Two-phase outcome-dependent sampling designs for sequential survival time analysis

by

© **Tzuemn-Renn Lin**

A thesis submitted to the School of Gradate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Department of Mathematics and Statistics

Memorial University of Newfoundland

March 2018

St. John's, Newfoundland and Labrador, Canada

Abstract

In some observational studies, the covariates of interest might be expensive to measure although the outcome variable could easily be obtained. In this situation, a cost-efficient two-phase outcome-dependent sampling design could be employed to measure the expensive covariate for more informative subjects. In phase one, all members of a random sample from a population or a cohort are measured for the outcome variable and inexpensive covariates. In phase two, a subset of the cohort is selected based on the outcome variable, and the expensive covariate is measured only for the selected individuals. Case-cohort design is a commonly used outcome-dependent sampling design in time-to-event analyses. In generalized case-cohort design, in which the selection probability depends only on the event indicator, a random subsample of individuals who experienced the event are selected, along with a random subsample of those with censored event times. It was previously shown that when the selection probability at phase two depends on observed event time and censoring time in addition to the event indicator, the efficiency of the design might increase. Efficient design has a lower variance of the coefficient estimate of the expensive covariate in the regression model. In this study, we consider bivariate sequential time-to-event data, which consists of gap times between two events observed in sequence, as the outcome variables. The objective of this study is to investigate efficient two-phase sampling designs for a predetermined phase two sample size. We consider sampling designs depending on the event indicators and gap times. A likelihood-based method is used to estimate the associations between the expensive covariate and the two gap times. We show that when the selection probability at phase two depends on the two observed gap times and censoring times in addition to their event indicators, the efficiency of the design might improve.

Lay summary

In some observational studies, the explanatory variable might be expensive to measure although the outcome variable could easily be obtained. It is prohibitive to assess the explanatory variable on all the subjects of a large study and cost-efficient study designs are desirable in this situation. One solution is two-phase outcome-dependent sampling design. In phase one, we measure the outcome variable for all the subjects. In phase two, we select a subset of the subjects based on the outcome variable and measure the expensive explanatory variable only for the selected subjects.

Case-cohort design is a commonly used outcome-dependent sampling design in survival analysis. Survival data usually consists of the time until an event of interest occurs and the censoring information for each subject. Generalized case-cohort design select a random subsample of the subjects who experienced the event along with a random subsample of those with censored event times. Its selection probability at phase two depends only on the event indicator. It was previously shown that when the selection probability at phase two depends on observed event time and censoring time in addition to the event indicator, the efficiency of the design might increase. Efficient design has a lower variance of the coefficient estimate of the expensive explanatory variable in the regression model.

In this study, we consider bivariate sequential time-to-event data as the outcome variables. It consists of gap times between two events observed in sequence. The objective of this study is to investigate efficient two-phase sampling designs for a predetermined phase two sample size. We consider sampling designs depending on the event indicators and gap times. A likelihood-based method is used to estimate the associations between the expensive explanatory variable and the two gap times. We show that when the selection probability at phase two depends on the two observed gap times and censoring times in addition to their event indicators, the efficiency of the design might improve.

Acknowledgements

I am greatly indebted to my supervisor Dr. Yildiz Yilmaz for leading me to the research field in this thesis. This work would not have been completed without her guidance, advice, and encouragement. It was indeed a great privilege for me to work under her supervision. I will remember her enduring patience and constant support.

I sincerely acknowledge the financial support provided by the School of Graduate Studies, the Department of Mathematics and Statistics, and Dr. Yildiz Yilmaz in the form of Graduate Fellowship, Teaching Assistantship, and Research Grant, respectively. Dr. Yildiz Yilmaz is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and the Faculty of Medicine, Memorial University of Newfoundland. Further, I would like to thank the Department of Mathematics and Statistics for providing me a friendly atmosphere and the necessary facilities to complete my Master of Science degree in Statistics.

Statement of contribution

Dr. Yildiz Yilmaz proposed the research question that was investigated throughout this thesis. The overall study was jointly designed by Dr. Tzuemn-Renn Lin and Dr. Yildiz Yilmaz. The algorithms were implemented, the simulation study was conducted and the manuscript was drafted by Dr. Tzuemn-Renn Lin. Dr. Yildiz Yilmaz supervised the study and contributed to the final manuscript.

Table of contents

List of tables

List of figures

List of abbreviations

BSS	Basic stratified sampling
p.d.f.	probability density function
SRS	Simple random sampling

Chapter 1

Introduction

In some observational studies, the covariates of interest might be expensive to measure although the outcome variable could easily be obtained. To reduce the cost and to achieve a pre-specified power of the test for association of the expensive covariate with the outcome variable, cost-efficient designs and procedures are desirable for studies with a limited budget. An outcome-dependent sampling scheme is a cost-efficient design in which a subset of the cohort is selected based on the outcome variable, which has been collected for the entire cohort. In a two-phase outcome-dependent sampling design, all members of the cohort are measured for the outcome variable and inexpensive covariates at phase one. Then at phase two, a subset of the cohort is selected based on the outcome variable (and inexpensive covariates) obtained at phase one and the expensive covariate is measured for the selected individuals (Neyman, 1938; Zhao and Lipsitz, 1992). The key advantage of outcome-dependent sampling designs is that it allows researchers to concentrate budgetary resources on observations with the greatest amount of information. In comparison to using the entire cohort, outcome-dependent sampling incurs some loss of efficiency to detect association between the outcome variable and the expensive covariate. However, by selecting an informative subset of individuals from an existing cohort, it is generally more efficient than simple random sampling (SRS) of the same number of individuals (Yilmaz and Bull, 2011; Zhou et al., 2002, 2007).

The outcome variable which is of interest in this study is a continuous time-to-event (i.e. survival time or failure time) subject to censoring. Consider a cohort of individuals followed up for an outcome of interest. The cases are those individuals who experienced the event of interest during the follow-up period. The non-cases are those individuals who did not experience the event of interest in the follow-up period and have a right censored time. Two commonly used outcome-dependent sampling designs for time-to-event data are nested case-control design (Thomas, 1977) and case-cohort design (Prentice, 1986). Case-cohort designs typically select all cases for phase two, along with a random subsample of non-cases. Thus, the case-cohort designs are useful for large-scale cohort studies with low event rate. When the event rate is not low, to reduce the cost, generalized case-cohort designs could be used where only a subsample of cases are selected for phase two, along with a random subsample of non-cases. Another design approach is outcome-dependent basic stratified sampling (BSS) for cases where all cases are partitioned into strata based on survival times (e.g. stratum of low, middle or high survival time) and a random sample of specified size is selected from each stratum (Ding et al., 2014). Related designs include outcome-dependent BSS for non-cases where all non-cases are partitioned into strata based on censoring times (e.g. stratum of low, middle or high censoring time) and a random sample of specified size is selected from each stratum (Lawless, 2018).

Sequential time-to-event data consists of a sequence of survival times $T_1, T_2, ...$ that represent the times between a specified series of events with T_1 being the time to the first event and T_j ($j = 2, 3, ..$) being the time between the $(j-1)$-th and j-th events. For a repairable system where maintenance actions can be taken to restore system components when they fail, for example, T_j $(j = 2, 3, ..)$ could be the time between the $(j-1)$-th and j-th failures. In these circumstances the survival time T_j $(j = 2, 3, ..)$ can be observed only if T_1,...,T_{j-1} have already been observed. Bivariate sequential time-to-event data consists of two gap times T_1 and T_2 observed in sequence, and a right censoring time (i.e. total followup time) C. For a cancer patient, for example, T_1 could be the time from cancer diagnosis to cancer recurrence, and T_2 be the time from cancer recurrence to death.

The objective of this study is to investigate efficient two-phase outcome-dependent sampling designs with bivariate sequential time-to-event data for a predetermined phase two sample size. We consider sampling designs depending on the event indicators and gap times. A likelihood-based method is used to estimate the associations between the expensive covariate and the two gap times. We show that when the selection probability at phase two depends on the two observed gap times and censoring times in addition to their event indicators, the efficiency of the design might improve compared to a generalized case-cohort design.

The layout of Chapter 1 is as follows. In Section 1.1, we first present some survival data notation. Some common parametric models, regression models and estimation methods for analysis of survival data are introduced. In Section 1.2, we set up the notation for bivariate sequential survival data. After giving the likelihood function of observed bivariate sequential data, we then introduce copula models for bivariate sequential survival data. In Section 1.3, we define what the two-phase outcome-dependent sampling is and introduce estimation methods for two-phase outcome-dependent sampling. We also define nested case-control design and case-cohort design, which are two examples of outcome-dependent sampling with the time to event of interest as the outcome variable. In Section 1.4, we set up the objectives of the study. Section 1.5 is the outline of the thesis.

1.1 Survival data analysis

Survival analysis considers methods for analyzing data where the outcome variable is a time-to-event. Examples of time-to-event are time from birth to cancer diagnosis, time from cancer diagnosis to cancer recurrence, time from cancer recurrence to death, time from disease onset to death, and time from entry to a study to relapse (Cox and Oakes, 1984; Fleming and Harrington, 1991; Kalbfleisch and Prentice, 2002; Lawless, 2003).

