left[-\frac{h_{1j}(j-1)\lambda_2 T e^{\beta}}{cm} + \frac{\lambda_2 T h_{1j}}{cm}\frac{e^{\beta}}{\exp(\lambda_2 T e^{\beta}/cm) - 1}\right],\end{aligned} \tag{2.10}$$

$$\frac{\partial \log(L)}{\partial \lambda_1} = -\left(n_1 - \sum_{i=1}^{m} d_{0i}\right)T - \left(n_1 - \sum_{i=1}^{m} d_{1i}\right)T e^{\beta}$$

$$+\sum_{i=1}^{m}\left[-\frac{d_{0i}(i-1)T}{m}+\frac{d_{0i}T}{m}\frac{1}{\exp(\lambda_1 T/m)-1}\right]$$
$$+\sum_{i=1}^{m}\left[-\frac{d_{1i}(i-1)Te^{\beta}}{m}+\frac{d_{1i}Te^{\beta}}{m}\frac{1}{\exp(\lambda_1 Te^{\beta}/m)-1}\right], \tag{2.11}$$

$$\frac{\partial \log(L)}{\partial \lambda_2} = -\left(n_2-\sum_{j=1}^{cm}h_{0j}\right)T-\left(n_2-\sum_{j=1}^{cm}h_{1i}\right)Te^{\beta}$$
$$+\sum_{j=1}^{cm}\left[-\frac{h_{0j}(j-1)T}{cm}+\frac{h_{0j}T}{cm}\frac{1}{\exp(\lambda_2 T/cm)-1}\right]$$
$$+\sum_{j=1}^{cm}\left[-\frac{h_{1j}(j-1)Te^{\beta}}{cm}+\frac{h_{1j}Te^{\beta}}{cm}\frac{1}{\exp(\lambda_2 Te^{\beta}/cm)-1}\right]. \tag{2.12}$$

The second-order partial derivatives are

$$\frac{\partial^2 \log(L)}{\partial \beta^2} = -\left(n_1-\sum_{i=1}^{m}d_{1i}\right)\lambda_1 Te^{\beta}-\left(n_2-\sum_{j=1}^{cm}h_{1j}\right)\lambda_2 Te^{\beta}$$
$$-\sum_{i=1}^{m}\frac{d_{1i}(i-1)\lambda_1 Te^{\beta}}{m}-\sum_{j=1}^{cm}\frac{h_{1j}(j-1)\lambda_2 Te^{\beta}}{cm}$$
$$+\sum_{i=1}^{m}\left[\frac{\lambda_1 Td_{1i}e^{\beta}}{m}\frac{\exp(\lambda_1 Te^{\beta}/m)-1-\lambda_1 Te^{\beta}\exp(\lambda_1 Te^{\beta}/m)/m}{(\exp(\lambda_1 Te^{\beta}/m)-1)^2}\right]$$
$$+\sum_{j=1}^{cm}\left[\frac{\lambda_2 Th_{1j}e^{\beta}}{cm}\frac{\exp(\lambda_2 Te^{\beta}/cm)-1-\lambda_2 Te^{\beta}\exp(\lambda_2 Te^{\beta}/cm)/cm}{(\exp(\lambda_2 Te^{\beta}/cm)-1)^2}\right], \tag{2.13}$$

$$\frac{\partial^2 \log(L)}{\partial \lambda_1^2} = -\sum_{i=1}^{m}\frac{d_{0i}T^2}{m^2}\frac{\exp(\lambda_1 T/m)}{[\exp(\lambda_1 T/m)-1]^2}-\sum_{i=1}^{m}\frac{d_{1i}T^2e^{2\beta}}{m^2}\frac{\exp(\lambda_1 Te^{\beta}/m)}{[\exp(\lambda_1 Te^{\beta}/m)-1]^2}, \tag{2.14}$$

$$\frac{\partial^2 \log(L)}{\partial \lambda_2^2} = -\sum_{j=1}^{cm}\frac{h_{0j}T^2}{(cm)^2}\frac{\exp(\lambda_2 T/cm)}{[\exp(\lambda_2 T/cm)-1]^2}-\sum_{j=1}^{cm}\frac{h_{1j}T^2e^{2\beta}}{(cm)^2}\frac{\exp(\lambda_2 Te^{\beta}/cm)}{[\exp(\lambda_2 Te^{\beta}/cm)-1]^2}, \tag{2.15}$$

$$\frac{\partial^2 \log(L)}{\partial \lambda_1 \partial \beta} = -\left(n_1 - \sum_{i=1}^{m} d_{1i}\right) Te^{\beta} - \sum_{i=1}^{m} \frac{d_{1i}(i-1)Te^{\beta}}{m} + \sum_{i=1}^{m} \left[\frac{d_{1i}Te^{\beta}}{m} \frac{1}{\exp(\lambda_1 Te^{\beta}/m) - 1} - \frac{d_{1i}\lambda_1 T^2 e^{2\beta}}{m^2} \frac{\exp(\lambda_1 Te^{\beta}/m)}{[\exp(\lambda_1 Te^{\beta}/m) - 1]^2}\right], \quad (2.16)$$

$$\frac{\partial^2 \log(L)}{\partial \lambda_2 \partial \beta} = -\left(n_2 - \sum_{j=1}^{cm} h_{1j}\right) Te^{\beta} - \sum_{j=1}^{cm} \frac{h_{1j}(j-1)Te^{\beta}}{cm} + \sum_{j=1}^{cm} \left[\frac{h_{1j}Te^{\beta}}{cm} \frac{1}{\exp(\lambda_2 Te^{\beta}/cm) - 1} - \frac{h_{1j}\lambda_2 T^2 e^{2\beta}}{(cm)^2} \frac{\exp(\lambda_2 Te^{\beta}/cm)}{[\exp(\lambda_2 Te^{\beta}/cm) - 1]^2}\right], \quad (2.17)$$

and

$$\frac{\partial^2 \log(L)}{\partial \lambda_1 \partial \lambda_2} = 0. \quad (2.18)$$

Since the event rate in each subgroup follows exponential distribution, the expectation for number of events between each consecutive assessment time point is

$$E(d_{0i}) = n_1 \exp\left(-\frac{(i-1)\lambda_1 T}{m}\right)\left[1 - \exp\left(-\frac{\lambda_1 T}{m}\right)\right], \quad (2.19)$$

$$E(d_{1i}) = n_1 \exp\left(-\frac{(i-1)\lambda_1 Te^{\beta}}{m}\right)\left[1 - \exp\left(-\frac{\lambda_1 Te^{\beta}}{m}\right)\right], \quad (2.20)$$

$$E(h_{0j}) = n_2 \exp\left(-\frac{(j-1)\lambda_2 T}{cm}\right)\left[1 - \exp\left(-\frac{\lambda_2 T}{cm}\right)\right], \quad (2.21)$$

$$E(h_{1j}) = n_2 \exp\left(-\frac{(j-1)\lambda_2 Te^{\beta}}{cm}\right)\left[1 - \exp\left(-\frac{\lambda_2 Te^{\beta}}{cm}\right)\right]. \quad (2.22)$$

By substitution of above expectations (2.19)-(2.22) into the equation (2.13), this equation can be written as

$$
\begin{aligned}
-E\frac{\partial^2 \log(L)}{\partial \beta^2} &= n_1\lambda_1 T e^{\beta}\left[1-\left(1-\exp\left(-\frac{\lambda_1 T e^{\beta}}{m}\right)\right)\sum_{i=1}^{m}\exp\left(-\frac{(i-1)\lambda_1 T e^{\beta}}{m}\right)\right] \\
&+ n_2\lambda_2 T e^{\beta}\left[1-\left(1-\exp\left(-\frac{\lambda_2 T e^{\beta}}{cm}\right)\right)\sum_{j=1}^{cm}\exp\left(-\frac{(j-1)\lambda_2 T e^{\beta}}{cm}\right)\right] \\
&+ \frac{n_1\lambda_1 T e^{\beta}}{m}\left[1-\exp\left(-\frac{\lambda_1 T e^{\beta}}{m}\right)\right]\sum_{i=1}^{m}(i-1)\exp\left(-\frac{(i-1)\lambda_1 T e^{\beta}}{m}\right) \\
&+ \frac{n_2\lambda_2 T e^{\beta}}{cm}\left[1-\exp\left(-\frac{\lambda_2 T e^{\beta}}{cm}\right)\right]\sum_{j=1}^{cm}(j-1)\exp\left(-\frac{(j-1)\lambda_2 T e^{\beta}}{cm}\right) \\
&- \frac{n_1\lambda_1 T e^{\beta}}{m}\frac{1-\lambda_1 T e^{\beta}/m-\exp(-\lambda_1 T e^{\beta}/m)}{\exp(\lambda_1 T e^{\beta}/m)-1}\sum_{i=1}^{m}\exp\left(-\frac{(i-1)\lambda_1 T e^{\beta}}{m}\right) \\
&- \frac{n_2\lambda_2 T e^{\beta}}{cm}\frac{1-\lambda_2 T e^{\beta}/cm-\exp(-\lambda_2 T e^{\beta}/cm)}{\exp(\lambda_2 T e^{\beta}/cm)-1}\sum_{j=1}^{cm}\exp\left(-\frac{(j-1)\lambda_2 T e^{\beta}}{cm}\right).
\end{aligned}
\tag{2.23}
$$

Since $\sum_{i=1}^{m}\exp(-(i-1)\lambda_1 T e^{\beta}/m)$ is a geometric progression with m terms, the first of which is 1, with ratio $\exp(-\lambda_1 T e^{\beta}/m)$, hence

$$
\sum_{i=1}^{m}\exp\left(-\frac{(i-1)\lambda_1 T e^{\beta}}{m}\right)=\frac{1-exp(-\lambda_1 T e^{\beta})}{1-exp(-\lambda_1 T e^{\beta}/m)}, \tag{2.24}
$$

and

$$
\sum_{j=1}^{cm}\exp\left(-\frac{(j-1)\lambda_2 T e^{\beta}}{cm}\right)=\frac{1-exp(-\lambda_2 T e^{\beta})}{1-exp(-\lambda_2 T e^{\beta}/cm)}. \tag{2.25}
$$

For $\sum_{i=1}^{m}(i-1)\exp(-(i-1)\lambda_1 T e^{\beta}/m)$, when $i=1$, the first term is 0, so it can be treated as a progression with $m-1$ terms and i start with 2. Let $k=i-1$, then

$$\sum_{i=1}^{m}(i-1)\exp\left(-\frac{(i-1)\lambda_1 Te^{\beta}}{m}\right)=\sum_{k=1}^{m-1}k\exp\left(-\frac{k\lambda_1 Te^{\beta}}{m}\right).$$

Since the sum of a finite series

$$\sum_{k=1}^{n}kz^k=z\frac{1-(n+1)z^n+nz^{n+1}}{(1-z)^2},$$

let $z=\exp(-\lambda_1 Te^{\beta}/m)$, then

$$\begin{aligned}\sum_{i=1}^{m}(i-1)\exp\left(-\frac{(i-1)\lambda_1 Te^{\beta}}{m}\right)&=\sum_{k=1}^{m-1}k\exp\left(-\frac{k\lambda_1 Te^{\beta}}{m}\right)\\&=\frac{1-m\exp(-(m-1)\lambda_1 Te^{\beta}/m)+(m-1)\exp(-\lambda_1 Te^{\beta})}{[\exp(\lambda_1 Te^{\beta}/m)-1][1-\exp(-\lambda_1 Te^{\beta}/m)]}.\end{aligned}\tag{2.26}$$

Similarly,

$$\begin{aligned}\sum_{j=1}^{cm}(j-1)\exp\left(-\frac{(j-1)\lambda_2 Te^{\beta}}{cm}\right)&=\sum_{k=1}^{cm-1}k\exp\left(-\frac{k\lambda_2 Te^{\beta}}{cm}\right)\\&=\frac{1-cm\exp(-(cm-1)\lambda_2 Te^{\beta}/cm)+(cm-1)\exp(-\lambda_2 Te^{\beta})}{[\exp(\lambda_2 Te^{\beta}/cm)-1][1-\exp(-\lambda_2 Te^{\beta}/cm)]}.\end{aligned}\tag{2.27}$$

Therefore, by substituting above (2.24)-(2.27) into equation (2.23), this equation can be written as

$$\begin{aligned}-E\frac{\partial^2\log(L)}{\partial\beta^2}&=n_1\lambda_1 Te^{\beta}\exp(-\lambda_1 Te^{\beta})+n_2\lambda_2 Te^{\beta}\exp(-\lambda_2 Te^{\beta})\\&\quad+\frac{n_1\lambda_1 Te^{\beta}}{m}\frac{1-m\exp[-\lambda_1 T(m-1)e^{\beta}/m]+(m-1)\exp(-\lambda_1 Te^{\beta})}{\exp(\lambda_1 Te^{\beta}/m)-1}\\&\quad-\frac{n_1\lambda_1 Te^{\beta}}{m}\frac{[1-\lambda_1 Te^{\beta}/m-\exp(-\lambda_1 Te^{\beta}/m)][1-\exp(-\lambda_1 Te^{\beta})]}{[\exp(\lambda_1 Te^{\beta}/m)-1][1-\exp(-\lambda_1 Te^{\beta}/m)]}\end{aligned}$$

$$+\frac{n_2\lambda_2 Te^{\beta}}{cm}\frac{1-cm\exp[-\lambda_2 T(cm-1)e^{\beta}/cm]+(cm-1)\exp(-\lambda_2 Te^{\beta})}{\exp(\lambda_2 Te^{\beta}/cm)-1}$$
$$-\frac{n_2\lambda_2 Te^{\beta}}{cm}\frac{[1-\lambda_2 Te^{\beta}/cm-\exp(-\lambda_2 Te^{\beta}/cm)][1-\exp(-\lambda_2 Te^{\beta})]}{[\exp(\lambda_2 Te^{\beta}/cm)-1][1-\exp(-\lambda_2 Te^{\beta}/cm)]}. \quad (2.28)$$

Through similar steps, I can also calculate the expectations of other second-order partial derivatives, and finally, the Fisher information matrix can be expressed below:

$$I=\begin{bmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{bmatrix}, \quad (2.29)$$

where

$$\begin{aligned} I_{11} &= -E\frac{\partial^2\log(L)}{\partial\beta^2} \\ &= n_1\lambda_1 Te^{\beta}\exp(-\lambda_1 Te^{\beta})+n_2\lambda_2 Te^{\beta}\exp(-\lambda_2 Te^{\beta}) \\ &+\frac{n_1\lambda_1 Te^{\beta}}{m}\frac{1-m\exp[-\lambda_1 T(m-1)e^{\beta}/m]+(m-1)\exp(-\lambda_1 Te^{\beta})}{\exp(\lambda_1 Te^{\beta}/m)-1} \\ &-\frac{n_1\lambda_1 Te^{\beta}}{m}\frac{[1-\lambda_1 Te^{\beta}/m-\exp(-\lambda_1 Te^{\beta}/m)][1-\exp(-\lambda_1 Te^{\beta})]}{[\exp(\lambda_1 Te^{\beta}/m)-1][1-\exp(-\lambda_1 Te^{\beta}/m)]} \\ &+\frac{n_2\lambda_2 Te^{\beta}}{cm}\frac{1-cm\exp[-\lambda_2 T(cm-1)e^{\beta}/cm]+(cm-1)\exp(-\lambda_2 Te^{\beta})}{\exp(\lambda_2 Te^{\beta}/cm)-1} \\ &-\frac{n_2\lambda_2 Te^{\beta}}{cm}\frac{[1-\lambda_2 Te^{\beta}/cm-\exp(-\lambda_2 Te^{\beta}/cm)][1-\exp(-\lambda_2 Te^{\beta})]}{[\exp(\lambda_2 Te^{\beta}/cm)-1][1-\exp(-\lambda_2 Te^{\beta}/cm)]}, \end{aligned}$$

$$\begin{aligned} I_{22} &= -E\frac{\partial^2\log(L)}{\partial\lambda_1^2} \\ &= \frac{n_1T^2[1-\exp(-\lambda_1 T)]}{m^2[\exp(\lambda_1 T/m)-1][1-\exp(-\lambda_1 T/m)]} \\ &+\frac{n_1T^2e^{2\beta}[1-\exp(-\lambda_1 Te^{\beta})]}{m^2[\exp(\lambda_1 Te^{\beta}/m)-1][1-\exp(-\lambda_1 Te^{\beta}/m)]}, \end{aligned}$$

$$
\begin{aligned}
I_{33} &= -E\frac{\partial^2 \log(L)}{\partial \lambda_2^2} \\
&= \frac{n_2 T^2[1-\exp(-\lambda_2 T)]}{(cm)^2[\exp(\lambda_2 T/cm)-1][1-\exp(-\lambda_2 T/cm)]} \\
&+ \frac{n_2 T^2 e^{2\beta}[1-\exp(-\lambda_2 T e^{\beta})]}{(cm)^2[\exp(\lambda_2 T e^{\beta}/cm)-1][1-\exp(-\lambda_2 T e^{\beta}/cm)]},
\end{aligned}
$$

$$
\begin{aligned}
I_{12} = I_{21} &= -E\frac{\partial^2 \log(L)}{\partial \lambda_1 \partial \beta} \\
&= n_1 T e^{\beta}\exp(-\lambda_1 T e^{\beta}) - \frac{n_1 T e^{\beta}}{m}\frac{1-\exp(-\lambda_1 T e^{\beta})}{\exp(\lambda_1 T e^{\beta}/m)-1} \\
&+ \frac{n_1 T e^{\beta}}{m}\frac{1-m\exp[-\lambda_1 T(m-1)e^{\beta}/m]+(m-1)\exp(-\lambda_1 T e^{\beta})}{\exp(\lambda_1 T e^{\beta}/m)-1} \\
&+ \frac{n_1 \lambda_1 T^2 e^{2\beta}}{m^2}\frac{1-\exp(-\lambda_1 T e^{\beta})}{[\exp(\lambda_1 T e^{\beta}/m)-1][1-\exp(-\lambda_1 T e^{\beta}/m)]},
\end{aligned}
$$

$$
\begin{aligned}
I_{13} = I_{31} &= -E\frac{\partial^2 \log(L)}{\partial \lambda_2 \partial \beta} \\
&= n_2 T e^{\beta}\exp(-\lambda_2 T e^{\beta}) - \frac{n_2 T e^{\beta}}{cm}\frac{1-\exp(-\lambda_2 T e^{\beta})}{\exp(\lambda_2 T e^{\beta}/cm)-1} \\
&+ \frac{n_2 T e^{\beta}}{cm}\frac{1-cm\exp[-\lambda_2 T(cm-1)e^{\beta}/cm]+(cm-1)\exp(-\lambda_2 T e^{\beta})}{\exp(\lambda_2 T e^{\beta}/cm)-1} \\
&+ \frac{n_2 \lambda_2 T^2 e^{2\beta}}{(cm)^2}\frac{1-\exp(-\lambda_2 T e^{\beta})}{[\exp(\lambda_2 T e^{\beta}/cm)-1][1-\exp(-\lambda_2 T e^{\beta}/cm)]},
\end{aligned}
$$

and $I_{23} = I_{32} = 0$.

From above formula for calculating Fisher information matrix, the sampling variance for $\hat{\beta}$ can be simply obtained by taking the corresponding component in the inversed Fisher information matrix I_{11}^{-1}. Therefore, the sampling variance of $\hat{\beta}$ can be expressed as a function of multiple parameters

$$
V(\hat{\beta}) = f(\lambda_1, \lambda_2, \hat{\beta}, m, c, n_1, n_2). \tag{2.30}
$$

When $c = 1$, this formula provides the sampling variance of $\hat{\beta}$ for regular balanced design. When $c > 1$, this formula provides the sampling variance of $\hat{\beta}$ for proposed unbalanced design.

2.2 Comparing Sampling Variance from Unbalanced Design to Balanced Design

2.2.1 Theoretical Results

To assess whether this unbalanced design with increased frequency of examinations in the high risk stratum can help to reduce the sampling variance of $\hat{\beta}$ when comparing to regular balanced design and to quantify the degree of reduction, I calculate the ratios of $V(\hat{\beta})$ for this unbalanced design vs. balanced design at different situations using the derived formula above.

In Figure 2.1, I assume fixed sample size $n_1 = n_2 = 50$ (total sample size 200), fixed study duration $T = 1$, and fixed baseline event rate in the low risk stratum $\lambda_1 = 1$, then I assess the the ratios of $V(\hat{\beta})$ for two unbalanced designs ($c = 2\&3$) vs. balanced design ($c = 1$) under various baseline event rates in the high risk stratum ($\lambda_2 = 2, 3, 4, 5$) and covariate effects ($\beta = -1, 1$)). In Figure 1A and 1B, $\beta = 1$, which denotes a positive covariate effect. In Figure 1C and 1D, $\beta = -1$, which denotes a negative covariate effect. In Figure 1A and 1C, the frequency of visits in the high risk stratum doubles the frequency of visits in the low risk stratum, and in Figure 1B and 1D, the frequency of visits in the high risk stratum triples the frequency of visits in the low risk stratum.