1.1.1 Basic concepts

Survival time

Survival time is the length of time that is measured from time origin to the time the event of interest occurred. It is important to precisely define the time origin and what the event is. Also, the scale for measuring the passage of time must be agreed. Survival time is also called failure time, or time-to-event.

Distribution functions of survival time

Let T be a continuous time-to-event. More precisely, T is a continuous nonnegative random variable from a homogeneous population. Let $f(t)$ denote the probability density function (p.d.f.) of T and let the cumulative distribution function be

$$F(t) = P(T \leq t) = \int_0^t f(u)du.$$

The probability of an individual experiencing the event after time t is given by the survival function

$$S(t) = P(T > t) = \int_t^\infty f(u)du. \tag{1.1}$$

Note that $S(t)$ is a monotone non-increasing continuous function with $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$.

A very important concept with time-to-event distributions is the hazard function $h(t)$, also known as the hazard rate,

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t \le T < t + \Delta t | T \ge t]}{\Delta t}. \tag{1.2}$$

The hazard function specifies the instantaneous rate of an individual experiencing the event at time t, given that the individual does not experience the event up to time t.

It is also useful to define the cumulative hazard function

$$H(t) = \int_0^t h(u)du, \tag{1.3}$$

which is the accumulated hazard up until time t.

The functions $f(t)$, $F(t)$, $S(t)$, $h(t)$, and $H(t)$ uniquely specify the distribution of T. The hazard function $h(t)$ in (1.2) could be written as

$$h(t) = \frac{f(t)}{S(t)}.$$

Then, the survival function could be written in terms of the hazard function as

$$S(t) = \exp[-\int_0^t h(u)du] = \exp[-H(t)]. \tag{1.4}$$

The above arguments also lead to the following expression of the p.d.f. $f(t)$ in terms of the hazard function $h(t)$ and the cumulative hazard function $H(t)$ as

$$f(t) = h(t)\exp[-H(t)]. \tag{1.5}$$

Right censoring

One important concept in survival analysis is censoring. There are various types of censoring, such as right censoring where the individual's time-to-event is known only to exceed a certain value, left censoring where all that is known is that the individual has experienced the event of interest prior to a certain value, and interval censoring where the only information is that the event occurs within some interval. Right censoring is the most common type of censoring. It can occur for various reasons. In life sciences, this might happen when the follow-up of individuals ends before the events of all individuals are observed, or due to a random process, for example, a person might drop out of a study, or for long-term studies, the patient might be lost to follow up.

Suppose that N individuals have survival times represented by random variables $T_1, ..., T_N$. The type I censoring mechanism is said to apply when each individual has a fixed potential censoring time $C_i > 0$, $i = 1, ..., N$, such that T_i is observed if $T_i \leq C_i$; otherwise, we know only that $T_i > C_i$. Type I censoring often arises when a study is conducted over a specified time period.

In medical datasets, in addition to type I censoring, random censoring is also commonly observed. Random censoring arises when other competing events not related with the event of interest cause subjects to be removed from the study. For example, patient withdrawal from a clinical trial, death due to some cause other than the one of interest, or migration. A random censoring mechanism is said to apply when each individual has a survival time T and a censoring time C, with T and C independent continuous random variables. All survival times $T_1, ..., T_N$ and censoring times $C_1, ..., C_N$ are assumed mutually independent. As in the case of type I censoring, for $i = 1, ..., N$, T_i is observed if $T_i \leq C_i$; otherwise, we know only that $T_i > C_i$.

Survival data

Survival data usually consists of the time until an event of interest occurs and the censoring information for each individual.

For a specific individual i, $i = 1, ..., N$, under study, we assume that there is a survival time T_i and a right censoring time C_i. The survival times $T_1, ..., T_N$ are assumed to be independent and identically distributed. The survival time T_i of an individual i, $i = 1, ..., N$, will be known if and only if the event is observed before the censoring time C_i (i.e., T_i is less than or equal to C_i). If T_i is greater than C_i, then the individual's survival time is censored at C_i.

The data from this experiment can be conveniently represented by pair of random variables (t_i, δ_i), $i = 1, ..., N$, where $t_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \le C_i)$. The event indicator δ_i indicates whether the observed survival time t_i corresponds to an event $(\delta_i = 1)$ or a censoring time $(\delta_i = 0)$. If the time-to-event is observed, then t_i is equal to T_i and if it is censored, then t_i is equal to C_i. Survival data might also include explanatory variables.

Likelihood function

Consider survival times T_i and right censoring times C_i for independent individuals $i = 1, ..., N$. Let $t_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \le C_i)$ be the observed survival times and their event indicators, respectively. Suppose the p.d.f. and survivor function of survival time T are $f(t)$ and $S(t)$, respectively, for $t \ge 0$. Assume that the censoring mechanism is non-informative. Then, the likelihood function of the data could be written as

$$L = \prod_{i=1}^{N} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}. \tag{1.6}$$

When there is a vector $Z' = (Z_1, ..., Z_p)$ of explanatory variables present, we denote the conditional survival time distributions given $Z = \mathbf{z}$ as $f(t|\mathbf{z})$, $S(t|\mathbf{z})$, and so on. The likelihood function L in (1.6) still apply with $f(t)$ and $S(t)$ replaced by $f(t|\mathbf{z})$ and $S(t|\mathbf{z})$, respectively.

1.1.2 Common Parametric Models for Survival Data

Various parametric families of models are available for the analysis of survival data. Among univariate models, a few distributions occupy a central position because of their demonstrated usefulness in a wide range of situations. Foremost in this category are the exponential, Weibull, log-normal, log-logistic, and gamma distributions. The Weibull distribution is the only continuous distribution that could be written in the form of an accelerated failure time model and a proportional hazards regression model.

Weibull distribution

If time-to-event variable T has a Weibull distribution, its hazard function is

$$h(t) = \lambda\gamma t^{(\gamma-1)}, \quad t > 0,$$

where $\lambda > 0$ is a scale parameter, and $\gamma > 0$ is a shape parameter. Its survival function is

$$S(t) = \exp[-\lambda t^{\gamma}], \quad t > 0,$$

and its p.d.f. is

$$f(t) = \lambda\gamma t^{(\gamma-1)} \exp[-\lambda t^{\gamma}], \quad t > 0.$$

The exponential distribution is a special case of the Weibull distribution when $\gamma = 1$.

It is sometimes useful to work with the logarithm of the survival times. If we take

$Y = \log(T)$, where T follows a Weibull distribution, then Y can be written as

$$Y = \mu + \sigma W,$$

where $\sigma = \gamma^{-1}$, $\mu = -(\log \lambda)/\gamma$ and W has the standard extreme value distribution.

1.1.3 Regression Models for Survival Data

Consider a survival time $T > 0$ and a vector $Z' = (Z_1, ..., Z_p)$ of explanatory variables associated with the survival time T. It is important to ascertain the relationship between the survival time T and the explanatory variables. Two modelling approaches to represent this relationship are commonly used: accelerated failure time model and proportional hazards regression model.

Accelerated failure time model

The first approach is analogous to the classical linear regression approach. In this approach, the natural logarithm of the survival time, $Y = \log(T)$, is modelled. This is the natural transformation made in linear models to convert positive variables to observations on the entire real line. A linear model is assumed for $Y = \log(T)$,

$$Y = \mu + \alpha' Z + \sigma W,$$

where μ is the intercept term, $\alpha' = (\alpha_1, ..., \alpha_p)$ is a vector of regression coefficients, $\sigma > 0$ is a scale parameter, and W is the error term. Common choices for the error term W include the standard normal distribution which yields a log-normal regression model, the extreme value distribution which yields a Weibull regression model, or a logistic distribution which yields a log-logistic regression model for the random variable T.

This model is called the accelerated failure time model. To see why this is so, let us define a baseline survival function $S_0(t)$ as the survival function of $\exp(\mu+\sigma W)$. That is, the survival function of $T = \exp(Y)$ when Z is a zero vector. Then, the survival function of T given Z becomes

$$\begin{aligned} S(t|Z) &= P[T > t|Z] \\ &= P[Y > \log(t)|Z] \\ &= P[\mu + \sigma W > \log(t) - \alpha' Z|Z] \\ &= P[\exp(\mu + \sigma W) > t\exp(-\alpha' Z)|Z] \\ &= S_0(t\exp(-\alpha' Z)). \end{aligned}$$

The effect of the explanatory variables in the original time scale is to change the time scale by a factor $\exp(-\alpha' Z)$. Depending on the sign of $\alpha' Z$, the time is either accelerated by a constant factor or degraded by a constant factor.

Note that the hazard function of an individual with covariate vector Z for this class of models is related to a baseline hazard function h_0, that is the hazard function of $T = \exp(Y)$ when Z is a zero vector, by

$$h(t|Z) = h_0[t\exp(-\alpha' Z)]\exp(-\alpha' Z). \tag{1.7}$$

Proportional hazards regression model

Another approach to modelling the effects of covariates on survival time is to model the conditional hazard function of time-to-event given the covariate vector Z as a product of a baseline hazard function $h_0(t)$ and a non-negative function of the covariates, $\phi(\beta' Z)$. That is,

$$h(t|Z) = h_0(t)\phi(\beta' Z), \tag{1.8}$$

where $\beta' = (\beta_1, ..., \beta_p)$ is a vector of regression coefficients. This model is called the multiplicative hazard function model. In applications of the model, $h_0(t)$ may have a specified parametric form or it may be left as an arbitrary nonnegative function. Any nonnegative function can be used for the link function $\phi(\cdot)$. Most applications use the proportional hazards regression model with $\phi(\beta' Z) = \exp(\beta' Z)$ which is chosen for its simplicity and for the fact that it is positive for any value of $\beta' Z$. The name proportional hazards comes from the fact that any two individuals have hazard functions that are constant multiples of one another over time.