As shown in Figure 2.1, using unbalanced design can reduce the sampling variance of $\hat{\beta}$ compared to using balanced design. The reduction of sampling variance from using unbalanced design tends to be greater when visits are relatively sparse (i.e. only 1 or 2 exams conducted in the low risk group), then this benefit begins to decline when m increases. The reduction of variance also tends to be greater when the event rate in the high risk group is much higher than the low risk group, and when the covariate effect ($\hat{\beta}$) is positive compared to negative (at fixed baseline event rate). Compared to the unbalanced design which doubles the frequency of visits in the high risk stratum (Figure 2.1A and 2.1C), the unbalanced design which triples the frequency of visits in

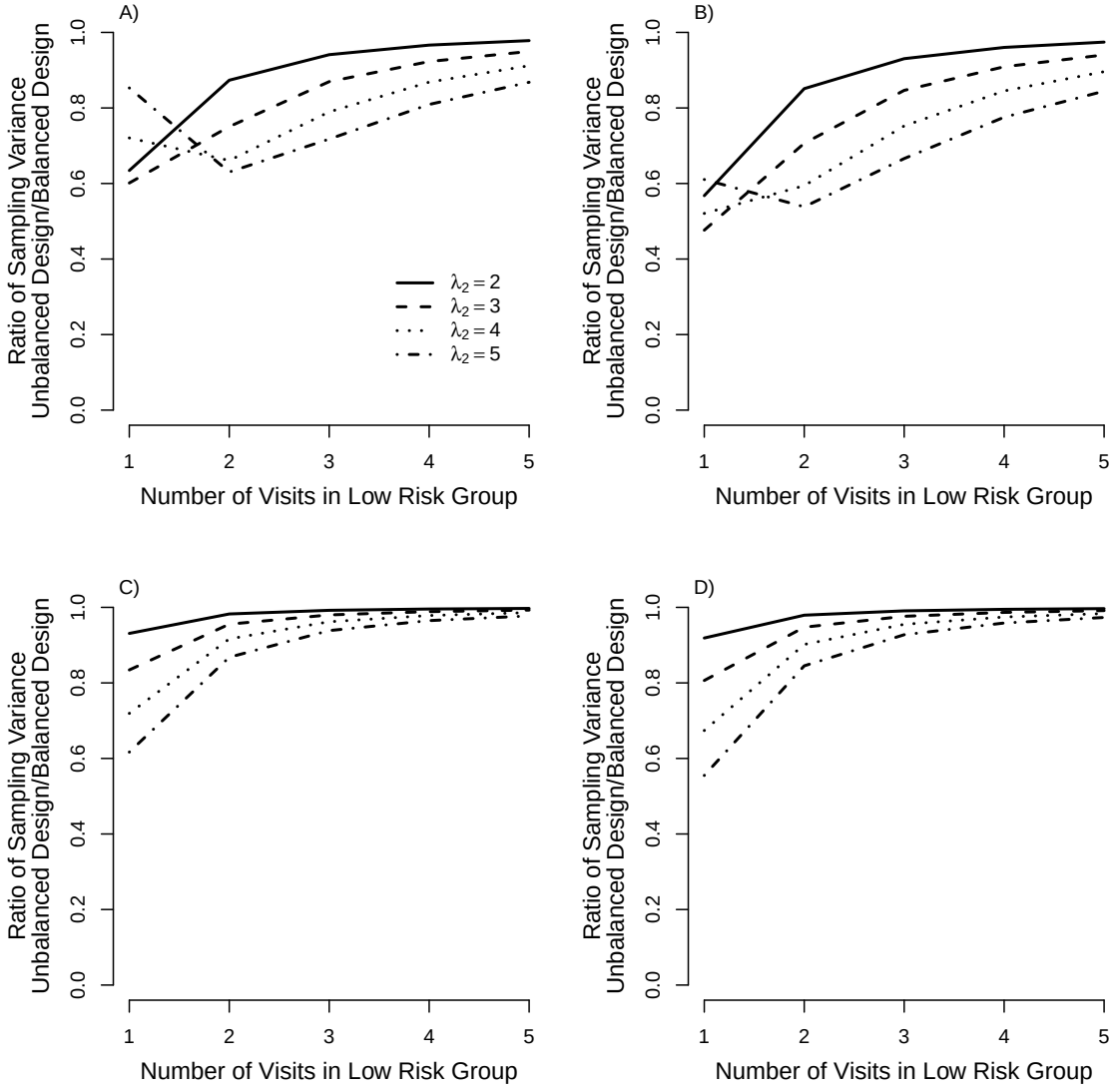

FIGURE 2.1: The precision for unbalanced design compared to balanced design in estimating a binary covariate effect. The vertical axis is sampling variance of $\hat{\beta}$ from unbalanced design $(c = 2, 3)$/balanced design $(c = 1)$. The horizontal axis is the number of evenly spaced visits in the low risk group throughout the study. In all 4 plots,assume $\lambda_1 = 1, n_1 = n_2 = 50, T = 1$, and the 4 lines in each plot represent different baseline event rates in the high risk stratum (λ_2). Plot A) $\beta = 1, c = 2; B) \beta = 1, c = 3; C) \beta = -1, c = 2; D) \beta = -1, c = 3$.

the high risk stratum (Figure 2.1B and 2.1D) leads to larger increase of precision for $\hat{\beta}$, especially when m is relatively small. The calculated variances from both balanced and unbalanced designs for selected data points are shown in Table 2.1.

Since in Figure 2.1, I assume equal sample size between the low risk stratum and high risk stratum, however, this may not be true in real data. Thus, in Figure 2.2, I evaluate the influence of different ratios of sample size between high risk stratum vs. low risk stratum on the ratio of $V(\hat{\beta})$ from unbalanced design vs. balanced design. Let $n_1 = 50$, and $n_2 = 25, 50, 100$ to represent different ratios of sample size between the two strata. As shown in Figure 2.2, higher ratio of subjects in high risk stratum vs. low risk stratum results in larger reduction of sampling variance for $\hat{\beta}$ from using unbalanced design compared to balanced design.

The unbalanced design evaluated in Figure 2.1 and 2.2 has increased frequency of visits in high risk stratum. Let m_H denote the number of visits in the high risk stratum, and let m_L denote the number of visits in the low risk stratum, in previously mentioned unbalanced design, $m_H = c * m_L$, where $c > 1$. However, another question naturally rising up is how about the efficiency of a "reversed unbalanced design" with increased frequency of visits in the low risk stratum ($m_L > m_H$).

In Figure 2.3, I compare the efficiency of covariate estimation from an unbalanced design with $m_H = 2m_L$ and a reversed unbalanced design with $m_L = 2m_H$ relative to that from balanced design. Same as previous figures, I assume $\lambda_1 = 1, \beta = 1, n_1 = n_2 = 50, T = 1$, and let $\lambda_2 = 2, 3, 4, 5$ for plot A to D. As shown in Figure 2.3, both unbalanced design with increased number of visits in the high risk stratum and reversed unbalanced design with increased number of visits in the low risk stratum can help to reduce the sampling variance of $\hat{\beta}$, however, unbalanced design generally leads to much larger reduction of sampling variance compared to the reversed unbalanced design. Especially when λ_2 is much bigger than λ_1, this difference tends to be greater. These results reveal that event rate plays an important role in the gained efficiency from using unbalanced design given the same number of visits are increased.

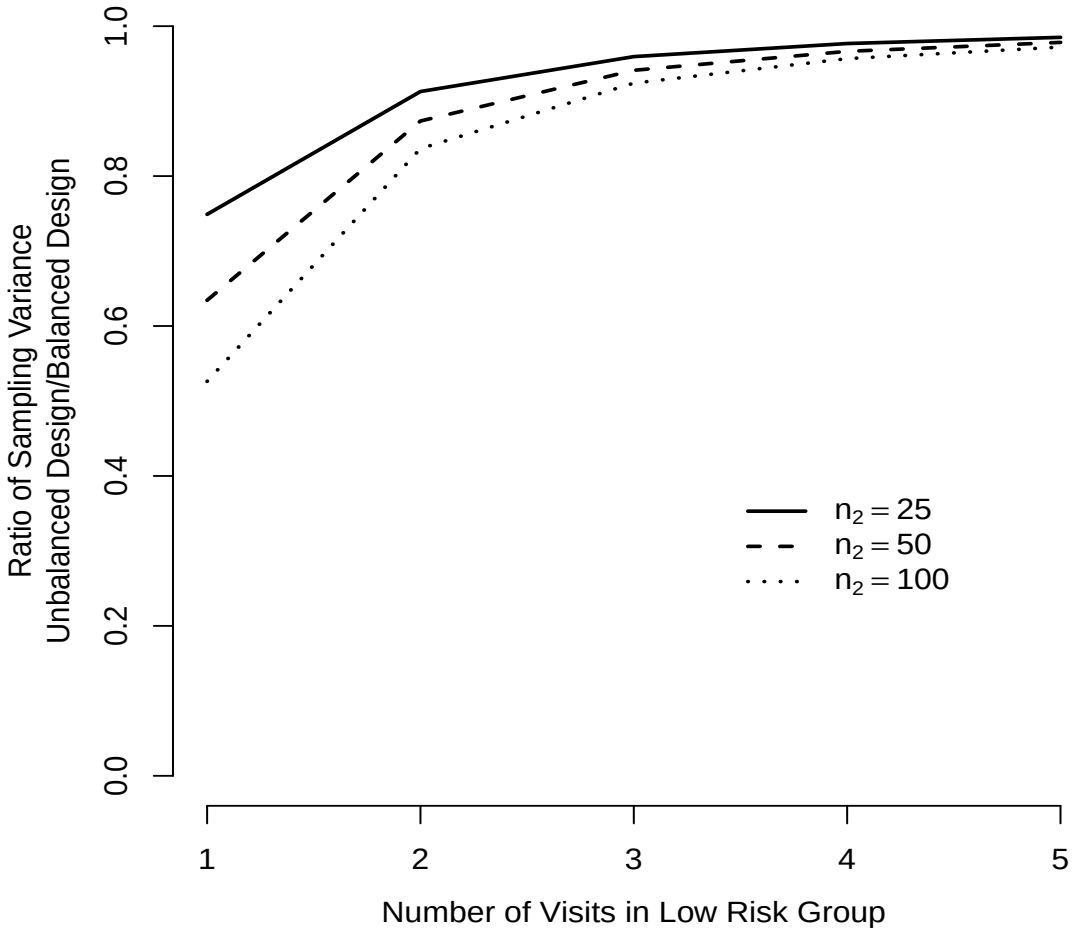

FIGURE 2.2: The precision of covariate effect estimation from unbalanced design compared to balanced design at different sample sizes. Assume $\lambda_1 = 1, \lambda_2 = 2, \beta = 1, n_1 = 50, T = 1$, and $c = 2$ in unbalanced design.

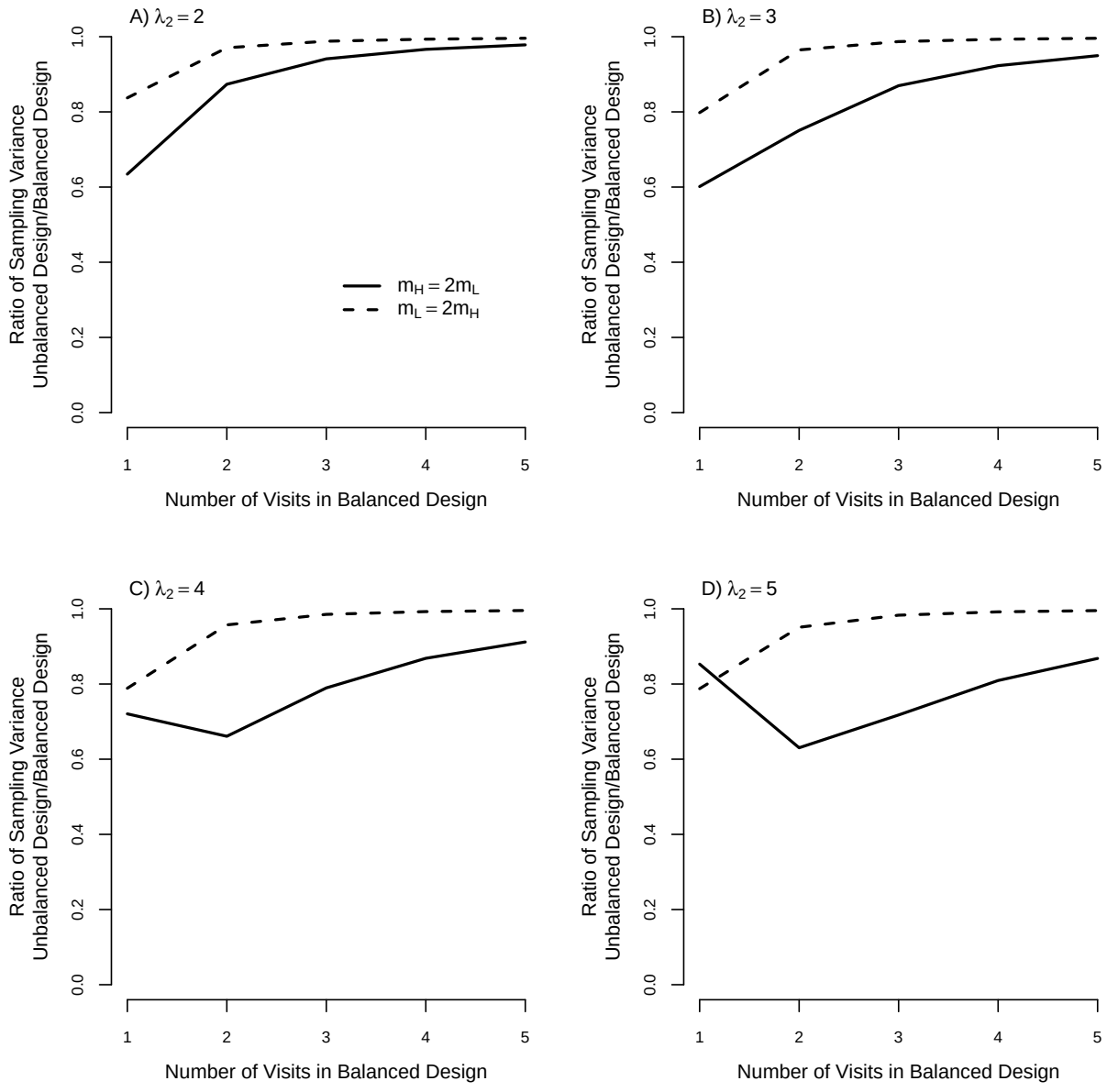

FIGURE 2.3: The efficiency of unbalanced design and reversed unbalanced design against balanced design. Assume $\lambda_1 = 1, \beta = 1, n_1 = n_2 = 50, T = 1$.

2.2.2 Numerical Results

Besides above results based on the derived theoretical formula for $V(\hat{\beta})$, I also conduct numerical simulation studies. The first objective of numerical studies is to evaluate the performance of above derived formula in section 2.1 on small samples, and the second objective of numerical studies is to evaluate the efficiency of the proposed unbalanced design against balanced design in more complicated situations where deriving of theoretical formula could be difficult.

In order to evaluate the performance of above derived formula on small samples, first, numerical simulation samples are generated and used to calculate sampling variances of $\hat{\beta}$ under each parameter setting for above theoretical results in Table 2.1. One thousand simulations are drawn on each parameter setting, and $\hat{\beta}$ is obtained using nlminb() function in R. The empirical sampling variances for $\hat{\beta}$ are also shown in Table 2.1 to compare with the asymptotic results from the derived theoretical formula.

As shown in Table 2.1, it is obvious that large discrepancy exists between asymptotic and empirical sampling variances when is m_L very small (i.e. $m_L = 1$) and the discrepancy is more pronounced when the event rate is higher. However, when $m_L \geq 2$, the discrepancies between asymptotic results and empirical results become much smaller. The plots for relative discrepancies between the asymptotic and empirical variances are shown in Appendix A Figure A.1 to A.6.

I also evaluate the discrepancy between asymptotic and empirical sampling variances (from 1,000 simulations) of $\hat{\beta}$ when total sample size ($N = 2(n_1 + n_2)$) varies. Let's define relative discrepancy between asymptotic and empirical sampling variances of $\hat{\beta}$ as $(V_e - V_a)/[(V_e + V_a)/2]$, where V_e is the empirical variance of $\hat{\beta}$ and V_a is asymptotic variance of $\hat{\beta}$. Assuming $\lambda_1 = 1, \lambda_2 = 4, \beta = 1, n_1 = n_2, T = 1, m_L = 3$ and $m_H = 2m_L$, the relative discrepancy between the two variances can be reduced to $< 10\%$ when total sample size is ≥ 100 (Figure 2.4).

TABLE 2.1: Asymptotic and empirical sampling variances of $\hat{\beta}$ for a binary covariate under balanced and unbalanced design. Assuming $n_1 = n_2 = 50, \lambda_1 = 1, T = 1$, and $m_H = 2m_L$ for unbalanced design.

m_L	Asymptotic Variance		Empirical Variance	
	Balanced Design	Unbalanced Design	Balanced Design	Unbalanced Design
$(\beta, \lambda_2) = (1, 2)$				
1	0.0523	0.0332	0.1471	0.0396
2	0.0295	0.0258	0.0304	0.0253
3	0.0262	0.0247	0.0274	0.0247
4	0.0251	0.0243	0.0266	0.0253
5	0.0247	0.0241	0.0269	0.0257
$(\beta, \lambda_2) = (1, 3)$				
1	0.0681	0.0410	0.2540	0.1234
2	0.0356	0.0267	0.0403	0.0280
3	0.0283	0.0247	0.0313	0.0256
4	0.0260	0.0240	0.0284	0.0263
5	0.0249	0.0237	0.0276	0.0264
$(\beta, \lambda_2) = (1, 4)$				
1	0.0721	0.0520	0.1975	0.1468
2	0.0435	0.0288	0.0454	0.0315
3	0.0321	0.0254	0.0330	0.0245
4	0.0280	0.0243	0.0294	0.0246
5	0.0261	0.0238	0.0267	0.0236
$(\beta, \lambda_2) = (-1, 2)$				
1	0.0419	0.0390	0.0415	0.0387
2	0.0386	0.0379	0.0390	0.0385
3	0.0380	0.0377	0.0391	0.0393
4	0.0378	0.0377	0.0392	0.0391
5	0.0378	0.0376	0.0383	0.0379
$(\beta, \lambda_2) = (-1, 3)$				
1	0.0430	0.0359	0.0511	0.0363
2	0.0356	0.0340	0.0338	0.0331
3	0.0343	0.0336	0.0350	0.0341
4	0.0339	0.0335	0.0345	0.0339
5	0.0337	0.0335	0.0362	0.0356
$(\beta, \lambda_2) = (-1, 4)$				
1	0.0495	0.0356	0.0564	0.0343
2	0.0353	0.0323	0.0374	0.0339
3	0.0330	0.0318	0.0345	0.0337
4	0.0323	0.0316	0.0341	0.0330
5	0.0319	0.0315	0.0342	0.0335

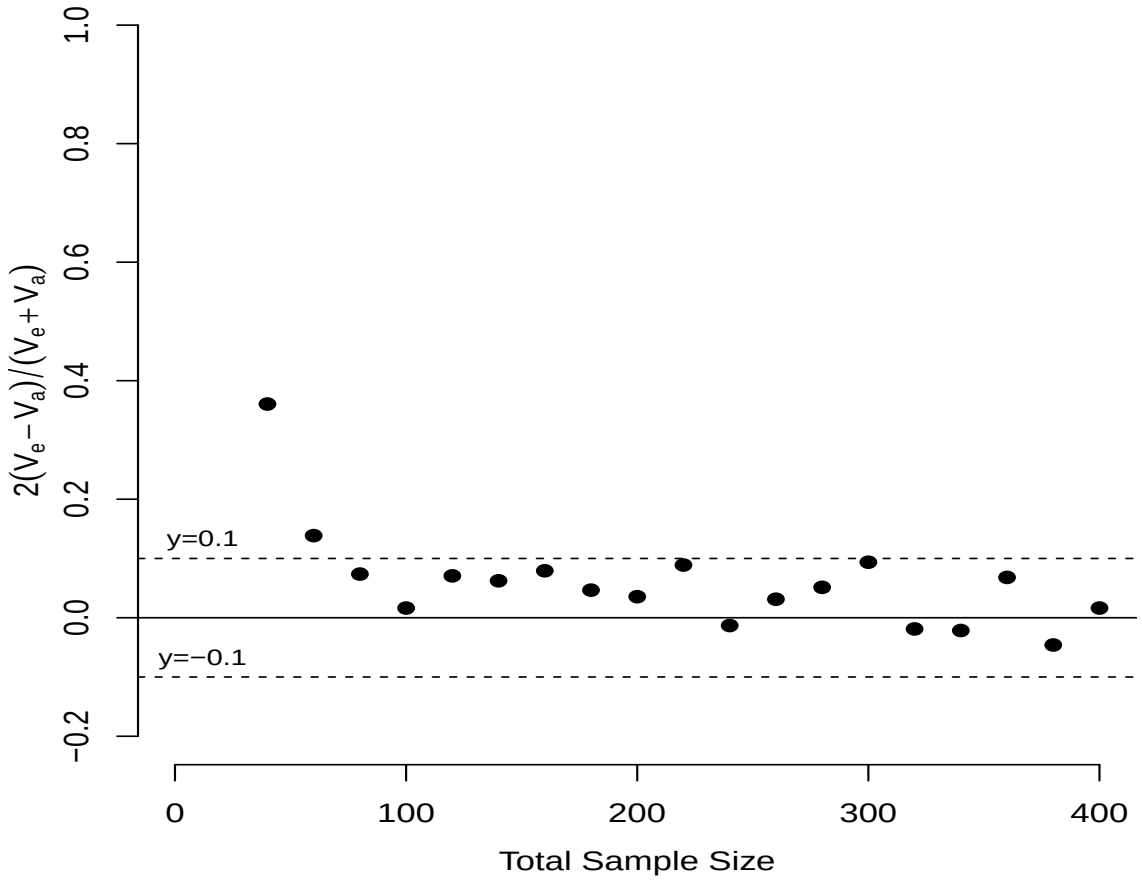

FIGURE 2.4: Relative discrepancy between asymptotic and empirical variances at different sample sizes. Assume $\lambda_1 = 1, \lambda_2 = 4, \beta = 1, n_1 = n_2, T = 1, m_L = 3$ and $m_H = 2m_L$.

When deriving the theoretical formula for sampling variance of $\hat{\beta}$, I assume a binary covariate of interest Z, here I also use numerical simulations to calculate the sampling variance of $\hat{\beta}$ for a continuous covariate under balanced design and unbalanced design. Assuming the continuous covariate Z is i.i.d. normal with mean 1 and SD 0.5, and the effect on event rate is also $\exp(\beta Z)$, other parameter settings are similar to Table 2.1, I evaluate the efficiency of unbalanced design vs. balanced design in estimating of $\hat{\beta}$ for the continuous covariate and present the results (from 1,000 simulations on each parameter setting) in Table 2.2. The results are similar to what I obtain above for the binary covariate (Table 2.1 and Figure 2.1). Clearly, using unbalanced design helps to reduce the sampling variance $\hat{\beta}$ compared to balanced design, and the relative reduction of

sampling variance is more pronounced when m_L is relatively small, when λ_2 is bigger, and when β is positive.

For all above results, I assume the exponential distribution of time to event outcome with constant event rate in each subgroup, however, this seldom happens in reality. In order to evaluate the performance of unbalanced design against balanced design in other situations in terms of improving precision, I also generate 1,000 numerical simulation samples from Weibull distribution of event times and calculate the sampling variance of $\hat{\beta}$. Two types of Weibull distribution are evaluated: $W(1.5)$ with event rate increasing with time and $W(0.8)$ with event rate decreasing with time. Assuming the baseline event rate in the low risk stratum $\lambda_1 = 1$, and the number of visits in the high risk stratum is twice as many as that in the low risk stratum $(c = 2)$, I calculate the sampling variance under balanced design $V(\hat{\beta})_1$ and unbalanced design $V(\hat{\beta})_2$ for several case scenarios: $\beta = (1, -1), \lambda_2 = (2, 4)$ and $m = (1, 2, 3, 4, 5)$ (Table 2.3). The ratios of the sampling variances from the two types of design $V(\hat{\beta})_2/V(\hat{\beta})_1$ under different parameter settings (Table 2.3) reveal similar trends to those observed from exponential distribution (Figure 2.1).

Another assumption I make when deriving theoretical formula is that visits are evenly spaced throughout the whole study duration T, however, this is not true in many situations. In a lot of clinical trials studies, scheduled visits are usually more intense in the early phase of the study than the later phase of the study, especially in those trials when the investigators expect quicker changes in the earlier phase of the trial. To assess the efficiency of parameter estimation using unbalanced design vs. balanced design in above situation, I generate 1,000 simulation samples based on Weibull distribution of event times. Assuming the event time follows $W(0.8)$ distribution (i.e. event rate decreases with time), and same as previous tables, let $n_1 = n_2 = 50, \lambda_1 = 1, T = 1$. For the balanced design, I assume there are three visits $(m_L = m_H = 3)$ at $t = (0.2, 0.5, 1.0)$; for unbalanced design, the number of visits in the low risk stratum is the same as balanced design, but the high risk stratum has six visits $(m_H = 2m_L = 6)$ at $t = (0.1, 0.2, 0.35, 0.5, 0.75, 1.0)$. Then I assess

TABLE 2.2: Empirical sampling variances of $\hat{\beta}$ for a continuous covariate under balanced and unbalanced design. Assuming total $N = 200$ with equal $N's$ in high risk and low risk groups,$\lambda_1 = 1, T = 1$, and $m_H = 2m_L$ for unbalanced design.

m_L	$V(\hat{\beta})_1^*$	$V(\hat{\beta})_2^{**}$	$V(\hat{\beta})_2^{**}/V(\hat{\beta})_1^*$
$(\beta, \lambda_2) = (1, 2)$			
1	0.1741	0.0817	0.47
2	0.0516	0.0356	0.69
3	0.0359	0.0280	0.78
4	0.0309	0.0267	0.86
5	0.0276	0.0250	0.90
$(\beta, \lambda_2) = (1, 3)$			
1	0.2499	0.1474	0.59
2	0.0644	0.0445	0.69
3	0.0454	0.0316	0.69
4	0.0337	0.0256	0.76
5	0.0304	0.0243	0.80
$(\beta, \lambda_2) = (1, 4)$			
1	0.2364	0.1674	0.71
2	0.0738	0.0508	0.69
3	0.0531	0.0371	0.70
4	0.0415	0.0301	0.73
5	0.0343	0.0266	0.78
$(\beta, \lambda_2) = (-1, 2)$			
1	0.0607	0.0540	0.89
2	0.0563	0.0556	0.99
3	0.0489	0.0492	1.00
4	0.0558	0.0555	0.99
5	0.0521	0.0521	1.00
$(\beta, \lambda_2) = (-1, 3)$			
1	0.0573	0.0478	0.83
2	0.0491	0.0468	0.95
3	0.0436	0.0433	0.99
4	0.0467	0.0463	0.99
5	0.0420	0.0418	1.00
$(\beta, \lambda_2) = (-1, 4)$			
1	0.0584	0.0486	0.83
2	0.0487	0.0448	0.92
3	0.0428	0.0411	0.96
4	0.0438	0.0424	0.97
5	0.0466	0.0459	0.98

*Variance of $\hat{\beta}$ for balanced design ($m_H = m_L$)

**Variance of $\hat{\beta}$ for unbalanced design ($m_H = 2m_L$)

TABLE 2.3: Empirical sampling variances of $\hat{\beta}$ under balanced and unbalanced design based on Weibull models. Assuming $n_1 = n_2 = 50, \lambda_1 = 1, T = 1$.

m_L	W(1.5)			W(0.8)		
	$V(\hat{\beta})_1^*$	$V(\hat{\beta})_2^{**}$	$V(\hat{\beta})_2^{**}/V(\hat{\beta})_1^*$	$V(\hat{\beta})_1^*$	$V(\hat{\beta})_2^{**}$	$V(\hat{\beta})_2^{**}/V(\hat{\beta})_1^*$
$(\beta, \lambda_2) = (1, 2)$						
1	0.1923	0.0369	0.19	0.1654	0.0585	0.35
2	0.0315	0.0287	0.91	0.0361	0.0302	0.84
3	0.0281	0.0267	0.95	0.0331	0.0306	0.92
4	0.0266	0.0258	0.97	0.0295	0.0281	0.95
5	0.0271	0.0261	0.96	0.0261	0.0254	0.97
$(\beta, \lambda_2) = (1, 4)$						
1	0.2029	0.0842	0.41	0.2358	0.2059	0.87
2	0.0419	0.0308	0.73	0.0597	0.0410	0.69
3	0.0299	0.0270	0.90	0.0429	0.0327	0.76
4	0.0281	0.0266	0.95	0.0378	0.0315	0.83
5	0.0279	0.0259	0.93	0.0372	0.0307	0.83
$(\beta, \lambda_2) = (-1, 2)$						
1	0.0420	0.0391	0.93	0.0441	0.0408	0.92
2	0.0428	0.0423	0.99	0.0422	0.0416	0.98
3	0.0391	0.0388	0.99	0.0382	0.0382	1.00
4	0.0403	0.0402	1.00	0.0403	0.0399	0.99
5	0.0394	0.0394	1.00	0.0400	0.0399	1.00
$(\beta, \lambda_2) = (-1, 4)$						
1	0.0584	0.0387	0.66	0.0590	0.0391	0.66
2	0.0388	0.0363	0.94	0.0391	0.0364	0.93
3	0.0384	0.0373	0.97	0.0369	0.0344	0.93
4	0.0378	0.0366	0.97	0.0353	0.0337	0.96
5	0.0345	0.0340	0.99	0.0355	0.0353	0.99

*Variance of $\hat{\beta}$ for balanced design ($m_H = m_L$)
**Variance of $\hat{\beta}$ for unbalanced design ($m_H = 2m_L$)

the empirical sampling variance of $\hat{\beta}$ for a binary covariate Z under balanced and unbalanced design when $\beta = (-1, 1)$ and $\lambda_2 = (2, 3, 4)$. As shown in Table 2.4, when examination times are unevenly spaced across the study, unbalanced design also can help to increase the precision of the parameter estimation.

TABLE 2.4: Empirical sampling variances of $\hat{\beta}$ under balanced and unbalanced design with unevenly spaced visits. Assuming $n_1 = n_2 = 50, \lambda_1 = 1, T = 1, m_L = 3$, and event time follows $W(0.8)$ distribution.

	$V(\hat{\beta})_1^*$	$V(\hat{\beta})_2^{**}$	$V(\hat{\beta})_2^{**}/V(\hat{\beta})_1^*$
$(\beta, \lambda_2) = (1, 2)$	0.0321	0.0302	0.94
$(\beta, \lambda_2) = (1, 3)$	0.0298	0.0274	0.92
$(\beta, \lambda_2) = (1, 4)$	0.0340	0.0287	0.84
$(\beta, \lambda_2) = (-1, 2)$	0.0400	0.0398	1.00
$(\beta, \lambda_2) = (-1, 3)$	0.0351	0.0343	0.98
$(\beta, \lambda_2) = (-1, 4)$	0.0354	0.0334	0.94

*Variance of $\hat{\beta}$ for balanced design ($m_H = m_L$)
**Variance of $\hat{\beta}$ for unbalanced design ($m_H = 2m_L$)

2.3 Power and Type I Error Estimation

An important application for above results is power estimation. Under $H_0 : \beta = 0$, under $H_A : \beta = \beta_A$. For a two-sided test with type I error α, power can be estimated by

$$\begin{aligned} &1 - P\left(-z_{\alpha/2} - \hat{\beta}_A/\sqrt{V(\hat{\beta}_A)} < Z < z_{\alpha/2} - \hat{\beta}_A/\sqrt{V(\hat{\beta}_A)} | H_A\right) \\ &= 1 - \Phi\left(z_{\alpha/2} - \hat{\beta}_A/\sqrt{V(\hat{\beta}_A)}\right) + \Phi\left(-z_{\alpha/2} - \hat{\beta}_A/\sqrt{V(\hat{\beta}_A)}\right) \end{aligned}$$

where $V(\hat{\beta}_A)$ is the estimated asymptotic variance of $\hat{\beta}_A$ using the formula derived in section 2.2.

For a hypothetic study with two groups parallel design, assuming total follow up time of 10 units, event times follow exponential distribution, baseline event rate in the low risk stratum $\lambda_1 = 0.3$, and baseline event rate in the high risk stratum $\lambda_2 = 0.6$, for a two-sided test at α level of 0.05, the asymptotic power and empirical power (from 1,000 simulations) for various sample sizes (N) and effect sizes (β) are shown in Table 2.5. For unbalanced design, I simply assume that the numbers of visits in the high risk group doubles the numbers of visits in the low risk group $(m_H = 2m_L)$. Due to the large discrepancy from theoretical results and numerical results when $m_L = 1$, this condition is not evaluated here.

TABLE 2.5: Theoretical power and empirical power at $\alpha = 0.05$ (2-sided) for balanced and unbalanced design based on exponential models. Assuming $T = 10$, $\lambda_1 = 0.3$, $\lambda_2 = 0.6, m_H = 2m_L$ for unbalanced design. The unit is percent.

m_L	Theoretical Power				Empirical Power			
	Balanced Design		Unbalanced Design		Balanced Design		Unbalanced Design	
	$\beta_A = 0.5$	$\beta_A = 0.6$	$\beta_A = 0.5$	$\beta_A = 0.6$	$\beta_A = 0.5$	$\beta_A = 0.6$	$\beta_A = 0.5$	$\beta_A = 0.6$
$N = 120$								
2	47	57	63	76	6	12	42	37
3	62	75	71	85	52	68	67	82
4	69	82	74	87	68	82	72	87
5	72	85	75	88	69	83	73	85
$N = 160$								
2	59	70	75	87	17	10	64	73
3	75	86	83	93	72	83	80	94
4	81	92	85	95	82	90	84	94
5	83	94	86	95	84	92	88	94

TABLE 2.6: Empirical Type I error rate at $\alpha = 0.05$ (2-sided) for balanced design and unbalanced design. Assuming $N = 200, T = 10$, $\lambda_1 = 0.3$, $\lambda_2 = 0.6, m_H = 2m_L$ for unbalanced design. The unit is percent.

m_L	Balanced Design	Unbalanced Design
Exp.		
2	4.6	4.3
3	5.3	4.6
4	4.9	4.9
5	5.1	5.0
$W(1.5)$		
2	6.4	2.4
3	5.3	4.0
4	5.4	5.2
5	5.7	5.3
$W(0.8)$		
2	5.2	5.1
3	4.7	4.5
4	5.4	5.6
5	5.2	5.4

Another important consideration is how this unbalanced design can influence the empirical Type I error rate compared to balanced design. Assuming total sample size $N = 200, T = 10, \lambda_1 = 0.3, \lambda_2 = 0.6$ and $m_H = 2m_L$, the empirical Type I error rates at $\alpha = 0.05$ from three parametric distributions: exponential, $W(1.5)$, and $W(0.8)$ are calculated from 1,000 simulation samples and the results are presented in Table 2.6. When comparing unbalanced design to balanced design, some improvement of empirical Type I error rate is observed when the number of visits are relatively small ($m_L = 2$ or 3) in each parametric model.

Chapter 3: Applications

In Chapter 2, both the theoretical and numerical studies results show that the unbalanced design with increased frequency of assessments in high risk stratum can help to reduce the sampling variance of covariate effect estimation, and thus to increase the power of the study for interval-censored data. In this chapter, the efficiency of unbalanced design vs. balanced design is compared using real data that were collected for a T1D study in order to further test the hypothesis in this dissertation.

3.1 Metabolic Control Study

Pancreatic islets are the regions of the pancreas that contain its hormone-producing cells. Among the different types of cells in the pancreas islets, β-cells are responsible for producing insulin, which has important effects on the metabolism of carbohydrates, fats and protein. The pancreatic β-cells are sensitive to the glucose concentration in the blood. When the glucose levels are high they secrete insulin into the blood; otherwise, when the glucose levels are low they stop producing insulin.

Among patients diagnosed with type 1 diabetes mellitus, the pancreatic β-cells are usually destroyed by an autoimmune process, and thus, insulin can no longer be synthesized or be secreted into the blood. Therefore, exogenous insulin is required for daily diabetes management among those diagnosed with T1D. However, at the clinical diagnosis of T1D, most patients still have residual pancreatic β-cells which can continue to secrete insulin for several additional years. Retention of β-cell function in patients with T1D has been associated with lower HbA1c levels and reductions in short-term and long-term complications (Steffes et al., 2003; The Diabetes Control

and Complications Trial Research Group, 1998). The Diabetes Control and Complications Trial (DCCT) also show that assignment to the intensive managed group reduced the risk for loss of C-peptide (a biomarker for β-cell function) by 57% over the mean 6.5 years of study (The Diabetes Control and Complications Trial Research Group, 1998).

Therefore, the DirecNet study group and Type 1 Diabetes TrialNet study group jointly embarked on a RCT study in 2009 to assess the effect of metabolic control at onset of diabetes on progression of T1D. The study protocol is listed on www.clinicaltrials.gov (NCT00891995). This study has been finished and the de-identified data are publicly available at http://direcnet.jaeb.org/Studies.aspx. The study was conducted at five clinical centers, and 71 participants were enrolled between May 2009 and October 2011. Major eligibility criteria included age 6 to <46 years, clinical diagnosis of type 1 diabetes and initiation of insulin therapy within the prior 7 days. Eligible participants were randomized to the intensive group or usual care group in a 2:1 ratio, stratified by clinical center and the presence of diabetic ketoacidosis.

Participants in the intensive-treatment group received hybrid closed-loop control using the Medtronic MiniMed system for 72-96 hours as inpatients followed by home use of sensor-augmented pump therapy. The Medtronic MiniMed system consists of a subcutaneous glucose sensor and insulin pump which communicate wirelessly with a bedside computer running a proportional-integral-derivative algorithm. The details of the HCLC treatment were described in previous publication (Diabetes Research in Children Network [DirecNet] and Type 1 Diabetes TrialNet Study Groups, 2013). Participants in the usual-care group received standard diabetes management as practiced at the participating centers.

Both treatment groups had a 90-min mixed-meal tolerance test (MMTT) at baseline, and then, 2-h MMTTs were performed at 2 and 6 weeks and at 3, 6, 9, 12, 18, and 24 months. MMTT is a method for stimulating C-peptide response. C-peptide is a byproduct when the pancreatic β-cells produce insulin. Measuring C-peptide can help to determine how much insulin a person can produce since C-peptide is secreted in equimolar amounts to insulin. C-peptide levels are measured

instead of insulin levels because by measuring C-peptide, we can assess a person's own insulin secretion even if the person receives insulin therapy, and because C-peptide is not metabolized by the liver which makes it a more stable measure of insulin secretion than insulin itself. Therefore, the measurement of C-peptide in response to a stimulus can provide a direct measure of the β-cell function in patients with T1D. During MMTT, a liquid meal (boost) is ingested by the patient in fasting state then C-peptide levels are measured over the subsequent 2-4 hours. In the 90-min abbreviate MMTT conducted in this study, C-peptide levels were measured at 0 and 90 minutes. In the 2-h MMTT conducted in this study, C-peptide levels were measured at -10 and/or 0 min, 15 min, 30 min, 60 min, 90 min, and 120 min.

The primary outcome of each participant in this study was area under the stimulated C-peptide curve from the 2 hour MMTT conducted at 12 month visit. The primary analysis results of this study were previously published with no significant difference found between the two treatment groups (Buckingham et al., 2013).

Among the secondary outcomes of this study, the incidences of the loss of the 2 hour peak C-peptide <0.2 pmol/mL were also assessed (data not published). Peak C-peptide is the maximum value of stimulated C-peptide levels measured over the 2 hour period. In the DCCT study (The Diabetes Control and Complications Trial Research Group, 1998), patients with stimulated peak C-peptide < 0.2 pmol/mL was defined as C-peptide nonresponders which is biomarker for loss of β-cell function. However, early works also used cutpoint of <0.3 pmol/mL to define insulin-requiring diabetes (Jones and Hattersley, 2013). Therefore, in this dissertation, when comparing the efficiency of unbalanced design vs. balanced design using data from this metabolic control study, both time until stimulated peak C-peptide < 0.2 pmol/mL and time until stimulated peak C-peptide <0.3 pmol/mL are used as study endpoints. Since MMTT was done repeatedly at each visit, there might be chances that stimulated peak C-peptide went below 0.2 pmol/mL then bounced back to above 0.2 pmol/mL at the next visit when MMTT was conducted again. In this case, only the time until first incidence of peak C-peptide < 0.2 pmol/mL or <0.3 pmol/mL is assessed.

In this metabolic control study, 68 out of 71 enrolled newly diagnosed T1D patients were tested as autoantibody positive (the factor that differentiates type 1 diabetes with type 2 diabetes) in the first year, only data from these 68 participants were reported in previous publications. In this dissertation, data from the same cohort are used.

3.2 Statistical Methods

The metabolic control study described above used balanced design where every subject had the same visit schedule. In order to create unbalanced design data out of this study, a baseline risk factor is needed. Previous study show that age at onset is a significant predictor for the time of disappearance of the β-cell function (Schiffrin et al., 1988), therefore, in this dissertation, two risk strata are created based on age at T1D onset. Stratum 1 is high risk stratum with age at onset <12 years (N=31); Stratum 2 is low risk stratum with age at onset $\geq$12 years (N=37).

For balanced design, it is assumed that every participant has stimulated C-peptide tested at 6, 12, and 24 months. Two types of unbalanced design are evaluated in this chapter: the first type is the unbalanced design that I proposed previously - for those age of onset <12 years, it is assumed that the participants have stimulated C-peptide tested at 3, 6, 9, 12, 18 and 24 months, otherwise, the participants have C-peptide tested at 6, 12, and 24 months; the second type is the reversed unbalanced design as mentioned in Chapter 2 – for those in the low risk group, assume the participants have stimulated C-peptide tested at 3, 6, 9, 12, 18 and 24 months, and for those in the high risk group, assume the participants have C-peptide tested at 6, 12, and 24 months. Since in the original study, the C-peptide data were collected at 2 and 6 weeks, 3, 6, 9, 12, 18, and 24 months after treatment started, the censor intervals can be manipulated based on original data. For example, if an event was observed at 12 month visit, based on the balanced design, the event is censored at (6, 12] months; based on the first unbalanced design, if the participant is in the high risk stratum, the event is censored at (9, 12] months, otherwise, the event is censored at (6, 12] months; for the reversed unbalanced design, the censor interval is reversed based on the risk factor.

To compare the efficiency of unbalanced designs vs. balanced design, both parametric methods and nonparametric methods are used in this chapter. Among the parametric methods, both exponential distribution and Weibull distribution of survival times are evaluated here. The main variable of interest for parameter estimation or comparing survival functions is the treatment group. Besides this, I also evaluate a few additional variables, including a binary variable Gender and a continuous variable HbA1c, in order to further testing the efficiency of the three types of designs. Stratum variable is adjusted as a covariate in the parametric regression models. Standard error (SE) for the covariates effect are compared among different models. For nonparametric methods, survival functions between the subgroups are compared within strata using the generalized log-rank statistics based on three different weight functions: Finkelstein (1986), Sun (1996) and Fay (1999). All the analyses are performed using SAS version 9.4 (SAS Institute, Cary, NC).

3.3 Results and Discussion

Among the 68 participants, 48 were assigned to the intensive treatment group and 20 were assigned to the usual-care group. Participants ranged in age from 7.8 to 45.7 years, with all but three <18 years old. Sixty-five percent were male and 92% were white.

The number of events for both stimulated peak C-peptide <0.2 pmol/mL and <0.3 pmol/mL and the number of censored subjects at the original observed time points are summarized in Table 3.1. Overall, 28 out of 68 (41%) ever had peak C-peptide drop below 0.2 pmol/mL and 39 out of 68 (57%) ever had peak C-peptide drop below 0.3 pmol/mL during the 24-month follow-up period. Among those without an event (either peak C-peptide dropping below 0.2 pmol/mL or below 0.3 pmol/mL), most participants were censored at the end of study (24 months).