Note that the conditional survival function of time-to-event given the covariate vector Z can be expressed in terms of a baseline survival function $S_0(t)$ as

$$S(t|Z) = S_0(t)^{\phi(\beta' Z)}.$$

Weibull regression model

Consider an accelerated failure time model

$$Y = \mu + \alpha' Z + \sigma W,$$

where μ is the intercept term, $\alpha' = (\alpha_1, ..., \alpha_p)$ is a vector of regression coefficients, $\sigma > 0$ is a scale parameter, and W has the extreme value distribution. When Z is zero, we obtain $Y = \mu + \sigma W$ and $T = \exp(Y) = \exp(\mu + \sigma W)$ has a Weibull distribution with the hazard function

$$h_0(t) = \lambda \gamma t^{(\gamma - 1)}, \quad t > 0,$$

where $\lambda = \exp(-\mu\gamma) > 0$ is a scale parameter, and $\gamma = \sigma^{-1} > 0$ is a shape parameter. From the equation (1.7), the hazard function of an individual with covariate vector Z for this class of models is related to a baseline hazard function h_0 by

$$\begin{aligned} h(t|Z) &= h_0[t\exp(-\alpha' Z)]\exp(-\alpha' Z) \\ &= \lambda\gamma[t\exp(-\alpha' Z)]^{(\gamma-1)}\exp(-\alpha' Z) \\ &= \lambda\gamma t^{(\gamma-1)}[\exp(-\alpha' Z)]^{\gamma} \\ &= h_0(t)\exp(-\gamma\alpha' Z) \end{aligned}$$

which is the proportional hazards regression model given in (1.8) with $\phi(\beta' Z) = \exp(\beta' Z) = \exp(-\gamma\alpha' Z)$, where $\beta' = -\gamma\alpha'$, and the baseline hazard function $h_0(t) = \lambda\gamma t^{(\gamma-1)}$ is the hazard function of the Weibull distribution. The Weibull distribution is the only continuous distribution which has the property of being both an accelerated failure time model and a proportional hazards regression model.

1.1.4 Estimation Methods for Survival Data

It is important to ascertain the relationship between the survival time T and explanatory variables $Z' = (Z_1, ..., Z_p)$. This can be achieved through modelling how $Z' = (Z_1, ..., Z_p)$ is associated with T through for example, the hazard function $h(t|Z)$. However, an initial analysis would typically employ nonparametric methods to estimate the survival function and summary statistics, and a comparison across several groups based on some explanatory variables.

Nonparametric Methods

When there is no covariate, survival data are conveniently summarized through the Kaplan-Meier estimate of the survival function $S(t)$ and the Nelson-Aalen estimate of the cumulative hazard function $H(t)$ (e.g. Lawless, 2003, Section 3.2). These methods are said to be nonparametric since they require no assumptions about the distribution of survival time.

Let (t_i, δ_i), $i = 1, ..., n$, be a sequence of survival data. Suppose that there are k $(k \le n)$ distinct times $t_{(1)} < t_{(2)} < ... < t_{(k)}$ at which events of interest occur. For $j = 1, ..., k$, let $d_j = \sum_{i=1}^{n} I(t_i = t_{(j)}, \delta_i = 1)$ be the number of events at $t_{(j)}$ and $r_j = \sum_{i=1}^{n} I(t_i \ge t_{(j)})$ be the number of individuals at risk at $t_{(j)}$. That is, r_j is the number of individuals who have not experienced the event and uncensored just prior to $t_{(j)}$.

The Kaplan-Meier estimate (Kaplan and Meier, 1958) of $S(t)$ is defined as

$$\hat{S}(t) = \prod_{j:t_{(j)}<t} \frac{r_j - d_j}{r_j}$$

which can be derived as a nonparametric maximum likelihood estimate of the survival function $S(t)$. An estimate of its variance is given by

$$\widehat{\text{Var}}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{j:t_{(j)}<t} \frac{d_j}{r_j(r_j - d_j)}$$

which is called the Greenwood's formula.

The Nelson-Aalen estimate of $H(t)$ is defined as

$$\tilde{H}(t) = \sum_{j:t_{(j)}<t} \frac{d_j}{r_j}$$

with an estimated variance

$$\widehat{\text{Var}}[\tilde{H}(t)] = \sum_{j:t_{(j)}<t} \frac{d_j}{r_j^2}.$$

An alternative variance estimate is given by

$$\widehat{\text{Var}}[\tilde{H}(t)] = \sum_{j:t_{(j)}<t} \frac{d_j(r_j - d_j)}{r_j^3}.$$

Parametric Methods

In the analysis of survival data, some modelling approaches such as accelerated failure time model and proportional hazards regression model are commonly used, and some specific time-to-event distributions such as exponential distribution, Weibull distribution, log-normal distribution, log-logistic, gamma distribution are frequently used. Statistical inference for parametric models are based on maximum likelihood methodology (e.g. Lawless, 2003).

Consider a parametric model for survival time T given $Z = \mathbf{z}$ with a $p \times 1$ parameter vector $\theta = (\theta_1, ..., \theta_p)'$. The likelihood function $L(\theta)$ of the observed data $\{(t_i, \delta_i, \mathbf{z}_i) : i = 1, ..., N\}$ could be written as in equation (1.6) with $f(t)$ and $S(t)$ replaced by $f(t|\mathbf{z};\theta)$ and $S(t|\mathbf{z};\theta)$, respectively. The maximum likelihood estimates $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_p)'$ of the unknown parameters $\theta = (\theta_1, ..., \theta_p)'$ are obtained simultaneously by maximizing the likelihood function $L(\theta)$. If $l(\theta)$ denotes the natural logarithm of $L(\theta)$, the score equations

$$U_{\theta_j}(\theta) = \frac{\partial l(\theta)}{\partial \theta_j} = 0, \quad j = 1, ..., p$$

are solved simultaneously to get the maximum likelihood estimates $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_p)'$ of $\theta = (\theta_1, ..., \theta_p)'$. Under regularity conditions and assuming that the model is correct, $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_p)'$ are consistent estimators of the true values $\theta = (\theta_1, ..., \theta_p)'$ and $\sqrt{N}(\hat{\theta} - \theta)$ is asymptotically distributed as $\mathrm{N}_p[\mathbf{0}, J^{-1}(\theta)]$ where

$$J(\theta) = E[-\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta'}]$$

is the Fisher information matrix.

Semiparametric Methods

The most frequently used semiparametric regression model for the analysis of survival data with covariates is the Cox proportional hazards regression model which takes the hazard function for survival time T given $p \times 1$ vector of fixed covariates $\mathbf{z}$ to be of the form

$$h(t|\mathbf{z}) = h_0(t)\exp(\beta'\mathbf{z}),$$

where $h_0(t)$ is an arbitrary baseline hazard function and β is a $p \times 1$ vector of regression coefficients. Note that the conditional survival function of time-to-event given covariate vector $\mathbf{z}$ can be expressed in terms of a baseline survival function $S_0(t)$ as

$$S(t|\mathbf{z}) = S_0(t)^{\exp(\beta'\mathbf{z})}.$$

Given the observed data $\{(t_i, \delta_i, \mathbf{z}_i) : i = 1, ..., N\}$, we want to estimate β and $S_0(t)$ (e.g. Lawless, 2003, Section 7.1).

Suppose there are k ($k \leq N$) distinct observed times $t_{(1)} < t_{(2)} < ... < t_{(k)}$. For $j = 1, ..., k$, let $R_j = R(t_{(j)})$ denote the risk set at $t_{(j)}$ which is the set of individuals who are at risk and uncensored just prior to time $t_{(j)}$. For $i = 1, ..., N$, let $Y_i(t) = I(t_i \geq t)$ be the risk indicator function which indicates whether individual i is at risk and uncensored just prior to time t. Notice that $Y_i(t_{(j)}) = 1$ if and only if $i \in R_j$.

Cox (1972) suggested the following partial likelihood function for estimating β:

$$L(\beta) = \prod_{i=1}^{N} \left(\frac{\exp(\beta'\mathbf{z}_i)}{\sum_{l=1}^{N} Y_l(t_i)\exp(\beta'\mathbf{z}_l)} \right)^{\delta_i}.$$

Although the likelihood function $L(\beta)$ is not a full likelihood in the usual sense, maximization of $L(\beta)$ yields an estimate $\hat{\beta}$ which is consistent and asymptotically normally distributed under suitable conditions, and score, information, and likelihood ratio statistics based on $L(\beta)$ behave as though it is an ordinary likelihood.

The Breslow estimate of baseline cumulative hazard function $H_0(t)$ is defined as

$$\hat{H}_0(t) = \sum_{i:t_i \le t} \left\{ \frac{\delta_i}{\sum_{l=1}^{N} Y_l(t_i) \exp(\hat{\beta}' \mathbf{z}_l)} \right\}$$

which becomes the Nelson-Aalen estimator $\tilde{H}_0(t)$ when $\hat{\beta} = 0$.

A simple way to estimate $S_0(t)$ is to exploit the relationship $S_0(t) = \exp[-H_0(t)]$ and define the Fleming-Harrington estimator of baseline survival function $S_0(t)$ as

$$\hat{S}_0(t) = \exp[-\hat{H}_0(t-)]$$

where $\hat{H}_0(t-) = \lim_{\Delta t \to 0^+} \hat{H}_0(t - \Delta t)$ is the left limit of $\hat{H}_0(t)$.

1.2 Sequential survival data analysis

Multivariate survival data arise commonly in biomedical research, clinical trials and epidemiological studies. Different from univariate survival data analysis, multivariate survival data analysis typically deals with various dependence structures among survival times within same subjects or clusters.

Multivariate survival data includes parallel clustered data in which each subject has more than one survival time which are observed in parallel or simultaneously and do not satisfy any order restrictions; for example, times to occurrence of a disease in paired organs within individual or times to disease onset or death in related individuals.

Multivariate survival data also arises when there is a sequence of survival times $T_1, T_2, ...$ that represent the times between a specified series of events with T_1 being the time to the first event and T_j $(j = 2, 3, ..)$ being the time between the $(j-1)$-th and

j-th events. For example, times between repeat admissions to a psychiatric facility or time to cancer recurrence from cancer diagnosis and time from cancer recurrence to death for cancer patients.

1.2.1 Bivariate survival time model

We focus for now on the case of bivariate survival times (e.g., Lawless, 2003; Yilmaz and Lawless, 2011). Suppose T_1 and T_2 are two survival times of an individual which may not be independent. The bivariate distribution function and survivor function for $T_1 \geq 0$ and $T_2 \geq 0$ are defined as

$$F(t_1, t_2) = P(T_1 \leq t_1, T_2 \leq t_2) \tag{1.9}$$

and

$$S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2), \tag{1.10}$$

respectively.

For continuous survival times T_1 and T_2, the bivariate survivor function can be expressed in terms of the distribution function as follow:

$$S(t_1, t_2) = 1 - F_1(t_1) - F_2(t_2) + F(t_1, t_2) \tag{1.11}$$

where $F_1(t_1) = F(t_1, \infty)$ and $F_2(t_2) = F(\infty, t_2)$ are the marginal distribution functions of T_1 and T_2, respectively. The marginal survivor functions of T_1 and T_2 are $S_1(t_1) = S(t_1, 0)$ and $S_2(t_2) = S(0, t_2)$, respectively.

1.2.2 Likelihood function

Likelihood function for parallel clustered data

In the case of parallel clustered data, for a specific individual or cluster under study, we assume that there are bivariate survival times (T_1, T_2) and potential right censoring times (C_1, C_2). There are four different types of observations:

1. neither T_1 nor T_2 is observed, i.e. $t_1 = C_1$ and $t_2 = C_2$;
2. $t_1 = T_1$ is observed but T_2 is not observed, i.e. $t_2 = C_2$;
3. $t_2 = T_2$ is observed but T_1 is not observed, i.e. $t_1 = C_1$;
4. both $t_1 = T_1$ and $t_2 = T_2$ are observed.

The data from this study can be conveniently represented by $(t_1, t_2) = (\min(T_1, C_1), \min(T_2, C_2))$ and $(\delta_1, \delta_2) = (I[T_1 = t_1], I[T_2 = t_2])$ which are the observed survival times and their event indicators for a cluster, respectively.

Suppose the sequence of bivariate survival times (T_{1i}, T_{2i}) of a random sample of independent clusters $i = 1, ..., N$ have common continuous joint survivor function $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$. Let (C_{1i}, C_{2i}) denote the potential right censoring times for cluster i, $i = 1, ..., N$. Assume that (C_{1i}, C_{2i}) is independent of the survival times (T_{1i}, T_{2i}), $i = 1, ..., N$. Let $(t_{1i}, t_{2i}) = (\min(T_{1i}, C_{1i}), \min(T_{2i}, C_{2i}))$ and $(\delta_{1i}, \delta_{2i}) = (I[T_{1i} = t_{1i}], I[T_{2i} = t_{2i}])$ be the observed survival times and their event indicators, respectively. Then the likelihood function is (Lawless, 2003)

$$L = \prod_{i=1}^{N} \left[\frac{\partial^2 S(t_{1i}, t_{2i})}{\partial t_{1i} \partial t_{2i}} \right]^{\delta_{1i}\delta_{2i}} \left[-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{1i}} \right]^{\delta_{1i}(1-\delta_{2i})} \times \left[-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{2i}} \right]^{(1-\delta_{1i})\delta_{2i}} [S(t_{1i}, t_{2i})]^{(1-\delta_{1i})(1-\delta_{2i})}. \quad (1.12)$$

Likelihood function for sequential survival data

In the case of sequential survival data, for a specific individual under study, we assume that there are two survival times T_1 and T_2 observed in sequence, and a right censoring time (total followup time) C. There are three different types of observations:

1. T_1 is not observed, i.e. $t_1 = C$;

2. $t_1 = T_1$ is observed but T_2 is not observed, i.e. $t_2 = C - t_1$;

3. both $t_1 = T_1$ and $t_2 = T_2$ are observed.

The observed sequential survival times and their event indicators for a subject are $(t_1, t_2) = (\min(T_1, C), \min(T_2, C - t_1))$ and $(\delta_1, \delta_2) = (I[T_1 = t_1], I[T_2 = t_2])$, respectively.

Suppose the sequence of survival times (T_{1i}, T_{2i}), observed in order, of a random sample of independent individuals $i = 1, ..., N$ have common continuous joint distribution function $F(t_1, t_2) = P(T_1 \leq t_1, T_2 \leq t_2)$. Let C_i denote the potential right censoring time (total followup time) for individual i, $i = 1, ..., N$. Assume that C_i is independent of the survival time $T_{1i}+T_{2i}$, $i = 1, ..., N$. Let $(t_{1i}, t_{2i}) = (\min(T_{1i}, C_i), \min(T_{2i}, C_i - t_{1i}))$ and $(\delta_{1i}, \delta_{2i}) = (I[T_{1i} = t_{1i}], I[T_{2i} = t_{2i}])$ be the observed survival times and their event indicators, respectively. Then, the likelihood function (Lawless, 2003) is

$$L = \prod_{i=1}^{N} \left[\frac{\partial^2 F(t_{1i}, t_{2i})}{\partial t_{1i} \partial t_{2i}} \right]^{\delta_{1i}\delta_{2i}} \left[\frac{\partial F_1(t_{1i})}{\partial t_{1i}} - \frac{\partial F(t_{1i}, t_{2i})}{\partial t_{1i}} \right]^{\delta_{1i}(1-\delta_{2i})} [1 - F_1(t_{1i})]^{1-\delta_{1i}} \quad (1.13)$$

where $F_1(t_1) = F(t_1, \infty)$ is the marginal distribution function of T_1.

When there is a vector $Z' = (Z_1, ..., Z_p)$ of explanatory variables present we denote the conditional survival time distributions given $Z = \mathbf{z}$ as $F(t_1, t_2|\mathbf{z})$, $S(t_1, t_2|\mathbf{z})$, $F_j(t_j|\mathbf{z})$, and so on. The likelihood functions (1.12) and (1.13) still apply when there are explanatory variables, with $F_1(t_{1i})$, $S(t_{1i}, t_{2i})$, and $F(t_{1i}, t_{2i})$ replaced by $F_1(t_{1i}|\mathbf{z})$, $S(t_{1i}, t_{2i}|\mathbf{z})$, and $F(t_{1i}, t_{2i}|\mathbf{z})$, respectively.

1.2.3 Copula models for sequential survival times

Copulas are functions used to construct a joint distribution function or survival function by combining the marginal distributions. Copula theory and different copula models are given in Joe (1997) and Nelsen (2006). A bivariate copula $C : [0,1]^2 \to [0,1]$ is a function $C(u_1, u_2)$ with the following properties. The margins of C are uniform: $C(u_1, 1) = u_1$, $C(1, u_2) = u_2$; C is a grounded function: $C(u_1, 0) = C(0, u_2) = 0$ and C is 2-increasing: $C(v_1, v_2) - C(v_1, u_2) - C(u_1, v_2) + C(u_1, u_2) \geq 0$ for all $(u_1, u_2) \in [0,1]^2$, $(v_1, v_2) \in [0,1]^2$ such that $0 \leq u_1 \leq v_1 \leq 1$ and $0 \leq u_2 \leq v_2 \leq 1$.

Sklar's theorem (Sklar, 1959) provides the theoretical foundation for the application of copulas. Let H be a two-dimensional distribution function with marginal distribution functions F and G. Then there exists a copula C such that

$$H(x, y) = C(F(x), G(y)). \tag{1.14}$$

Conversely, for any univariate distribution functions F and G and any copula C, the function H in (1.14) is a two-dimensional distribution function with marginals F and G. Furthermore, if F and G are continuous, then C is unique.

Copula models have some attractive properties such as the marginal distributions can come from any and different families, the dependence structure can be investigated separately from the marginal distributions since the measures of dependence do not appear in the marginal distributions, and copulas are invariant under strictly increasing transformations of the margins.