TABLE 3.1: Summary of events for peak C-peptide <0.2 pmol/mL and <0.3 pmol/mL in the metabolic control study (N=68)

Months	Peak C-peptide <0.2 pmol/mL		Peak C-peptide <0.3 pmol/mL	
	Events	Censors	Events	Censors
3	1	0	1	0
6	2	0	7	0
9	7	0	4	0
12	5	2	10	1
18	10	0	11	0
24	3	38	6	28
Total	28	40	39	29

TABLE 3.2: Summary of events for peak C-peptide <0.2 pmol/mL by age groups

Months	Age 7-11 years (N=31)		Age 12 years and above (N=37)	
	Events	Censors	Events	Censors
3	0	0	1	0
6	2	0	0	0
9	5	0	2	0
12	4	1	1	1
18	4	0	6	0
24	2	13	1	25
Total	17	14	11	26

TABLE 3.3: Summary of events for peak C-peptide <0.3 pmol/mL by age groups

Months	Age 7-11 years (N=31)		Age 12 years and above (N=37)	
	Events	Censors	Events	Censors
3	0	0	1	0
6	7	0	0	0
9	1	0	3	0
12	6	0	4	1
18	7	0	4	0
24	2	8	4	20
Total	23	8	16	21

TABLE 3.4: Covariate effect estimation using parametric models under balanced and unbalanced design for peak C-peptide <0.2 pmol/mL event time

Effect	Model	Number of Tests	Estimate	SE	Chi-square
Treatment Intensive vs. Standard	Exp	$m_L = m_H = 3$	-0.3039	0.4666	0.42
		$m_H = 6, m_L = 3$	-0.3245	0.4658	0.49
		$m_H = 3, m_L = 6$	-0.2963	0.4663	0.40
	Weibull	$m_L = m_H = 3$	-0.2183	0.3045	0.51
		$m_H = 6, m_L = 3$	-0.2290	0.2978	0.59
		$m_H = 3, m_L = 6$	-0.2274	0.3309	0.47
Gender Female vs. Male	Exp	$m_L = m_H = 3$	0.1382	0.4061	0.12
		$m_H = 6, m_L = 3$	0.1472	0.4053	0.13
		$m_H = 3, m_L = 6$	0.1431	0.4060	0.12
	Weibull	$m_L = m_H = 3$	0.0735	0.2679	0.08
		$m_H = 6, m_L = 3$	0.0764	0.2619	0.09
		$m_H = 3, m_L = 6$	0.0893	0.2909	0.09
HbA1c at 3 Months	Exp	$m_L = m_H = 3$	-0.7339	0.3080	5.68
		$m_H = 6, m_L = 3$	-0.7257	0.3035	5.72
		$m_H = 3, m_L = 6$	-0.7540	0.3083	5.98
	Weibull	$m_L = m_H = 3$	-0.5099	0.2116	5.80
		$m_H = 6, m_L = 3$	-0.5099	0.2034	6.28
		$m_H = 3, m_L = 6$	-0.5631	0.2288	6.06

m_L: number of MMTT tests in the low risk stratum; m_H: number of MMTT tests in the high risk stratum

The number of events for stimulated peak C-peptide <0.2 pmol/mL and <0.3 pmol/mL by age strata are summarized in Table 3.2 and 3.3 respectively. Among those participants of age 7-11 years at diagnosis (Stratum 1, N=31), 17 (55%) had peak C-peptide <0.2 pmol/mL and 23 (74%) had peak C-peptide <0.3 pmol/mL during the 2-year follow-up period. Among those age 12 years and above at diagnosis (Stratum 2, N=37), 11 (30%) had peak C-peptide <0.2 pmol/mL and 16 (43%) had peak C-peptide <0.3 pmol/mL during the 2-year follow-up period. Apparently, the event rates in Stratum 1 are higher than those in Stratum 2, which agree with the findings in Schiffrin et al. (1988).

TABLE 3.5: Covariate effect estimation using parametric models under balanced and unbalanced design for peak C-peptide <0.3 pmol/mL event time

Effect	Model	Number of Tests	Estimate	SE	Chi-square
Treatment Intensive vs. Standard	Exp	$m_L = m_H = 3$	-0.2801	0.3865	0.53
		$m_H = 6, m_L = 3$	-0.2930	0.3851	0.58
		$m_H = 3, m_L = 6$	-0.2745	0.3861	0.51
	Weibull	$m_L = m_H = 3$	-0.2175	0.2839	0.59
		$m_H = 6, m_L = 3$	-0.2121	0.2454	0.75
		$m_H = 3, m_L = 6$	-0.2173	0.2889	0.57
Gender Female vs. Male	Exp	$m_L = m_H = 3$	0.5309	0.3688	2.07
		$m_H = 6, m_L = 3$	0.5364	0.3677	2.13
		$m_H = 3, m_L = 6$	0.5285	0.3687	2.06
	Weibull	$m_L = m_H = 3$	0.3978	0.2754	2.09
		$m_H = 6, m_L = 3$	0.3535	0.2392	2.18
		$m_H = 3, m_L = 6$	0.4032	0.2802	2.07
HbA1c at 3 Months	Exp	$m_L = m_H = 3$	-0.6896	0.2670	6.67
		$m_H = 6, m_L = 3$	-0.6374	0.2641	5.82
		$m_H = 3, m_L = 6$	-0.6900	0.2659	6.73
	Weibull	$m_L = m_H = 3$	-0.5487	0.1948	7.94
		$m_H = 6, m_L = 3$	-0.4783	0.1689	8.02
		$m_H = 3, m_L = 6$	-0.5562	0.1972	7.95

m_L: number of MMTT tests in the low risk stratum; m_H: number of MMTT tests in the high risk stratum

The results from parametric analyses of selected covariates effect under balanced design and unbalanced designs are shown in Table 3.4 and 3.5. In both tables, m_H represents the number of MMTT tests in the high risk stratum and m_L represents the number of MMTT tests in the low risk stratum. The rows with $m_H = m_L = 3$ represent the balanced design with all individuals having 3 MMTTs in 24 months; the rows with $m_H = 6, m_L = 3$ represents the unbalanced design with increased testing frequency in the high risk stratum; and the rows with $m_H = 3, m_L = 6$ represents the reversed unbalanced design with increased testing frequency in the low risk stratum. As shown in both tables, the unbalanced design with increased testing frequency in the high risk stratum generates the smallest standard error (SE) for the covariate effect among all three designs evaluated here. The reduction of SE tends to be greater when Weibull model is used rather than when exponential model is used, also tends to be greater when the study endpoint is failure time

for peak C-peptide drop below 0.3 pmol/mL compared to drop below 0.2 pmol/mL, which means that the event rate is relatively higher.

TABLE 3.6: Summary of results from nonparametric comparisons under balanced and unbalanced design for peak C-peptide <0.2 pmol/mL event time

Effect	Weight	Number of Tests	Generalized Log-rank	Variance	Chi-square
Treatment: Intensive vs. Standard	Finkelstein	$m_L = m_H = 3$	1.4602	5.2487	0.4063
		$m_H = 6, m_L = 3$	1.5636	5.4335	0.4499
		$m_H = 3, m_L = 6$	1.4094	5.2530	0.3782
	Sun	$m_L = m_H = 3$	1.0699	4.0980	0.2793
		$m_H = 6, m_L = 3$	1.3259	4.5622	0.3853
		$m_H = 3, m_L = 6$	1.0642	4.2101	0.2690
	Fay	$m_L = m_H = 3$	0.5400	3.2694	0.0892
		$m_H = 6, m_L = 3$	0.7012	3.3705	0.1459
		$m_H = 3, m_L = 6$	0.4572	3.2778	0.0638
Gender: Female vs. Male	Finkelstein	$m_L = m_H = 3$	-0.9806	6.6312	0.1450
		$m_H = 6, m_L = 3$	-1.2079	6.6637	0.2189
		$m_H = 3, m_L = 6$	-0.9479	6.6445	0.1352
	Sun	$m_L = m_H = 3$	-1.0779	5.1000	0.2278
		$m_H = 6, m_L = 3$	-1.1291	5.5895	0.2281
		$m_H = 3, m_L = 6$	-1.1178	5.2242	0.2392
	Fay	$m_L = m_H = 3$	-1.3300	4.0564	0.4361
		$m_H = 6, m_L = 3$	-1.6848	4.1258	0.6880
		$m_H = 3, m_L = 6$	-1.2751	4.0696	0.3995

m_L: number of MMTT tests in the low risk stratum; m_H: number of MMTT tests in the high risk stratum

The results from nonparametric comparison of survival curves between the subgroups of selected binary variables under balanced and unbalanced designs are summarized in Table 3.6 and 3.7 (same notations used for m_L and m_H as in Table 3.4 and 3.5). Since the theoretical results derived in Chapter 2 as well as all the simulation results are based on assumption for parametric models, and the hypothesis of this dissertation is that unbalanced design can help to improve the precision of the parameter estimation, therefore, the nonparametric analysis results in this Chapter are more exploratory rather than confirmatory in nature.

TABLE 3.7: Summary of results from nonparametric comparisons under balanced and unbalanced design for peak C-peptide <0.3 pmol/mL event time

Effect	Weight	Number of Tests	Generalized Log-rank	Variance	Chi-square
Treatment: Intensive vs. Standard	Finkelstein	$m_L = m_H = 3$	2.0777	7.7424	0.5576
		$m_H = 6, m_L = 3$	2.3213	7.9114	0.6811
		$m_H = 3, m_L = 6$	2.0624	7.7924	0.5458
	Sun	$m_L = m_H = 3$	1.6941	5.4724	0.5245
		$m_H = 6, m_L = 3$	1.8779	5.9237	0.5953
		$m_H = 3, m_L = 6$	1.7756	5.8976	0.5346
	Fay	$m_L = m_H = 3$	1.3379	3.9677	0.4511
		$m_H = 6, m_L = 3$	1.5314	4.0214	0.5832
		$m_H = 3, m_L = 6$	1.3109	4.0200	0.4275
Gender: Female vs. Male	Finkelstein	$m_L = m_H = 3$	-4.4883	9.0834	2.2178
		$m_H = 6, m_L = 3$	-4.4065	9.1948	2.1118
		$m_H = 3, m_L = 6$	-4.4730	9.1545	2.1855
	Sun	$m_L = m_H = 3$	-3.7779	6.4072	2.2276
		$m_H = 6, m_L = 3$	-3.8875	6.8806	2.1964
		$m_H = 3, m_L = 6$	-4.0499	6.8588	2.3913
	Fay	$m_L = m_H = 3$	-3.1024	4.6520	2.0689
		$m_H = 6, m_L = 3$	-3.1024	4.7046	2.0458
		$m_H = 3, m_L = 6$	-3.0754	4.7129	2.0068

m_L: number of MMTT tests in the low risk stratum; m_H: number of MMTT tests in the high risk stratum

From Table 3.6 and 3.7 we can see that results from three types of weight functions depend on particular variable and endpoint. When the endpoint is time until stimulated C-peptide drop below 0.3 pmol/mL, Finkelstein's method tends to give the highest score and Fay's method tend to give the lowest score for both variables evaluated (Table 3.7). However, when the endpoint is time until stimulated C-peptide drop below 0.2 pmol/mL, the direction is the same for Treatment variable, but is reversed for the Gender variable. When comparing balanced design vs. unbalanced designs, in Table 3.6, the unbalanced design with increased testing frequency in the high risk stratum tends to give the largest score in generalized log-rank statistics (in absolute value) among all three types of designs, and thus have the largest power. However, when comparing the reversed unbalanced design with increased testing frequency in the low risk stratum with the balanced design, it does not generate higher score for generalized log-rank statistic. In Table 3.7, the same trends are

observed for comparing treatment groups, however, when comparing female vs. male, the benefit of unbalanced design previously shown is not observed here. It is unclear why this happens for the peak C-peptide <0.3 pmol/mL end point but not for the peak C-peptide <0.2 pmol/mL end point. One possible explanation is the relatively small sample size and large variation. And more studies are warranted for the nonparametric comparisons of these different designs.

Chapter 4: Discussions

In longitudinal studies, especially those study endpoints can be influenced by the follow-up time and assessment frequency, how to decide and balance the number of follow-up visits/assessments, the total length of the follow-up, and the number of subjects recruited is always a difficult task for researchers. These choices can influence both the statistical power and the study cost.

As an important aspect of longitudinal study design, however, the sampling time of measurements, or *temporal design*, usually only receive the briefest attention in research reports, such as "measurements took yearly (or semi-annually) from 2010 to 2015". The scientific or theoretical reasons, as opposed to the logistic reasons, for the choice of temporal design are rarely mentioned. And the discussion of how this choice might have affected the study results are equally rare.

The temporal design of a longitudinal study includes two main factors. One is the duration of the total follow-up time. The study must extend long enough in duration to allow the effect of interest to occur. For example, in a clinical trial for investigation drugs, the study must follow enough duration for the drugs to take effect (if they are slow reaction drugs). In a longitudinal study with time to event outcome, we often need to follow enough duration in order to obtain adequate number of events for enough statistical power.

The second main factor of temporal design is the measurement interval, which is the time allowed to elapse between two consecutive measurements in a longitudinal study. There have been only a few studies which evaluated the measurement interval in different types of longitudinal studies. Gollub and Reichardt (1987) evaluated the influence of time lags when using latent longitudinal approach on casual modeling. They stated that: "...effect sizes can vary as a function of

the length of the time lag between a cause and the time for which its effect is assessed. That is, different time lags typically have different effect sizes." (p.82)

Other authors such as Cohen (1991) and Collins (1996) have pointed out that when changing phenomena are the object of a longitudinal study, the relation between two variables can be different with different measurement intervals.

The measurement interval in the temporal design is usually evaluated in terms of the planned number of measurements in each subject (like this dissertation). A few studies have been published on balancing the number of subjects (n) and number of measurements (m) in terms of power and total costs in longitudinal studies with continuous outcome. Galbraith and Marschner (2002) derive theoretical formula for the relationship between number of subjects and number of measurements, based on mixed-effects linear models, in terms of power and total study costs when the objective of the longitudinal study is to compare rates of changes in a continuous response variable between two groups. Based on their derived formula, they provide some practical guidelines on how to minimize the cost of the study by balancing n and m when both quantities are flexible.

Tekle, Tan, and Berger (2008) evaluate optimal design problems for logistic mixed effects models for binary longitudinal responses. Unlike the study of Galbraith and Marschner (2002) and this dissertation which assume evenly spaced measurements, this study try to optimize the timing of measurements when both number of measurements and total study duration are fixed, which is another aspect of the temporal design. There are also a wide range of literature on optimal designs for longitudinal studies with continuous responses.

Given that some studies have been published on the timing and spacing of measurements in longitudinal studies with continuous responses or even binary responses, the publication on the timing and spacing of measurement intervals in longitudinal study with survival outcome is lacking. One main reason is that most survival endpoints studied are known exactly, such as death, diagnosis of a disease based on symptoms, in this case, timing and spacing of repeated

measurements are not related to the study endpoint (although maybe related to other secondary outcomes). However, the timing and spacing of measurements issue can be related to those survival endpoints which are interval censored by two consecutive measurements.

Alexander (2008) is the first and the only person, to the author's knowledge, to systematically and theoretically assess how the temporal design (in terms of number of measurements in fixed study duration) can influence the precision of the risk estimator from interval-censored survival data. The most important reason for the lacking of publications on this temporal study design issue in interval-censored survival data is probably the under-development of methodology and lagged-behind application for the analysis methods. Currently, when analyzing interval-censored survival data, most statisticians still use the simple imputation based methods (e.g. right-point imputation or mid-point imputation) then apply standard methods for right-censored survival data. Although the likelihood based parametric methods for interval-censored survival data are easy to understand and implement, unlike other type of response variables, parametric methods are much less used when analyzing survival data since it is difficult to know which parametric distribution the survival function will follow in advance. On the other hand, despite of some development on the nonparametric comparison and regressions having been made in recent years, the application of such methods has been lagged far behind.

The publication from Alexander (2008) only evaluates the relationship between number of measurements in fixed study duration and the precision of event rate estimation from interval-censored data, this dissertation extends that research by evaluating the influence of changing number of measurements on precision of the covariate effect estimation from interval-censored survival data, and further proposes a new unbalanced design in which the number of measurements in each individual varies according to certain risk factor(s) measured from the study. To the author's knowledge, this is the first study which proposes and systematically evaluates this type of design when collecting interval-censored survival data. As a pioneer work in this particular field, I try to keep the theoretical derivation and numerical studies within parametric scope.

Although the parametric results are within expectation and prove the hypothesis of this dissertation, the nonparametric analyses in Chapter 3 show some contradictory results. Since the nonparametric methods are not within the main scope of this study, it is unclear why the patterns are quite different from parametric results. Obviously, more studies are needed in this field.

When applying the unbalanced design proposed by this dissertation, one important issue need be emphasized here. Since one of major assumptions for current survival analysis methods is that the censor time is not related to the data (non-informative censoring), using of unbalanced design will result in violation of this important assumption since the censor intervals depend on observed data. In order to correct this, the stratum variable which is used to determine the measurement schedule will need to be adjusted in the regression model. This can be easily achieved in parametric regression models by adding a covariate for stratum, and the nonparametric comparison can be performed within stratum.

One limitation of this study is that losses to follow-up and accrual times are not considered. For the simplicity of formula derivation, I assume every subject has the same fixed total follow-up time. But this is seldom true for many studies. Another limitation is that I only consider a risk factor measured at baseline which is used to determine the assessments frequency, however, risk factor may changes during the follow-up, therefore, this stratum variable can also be allowed to change when designing the study. In these kind of complicated conditions, developing a theoretical formula is not feasible and we need to rely on extensive simulations to estimate the sampling error and power under unbalanced design.

The unbalanced design evaluated in this dissertation can be applied in both longitudinal observational studies and clinical trials. Based on the results, this design can help to improve the efficiency and power of the study. In another perspective, adopting this type of design can help to reduce the number of subjects needed at fixed power. This is particularly valuable for some RCTs which involve tremendous cost on the treatment of each subject.

Chapter 5: Concluding Remarks and Future Research

5.1 Summary and Conclusions

In Chapter 1 of this dissertation, I first introduced the definition of "balanced design" vs. "unbalanced design" in longitudinal studies. In this chapter, the longitudinal study design in which all the participants have the same visit/assessment schedule is defined as "balanced design"; the opposite is "unbalanced design", in which different participants have different visit/assessment schedule depending on certain data collected from the study. In this chapter, I provided some examples for different types of balanced design and unbalanced design.

Next, I provided a brief review on the theoretical background for the interval-censored time to event data and associated analysis methods. Last, I reviewed the current limited literatures which study the influence of study design on parameter estimation from interval-censored time to event time data, then based on these findings, I proposed an unbalanced design with only increasing the frequency of assessments in the high risk group based certain baseline risk factor(s) when collecting interval-censored time to event data.

In Chapter 2, first, I provided theoretical proof in parametric scope (assuming the actual event time follows exponential distribution) that the unbalanced design I proposed has better efficiency than the common balanced design in terms of parameter estimation. In the second part, the results from numerical studies under different parameter settings were presented.

As previous literature show that the sampling error of the event rate could be reduced by increasing the number of assessments in fixed study period when collecting interval-censored time

to event data (Alexander, 2008), the results from Chapter 2 show similar trends when estimating the sampling variance of covariate effect on event rate. This study also show that the reduction of sampling variance can be mostly achieved by simply increasing the number of assessments in the high risk stratum based on a baseline risk factor (unbalanced design) without having to increase the number of assessments in the whole study cohort. I also show that this unbalanced design has better efficiency than a "reversed unbalanced design" in which the number of assessments in the low risk stratum rather than the high risk stratum are increased.

In Chapter 3, I applied this unbalanced design into the data which were collected for a new onset T1D metabolic control study by DirecNet and TrialNet study groups, and compared the efficiency of the balanced design with two types of unbalanced designs (one with increased testing frequency in the high risk stratum, and the other in reversed direction with increased testing frequency in the low risk stratum) on two study endpoints: the failure time of stimulated peak C-peptide drop below 0.2 pmol/mL and below 0.3 pmol/mL in the first time as measured by lab data. In the unbalanced designs, age at T1D diagnosis was used to separate the study cohort into two risk strata. In this Chapter, when using the parametric analysis methods for analyzing selected covariate effects, the results show that generally, the unbalanced design with increased testing frequency in the high risk stratum gives the smallest SE of the covariate effect estimate ($\hat{\beta}$) and largest power (as reflected by Chi-squares in Table 3.4 and 3.5) among all three designs evaluated. The results from nonparametric comparison of the survival curves were also presented for exploratory purpose, since the generalized log-rank statistics calculated from the nonparametric methods are in totally different scope with the parametric parameter estimation in Chapter 2. The results show that mostly the calculated generalized log-rank statistics from the unbalanced design with increased testing frequency in the high risk stratum is higher than the generalized log-rank statistics calculated from the balanced design, however, in some occasions, some minor reduction in the calculated generalized log-rank statistics from this unbalanced design was also observed. Meanwhile, it was also observed that the reversed unbalanced design with increased testing frequency in the low risk stratum tends to generate generalized log-rank statistics even lower than

that from the balanced design quite often. The reason for this behavior is unclear and need further investigation.

In Chapter 4, first some discussions on the current status of researches related to temporal design of longitudinal studies are provided. Second, the contributions and limitations of this particular research are discussed.

5.2 Future Research

This dissertation mainly points to a new direction when designing longitudinal studies with interval-censored time to event outcomes, more studies are needed in order to apply this type of design when collecting longitudinal data. 1) As I discussed before, more studies need be done to evaluate the efficiency of this unbalanced design in the nonparametric scope. 2) Since the sample size of the metabolic control study that I use in Chapter 3 is relatively small, it is better to able to evaluate this proposed design in larger samples. 3) In future research studies on this topic, some factors not considered in this dissertation can be incorporated, such as accrual time, early drop-out, and time-dependent risk factor for defining strata. 4) Since increasing the testing frequency in the high risk group still lead to increased total cost, another direction for future researches can be that, while keeping the total number of testing for all subjects constant, evaluating the efficiency of an unbalanced design with decreased testing frequency in the low risk stratum and increased testing frequency in the high risk stratum at the same time. 5) Study on either optimizing the power by adjusting different assessment schedules based on risk group at fixed cost or optimize the cost by adjusting different assessment schedules based on risk group at fixed power.

Bibliography

Alexander, N. (2008). "Precision of rate estimation under uniform interval censoring". In: *Statistics in Medicine* 27, pp. 3442–3445.

Betensky, R. A. et al. (2002). "A local likelihood proportional hazards model for interval censored data". In: *Statistics in Medicine* 21, 263–275.