Archimedean copulas are commonly used. Copulas are called Archimedean when they are of the form

$$C(u_1, u_2) = \psi^{-1}[\psi(u_1) + \psi(u_2)]$$

where ψ is a decreasing convex function on $[0,1]$ satisfying $\psi(1) = 0$. The most important characteristic of bivariate Archimedean copulas is that all the information about the 2-dimensional dependence structure is contained in a univariate generator, ψ. Some fundamental properties of Archimedean copulas are given in Joe (1997, Section 4.2) and Nelson (2006, Section 4.3).

One frequently used one-parameter Archimedean copula is the Clayton copula which has the form

$$C_\phi(u_1, u_2) = (u_1^{-\phi} + u_2^{-\phi} - 1)^{-1/\phi}, \quad \phi > 0, \tag{1.15}$$

where ϕ is the dependence parameter. Its generator function is

$$\psi_\phi(t) = t^{-\phi} - 1.$$

We focus for now on the analysis of sequential survival data. For each individual under study, we assume that there are two survival times T_1 and T_2 observed in sequence. Then, by Sklar's theorem (Sklar, 1959), there exists a unique copula C such that for all $t_1, t_2 \geq 0$, the bivariate distribution function (1.9) becomes

$$F(t_1, t_2) = C(F_1(t_1), F_2(t_2)), \tag{1.16}$$

where $F_1(t_1) = F(t_1, \infty)$ and $F_2(t_2) = F(\infty, t_2)$ are the marginal distribution functions of T_1 and T_2, respectively. The likelihood function (1.13) is then written in terms of

$C(F_1(t_1), F_2(t_2))$ as

$$L = \prod_{i=1}^{N} \left[\frac{\partial^2 C(F_1(t_{1i}), F_2(t_{2i}))}{\partial t_{1i} \partial t_{2i}} \right]^{\delta_{1i}\delta_{2i}} \times \left[\frac{\partial F_1(t_{1i})}{\partial t_{1i}} - \frac{\partial C(F_1(t_{1i}), F_2(t_{2i}))}{\partial t_{1i}} \right]^{\delta_{1i}(1-\delta_{2i})} [1 - F_1(t_{1i})]^{1-\delta_{1i}}. \quad (1.17)$$

When there is a vector $Z' = (Z_1, ..., Z_p)$ of explanatory variables present we denote the marginal distribution functions of T_1 and T_2 given $Z = \mathbf{z}$ as $F_1(t_1|\mathbf{z})$ and $F_2(t_2|\mathbf{z})$, respectively. The likelihood function in (1.17) still apply with $F_1(t_1)$ and $F_2(t_2)$ replaced by $F_1(t_1|\mathbf{z})$ and $F_2(t_2|\mathbf{z})$, respectively.

Parametric Estimation

Suppose the marginal distribution functions of T_1 and T_2 given $Z = \mathbf{z}$ are $F_1(t_1|\mathbf{z}; \beta_1)$ and $F_2(t_2|\mathbf{z}; \beta_2)$, respectively, and the bivariate distribution function of (T_1, T_2) given $Z = \mathbf{z}$ is $F(t_1, t_2|\mathbf{z}) = C_\alpha(F_1(t_1|\mathbf{z}; \beta_1), F_2(t_2|\mathbf{z}; \beta_2))$, where β_1, β_2 and α are vectors of parameters. Let $\theta = (\beta_1', \beta_2', \alpha')'$. Then, the likelihood function $L(\theta)$ of the observed data $\{(t_{1i}, t_{2i}, \delta_{1i}, \delta_{2i}, \mathbf{z}) : i = 1, ..., N\}$ is written as in (1.17) with $F_1(t_1)$, $F_2(t_2)$ and $C(F_1(t_1), F_2(t_2))$ replaced by $F_1(t_1|\mathbf{z}; \beta_1)$, $F_2(t_2|\mathbf{z}; \beta_2)$ and $C_\alpha(F_1(t_1|\mathbf{z}; \beta_1), F_2(t_2|\mathbf{z}; \beta_2))$, respectively.

When analyzing the given observed data $\{(t_{1i}, t_{2i}, \delta_{1i}, \delta_{2i}, \mathbf{z}) : i = 1, ..., N\}$, the maximum likelihood estimate $\hat{\theta} = (\hat{\beta}_1{}', \hat{\beta}_2{}', \hat{\alpha}')'$ of the unknown parameters $\theta = (\theta_1, ..., \theta_p)' = (\beta_1', \beta_2', \alpha')'$ are obtained simultaneously by maximizing the likelihood function $L(\theta)$. Suppose $l(\theta)$ denotes the natural logarithm of $L(\theta)$, then the score equations

$$U_{\theta_j}(\theta) = \frac{\partial l(\theta)}{\partial \theta_j} = 0, \quad j = 1, ..., p$$

are solved simultaneously to get the maximum likelihood estimates $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_p)'$ of

$\theta = (\theta_1, ..., \theta_p)'$. Under regularity conditions and assuming that the model is correct, $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_p)'$ are consistent estimators of the true values $\theta = (\theta_1, ..., \theta_p)'$ and $\sqrt{N}(\hat{\theta} - \theta)$ is asymptotically distributed as $\mathrm{N}_p[\mathbf{0}, J^{-1}(\theta)]$ where

$$J(\theta) = E[-\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}]$$

is the Fisher information matrix.

1.3 Two-phase outcome-dependent sampling

1.3.1 Two-phase sampling

Two-phase sampling is a sampling technique that aims to reduce the cost of the study. It was originally introduced in survey sampling by Neyman (1938) for estimation of the finite population mean of a variable.

At phase one, a large sample is drawn from a population, and information on variables that are easier to measure is collected. These phase one variables may be important variables such as exposure in a regression model, or simply may be auxiliary variables that are correlated with unavailable variables at phase one. At phase two, a subsample is selected based on the values of the collected variables to obtain phase two variables that are costly or difficult to measure.

For example, the phase one sample can be stratified based on the values of the collected variables. At phase two, a subsample is drawn without replacement from each stratum to obtain phase two variables that are costly or difficult to measure. Strata formation seeks either to oversample subjects with important phase one variables, or to effectively sample subjects with targeted phase two variables, or both. This way, two-phase sampling achieves effective access to important variables with less cost.

1.3.2 Outcome-dependent sampling

An outcome-dependent sampling scheme is a retrospective sampling scheme where the expensive covariates are observed with a probability depending on the outcome variable. The principal idea of an outcome-dependent sampling design is to concentrate resources where there is the greatest amount of information. By allowing the selection probability of each individual in the outcome-dependent sample to depend on the outcome, the investigators attempt to enhance the efficiency and reduce the cost of the study (Zhou et al., 2002).

Nested case-control design and case-cohort design are two examples of outcome-dependent sampling designs which could be applied to survival data where the outcome variable is a time-to-event.

1.3.3 Estimation methods

Consider a two-phase outcome-dependent design to collect an expensive covariate data. Suppose that a finite population of N individuals has outcome values y_i, $i = 1, ..., N$ generated as independent realizations from a model $f(y|x;\theta)g(x)$. Here, Y is the outcome variable, X is the expensive covariate, $f(y|x;\theta)$ is the conditional p.d.f. of Y given $X = x$ and $g(x)$ is the marginal distribution of X. Let $G(x)$ denote the distribution function corresponding to $g(x)$. Since the covariate X is expensive to measure, two-phase sampling technique is used to reduce the cost. The observed data at phase one is $\{y_i : i = 1, ..., N\}$. At phase two, a subsample of size n is selected based on the values of the collected variables to obtain phase two variable that are costly or difficult to measure. An outcome-dependent sampling scheme is used at phase two to allow the selection probability of each individual in the finite population of N individuals to depend on the outcome variable. The estimation of θ is based on the fully observed data (y_i, x_i) of n individuals selected at phase two and might also be based on the not fully observed data y_i of $N - n$ individuals not selected for phase two. For a fixed given phase two sample size n, the goal is to enhance the efficiency by concentrating resources where there is the greatest amount of information.

Let $R_i = I(\text{individual } i \text{ is selected})$ be the indicator function for individual i being selected at phase two and let π_i denote the conditional inclusion probability $P(R_i = 1|x_i, y_i)$. We assume that the probability that individual i is selected at phase two does not depend on the expensive covariate. Therefore, $\pi_i = P(R_i = 1|x_i, y_i) = P(R_i = 1|y_i)$ and the expensive covariate X is missing at random for individuals that are not selected for phase two (Rubin, 1976). Suppose $V = \{i : R_i = 1, i = 1, ..., N\}$ denotes the set of individuals selected at phase two, where the size of V is n. Then $\bar{V} = \{i : R_i = 0, i = 1, ..., N\}$ is the set of individuals who are not selected, where the size of $\bar{V}$ is $N - n$.

Various estimating procedures have been proposed for data collected through a case-cohort study design. These have proceeded mainly along two lines, likelihood-based approaches and pseudolikelihood-based approaches (Lawless et al., 1999). Likelihood-based approaches can handle certain sampling schemes that other approaches may not, for example, schemes where some individuals have zero probability of selection for the phase two sample.

Full likelihood

The full likelihood function of the observed data $\{(y_i, x_i) : i \in V\} \cup \{y_i : i \in \bar{V}\}$ for the unknown parameters θ and G is proportional to

$$L_F(\theta, G) = \left(\prod_{i \in V} f(y_i|x_i; \theta) dG(x_i)\right)\left(\prod_{i \in \bar{V}} \int_x f(y_i|x; \theta) dG(x)\right). \tag{1.18}$$

Semiparametric maximum likelihood estimation based on (1.18) has been discussed by many authors (Lawless et al., 1999; Lawless, 2018; Zeng and Lin, 2014; Zhang and Rockette, 2005; Zhao et al., 2009). One approach is to maximize the likelihood function in (1.18) jointly with respect to θ and G. The estimation method becomes parametric when X is categorical (Wild 1991, Scott and Wild 1997) or when G is discrete with relatively few points of support (Hsieh et al., 1985). In these cases, maximum likelihood estimates of θ from the full likelihood L_F are regular maximum likelihood estimates and the usual large sample theory for maximum likelihood estimates applies subject to some regularity conditions (Lawless et al., 1999).

Conditional likelihood

Conditional likelihood is an alternative to the full likelihood. It is based on the conditional p.d.f. $f(y_i|x_i, R_i = 1; \theta)$ of Y given $X = x_i$ and being selected at phase two. Thus, the conditional likelihood is

$$L_C(\theta) = \prod_{i \in V} f(y_i|x_i, R_i = 1; \theta). \tag{1.19}$$

Weighted pseudolikelihood

Weighted pseudolikelihood is a pseudolikelihood-based method. It employs the Horvitz-Thompson approach in which we use the completely observed individuals only and weight their contributions inversely according to their probability of selection to give the log-pseudolikelihood function

$$l_W(\theta) = \sum_{i \in V} w_i \log f(y_i|x_i; \theta), \tag{1.20}$$

where $w_i = \pi_i^{-1}$ is the weight of individual i being selected at phase two. This approach should not be used under a sampling design where a selection probability is zero or close to zero for individual i. The Horvitz-Thompson approach is known to be inefficient (Robins et al., 1994). One reason is that it often ignores much of the information available for the cohort. One option is to modify the weights $w_i = \pi_i^{-1}$ using the double weighting method of Kulish and Lin (2004) or the calibration technique of Breslow et al. (2009) so that they better reflect the full cohort information.

1.3.4 Estimation methods for outcome-dependent BSS

Outcome-dependent BSS was considered by Imbens and Lancastes (1996) and Lawless et al. (1999). In a two-phase outcome-dependent sampling scheme, suppose that the phase one data y_i, $i = 1, ..., N$ is partitioned into K strata $S_1, ..., S_K$ based on continuous outcome variable Y using $(K-1)$ cut-off values $c_1 < c_2 < ... < c_{K-1}$ as shown in the following:

$$\begin{aligned} &\underbrace{y_{(1)} < ... < y_{(N_1)}}_{S_1} < c_1 < \underbrace{y_{(N_1+1)} < ... < y_{(N_1+N_2)}}_{S_2} \\ &\qquad < c_2 < ... < c_{K-1} < \underbrace{y_{(N_1+...+N_{K-1}+1)} < ... < y_{(N)}}_{S_K}, \end{aligned} \tag{1.21}$$

where N_j is the size of stratum S_j obtained under the defined cut-off values, $j = 1, ..., K$, and $\sum_{j=1}^{K} N_j = N$.

At phase two, a subsample is drawn without replacement from each stratum to obtain phase two variables that are costly or difficult to measure. BSS is a sampling scheme where a simple random sample of specified size n_j is selected from stratum

S_j, $j = 1, ..., K$ as shown in the following:

$$\underbrace{y_{(1)} < ... < y_{(N_1)}}_{n_1} < c_1 < \underbrace{y_{(N_1+1)} < ... < y_{(N_1+N_2)}}_{n_2}$$
$$< c_2 < ... < c_{K-1} < \underbrace{y_{(N_1+...+N_{K-1}+1)} < ... < y_{(N)}}_{n_K}, \tag{1.22}$$

where $\sum_{j=1}^{K} n_j = n$. The probability that individual i is sampled (selected) and fully observed is $p_j = n_j / N_j$, $j = 1, ..., K$.

Suppose $D_j = \{i : R_i = 1, i \in S_j\}$ denotes the set of individuals selected from stratum S_j, where the size of D_j is n_j. Then $\bar{D}_j = \{i : R_i = 0, i \in S_j\}$ is the set of individuals who are not selected from stratum S_j.

Under the outcome-dependent BSS, the full likelihood (1.18) becomes

$$L_F(\theta, G) = \prod_{j=1}^{K} \left[\left(\prod_{i \in D_j} f(y_i | x_i; \theta) dG(x_i) \right) \left(\prod_{i \in \bar{D}_j} \int_x f(y_i | x; \theta) dG(x) \right) \right]. \tag{1.23}$$

The weighted pseudolikelihood (1.20) becomes

$$l_W(\theta) = \sum_{j=1}^{K} p_j^{-1} \sum_{i \in D_j} \log(f(y_i | x_i; \theta)). \tag{1.24}$$

The use of $p_j = n_j / N_j$ provides an unbiased estimating equation for θ.

The conditional likelihood (1.19) becomes

$$L_C(\theta) = \prod_{j=1}^{K} \prod_{i \in D_j} \left[\frac{p_j f(y_i | x_i; \theta)}{\sum_{l=1}^{K} p_l Q_l^*(x_i; \theta)} \right], \tag{1.25}$$

where

$$Q_l^*(x, \theta) = P(Y \in S_l | x; \theta).$$

The log-pseudolikelihood function arising from equation (1.25) is

$$l_C(\theta) = \sum_{j=1}^{K} \sum_{i \in D_j} \left[\log\{f(y_i|x_i;\theta)\} - \log\left\{ \sum_{l=1}^{K} p_l Q_l^*(x_i;\theta) \right\} \right]. \tag{1.26}$$

The use of $p_j = n_j/N_j$ provides an unbiased estimating equation for θ. In other words, under BSS with the stratum-specific sampling probabilities $p_j = n_j/N_j$ pre-specified, it can be shown that the score function corresponding to equation (1.26),

$$S_C(\theta) = \frac{\partial l_C}{\partial \theta},$$

provides an unbiased estimating equation for θ.

1.3.5 Nested case-control design

The nested case-control design was originally suggested by Thomas (1977). See also Prentice and Breslow (1978). The nested case-control design is an extension of a case-control study to a survival analysis setting in which the outcome of interest is a time-to-event, and in general, the focus is on making inference on whether the time-to-event is associated with exposures of interest (e.g. Keogh and Cox, 2014, Chapter 7).

Consider a cohort of individuals followed up for an outcome of interest. The cases are those individuals who experienced the event of interest during the follow-up period. Individuals who did not experience the event of interest have a right censored time. The main steps for selecting a nested case-control sample are as follows:

1. Cases are identified within the cohort at the time at which they are observed to experience the event of interest. Often all cases observed during a particular period of follow-up are selected.

2. At a given event time, the risk set is the set of individuals who were eligible to experience the event at that time, that is, who will remain in the cohort, have not yet experienced the event just prior to the observed event time and have not been censored.

3. We identify the risk set at each case's event time and take a sample of one or more individuals from the corresponding risk set. We refer to these individuals the control set for that case. Under the standard nested case-control design, at each event time the controls are selected randomly from the risk set, excluding the case itself.

1.3.6 Case-cohort design

The case-cohort design was originally suggested by Prentice (1986). The case-cohort design is an alternative to the nested case-control design.

Consider a cohort of individuals followed up for an event of interest. The cases are those individuals who experienced the event of interest during the follow-up period. The main steps for selecting a case-cohort sample are as follows:

1. A set S of individuals called the subcohort is sampled at random and without replacement from the cohort at the start of the follow-up period.

2. Because the subcohort S is a random sample from the cohort, it will typically contain some cases of the event of interest.

3. A case-cohort sample thus consists of the subcohort S plus all additional cases observed in the cohort.

The key idea of this study design is to obtain the measurements of primary exposure variables only on a subset of the entire cohort (subcohort) and all the individuals who experienced the event of interest (cases) in the cohort. Thus, the case-cohort study design is particularly useful for large-scale cohort studies with a low event (e.g. disease) rate if a limited number of individuals is needed to be selected.

The requirement of sampling all the cases in the original case-cohort design will limit the application of case-cohort study designs if the event rate is not rare. To reduce the cost, a generalized case-cohort design is used where only a random sample from cases and a random sample from non-cases are selected.

1.4 Objectives of the study

P. Judd (2016) explored extensions of case-cohort sampling designs that result in more efficient sampling designs for univariate survival analysis. She found that balancing the number of cases and non-cases given a phase two sample size produce more efficient estimates under a generalized case-cohort design which is based on event indicator. When comparing sampling designs dependent on both survival time and event indicator, sample design efficiency improves if the cases with short survival times are assigned a higher selection probability. Similarly, sample design efficiency improves if the non-cases with long censoring times are assigned a higher selection probability (Judd, 2016).