Buckingham, B. et al. (2013). "Effectiveness of early intensive therapy on β-cell preservation in type 1 diabetes". In: *Diabetes Care* 36, pp. 4030–4035.

ClinicalTrials.gov (2012). *HIV Vaccine trial in Thai adults*. URL: https://clinicaltrials.gov/ct2/show/NCT00223080.

– (2016). *Hormone therapy with or without Everolimus in treating patients with breast cancer*. URL: https://clinicaltrials.gov/show/NCT01674140.

Cohen, P. (1991). "A source of bias in longitudinal investigations of change". In: *Best Methods for the Analysis of Change*. Ed. by L. M. Collins and J. L. Horn. Washington, DC: American Psychological Association, pp. 18–25.

Collins, L. M. (1996). "Measurement of change in research on aging: old and new issues from an individual growth perspective". In: *Handbook of the Psychology of Aging*. Ed. by J. E. Birren and K. W. Schaie. San Diego, CA: Academic Press, pp. 38–56.

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

Diabetes Research in Children Network [DirecNet] and Type 1 Diabetes TrialNet Study Groups (2013). "The effects of inpatient hybrid closed-loop therapy initiated within 1 week of type 1 diabetes diagnosis". In: *Diabetes Technology & Therapeutics* 15, pp. 401–408.

Diabetic Retinopathy Clinical Research Network [DRCR.net] (2015). "Panretinal photocoagulation vs intravitreous ranibizumab for proliferative diabetic retinopathy: a randomized clinical trial". In: *JAMA* 314, pp. 2137–2146.

DRCR.net (2010). "Randomized trial evaluating ranibizumab plus prompt or deferred laser or triamcinolone plus prompt laser for diabetic macular edema". In: *Ophthalmology* 117, pp. 1064–1077.

Dunedin Multidisciplinary Health & Development Research Unit (n.d.). *The Dunedin multidisciplinary health and development study*. URL: http://dunedinstudy.otago.ac.nz/studies.

Fay, M. P. (1996). "Rank invariant tests for interval censored data under the grouped continuous model". In: *Biometics* 52, pp. 811–822.

– (1999). "Comparing several score tests for interval censored data". In: *Statistics in Medicine* 18, pp. 273–285.

Finkelstein, D. M. (1986). "A proportional hazards model for interval-censored failure time data". In: *Biometrics* 42, pp. 845–854.

Framingham Heart Study (n.d.). *History of the Framingham heart study*. URL: https://www.framinghamheartstudy.org.

Galbraith, S. and I. C. Marschner (2002). "Guidelines for the design of clinical trials with longitudinal outcomes". In: *Controlled Clinical Trials* 23, pp. 257–273.

Glenn, H. (2011). "Proportional hazards regression with interval censored data using an inverse probability weight". In: *Lifetime Data Analysis* 17, pp. 373–385.

Gollub, H. F. and C. S. Reichardt (1987). “Taking account of time lags in causal models”.
In: *Child Development* 58, pp. 80–92.

Groeneboom, P. and J. A. Wellner (1992).
Information bounds and nonparametric maximum likelihood estimation.
New York: Birkhauser.

Jones, A. G. and A. T. Hattersley (2013).
“The clinical utility of C-peptide measurement in the care of patients with diabetes”.
In: *Diabetic Medicine* 30, pp. 803–817.

Kaplan, E. L. and P. Meier (1958). “Nonparametric estimation from incomplete observations”.
In: *Journal of the American Statistical Society*. A 53, pp. 457–481.

Klein, J. P. and M. L. Moeschberger (2003).
Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer-Verlag.

Peto, R. (1973). “Experimental survival curves for interval-censored data”.
In: *Applied Statistics* 22, pp. 86–91.

Schiffrin, A. et al. (1988).
“Prospective study of predictors of beta-cell survival in type 1 diabetes”.
In: *Diabetes* 37, pp. 920–925.

Steffes, M. W. et al. (2003). “β-cell function and the development of diabetes-related
complications in the diabetes control and complications trial”.
In: *Diabetes Care* 26, pp. 832–836.

Sun, J. (1996).
“A nonparametric test for interval-censored failure time data with application to AIDS studies”.
In: *Statistics in Medicine* 15, pp. 1387–1395.

Sun, J., Y. Feng, and H. Zhao (2015). “Simple estimation procedures for regression analysis of
interval-censored failure time data under the proportional hazards model”.
In: *Lifetime Data Analysis* 21, pp. 138–155.

Sun, X. and C. Chen (2010). "Comparison of Finkelstein's method with the conventional approach for interval-censored data analysis".
In: *Statistics in Biopharmaceutical Research* 2, pp. 97–108.

Tang, S., C. Holland, and R. Sridhara (2008).
"Consequences of asymmetry in progression assessments".
In: *Proceedings of American Statistics Association*.

Tanner, M. A. (1991).
Tools for Statistical Inference: Observed Data and Data Augmentation Methods.
New York: Springer-Verlag.

Tanner, M. A. and W. H. Wong (1987).
"The application of imputation to an estimation problem in grouped lifetime analysis".
In: *Technometrics* 29, pp. 23–32.

Tekle, F. B., F. E. S. Tan, and M. P. F. Berger (2008).
"Maximin D-optimal designs for binary longitudinal responses".
In: *Computational Statistics and Data Analysis* 52, pp. 5253–5262.

The Diabetes Control and Complications Trial Research Group (1998).
"Effect of intensive therapy on residual β-cell function in patients with type 1 diabetes in the diabetes control and complications trial. A randomized, controlled trial".
In: *Annals of Internal Medicine* 128, pp. 517–523.

Turnbull, B. W. (1976).
"The empirical distribution function with arbitrarily grouped, censored and truncated data".
In: *Journal of the Royal Statistical Society*. B 38, pp. 290–295.

Type 1 Diabetes TrialNet (2009).
TN-01 natural history – protocol synopsis & specimen collection schedule.
URL: https://www.diabetestrialnet.org/documents/ancillary/6aNaturalHistoryProtocolSynopsisSpecimenCollectionSchedule.pdf.

Type 1 Diabetes TrialNet (2011).

TN-01 natural history – protocol synopsis & assessment schedule.

URL: https://www.diabetestrialnet.org/documents/biobank/6NaturalHistoryProtocolSynopsisAssessmentSchedule.pdf.

Wellner, J. A. and Y. Zhan (1997). "A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data".

In: *Journal of the American Statistical Association* 92, pp. 945–959.

Winkels, R. M. et al. (2014).

"The COLON study: Colorectal cancer: Longitudinal, Observation study on Nutritional and lifestyle factors that may influence colorectal tumour recurrence, survial and quality of life".

In: *BMC Cancer* 14, p. 374.

Appendix A: Supplemental Figures

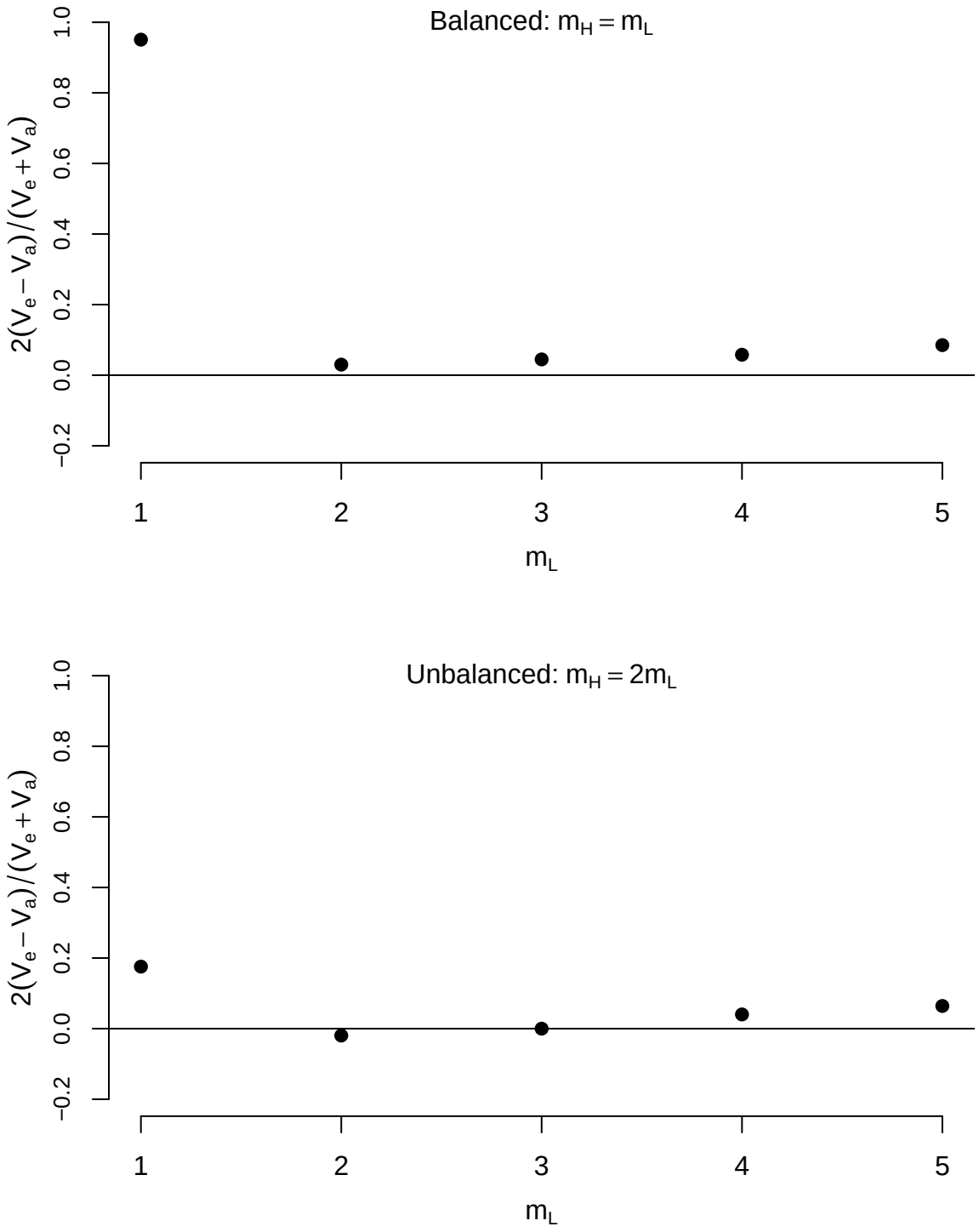

FIGURE A.1: Relative discrepancy between asymptotic and empirical variances. Assuming $\lambda_1 = 1, \lambda_2 = 2, \beta = 1, n_1 = n_2 = 50, T = 1$.

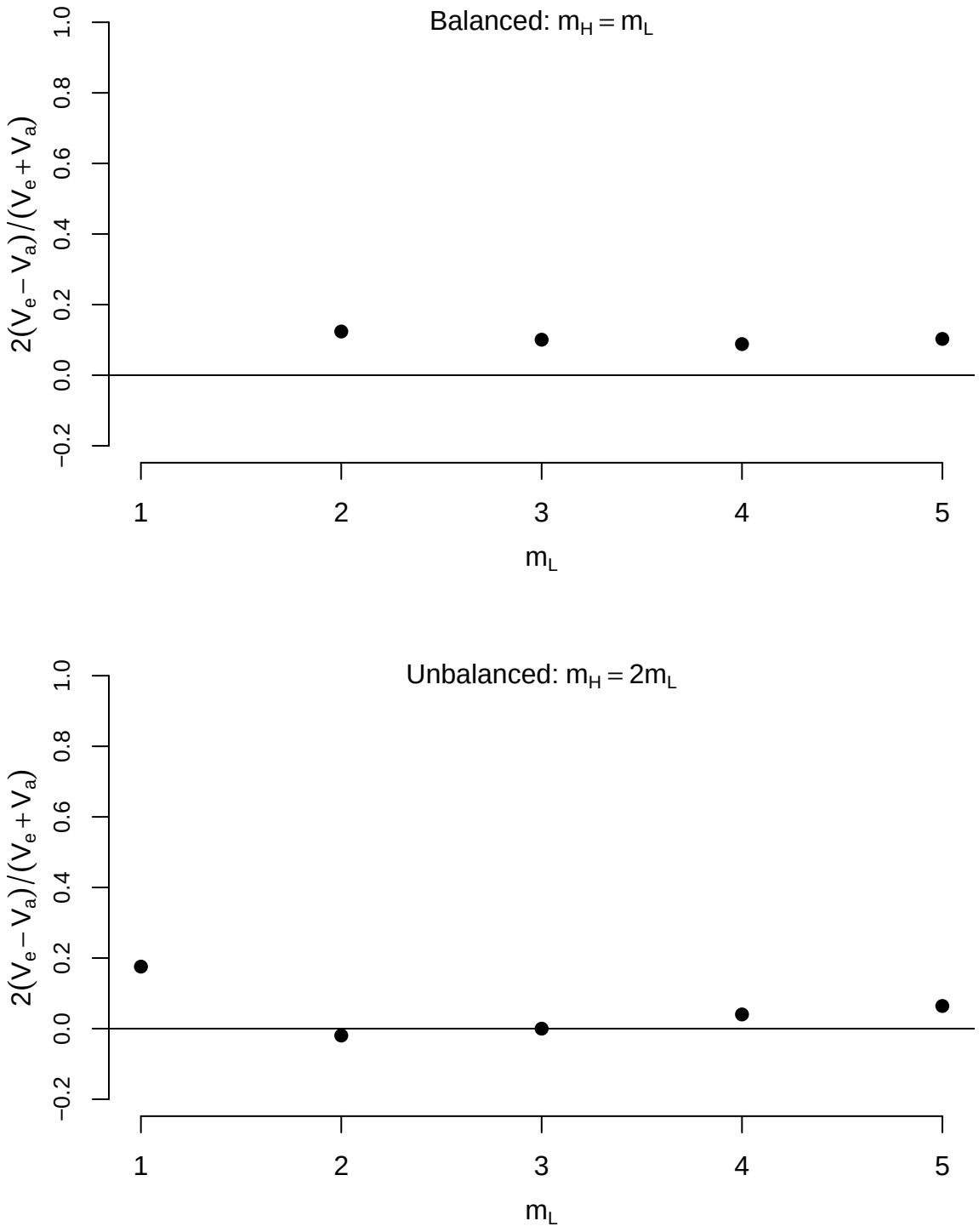

FIGURE A.2: Relative discrepancy between asymptotic and empirical variances. Assuming $\lambda_1 = 1, \lambda_2 = 3, \beta = 1, n_1 = n_2 = 50, T = 1$.

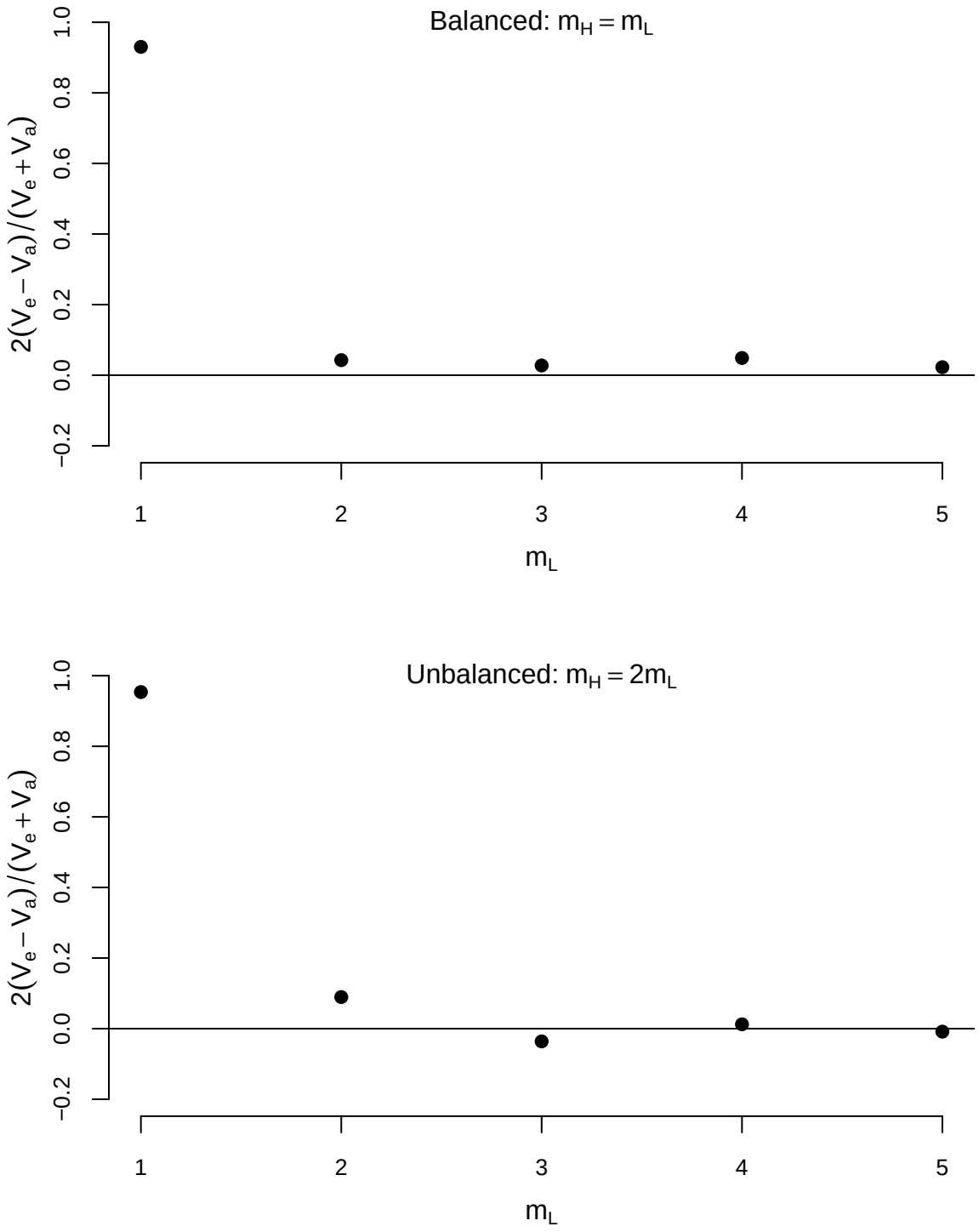

FIGURE A.3: Relative discrepancy between asymptotic and empirical variances. Assuming $\lambda_1 = 1, \lambda_2 = 4, \beta = 1, n_1 = n_2 = 50, T = 1$.

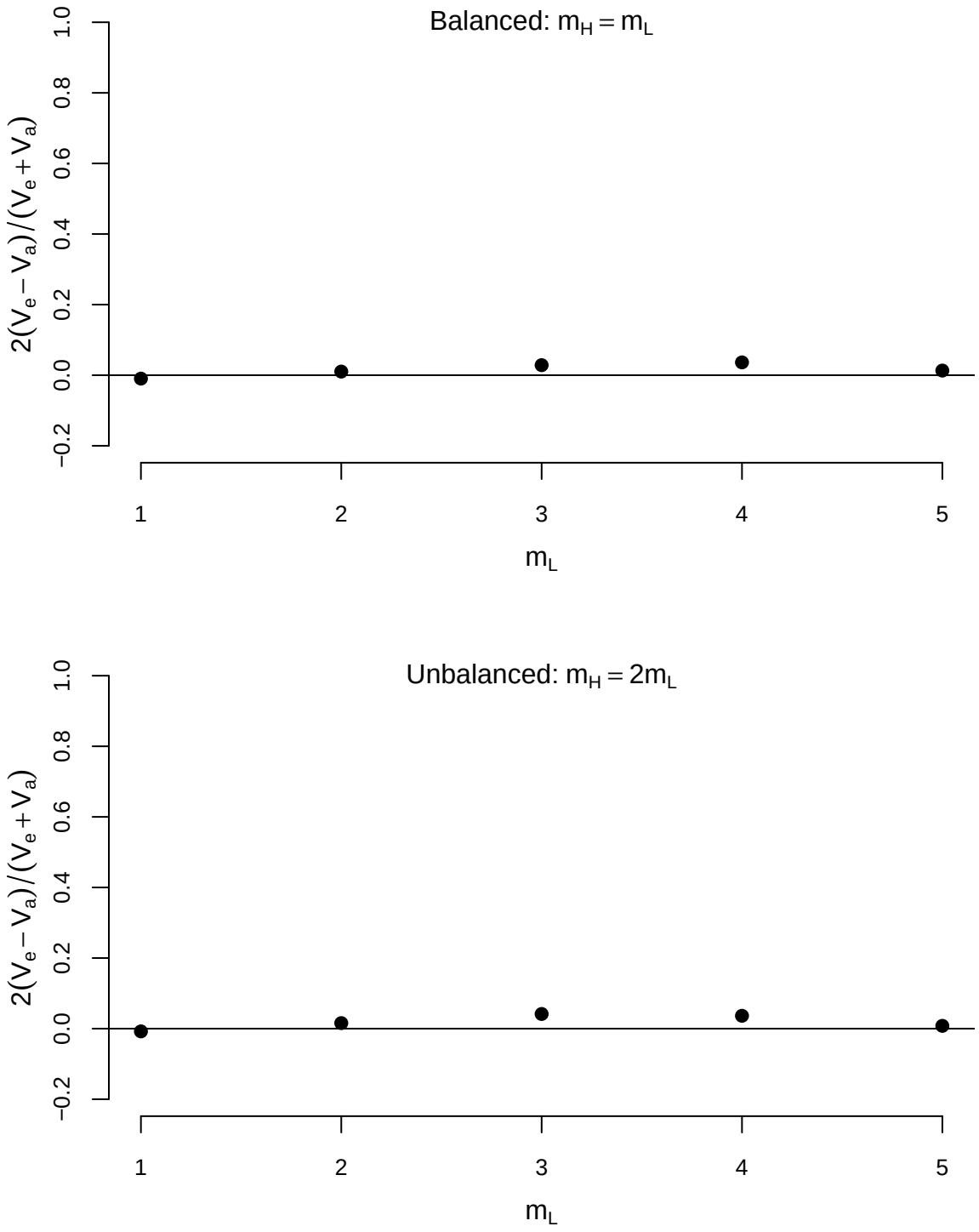

FIGURE A.4: Relative discrepancy between asymptotic and empirical variances. Assuming $\lambda_1 = 1, \lambda_2 = 2, \beta = -1, n_1 = n_2 = 50, T = 1$.