Compared to other designs, efficient design has a lower variance of the coefficient estimate of the expensive covariate in the regression model. The objective of this study is to investigate efficient two-phase sampling designs with bivariate sequential survival data for a predetermined phase two sample size under the likelihood-based approach. Suppose we observed a cohort of bivariate sequential survival data of size N at phase one. A subsample of fixed size (n) will be drawn at phase two in order to obtain measurement of covariate X which is costly or difficult to measure. In Chapter 2, we will describe how to explore generalized case-cohort design and outcome-dependent BSS design that result in more efficient sampling designs using bivariate sequential survival data. In this study, we assume that the assumed model is correct and there is only one expensive covariate and no other covariates.

1.5 Outline of the thesis

The thesis is organized as follows.

In Chapter 2, we describe generalized case-cohort design and outcome-dependent BSS design for bivariate sequential survival data. A generalized case-cohort design can either based on first event indicator only or based on both first and second event indicators. An outcome-dependent BSS design can either based on time-to-first event and its event indicator only or based on both time-to-events and their event indicators. We will describe stratifications considered under outcome-dependent BSS designs.

In Chapter 3, we investigate the efficiency of the sampling designs described in Chapter 2 when there is a moderate dependence between the two gap times.

In Chapter 4, we investigate the efficiency of the sampling designs described in Chapter 2 when there is a high dependence between the two gap times.

Chapter 5 summarizes the study and give a brief discussion.

Chapter 2

Two-phase outcome-dependent sampling designs for bivariate sequential time-to-event data

Bivariate sequential time-to-event data consists of two gap times T_1 and T_2 observed in sequence, and a right censoring time (total followup time) C. In a cancer study, for example, T_1 could be the time from cancer diagnosis to cancer recurrence and T_2 be the time from cancer recurrence to death.

In some observational studies, the covariates of interest might be expensive to measure although the outcome variable could easily be obtained. Two-phase sampling is a sampling technique that aims to reduce the cost of the study. At phase one, a large sample is drawn from a population, and information on variables that are easier to measure is collected. At phase two, a subsample is selected based on the values of the collected variables to obtain phase two variables that are costly or difficult to measure. An outcome-dependent sampling scheme is a retrospective sampling scheme where the expensive exposure variables/covariates are observed with a probability depending on the outcome variable. The principal idea of an outcome-dependent sampling design is to concentrate resources where there is the greatest amount of information in order to enhance the efficiency of the design.

In this chapter, we will describe some two-phase outcome-dependent sampling designs for bivariate sequential survival data with a covariate which is costly or difficult to measure. Phase one data consists of bivariate sequential time-to-event data for a random sample or cohort of N individuals from a population. This phase one data can be stratified based on the event indicators and the survival times. A phase two sample of fixed size (n) is drawn based on the strata of phase one in order to obtain a covariate which is costly or difficult to measure. We will adopt the full likelihood-based approach to analyze the survival data which includes observations with complete and incomplete covariate data. The objective of this study is to investigate efficient two-phase sampling designs with bivariate sequential survival data for a predetermined phase two sample size. Compared to other designs, efficient design has a lower variance of the coefficient estimate of the expensive covariate in the regression model.

The layout of Chapter 2 is as follows. In Section 2.1, we describe four phase two sampling designs: (1) design based on first event indicator; (2) design based on time-to-first event and its event indicator; (3) design based on first and second event indicators; and (4) design based on first and second gap times and their event indicators. In Section 2.2, we first describe how to generate the phase one data. Using the generated data, we then describe the stratification based on time-to-event T_1 and its event indicator. Finally we describe the stratification based on first and second gap times and their event indicators.

2.1 Outcome-dependent sampling design

Suppose the gap times (T_{1i}, T_{2i}), observed in order for a random sample of independent individuals, $i = 1, ..., N$, have common joint continuous distribution function $F(t_1, t_2)$ and joint survivor function $S(t_1, t_2)$. Let C_i denote the potential right censoring time for individual i, $i = 1, ..., N$. Let X_i be a covariate for individual i. Assume that C_i is conditionally independent of the survival time $T_{1i} + T_{2i}$ given X_i. Let $(t_{1i}, t_{2i}) = (\min(T_{1i}, C_i), \min(T_{2i}, C_i - t_{1i}))$ and $(\delta_{1i}, \delta_{2i}) = (I[T_{1i} = t_{1i}], I[T_{2i} = t_{2i}])$ be the observed gap times and their event indicators, respectively.

When the covariate X_i is collected for each individual i, the observed data is $\{(t_{1i}, \delta_{1i}, t_{2i}, \delta_{2i}, x_i) : i = 1, ..., N\}$, and the likelihood function is given in (1.13) with $F_1(t_{1i})$ and $F(t_{1i}, t_{2i})$ replaced by $F_1(t_{1i}|x_i)$ and $F(t_{1i}, t_{2i}|x_i)$, respectively.

When the covariate X_i is expensive to measure, two-phase sampling technique could be used to reduce the cost. Then, the observed data at phase one is $\{(t_{1i}, \delta_{1i}, t_{2i}, \delta_{2i}) : i = 1, ..., N\}$. At phase two, in the outcome-dependent sampling, the phase one sample is then stratified based on these phase one variables. A generalized case-cohort design would be either based on the first event indicator only or the second event indicator only depending on the event of interest. In this study, we consider sampling design depending on both event indicators. An outcome-dependent BSS design can either be based only on time-to-first event and its event indicator or based on first and second gap times and their event indicators.

We will adopt the full likelihood-based approach to estimate the regression coefficient of the expensive covariate. For $i = 1, ..., N$, let us denote $L_i(x)$ as the contribution of the ith individual data $(t_{1i}, \delta_{1i}, t_{2i}, \delta_{2i}, x)$ in the likelihood function L in (1.13):

$$L_i(x) = \left[\frac{\partial^2 F(t_{1i}, t_{2i}|x)}{\partial t_{1i} \partial t_{2i}}\right]^{\delta_{1i}\delta_{2i}} \left[\frac{\partial F_1(t_{1i}|x)}{\partial t_{1i}} - \frac{\partial F(t_{1i}, t_{2i}|x)}{\partial t_{1i}}\right]^{\delta_{1i}(1-\delta_{2i})} [1 - F_1(t_{1i}|x)]^{1-\delta_{1i}}.$$

Let $g(x)$ be the marginal distribution of X, and $G(x)$ denote the distribution function corresponding to $g(x)$. Then the full likelihood function is defined by (1.18) with $f(y_i|x_i;\theta)$ and $f(y_i|x;\theta)$ replaced by $L_i(x_i)$ and $L_i(x)$, respectively. In particular, if the covariate X is binary following the Bernoulli distribution with probability of success p, then (1.18) becomes

$$L_F = \left(\prod_{i\in V} L_i(x_i)g(x_i)\right)\left(\prod_{i\in\bar{V}}\sum_{x=0}^{1} L_i(x)g(x)\right), \tag{2.1}$$

where $g(1) = p$ and $g(0) = 1 - p$.

2.1.1 Generalized case-cohort design based on the event indicator of the first gap time

Suppose the phase one cohort is stratified based on the event indicators δ_{1i}, $i = 1, ..., N$, of the first gap time T_1. The resulting strata are $S_{\text{cases}} = \{i : \delta_{1i} = 1, 1 \le i \le N\}$ and $S_{\text{noncases}} = \{i : \delta_{1i} = 0, 1 \le i \le N\}$ with size N_{cases} and N_{noncases}, respectively, where $N_{\text{cases}} + N_{\text{noncases}} = N$. A subsample of fixed size n is drawn at phase two in order to obtain the covariate X which is costly or difficult to measure. Suppose the size of the subsample from the case stratum S_{cases} is denoted by n_{cases} and the size of the subsample from the non-case stratum S_{noncases} is denoted by n_{noncases}, where $n_{\text{cases}} + n_{\text{noncases}} = n$. Given the fixed size n of subsample, different allocations $(n_{\text{cases}}, n_{\text{noncases}})$ define different generalized case-cohort designs based on T_1 event indicator. The aim is to identify the allocation $(n_{\text{cases}}, n_{\text{noncases}})$ which is the most efficient sampling design under the likelihood-based method. Efficient sampling design minimizes the variance of the coefficient estimate of the expensive covariate for the survival time T_1.

Let $R_i = I(\text{individual } i \text{ is selected})$ be the indicator function for individual i being selected at phase two. Suppose $D_{\text{cases}} = \{i : R_i = 1, i \in S_{\text{cases}}\}$ denotes the set of individuals selected from stratum S_{cases}, where the size of D_{cases} is n_{cases}. Similarly, suppose $D_{\text{noncases}} = \{i : R_i = 1, i \in S_{\text{noncases}}\}$ denotes the set of individuals selected from stratum S_{noncases}, where the size of D_{noncases} is n_{noncases}. Then $\bar{D}_{\text{cases}} = \{i : R_i = 0, i \in S_{\text{cases}}\}$ is the set of individuals who are not selected from stratum S_{cases} and $\bar{D}_{\text{noncases}} = \{i : R_i = 0, i \in S_{\text{noncases}}\}$ is the set of individuals who are not selected from stratum S_{noncases}. Therefore, the full likelihood function is (2.1) with $V = D_{\text{cases}} \cup D_{\text{noncases}}$ which is the set of individuals who are selected at phase two and $\bar{V} = \bar{D}_{\text{cases}} \cup \bar{D}_{\text{noncases}}$ which is the set of individuals who are not selected at phase two.