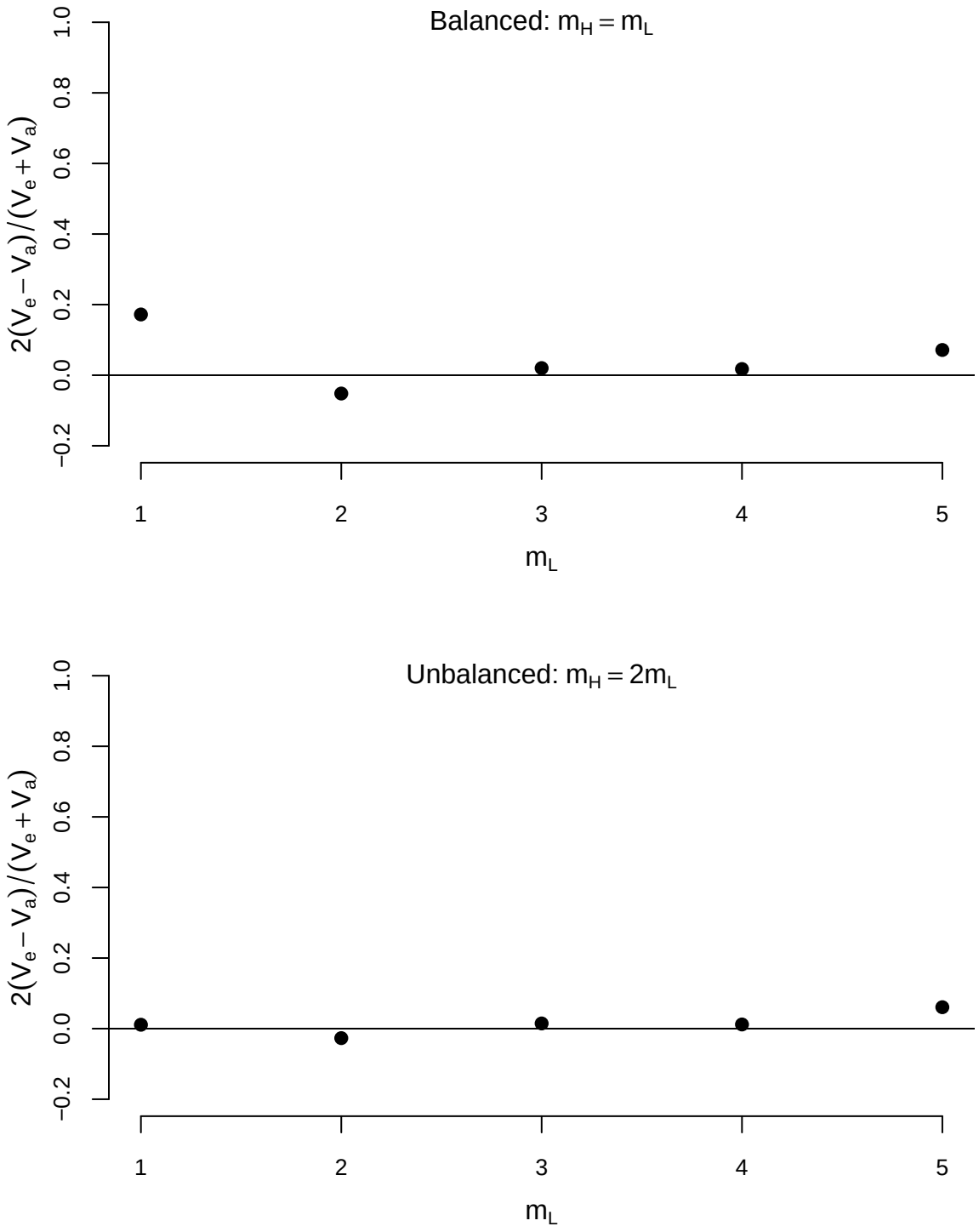

FIGURE A.5: Relative discrepancy between asymptotic and empirical variances. Assuming $\lambda_1 = 1, \lambda_2 = 3, \beta = -1, n_1 = n_2 = 50, T = 1$.

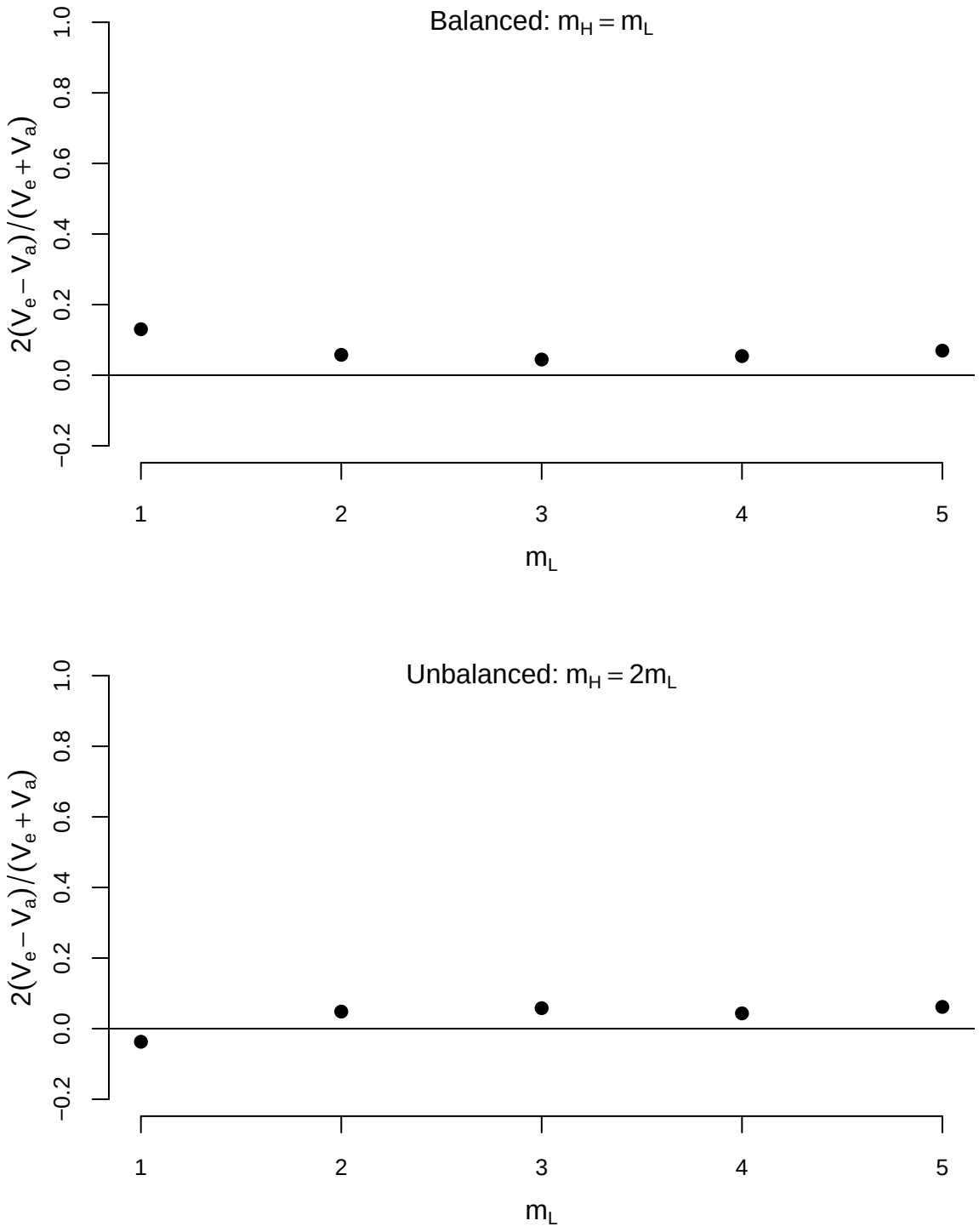

FIGURE A.6: Relative discrepancy between asymptotic and empirical variances. Assuming $\lambda_1 = 1, \lambda_2 = 4, \beta = -1, n_1 = n_2 = 50, T = 1$.

Appendix B: Selected R Codes for Chapter 3

```
#############################################################
# Table 2.1 Asymptotic variance (Exponential distribution) #
#############################################################

rm(list=ls(all=T))

#calculate variance ratio of inbalanced design:balanced design
n1<-50
n2<-50
beta<--1  #change between 1 and -1
T<-1
lambda1<-1
lambda2<-5  #change among 2,3,4
#create an empty vector for variances and ratio
v1<-rep(NA,5)
v2<-rep(NA,5)
v3<-rep(NA,5)
ratio2.1<-rep(NA,5)
ratio3.1<-rep(NA,5)


for (m in 1:5){
#variance from balanced design
c<-1
temp1<-n1*lambda1*T*exp(beta)*exp(-lambda1*T*exp(beta))
temp2<-n2*lambda2*T*exp(beta)*exp(-lambda2*T*exp(beta))

temp3<-n1*lambda1*T/m*exp(beta)

temp4<-1-m*exp(-lambda1*T*(m-1)/m*exp(beta))+(m-1)*exp(-lambda1*T
    ↪ *exp(beta))
temp5<-exp(lambda1*T/m*exp(beta))-1

temp6<-(1-lambda1*T/m*exp(beta)-exp(-lambda1*T/m*exp(beta)))*(1-
    ↪ exp(-lambda1*T*exp(beta)))
```

```
temp7<-(exp(lambda1*T/m*exp(beta))-1)*(1-exp(-lambda1*T/m*exp(
    ↪ beta)))

temp8<-n2*lambda2*T/(c*m)*exp(beta)

temp9<-1-c*m*exp(-lambda2*T*(c*m-1)/(c*m)*exp(beta))+(c*m-1)*exp
    ↪ (-lambda2*T*exp(beta))
temp10<-exp(lambda2*T/(c*m)*exp(beta))-1

temp11<-(1-lambda2*T/(c*m)*exp(beta)-exp(-lambda2*T/(c*m)*exp(
    ↪ beta)))*(1-exp(-lambda2*T*exp(beta)))
temp12<-(exp(lambda2*T/(c*m)*exp(beta))-1)*(1-exp(-lambda2*T/(c*m
    ↪ )*exp(beta)))

#fisher information matrix
I.11<-temp1+temp2+temp3*temp4/temp5-temp3*temp6/temp7+temp8*temp9
    ↪ /temp10-temp8*temp11/temp12
I.12<-temp1/lambda1+(temp3/lambda1)*(temp4/temp5)-(temp3/lambda1)
    ↪ *(1-exp(-lambda1*T*exp(beta)))/temp5+temp3*T*exp(beta)/m*
    ↪ ((1-exp(-lambda1*T*exp(beta)))/temp7)
I.13<-temp2/lambda2+(temp8/lambda2)*(temp9/temp10)-(temp8/lambda2
    ↪ )*(1-exp(-lambda2*T*exp(beta)))/temp10+temp8*T*exp(beta)/(c
    ↪ *m)*((1-exp(-lambda2*T*exp(beta)))/temp12)
I.21<-I.12
I.22<-n1*T^2*(1-exp(-lambda1*T))/(m^2*(exp(lambda1*T/m)-1)*(1-exp
    ↪ (-lambda1*T/m)))+n1*T^2*exp(2*beta)*(1-exp(-lambda1*T*exp(
    ↪ beta)))/(m^2*temp7)
I.23<-0
I.31<-I.13
I.32<-I.23
I.33<-n2*T^2*(1-exp(-lambda2*T))/((c*m)^2*(exp(lambda2*T/(c*m))
    ↪ -1)*(1-exp(-lambda2*T/(c*m))))+n2*T^2*exp(2*beta)*(1-exp(-
    ↪ lambda2*T*exp(beta)))/((c*m)^2*temp12)

I<-matrix(c(I.11,I.12,I.13,I.21,I.22,I.23,I.31,I.32,I.33),3,3)
I.inv<-solve(I)
v1[m]<-I.inv[1,1]

#for unbalanced design 2:1
c<-2
temp1<-n1*lambda1*T*exp(beta)*exp(-lambda1*T*exp(beta))
temp2<-n2*lambda2*T*exp(beta)*exp(-lambda2*T*exp(beta))

temp3<-n1*lambda1*T/m*exp(beta)
```

```
temp4<-1-m*exp(-lambda1*T*(m-1)/m*exp(beta))+(m-1)*exp(-lambda1*T
    ↪ *exp(beta))
temp5<-exp(lambda1*T/m*exp(beta))-1

temp6<-(1-lambda1*T/m*exp(beta)-exp(-lambda1*T/m*exp(beta)))*(1-
    ↪ exp(-lambda1*T*exp(beta)))
temp7<-(exp(lambda1*T/m*exp(beta))-1)*(1-exp(-lambda1*T/m*exp(
    ↪ beta)))

temp8<-n2*lambda2*T/(c*m)*exp(beta)

temp9<-1-c*m*exp(-lambda2*T*(c*m-1)/(c*m)*exp(beta))+(c*m-1)*exp
    ↪ (-lambda2*T*exp(beta))
temp10<-exp(lambda2*T/(c*m)*exp(beta))-1

temp11<-(1-lambda2*T/(c*m)*exp(beta)-exp(-lambda2*T/(c*m)*exp(
    ↪ beta)))*(1-exp(-lambda2*T*exp(beta)))
temp12<-(exp(lambda2*T/(c*m)*exp(beta))-1)*(1-exp(-lambda2*T/(c*m
    ↪ )*exp(beta)))

#fisher information matrix
I.11<-temp1+temp2+temp3*temp4/temp5-temp3*temp6/temp7+temp8*temp9
    ↪ /temp10-temp8*temp11/temp12
I.12<-temp1/lambda1+(temp3/lambda1)*(temp4/temp5)-(temp3/lambda1)
    ↪ *(1-exp(-lambda1*T*exp(beta)))/temp5+temp3*T*exp(beta)/m*
    ↪ ((1-exp(-lambda1*T*exp(beta)))/temp7)
I.13<-temp2/lambda2+(temp8/lambda2)*(temp9/temp10)-(temp8/lambda2
    ↪ )*(1-exp(-lambda2*T*exp(beta)))/temp10+temp8*T*exp(beta)/(c
    ↪ *m)*((1-exp(-lambda2*T*exp(beta)))/temp12)
I.21<-I.12
I.22<-n1*T^2*(1-exp(-lambda1*T))/(m^2*(exp(lambda1*T/m)-1)*(1-exp
    ↪ (-lambda1*T/m)))+n1*T^2*exp(2*beta)*(1-exp(-lambda1*T*exp(
    ↪ beta)))/(m^2*temp7)
I.23<-0
I.31<-I.13
I.32<-I.23
I.33<-n2*T^2*(1-exp(-lambda2*T))/((c*m)^2*(exp(lambda2*T/(c*m))
    ↪ -1)*(1-exp(-lambda2*T/(c*m))))+n2*T^2*exp(2*beta)*(1-exp(-
    ↪ lambda2*T*exp(beta)))/((c*m)^2*temp12)

I<-matrix(c(I.11,I.12,I.13,I.21,I.22,I.23,I.31,I.32,I.33),3,3)
I.inv<-solve(I)
v2[m]<-I.inv[1,1]

ratio2.1[m]<-v2[m]/v1[m]
```

```
}

v1
v2
ratio2.1


##############################################
# Table 2.1 Empirical Results                #
##############################################
rm(list=ls())
#log likelihood function for unbalance design, use -logL for
   ↪ nlminb
likelihood1<-function(par){
    beta1<-par[1]
    beta2<-par[2]
    beta3<-par[3]
    temp1<-exp(-L1[1:100]*exp(beta1+beta3*Z[1:100]))
    temp2<-exp(-R1[1:100]*exp(beta1+beta3*Z[1:100]))
    temp3<-exp(-L1[101:200]*exp(beta2+beta3*Z[101:200]))
    temp4<-exp(-R1[101:200]*exp(beta2+beta3*Z[101:200]))

   -sum(log(temp1 - temp2))-sum(log(temp3 - temp4))
}

#log likelihood function for balance design, use -logL for nlminb
likelihood2<-function(par){
    beta1<-par[1]
    beta2<-par[2]
    beta3<-par[3]
    temp1<-exp(-L2[1:100]*exp(beta1+beta3*Z[1:100]))
    temp2<-exp(-R2[1:100]*exp(beta1+beta3*Z[1:100]))
    temp3<-exp(-L2[101:200]*exp(beta2+beta3*Z[101:200]))
    temp4<-exp(-R2[101:200]*exp(beta2+beta3*Z[101:200]))

   -sum(log(temp1 - temp2))-sum(log(temp3 - temp4))
}


#simulation
n1<-50
n2<-50
beta<--1 #change between 1 and -1
lambda1<-1
lambda2<-5 #change among 2,3,4
```

```
c<-2
m<-seq(1:5)
v1<-rep(NA,5)
v2<-rep(NA,5)

#variable for group
group1.1<-rep(0,n1)
group1.2<-rep(1,n1)
group2.1<-rep(0,n2)
group2.2<-rep(1,n2)
Z<-c(group1.1,group1.2,group2.1,group2.2)

#create an empty vector for ratio
ratio<-rep(NA,5)
for (m in 1:5){

result.u<-rep(NA,1000)
result.b<-rep(NA,1000)

for (sim in 1:1000) {
   #create exponential distribution with different rates
   exp1.1<-rexp(n1, rate=lambda1)
   exp1.2<-rexp(n1, rate=lambda1*exp(beta))
   exp2.1<-rexp(n2, rate=lambda2)
   exp2.2<-rexp(n2, rate=lambda2*exp(beta))
   t<-c(exp1.1,exp1.2,exp2.1,exp2.2)

   #create empty vector for left and right bound of the visit
      ↪ window
   L1<-rep(NA, 2*n1+2*n2)
   R1<-rep(NA, 2*n1+2*n2)
   L2<-rep(NA, 2*n1+2*n2)
   R2<-rep(NA, 2*n1+2*n2)

   #for strata 1 with low risk, let L1=L2 R1=R2
   for (i in 1:100){
      if (t[i]>1) {
         L1[i]<-1
         R1[i]<-Inf
         L2[i]<-1
         R2[i]<-Inf }
     else {for (j in 1:m) {
         if (t[i]>(j-1)/m & t[i]<= j/m) {
         L1[i]<-(j-1)/m
         R1[i]<-j/m
```

```
            L2[i]<-(j-1)/m
            R2[i]<-j/m }}}
      } #close for (i in 1:100)

    #for strata 2 with high risk, use c*m visits for L1 & R1, use
       ↪ m visits for L2 & R2
    for (i in 101:200){
        if (t[i]>1) {
            L1[i]<-1
            R1[i]<-Inf
            L2[i]<-1
            R2[i]<-Inf }
      else {
            for (j in 1:(c*m)) {
              if (t[i]>(j-1)/(c*m) & t[i]<= j/(c*m)) {
              L1[i]<-(j-1)/(c*m)
              R1[i]<-j/(c*m)}}
            for (j in 1:m) {
              if (t[i]>(j-1)/m & t[i]<= j/m) {
              L2[i]<-(j-1)/m
              R2[i]<-j/m}}} #close else
      } #close for (i in 101:200)
    result.sim1<-nlminb(start=c(0.8,1.8,1.5), obj=likelihood1)
    result.u[sim]<-result.sim1$par[3]

    result.sim2<-nlminb(start=c(0.8,1.8,1.5), obj=likelihood2)
    result.b[sim]<-result.sim2$par[3]

  } #close for (sim in ...)

v1[m]<-var(result.b)
v2[m]<-var(result.u)
ratio[m]<-v2[m]/v1[m] } # close for (m in 1:5)

#print results
v1
v2
ratio
```

```
##########################################
# Table 2.3 Empirical Results - Weibull  #
##########################################

rm(list=ls())
#log likelihood function for unbalance design, use -logL for
   ↪ nlminb
likelihood1.w<-function(par){
    beta1<-par[1]
    beta2<-par[2]
    beta3<-par[3]
    beta4<-par[4]
    temp1<-exp(-(L1[1:100]**beta4)*exp(beta1+beta3*Z[1:100]))
    temp2<-exp(-(R1[1:100]**beta4)*exp(beta1+beta3*Z[1:100]))
    temp3<-exp(-(L1[101:200]**beta4)*exp(beta2+beta3*Z[101:200]))
    temp4<-exp(-(R1[101:200]**beta4)*exp(beta2+beta3*Z[101:200]))

   -sum(log(temp1 - temp2))-sum(log(temp3 - temp4))
}

#log likelihood function for balance design, use -logL for nlminb
likelihood2.w<-function(par){
    beta1<-par[1]
    beta2<-par[2]
    beta3<-par[3]
    beta4<-par[4]
    temp1<-exp(-(L2[1:100]**beta4)*exp(beta1+beta3*Z[1:100]))
    temp2<-exp(-(R2[1:100]**beta4)*exp(beta1+beta3*Z[1:100]))
    temp3<-exp(-(L2[101:200]**beta4)*exp(beta2+beta3*Z[101:200]))
    temp4<-exp(-(R2[101:200]**beta4)*exp(beta2+beta3*Z[101:200]))

   -sum(log(temp1 - temp2))-sum(log(temp3 - temp4))
}

#simulation
n1<-50
n2<-50
beta<--1 #change between 1 and -1
lambda1<-1
lambda2<-4 #change between 2 and 4
alpha<-0.8 #change between 1.5 and 0.8
c<-2
m<-seq(1:5)
v1<-rep(NA,5)
```

```
v2<-rep(NA,5)

#variable for group
group1.1<-rep(0,n1)
group1.2<-rep(1,n1)
group2.1<-rep(0,n2)
group2.2<-rep(1,n2)
Z<-c(group1.1,group1.2,group2.1,group2.2)

#create an empty vector for ratio
ratio<-rep(NA,5)
for (m in 1:5){

result.u<-rep(NA,1000)
result.b<-rep(NA,1000)

for (sim in 1:1000) {
   #create exponential distribution with different rates
   w1.1<-rweibull(n1, shape=alpha, scale=lambda1**(-1/alpha))
   w1.2<-rweibull(n1, shape=alpha, scale=(lambda1*exp(beta))**(-1
      ↪ /alpha))
   w2.1<-rweibull(n2, shape=alpha, scale=lambda2**(-1/alpha))
   w2.2<-rweibull(n2, shape=alpha, scale=(lambda2*exp(beta))**(-1
      ↪ /alpha))
   t<-c(w1.1,w1.2,w2.1,w2.2)

   #create empty vector for left and right bound of the visit
      ↪ window
   L1<-rep(NA, 2*n1+2*n2)
   R1<-rep(NA, 2*n1+2*n2)
   L2<-rep(NA, 2*n1+2*n2)
   R2<-rep(NA, 2*n1+2*n2)

   #for strata 1 with low risk, let L1=L2 R1=R2
   for (i in 1:100){
      if (t[i]>1) {
         L1[i]<-1
         R1[i]<-Inf
         L2[i]<-1
         R2[i]<-Inf }
     else {for (j in 1:m) {
         if (t[i]>(j-1)/m & t[i]<= j/m) {
         L1[i]<-(j-1)/m
         R1[i]<-j/m
         L2[i]<-(j-1)/m
```

```
        R2[i]<-j/m }}}
    } #close for (i in 1:100)

  #for strata 2 with high risk, use c*m visits for L1 & R1, use
     ↪ m visits for L2 & R2
  for (i in 101:200){
     if (t[i]>1) {
        L1[i]<-1
        R1[i]<-Inf
        L2[i]<-1
        R2[i]<-Inf }
    else {
        for (j in 1:(c*m)) {
          if (t[i]>(j-1)/(c*m) & t[i]<= j/(c*m)) {
          L1[i]<-(j-1)/(c*m)
          R1[i]<-j/(c*m)}}
        for (j in 1:m) {
          if (t[i]>(j-1)/m & t[i]<= j/m) {
          L2[i]<-(j-1)/m
          R2[i]<-j/m}}} #close else
    } #close for (i in 101:200)
  result.sim1<-nlminb(start=c(0.8,1.8,1.5,1), obj=likelihood1.w)
  result.u[sim]<-result.sim1$par[3]

  result.sim2<-nlminb(start=c(0.8,1.8,1.5,1), obj=likelihood2.w)
  result.b[sim]<-result.sim2$par[3]

 } #close for (sim in ...)
v1[m]<-var(result.b)
v2[m]<-var(result.u)
ratio[m]<-v2[m]/v1[m] } # close for (m in 1:5)

#print ratio
v1
v2
ratio


################################################
#  Table 2.4 Unevenly spaced visits            #
################################################

rm(list=ls(all=T))
```

```
#log likelihood function for unbalance design, use -logL for
   ↪ nlminb
likelihood1.w<-function(par){
    beta1<-par[1]
    beta2<-par[2]
    beta3<-par[3]
    beta4<-par[4]
    temp1<-exp(-(L1[1:100]**beta4)*exp(beta1+beta3*Z[1:100]))
    temp2<-exp(-(R1[1:100]**beta4)*exp(beta1+beta3*Z[1:100]))
    temp3<-exp(-(L1[101:200]**beta4)*exp(beta2+beta3*Z[101:200]))
    temp4<-exp(-(R1[101:200]**beta4)*exp(beta2+beta3*Z[101:200]))

   -sum(log(temp1 - temp2))-sum(log(temp3 - temp4))
}

#log likelihood function for balance design, use -logL for nlminb
likelihood2.w<-function(par){
    beta1<-par[1]
    beta2<-par[2]
    beta3<-par[3]
    beta4<-par[4]
    temp1<-exp(-(L2[1:100]**beta4)*exp(beta1+beta3*Z[1:100]))
    temp2<-exp(-(R2[1:100]**beta4)*exp(beta1+beta3*Z[1:100]))
    temp3<-exp(-(L2[101:200]**beta4)*exp(beta2+beta3*Z[101:200]))
    temp4<-exp(-(R2[101:200]**beta4)*exp(beta2+beta3*Z[101:200]))

   -sum(log(temp1 - temp2))-sum(log(temp3 - temp4))
}

#simulation
n1<-50
n2<-50
beta<--1 #change between 1 and -1
lambda1<-1
lambda2<-4 #change among 2,3,4
alpha<-0.8

#variable for group
group1.1<-rep(0,n1)
group1.2<-rep(1,n1)
group2.1<-rep(0,n2)
group2.2<-rep(1,n2)
Z<-c(group1.1,group1.2,group2.1,group2.2)

result.u<-rep(NA,1000)
```

```
result.b<-rep(NA,1000)

for (sim in 1:1000) {
   #create exponential distribution with different rates
   w1.1<-rweibull(n1, shape=alpha, scale=lambda1**(-1/alpha))
   w1.2<-rweibull(n1, shape=alpha, scale=(lambda1*exp(beta))**(-1
      ↪ /alpha))
   w2.1<-rweibull(n2, shape=alpha, scale=lambda2**(-1/alpha))
   w2.2<-rweibull(n2, shape=alpha, scale=(lambda2*exp(beta))**(-1
      ↪ /alpha))
   t<-c(w1.1,w1.2,w2.1,w2.2)

   #create empty vector for left and right bound of the visit
      ↪ window
   L1<-rep(NA, 2*n1+2*n2)
   R1<-rep(NA, 2*n1+2*n2)
   L2<-rep(NA, 2*n1+2*n2)
   R2<-rep(NA, 2*n1+2*n2)

   #Assume there are 3 visits in balanced design: 0.2,0.5,1.0 and
      ↪  in unbalanced
   #design there are 6 visits in high risk group:
      ↪ 0.1,0.2,0.35,0.5,0.75,1.0
   #for strata 1 with low risk, let L1=L2 R1=R2
   for (i in 1:100){
      if (t[i]>0 & t[i]<= 0.2){
         L1[i]<-0
         R1[i]<-0.2}
      else if (t[i]>0.2 & t[i]<= 0.5){
         L1[i]<-0.2
         R1[i]<-0.5}
      else if (t[i]>0.5 & t[i]<= 1.0){
         L1[i]<-0.5
         R1[i]<-1.0}
      else if (t[i]>1.0){
         L1[i]<-1.0
         R1[i]<-Inf}
      L2[i]<-L1[i]
      R2[i]<-R1[i]} #close for (i in 1:100)

   #for strata 2 with high risk, use 6 visits for L1&R1, use 3
      ↪ visits for L2&R2
   for (i in 101:200){
      if (t[i]>0 & t[i]<= 0.2){
         L2[i]<-0
```

```
            R2[i]<-0.2
            if (t[i]>0 & t[i]<= 0.1){
               L1[i]<-0
               R1[i]<-0.1}
            else {
               L1[i]<-0.1
               R1[i]<-0.2}}
         else if (t[i]>0.2 & t[i]<= 0.5){
            L2[i]<-0.2
            R2[i]<-0.5
            if (t[i]>0.2 & t[i]<= 0.35){
               L1[i]<-0.2
               R1[i]<-0.35}
            else {
               L1[i]<-0.35
               R1[i]<-0.5}}
         else if (t[i]>0.5 & t[i]<= 1.0){
            L2[i]<-0.5
            R2[i]<-1.0
            if (t[i]>0.5 & t[i]<= 0.75){
               L1[i]<-0.5
               R1[i]<-0.75}
            else {
               L1[i]<-0.75
               R1[i]<-1.0}}
         else if (t[i]>1.0){
            L1[i]<-1.0
            R1[i]<-Inf
            L2[i]<-1.0
            R2[i]<-Inf}} #close for (i in 101:200)

      result.sim1<-nlminb(start=c(0.8,1.8,1.5,1), obj=likelihood1.w)
      result.u[sim]<-result.sim1$par[3]

      result.sim2<-nlminb(start=c(0.8,1.8,1.5,1), obj=likelihood2.w)
      result.b[sim]<-result.sim2$par[3]

    } #close for (sim in ...)

v1<-var(result.b)
v2<-var(result.u)
ratio<-v2/v1

#print results
v1
```

```
v2
ratio


##############################################
# Table 2.5. Theoretical Power               #
##############################################
rm(list=ls(all=T))

#calculate variance ratio of inbalanced design:balanced design
n1<-30 #change between 30 and 40
n2<-30 #change between 30 and 40
beta<-0.6 #change between 0.5 and 0.6
T<-10
lambda1<-0.3
lambda2<-0.6
#create an empty vector for variance and power
v1<-rep(NA,5)
v2<-rep(NA,5)
p1<-rep(NA,5)
p2<-rep(NA,5)


for (m in 2:5){
#variance from balanced design
c<-1
temp1<-n1*lambda1*T*exp(beta)*exp(-lambda1*T*exp(beta))
temp2<-n2*lambda2*T*exp(beta)*exp(-lambda2*T*exp(beta))

temp3<-n1*lambda1*T/m*exp(beta)

temp4<-1-m*exp(-lambda1*T*(m-1)/m*exp(beta))+(m-1)*exp(-lambda1*T
    ↪ *exp(beta))
temp5<-exp(lambda1*T/m*exp(beta))-1

temp6<-(1-lambda1*T/m*exp(beta)-exp(-lambda1*T/m*exp(beta)))*(1-
    ↪ exp(-lambda1*T*exp(beta)))
temp7<-(exp(lambda1*T/m*exp(beta))-1)*(1-exp(-lambda1*T/m*exp(
    ↪ beta)))

temp8<-n2*lambda2*T/(c*m)*exp(beta)

temp9<-1-c*m*exp(-lambda2*T*(c*m-1)/(c*m)*exp(beta))+(c*m-1)*exp
    ↪ (-lambda2*T*exp(beta))
temp10<-exp(lambda2*T/(c*m)*exp(beta))-1
```

```
temp11<-(1-lambda2*T/(c*m)*exp(beta)-exp(-lambda2*T/(c*m)*exp(
    ↪ beta)))*(1-exp(-lambda2*T*exp(beta)))
temp12<-(exp(lambda2*T/(c*m)*exp(beta))-1)*(1-exp(-lambda2*T/(c*m
    ↪ )*exp(beta)))

#fisher information matrix
I.11<-temp1+temp2+temp3*temp4/temp5-temp3*temp6/temp7+temp8*temp9
    ↪ /temp10-temp8*temp11/temp12
I.12<-temp1/lambda1+(temp3/lambda1)*(temp4/temp5)-(temp3/lambda1)
    ↪ *(1-exp(-lambda1*T*exp(beta)))/temp5+temp3*T*exp(beta)/m*
    ↪ ((1-exp(-lambda1*T*exp(beta)))/temp7)
I.13<-temp2/lambda2+(temp8/lambda2)*(temp9/temp10)-(temp8/lambda2
    ↪ )*(1-exp(-lambda2*T*exp(beta)))/temp10+temp8*T*exp(beta)/(c
    ↪ *m)*((1-exp(-lambda2*T*exp(beta)))/temp12)
I.21<-I.12
I.22<-n1*T^2*(1-exp(-lambda1*T))/(m^2*(exp(lambda1*T/m)-1)*(1-exp
    ↪ (-lambda1*T/m)))+n1*T^2*exp(2*beta)*(1-exp(-lambda1*T*exp(
    ↪ beta)))/(m^2*temp7)
I.23<-0
I.31<-I.13
I.32<-I.23
I.33<-n2*T^2*(1-exp(-lambda2*T))/((c*m)^2*(exp(lambda2*T/(c*m))
    ↪ -1)*(1-exp(-lambda2*T/(c*m))))+n2*T^2*exp(2*beta)*(1-exp(-
    ↪ lambda2*T*exp(beta)))/((c*m)^2*temp12)

I<-matrix(c(I.11,I.12,I.13,I.21,I.22,I.23,I.31,I.32,I.33),3,3)
I.inv<-solve(I)
v1[m]<-I.inv[1,1]

#for unbalanced design 2:1
c<-2
temp1<-n1*lambda1*T*exp(beta)*exp(-lambda1*T*exp(beta))
temp2<-n2*lambda2*T*exp(beta)*exp(-lambda2*T*exp(beta))

temp3<-n1*lambda1*T/m*exp(beta)

temp4<-1-m*exp(-lambda1*T*(m-1)/m*exp(beta))+(m-1)*exp(-lambda1*T
    ↪ *exp(beta))
temp5<-exp(lambda1*T/m*exp(beta))-1

temp6<-(1-lambda1*T/m*exp(beta)-exp(-lambda1*T/m*exp(beta)))*(1-
    ↪ exp(-lambda1*T*exp(beta)))
temp7<-(exp(lambda1*T/m*exp(beta))-1)*(1-exp(-lambda1*T/m*exp(
    ↪ beta)))
```

```
temp8<-n2*lambda2*T/(c*m)*exp(beta)

temp9<-1-c*m*exp(-lambda2*T*(c*m-1)/(c*m)*exp(beta))+(c*m-1)*exp
    ↪ (-lambda2*T*exp(beta))
temp10<-exp(lambda2*T/(c*m)*exp(beta))-1

temp11<-(1-lambda2*T/(c*m)*exp(beta)-exp(-lambda2*T/(c*m)*exp(
    ↪ beta)))*(1-exp(-lambda2*T*exp(beta)))
temp12<-(exp(lambda2*T/(c*m)*exp(beta))-1)*(1-exp(-lambda2*T/(c*m
    ↪ )*exp(beta)))

#fisher information matrix
I.11<-temp1+temp2+temp3*temp4/temp5-temp3*temp6/temp7+temp8*temp9
    ↪ /temp10-temp8*temp11/temp12
I.12<-temp1/lambda1+(temp3/lambda1)*(temp4/temp5)-(temp3/lambda1)
    ↪ *(1-exp(-lambda1*T*exp(beta)))/temp5+temp3*T*exp(beta)/m*
    ↪ ((1-exp(-lambda1*T*exp(beta)))/temp7)
I.13<-temp2/lambda2+(temp8/lambda2)*(temp9/temp10)-(temp8/lambda2
    ↪ )*(1-exp(-lambda2*T*exp(beta)))/temp10+temp8*T*exp(beta)/(c
    ↪ *m)*((1-exp(-lambda2*T*exp(beta)))/temp12)
I.21<-I.12
I.22<-n1*T^2*(1-exp(-lambda1*T))/(m^2*(exp(lambda1*T/m)-1)*(1-exp
    ↪ (-lambda1*T/m)))+n1*T^2*exp(2*beta)*(1-exp(-lambda1*T*exp(
    ↪ beta)))/(m^2*temp7)
I.23<-0
I.31<-I.13
I.32<-I.23
I.33<-n2*T^2*(1-exp(-lambda2*T))/((c*m)^2*(exp(lambda2*T/(c*m))
    ↪ -1)*(1-exp(-lambda2*T/(c*m))))+n2*T^2*exp(2*beta)*(1-exp(-
    ↪ lambda2*T*exp(beta)))/((c*m)^2*temp12)

I<-matrix(c(I.11,I.12,I.13,I.21,I.22,I.23,I.31,I.32,I.33),3,3)
I.inv<-solve(I)
v2[m]<-I.inv[1,1]

p1[m]<-1-(pnorm(1.96-beta/sqrt(v1[m]))-pnorm(-1.96-beta/sqrt(v1[m
    ↪ ])))
p2[m]<-1-(pnorm(1.96-beta/sqrt(v2[m]))-pnorm(-1.96-beta/sqrt(v2[m
    ↪ ])))
}

p1
p2
```

```
#####################################################
#          Table 2.5    Empirical Power             #
#####################################################

rm(list=ls(all=T))

#log likelihood function for unbalance design, use -logL for
   ↪ nlminb
likelihood1<-function(par){
    beta1<-par[1]
    beta2<-par[2]
    beta3<-par[3]
    temp1<-exp(-L1[1:(2*n1)]*exp(beta1+beta3*Z[1:(2*n1)]))
    temp2<-exp(-R1[1:(2*n1)]*exp(beta1+beta3*Z[1:(2*n1)]))
    temp3<-exp(-L1[(2*n1+1):(2*n1+2*n2)]*exp(beta2+beta3*Z[(2*n1
       ↪ +1):(2*n1+2*n2)]))
    temp4<-exp(-R1[(2*n1+1):(2*n1+2*n2)]*exp(beta2+beta3*Z[(2*n1
       ↪ +1):(2*n1+2*n2)]))

    -sum(log(temp1 - temp2))-sum(log(temp3 - temp4))
}

#log likelihood function for balance design, use -logL for nlminb
likelihood2<-function(par){
    beta1<-par[1]
    beta2<-par[2]
    beta3<-par[3]
    temp1<-exp(-L2[1:(2*n1)]*exp(beta1+beta3*Z[1:(2*n1)]))
    temp2<-exp(-R2[1:(2*n1)]*exp(beta1+beta3*Z[1:(2*n1)]))
    temp3<-exp(-L2[(2*n1+1):(2*n1+2*n2)]*exp(beta2+beta3*Z[(2*n1
       ↪ +1):(2*n1+2*n2)]))
    temp4<-exp(-R2[(2*n1+1):(2*n1+2*n2)]*exp(beta2+beta3*Z[(2*n1
       ↪ +1):(2*n1+2*n2)]))

    -sum(log(temp1 - temp2))-sum(log(temp3 - temp4))
}



#simulation
T<-10
n1<-30 #change between 30 and 40
n2<-30 #change between 30 and 40
beta<-0.6 #change between 0.5 and 0.6
```

```
lambda1<-0.3
lambda2<-0.6
p.b<-rep(NA,5)
p.u<-rep(NA,5)

#variable for group
group1.1<-rep(0,n1)
group1.2<-rep(1,n1)
group2.1<-rep(0,n2)
group2.2<-rep(1,n2)
Z<-c(group1.1,group1.2,group2.1,group2.2)

for (m in 2:5){

beta.hat.u<-rep(NA,1000)
beta.hat.b<-rep(NA,1000)

for (sim in 1:1000){
   #create exponential distribution with different rates
   exp1.1<-rexp(n1, rate=lambda1)
   exp1.2<-rexp(n1, rate=lambda1*exp(beta))
   exp2.1<-rexp(n2, rate=lambda2)
   exp2.2<-rexp(n2, rate=lambda2*exp(beta))
   t<-c(exp1.1,exp1.2,exp2.1,exp2.2)

   #create empty vector for left and right bound of the visit
      ↪ window
   L1<-rep(NA, 2*n1+2*n2)
   R1<-rep(NA, 2*n1+2*n2)
   L2<-rep(NA, 2*n1+2*n2)
   R2<-rep(NA, 2*n1+2*n2)
   c<-2
   #for strata 1 with low risk, let L1=L2 R1=R2
   for (i in 1:(2*n1)){
      if (t[i]>T) {
         L1[i]<-T
         R1[i]<-Inf
         L2[i]<-T
         R2[i]<-Inf }
     else {for (j in 1:m) {
         if (t[i]>T*(j-1)/m & t[i]<= T*j/m) {
         L1[i]<-T*(j-1)/m
         R1[i]<-T*j/m
         L2[i]<-T*(j-1)/m
         R2[i]<-T*j/m }}}
```

```
    } #close for (i in 1:(2*n1))

  #for strata 2 with high risk, use c*m visits for L1 & R1, use
     ↪ m visits for L2 & R2
  for (i in (2*n1+1):(2*n1+2*n2)){
        if (t[i]>T) {
        L1[i]<-T
        R1[i]<-Inf
        L2[i]<-T
        R2[i]<-Inf }
    else {
        for (j in 1:(c*m)) {
          if (t[i]>T*(j-1)/(c*m) & t[i]<= T*j/(c*m)) {
          L1[i]<-T*(j-1)/(c*m)
          R1[i]<-T*j/(c*m)}}
        for (j in 1:m) {
          if (t[i]>T*(j-1)/m & t[i]<= T*j/m) {
          L2[i]<-T*(j-1)/m
          R2[i]<-T*j/m}}} #close else
    } #close for (i in (2*n1+1):(2*n1+2*n2))
  result.sim1<-nlminb(start=c(0.8,1.8,1.5), obj=likelihood1)
  lambda1.hat.u<-exp(result.sim1$par[1])
  lambda2.hat.u<-exp(result.sim1$par[2])
  beta.hat.u[sim]<-result.sim1$par[3]

  result.sim2<-nlminb(start=c(0.8,1.8,1.5), obj=likelihood2)
  lambda1.hat.b<-exp(result.sim2$par[1])
  lambda2.hat.b<-exp(result.sim2$par[2])
  beta.hat.b[sim]<-result.sim2$par[3]
 } #close for (sim in ...)

v.u<-var(beta.hat.u)
v.b<-var(beta.hat.b)
z.b<-beta.hat.b/sqrt(v.b)
z.u<-beta.hat.u/sqrt(v.u)

p.b[m]<-mean(z.b< -1.96 | z.b>1.96)
p.u[m]<-mean(z.u< -1.96 | z.u>1.96)
} # close for (m in 2:5)

#print results
p.b
p.u
```

```
##############################################
#  Table 2.6 Empirical type 1 error          #
##############################################


#########Exponential distribution##############

rm(list=ls(all=T))

#log likelihood function for unbalance design, use -logL for
   ↪ nlminb
likelihood1<-function(par){
    beta1<-par[1]
    beta2<-par[2]
    beta3<-par[3]
    temp1<-exp(-L1[1:(2*n1)]*exp(beta1+beta3*Z[1:(2*n1)]))
    temp2<-exp(-R1[1:(2*n1)]*exp(beta1+beta3*Z[1:(2*n1)]))
    temp3<-exp(-L1[(2*n1+1):(2*n1+2*n2)]*exp(beta2+beta3*Z[(2*n1
       ↪ +1):(2*n1+2*n2)]))
    temp4<-exp(-R1[(2*n1+1):(2*n1+2*n2)]*exp(beta2+beta3*Z[(2*n1
       ↪ +1):(2*n1+2*n2)]))

   -sum(log(temp1 - temp2))-sum(log(temp3 - temp4))
}

#log likelihood function for balance design, use -logL for nlminb
likelihood2<-function(par){
    beta1<-par[1]
    beta2<-par[2]
    beta3<-par[3]
    temp1<-exp(-L2[1:(2*n1)]*exp(beta1+beta3*Z[1:(2*n1)]))
    temp2<-exp(-R2[1:(2*n1)]*exp(beta1+beta3*Z[1:(2*n1)]))
    temp3<-exp(-L2[(2*n1+1):(2*n1+2*n2)]*exp(beta2+beta3*Z[(2*n1
       ↪ +1):(2*n1+2*n2)]))
    temp4<-exp(-R2[(2*n1+1):(2*n1+2*n2)]*exp(beta2+beta3*Z[(2*n1
       ↪ +1):(2*n1+2*n2)]))

   -sum(log(temp1 - temp2))-sum(log(temp3 - temp4))
}



#simulation
```

```
T<-10
n1<-50
n2<-50
beta<-0
lambda1<-0.3
lambda2<-0.6
alpha.b<-rep(NA,5)
alpha.u<-rep(NA,5)

#variable for group
group1.1<-rep(0,n1)
group1.2<-rep(1,n1)
group2.1<-rep(0,n2)
group2.2<-rep(1,n2)
Z<-c(group1.1,group1.2,group2.1,group2.2)

for (m in 2:5){

beta.hat.u<-rep(NA,1000)
beta.hat.b<-rep(NA,1000)

for (sim in 1:1000){
   #create exponential distribution with different rates
   exp1.1<-rexp(n1, rate=lambda1)
   exp1.2<-rexp(n1, rate=lambda1*exp(beta))
   exp2.1<-rexp(n2, rate=lambda2)
   exp2.2<-rexp(n2, rate=lambda2*exp(beta))
   t<-c(exp1.1,exp1.2,exp2.1,exp2.2)

   #create empty vector for left and right bound of the visit
      ↪ window
   L1<-rep(NA, 2*n1+2*n2)
   R1<-rep(NA, 2*n1+2*n2)
   L2<-rep(NA, 2*n1+2*n2)
   R2<-rep(NA, 2*n1+2*n2)
   c<-3
   #for strata 1 with low risk, let L1=L2 R1=R2
   for (i in 1:(2*n1)){
      if (t[i]>T) {
         L1[i]<-T
         R1[i]<-Inf
         L2[i]<-T
         R2[i]<-Inf }
     else {for (j in 1:m) {
         if (t[i]>T*(j-1)/m & t[i]<= T*j/m) {
```

```
            L1[i]<-T*(j-1)/m
            R1[i]<-T*j/m
            L2[i]<-T*(j-1)/m
            R2[i]<-T*j/m }}}
      } #close for (i in 1:(2*n1))

    #for strata 2 with high risk, use c*m visits for L1 & R1, use
       ↪ m visits for L2 & R2
    for (i in (2*n1+1):(2*n1+2*n2)){
          if (t[i]>T) {
          L1[i]<-T
          R1[i]<-Inf
          L2[i]<-T
          R2[i]<-Inf }
      else {
          for (j in 1:(c*m)) {
            if (t[i]>T*(j-1)/(c*m) & t[i]<= T*j/(c*m)) {
            L1[i]<-T*(j-1)/(c*m)
            R1[i]<-T*j/(c*m)}}
          for (j in 1:m) {
            if (t[i]>T*(j-1)/m & t[i]<= T*j/m) {
            L2[i]<-T*(j-1)/m
            R2[i]<-T*j/m}}} #close else
      } #close for (i in (2*n1+1):(2*n1+2*n2))
    result.sim1<-nlminb(start=c(0.8,1.8,1.5), obj=likelihood1)
    lambda1.hat.u<-exp(result.sim1$par[1])
    lambda2.hat.u<-exp(result.sim1$par[2])
    beta.hat.u[sim]<-result.sim1$par[3]

    result.sim2<-nlminb(start=c(0.8,1.8,1.5), obj=likelihood2)
    lambda1.hat.b<-exp(result.sim2$par[1])
    lambda2.hat.b<-exp(result.sim2$par[2])
    beta.hat.b[sim]<-result.sim2$par[3]
  } #close for (sim in ...)

v.u<-var(beta.hat.u)
v.b<-var(beta.hat.b)
z.b<-beta.hat.b/sqrt(v.b)
z.u<-beta.hat.u/sqrt(v.u)

alpha.b[m]<-mean(z.b< -1.96 | z.b>1.96)
alpha.u[m]<-mean(z.u< -1.96 | z.u>1.96)
} # close for (m in 2:5)

#print results
```

```
alpha.b
alpha.u


##############Weibull distribution###################

rm(list=ls())
#log likelihood function for unbalance design, use -logL for
   ↪ nlminb
likelihood1.w<-function(par){
    beta1<-par[1]
    beta2<-par[2]
    beta3<-par[3]
    beta4<-par[4]
    temp1<-exp(-(L1[1:(2*n1)]**beta4)*exp(beta1+beta3*Z[1:(2*n1)])
       ↪ )
    temp2<-exp(-(R1[1:(2*n1)]**beta4)*exp(beta1+beta3*Z[1:(2*n1)])
       ↪ )
    temp3<-exp(-(L1[(2*n1+1):(2*n1+2*n2)]**beta4)*exp(beta2+beta3*
       ↪ Z[(2*n1+1):(2*n1+2*n2)]))
    temp4<-exp(-(R1[(2*n1+1):(2*n1+2*n2)]**beta4)*exp(beta2+beta3*
       ↪ Z[(2*n1+1):(2*n1+2*n2)]))

   -sum(log(temp1 - temp2))-sum(log(temp3 - temp4))
}

#log likelihood function for balance design, use -logL for nlminb
likelihood2.w<-function(par){
    beta1<-par[1]
    beta2<-par[2]
    beta3<-par[3]
    beta4<-par[4]
    temp1<-exp(-(L2[1:(2*n1)]**beta4)*exp(beta1+beta3*Z[1:(2*n1)])
       ↪ )
    temp2<-exp(-(R2[1:(2*n1)]**beta4)*exp(beta1+beta3*Z[1:(2*n1)])
       ↪ )
    temp3<-exp(-(L2[(2*n1+1):(2*n1+2*n2)]**beta4)*exp(beta2+beta3*
       ↪ Z[(2*n1+1):(2*n1+2*n2)]))
    temp4<-exp(-(R2[(2*n1+1):(2*n1+2*n2)]**beta4)*exp(beta2+beta3*
       ↪ Z[(2*n1+1):(2*n1+2*n2)]))

   -sum(log(temp1 - temp2))-sum(log(temp3 - temp4))
}
```

```
#simulation
T<-10
n1<-50
n2<-50
beta<-0
lambda1<-0.3
lambda2<-0.6
alpha<-0.8 #change between 1.5 and 0.8
c<-2
m<-seq(1:5)
alpha.b<-rep(NA,5)
alpha.u<-rep(NA,5)

#variable for group
group1.1<-rep(0,n1)
group1.2<-rep(1,n1)
group2.1<-rep(0,n2)
group2.2<-rep(1,n2)
Z<-c(group1.1,group1.2,group2.1,group2.2)

#create an empty vector for ratio
ratio<-rep(NA,5)
for (m in 2:5){

beta.hat.u<-rep(NA,1000)
beta.hat.b<-rep(NA,1000)

for (sim in 1:1000) {
   #create exponential distribution with different rates
   w1.1<-rweibull(n1, shape=alpha, scale=lambda1**(-1/alpha))
   w1.2<-rweibull(n1, shape=alpha, scale=(lambda1*exp(beta))**(-1
      ↪ /alpha))
   w2.1<-rweibull(n2, shape=alpha, scale=lambda2**(-1/alpha))
   w2.2<-rweibull(n2, shape=alpha, scale=(lambda2*exp(beta))**(-1
      ↪ /alpha))
   t<-c(w1.1,w1.2,w2.1,w2.2)

   #create empty vector for left and right bound of the visit
      ↪ window
   L1<-rep(NA, 2*n1+2*n2)
   R1<-rep(NA, 2*n1+2*n2)
   L2<-rep(NA, 2*n1+2*n2)
   R2<-rep(NA, 2*n1+2*n2)

   #for strata 1 with low risk, let L1=L2 R1=R2
```

```
    for (i in 1:(2*n1)){
       if (t[i]>T) {
          L1[i]<-T
          R1[i]<-Inf
          L2[i]<-T
          R2[i]<-Inf }
      else {for (j in 1:m) {
          if (t[i]>T*(j-1)/m & t[i]<= T*j/m) {
          L1[i]<-T*(j-1)/m
          R1[i]<-T*j/m
          L2[i]<-T*(j-1)/m
          R2[i]<-T*j/m }}}
      }

    #for strata 2 with high risk, use c*m visits for L1 & R1, use
       ↪ m visits for L2 & R2
    for (i in (2*n1+1):(2*n1+2*n2)){
       if (t[i]>T) {
          L1[i]<-T
          R1[i]<-Inf
          L2[i]<-T
          R2[i]<-Inf }
      else {
          for (j in 1:(c*m)) {
            if (t[i]>T*(j-1)/(c*m) & t[i]<= T*j/(c*m)) {
            L1[i]<-T*(j-1)/(c*m)
            R1[i]<-T*j/(c*m)}}
          for (j in 1:m) {
            if (t[i]>T*(j-1)/m & t[i]<= T*j/m) {
            L2[i]<-T*(j-1)/m
            R2[i]<-T*j/m}}} #close else
      }
    result.sim1<-nlminb(start=c(0.8,1.8,1.5,1), obj=likelihood1.w)
    beta.hat.u[sim]<-result.sim1$par[3]

    result.sim2<-nlminb(start=c(0.8,1.8,1.5,1), obj=likelihood2.w)
    beta.hat.b[sim]<-result.sim2$par[3]

  } #close for (sim in ...)
v.u<-var(beta.hat.u)
v.b<-var(beta.hat.b)
z.b<-beta.hat.b/sqrt(v.b)
z.u<-beta.hat.u/sqrt(v.u)

alpha.b[m]<-mean(z.b< -1.96 | z.b>1.96)
```

```
alpha.u[m]<-mean(z.u< -1.96 | z.u>1.96)
} # close for (m in 2:5)

alpha.b
alpha.u
```

Appendix C: SAS Codes for Chapter 3

```
/*************************************************
     Analysis for Chapter 3 in Dissertation
*************************************************/

options ls=100 nodate nonumber orientation=portrait;
libname cpep "H:\Met cont";
ods listing gpath='H:\Figures';
ods graphics on;

data subjects;
    set cpep.subjects;
run;

proc tabulate data=subjects format=6.0 missing;
    class eventMos03 eventMos02 event03 event02;
    table all eventMos02, event02*n;
        table all eventMos03, event03*n;
    title  'Table 1';
run;

proc tabulate data=subjects format=6.0 missing;
    class eventMos02 event02 ageGrp;
    table all eventMos02, ageGrp*event02*n;
    title  'Table 2';
run;

proc tabulate data=subjects format=6.0 missing;
    class eventMos03 event03 ageGrp;
    table all eventMos03, ageGrp*event03*n;
    title  'Table 3';
run;

data subjects;
    set subjects(keep=PtID ageGrp TxGroup Gender a1c3 event03
      ↪ eventMos03
```

```
                    event02 eventMos02);
  /*right-censored subjects*/
  if event03=0 then do;
     L1=eventMos03; R1=.;
     L2=eventMos03; R2=.;
     L3=eventMos03; R3=.;
     L3r=eventMos03; R3r=.;
  end;
  /*subjects with events*/
  else if event03=1 then do;
     /*assume every subject has 6 visits, create intervals
       ↪ based on 6 visits*/
     R1=eventMos03;
     if eventMos03=3 then L1=0.01;
     else if eventMos03 in (6,9,12) then L1=R1-3;
     else if eventMos03 in (18,24) then L1=R1-6;

     /*assume every subject has 3 visits, create intervals
       ↪ based on 3 visits*/
     if eventMos03 in (3,6) then do;
        L2=0.01; R2=6;
     end;
     else if eventMos03 in (9,12) then do;
        L2=6; R2=12;
     end;
     else if eventMos03 in (18,24) then do;
        L2=12; R2=24;
     end;
     /*create intervals based on age group*/
     if ageGrp='A)7-<12␣yrs' then do; *high risk group;
        L3=L1; R3=R1;
     end;
     else if ageGrp='B)>=12␣years' then do; *low risk group
       ↪ ;
        L3=L2; R3=R2;
     end;
/*reverse the direction of unbalanced design*/
     if ageGrp='A)7-<12␣yrs' then do; *high risk group;
        L3r=L2; R3r=R2;
     end;
     else if ageGrp='B)>=12␣years' then do; *low risk group
       ↪ ;
        L3r=L1; R3r=R1;
     end;
```

```
    end;

    /*Repeat for event02 outcome*/
if event02=0 then do;
       L4=eventMos02; R4=.;
       L5=eventMos02; R5=.;
       L6=eventMos02; R6=.;
       L6r=eventMos02; R6r=.;
    end;
    /*subjects with events*/
    else if event02=1 then do;
       /*assume every subject has 6 visits, create intervals
          ↪ based on 6 visits*/
       R4=eventMos02;
       if eventMos02=3 then L4=0.01;
       else if eventMos02 in (6,9,12) then L4=R4-3;
       else if eventMos02 in (18,24) then L4=R4-6;

       /*assume every subject has 3 visits, create intervals
          ↪ based on 3 visits*/
       if eventMos02 in (3,6) then do;
          L5=0.01; R5=6;
       end;
       else if eventMos02 in (9,12) then do;
          L5=6; R5=12;
       end;
       else if eventMos02 in (18,24) then do;
          L5=12; R5=24;
       end;
       /*create intervals based on age group*/
       if ageGrp='A)7-<12␣yrs' then do; *high risk group;
          L6=L4; R6=R4;
       end;
       else if ageGrp='B)>=12␣years' then do; *low risk group
          ↪ ;
          L6=L5; R6=R5;
       end;

       /*reverse the direction of unbalanced design*/
       if ageGrp='A)7-<12␣yrs' then do; *high risk group;
          L6r=L5; R6r=R5;
       end;
       else if ageGrp='B)>=12␣years' then do; *low risk group
          ↪ ;
          L6r=L4; R6r=R4;
```

```
            end;
         end;
run;

*macro for categorical variables;
%macro lifereg1 (L=, R=, var=, D=);
   proc lifereg data=subjects;
      class &var ageGrp;
      model (&L, &R)= &var ageGrp /D= &D;
   run;
%mend;

%lifereg1(L=L5, R=R5, var=TxGroup, D=Exponential);
%lifereg1(L=L6, R=R6, var=TxGroup, D=Exponential);
%lifereg1(L=L6r, R=R6r, var=TxGroup, D=Exponential);

%lifereg1(L=L5, R=R5, var=TxGroup, D=Weibull);
%lifereg1(L=L6, R=R6, var=TxGroup, D=Weibull);
%lifereg1(L=L6r, R=R6r, var=TxGroup, D=Weibull);

%lifereg1(L=L5, R=R5, var=Gender, D=Exponential);
%lifereg1(L=L6, R=R6, var=Gender, D=Exponential);
%lifereg1(L=L6r, R=R6r, var=Gender, D=Exponential);

%lifereg1(L=L5, R=R5, var=Gender, D=Weibull);
%lifereg1(L=L6, R=R6, var=Gender, D=Weibull);
%lifereg1(L=L6r, R=R6r, var=Gender, D=Weibull);

%lifereg1(L=L2, R=R2, var=TxGroup, D=Exponential);
%lifereg1(L=L3, R=R3, var=TxGroup, D=Exponential);
%lifereg1(L=L3r, R=R3r, var=TxGroup, D=Exponential);

%lifereg1(L=L2, R=R2, var=TxGroup, D=Weibull);
%lifereg1(L=L3, R=R3, var=TxGroup, D=Weibull);
%lifereg1(L=L3r, R=R3r, var=TxGroup, D=Weibull);

%lifereg1(L=L2, R=R2, var=Gender, D=Exponential);
%lifereg1(L=L3, R=R3, var=Gender, D=Exponential);
%lifereg1(L=L3r, R=R3r, var=Gender, D=Exponential);

%lifereg1(L=L2, R=R2, var=Gender, D=Weibull);
%lifereg1(L=L3, R=R3, var=Gender, D=Weibull);
%lifereg1(L=L3r, R=R3r, var=Gender, D=Weibull);
```

```
%macro lifereg2 (L=, R=, var=, D=);
    proc lifereg data=subjects;
       class ageGrp;
       model (&L, &R)= &var ageGrp /D= &D;
    run;
%mend;

%lifereg2(L=L5, R=R5, var=a1c3, D=Exponential);
%lifereg2(L=L6, R=R6, var=a1c3, D=Exponential);
%lifereg2(L=L6r, R=R6r, var=a1c3, D=Exponential);
%lifereg2(L=L5, R=R5, var=a1c3, D=Weibull);
%lifereg2(L=L6, R=R6, var=a1c3, D=Weibull);
%lifereg2(L=L6r, R=R6r, var=a1c3, D=Weibull);

%lifereg2(L=L2, R=R2, var=a1c3, D=Exponential);
%lifereg2(L=L3, R=R3, var=a1c3, D=Exponential);
%lifereg2(L=L3r, R=R3r, var=a1c3, D=Exponential);

%lifereg2(L=L2, R=R2, var=a1c3, D=Weibull);
%lifereg2(L=L3, R=R3, var=a1c3, D=Weibull);
%lifereg2(L=L3r, R=R3r, var=a1c3, D=Weibull);


%macro ictest(var=, weight=, L=, R=);
     proc iclifetest data=subjects plots=survival(cl) impute(seed
        ↪ =1234);
            strata ageGrp;
        test &var / weight=&weight;
        time (&L, &R);
     run;
%mend;

%ictest(var=TxGroup, weight=FINKELSTEIN, L=L5, R=R5);
%ictest(var=TxGroup, weight=FINKELSTEIN, L=L6, R=R6);
%ictest(var=TxGroup, weight=FINKELSTEIN, L=L6r, R=R6r);

%ictest(var=TxGroup, weight=SUN, L=L5, R=R5);
%ictest(var=TxGroup, weight=SUN, L=L6, R=R6);
%ictest(var=TxGroup, weight=SUN, L=L6r, R=R6r);

%ictest(var=TxGroup, weight=FAY, L=L5, R=R5);
%ictest(var=TxGroup, weight=FAY, L=L6, R=R6);
%ictest(var=TxGroup, weight=FAY, L=L6r, R=R6r);

%ictest(var=Gender, weight=FINKELSTEIN, L=L5, R=R5);
```

```
%ictest(var=Gender, weight=FINKELSTEIN, L=L6, R=R6);
%ictest(var=Gender, weight=FINKELSTEIN, L=L6r, R=R6r);

%ictest(var=Gender, weight=SUN, L=L5, R=R5);
%ictest(var=Gender, weight=SUN, L=L6, R=R6);
%ictest(var=Gender, weight=SUN, L=L6r, R=R6r);

%ictest(var=Gender, weight=FAY, L=L5, R=R5);
%ictest(var=Gender, weight=FAY, L=L6, R=R6);
%ictest(var=Gender, weight=FAY, L=L6r, R=R6r);


%ictest(var=TxGroup, weight=FINKELSTEIN, L=L2, R=R2);
%ictest(var=TxGroup, weight=FINKELSTEIN, L=L3, R=R3);
%ictest(var=TxGroup, weight=FINKELSTEIN, L=L3r, R=R3r);

%ictest(var=TxGroup, weight=SUN, L=L2, R=R2);
%ictest(var=TxGroup, weight=SUN, L=L3, R=R3);
%ictest(var=TxGroup, weight=SUN, L=L3r, R=R3r);

%ictest(var=TxGroup, weight=FAY, L=L2, R=R2);
%ictest(var=TxGroup, weight=FAY, L=L3, R=R3);
%ictest(var=TxGroup, weight=FAY, L=L3r, R=R3r);

%ictest(var=Gender, weight=FINKELSTEIN, L=L2, R=R2);
%ictest(var=Gender, weight=FINKELSTEIN, L=L3, R=R3);
%ictest(var=Gender, weight=FINKELSTEIN, L=L3r, R=R3r);

%ictest(var=Gender, weight=SUN, L=L2, R=R2);
%ictest(var=Gender, weight=SUN, L=L3, R=R3);
%ictest(var=Gender, weight=SUN, L=L3r, R=R3r);

%ictest(var=Gender, weight=FAY, L=L2, R=R2);
%ictest(var=Gender, weight=FAY, L=L3, R=R3);
%ictest(var=Gender, weight=FAY, L=L3r, R=R3r);
```

Appendix D: IRB Letter

USF
UNIVERSITY OF
SOUTH FLORIDA

RESEARCH INTEGRITY AND COMPLIANCE
Institutional Review Boards, FWA No. 00001669
12901 Bruce B. Downs Blvd., MDC035 • Tampa, FL 33612-4799
(813) 974-5638 • FAX (813) 974-7091

11/2/2016

Yougui Wu, PhD
Epidemiology and Biostatistics
13201 Bruce B. Downs Blvd. MDC56
Tampa, FL 33612

RE: **Not Human Subjects Research Determination**
IRB#: Pro00028176
Title: Efficiency of an Unbalanced Design in Collecting Time to Event Data with Interval Censoring

Dear Dr. Wu:

The Institutional Review Board (IRB) has reviewed your application and determined the activities do not meet the definition of human subjects research. Therefore, this project is not under the purview of the USF IRB and approval is not required. If the scope of your project changes in the future, please contact the IRB for further guidance.

All research activities, regardless of the level of IRB oversight, must be conducted in a manner that is consistent with the ethical principles of your profession. Please note that there may be requirements under the HIPAA Privacy Rule that apply to the information/data you will utilize. For further information, please contact a HIPAA Program administrator at 813-974-5638.

We appreciate your dedication to the ethical conduct of research at the University of South Florida. If you have any questions regarding this matter, please call 813-974-5638.

Sincerely,

V Jorgensen MD

E. Verena Jorgensen, M.D., Chairperson
USF Institutional Review Board