After obtaining the most efficient sampling design which is based on T_1 event indicator, we will next stratify the case stratum S_{cases} based on time-to-event T_1 and stratify the non-case stratum S_{noncases} based on censoring time C.

2.1.2 Outcome-dependent BSS design based on the first gap time and its event indicator

Recall that the phase one cohort can be stratified into the strata $(S_{\text{cases}}, S_{\text{noncases}})$ based on the event indicator of the first gap time T_1. For a fixed phase two sample size n, we can obtain the most efficient sampling design (n_{cases}, n_{noncases}) for the strata (S_{cases}, S_{noncases}) under the full likelihood-based approach for a given phase two sample size $n = n_{\text{cases}} + n_{\text{noncases}}$. A more efficient design could be achieved by selecting a more informative sample. In genetic association studies, budgetary constraints prevent genotyping all individuals in a cohort (Huang and Lin, 2007; Lin et al., 2013) and extreme sampling designs are being used since it is more efficient than simple random sampling of the same number of individuals (Yilmaz and Bull, 2011). For example, in Lin et al. (2013) in the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project, subjects with the highest or lowest values of body mass index, LDL, or blood pressure were selected for whole exome sequencing, and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) resequencing project adopted a one-tailed sampling design by selecting subjects with the highest values of a quantitative trait, along with a random sample. Also, Lawless (2018) compared extreme strata sampling designs with some others. Based on such studies, in this thesis we assessed the efficiency of different designs and tried to understand the efficiency gain under extreme strata sampling.

We can stratify all T_1 cases in S_{cases} into strata ($S_{\text{cases},1}$, $S_{\text{cases},2}$, $S_{\text{cases},3}$) based on time-to-event T_1 using two cut-off values $c_{L1} < c_{U2}$ which are defined in Section 2.2.2:

$$\underbrace{T_{1(1)} < \ldots < T_{1(N_{\text{cases},1})}}_{S_{\text{cases},1}} < c_{L1} < \underbrace{T_{1(N_{\text{cases},1}+1)} < \ldots < T_{1(N_{\text{cases},1}+N_{\text{cases},2})}}_{S_{\text{cases},2}} < c_{U1} < \underbrace{T_{1(N_{\text{cases},1}+N_{\text{cases},2}+1)} < \ldots < T_{1(N_{\text{cases},1}+N_{\text{cases},2}+N_{\text{cases},3})}}_{S_{\text{cases},3}}, \quad (2.2)$$

where $N_{\text{cases},j}$ is the size of stratum $S_{\text{cases},j}$, $j = 1, 2, 3$, and $\sum_{j=1}^{3} N_{\text{cases},j} = N_{\text{cases}}$.

Similarly, we can stratify all T_1 non-cases in S_{noncases} into strata ($S_{\text{noncases},1}$, $S_{\text{noncases},2}$, $S_{\text{noncases},3}$) based on their censoring times C_i using two cut-off values $c^*_{L1} < c^*_{U1}$ which are defined in Section 2.2.2:

$$\underbrace{C_{(1)} < \ldots < C_{(N_{\text{noncases},1})}}_{S_{\text{noncases},1}} < c^*_{L1} < \underbrace{C_{(N_{\text{noncases},1}+1)} < \ldots < C_{(N_{\text{noncases},1}+N_{\text{noncases},2})}}_{S_{\text{noncases},2}} < c^*_{U1} < \underbrace{C_{(N_{\text{noncases},1}+N_{\text{noncases},2}+1)} < \ldots < C_{(N_{\text{noncases},1}+N_{\text{noncases},2}+N_{\text{noncases},3})}}_{S_{\text{noncases},3}}, \quad (2.3)$$

where $N_{\text{noncases},j}$ is the size of stratum $S_{\text{noncases},j}$, $j = 1, 2, 3$, and $\sum_{j=1}^{3} N_{\text{noncases},j} = N_{\text{noncases}}$.

Section 2.2.2 gives more details on finding two cut-off values $c_{L1} < c_{U1}$ for T_1 cases S_{cases} and $c^*_{L1} < c^*_{U1}$ for T_1 non-cases S_{noncases}. We consider a small c_{L1} and c^*_{L1} values and a high c_{U1} and c^*_{U1} values so that there are less number of individuals in the extreme strata since the data in the extreme strata might be more informative, and our aim is to understand the importance of sampling from extreme strata.

After obtaining the most efficient sampling design (n_{cases}, n_{noncases}) for the strata (S_{cases}, S_{noncases}), we do outcome-dependent BSS on the strata ($S_{\text{cases},1}$, $S_{\text{cases},2}$, $S_{\text{cases},3}$) and ($S_{\text{noncases},1}$, $S_{\text{noncases},2}$, $S_{\text{noncases},3}$). A sample of fixed size n_{cases} is drawn from S_{cases} at phase two in order to obtain the covariate X which is costly or difficult to measure. From the stratum $S_{\text{cases},j}$, $n_{\text{cases},j}$ is selected ($j = 1, 2, 3$) as shown below:

$$\underbrace{T_{1(1)} < \dots < T_{1(N_{\text{cases},1})}}_{n_{\text{cases},1}} < c_{L1} < \underbrace{T_{1(N_{\text{cases},1}+1)} < \dots < T_{1(N_{\text{cases},1}+N_{\text{cases},2})}}_{n_{\text{cases},2}}$$

$$< c_{U1} < \underbrace{T_{1(N_{\text{cases},1}+N_{\text{cases},2}+1)} < \dots < T_{1(N_{\text{cases},1}+N_{\text{cases},2}+N_{\text{cases},3})}}_{n_{\text{cases},3}},$$

and $\sum_{j=1}^{3} n_{\text{cases},j} = n_{\text{cases}}$. Similarly, a sample of fixed size n_{noncases} is drawn from S_{noncases} and $n_{\text{noncases},j}$ is selected from the stratum $S_{\text{noncases},j}$ as shown below:

$$\underbrace{C_{(1)} < \dots < C_{(N_{\text{noncases},1})}}_{n_{\text{noncases},1}} < c^*_{L1} < \underbrace{C_{(N_{\text{noncases},1}+1)} < \dots < C_{(N_{\text{noncases},1}+N_{\text{noncases},2})}}_{n_{\text{noncases},2}}$$

$$< c^*_{U1} < \underbrace{C_{(N_{\text{noncases},1}+N_{\text{noncases},2}+1)} < \dots < C_{(N_{\text{noncases},1}+N_{\text{noncases},2}+N_{\text{noncases},3})}}_{n_{\text{noncases},3}},$$

and $\sum_{j=1}^{3} n_{\text{noncases},j} = n_{\text{noncases}}$. Given the fixed sizes (n_{cases}, n_{noncases}) of cases and non-cases to be selected, one may choose how to allocate it among the strata (($S_{\text{cases},j} : j = 1, 2, 3$), ($S_{\text{noncases},j} : j = 1, 2, 3$)). Different allocations (($n_{\text{cases},j} : j = 1, 2, 3$), ($n_{\text{noncases},j} : j = 1, 2, 3$)) define different outcome-dependent BSS designs based on the first gap time T_1 and its event indicator.

Given the fixed sizes (n_{cases}, n_{noncases}) of cases and non-cases to be selected, there is an allocation $((n_{\text{cases},j} : j = 1,2,3), (n_{\text{noncases},j} : j = 1,2,3))$ among the strata $((S_{\text{cases},j} : j = 1,2,3), (S_{\text{noncases},j} : j = 1,2,3))$ satisfying

$$\frac{n_{\text{cases},1}}{N_{\text{cases},1}} = \frac{n_{\text{cases},2}}{N_{\text{cases},2}} = \frac{n_{\text{cases},3}}{N_{\text{cases},3}} \tag{2.4}$$

and

$$\frac{n_{\text{noncases},1}}{N_{\text{noncases},1}} = \frac{n_{\text{noncases},2}}{N_{\text{noncases},2}} = \frac{n_{\text{noncases},3}}{N_{\text{noncases},3}}. \tag{2.5}$$

Thus, (2.4) implies that the sampling probability is the same for all T_1 cases in S_{cases} and (2.5) implies that the sampling probability is the same for all T_1 non-cases in S_{noncases}. Therefore, the outcome-dependent BSS design defined by the allocation ($n_{\text{cases},j}$, $n_{\text{noncases},j}$) satisfying (2.4) and (2.5) is a SRS in S_{cases} and S_{noncases}, respectively. It is actually a generalized case-cohort design defined by the allocation (n_{cases}, n_{noncases}) among the strata (S_{cases}, S_{noncases}).

We will adopt the full likelihood-based approach to estimate the regression coefficient of the expensive covariate and to obtain the most efficient sampling design $((n_{\text{cases},j} : j = 1,2,3), (n_{\text{noncases},j} : j = 1,2,3))$ for the strata $((S_{\text{cases},j} : j = 1,2,3), (S_{\text{noncases},j} : j = 1,2,3))$ which is based on time-to-event T_1 and its event indicator. Efficient sampling design minimizes the variance of the coefficient estimate of the expensive covariate for the survival time T_1.

Let $R_i = I(\text{individual } i \text{ is selected})$ be the indicator function for individual i being selected at phase two. Suppose $D_{\text{cases},j} = \{i : R_i = 1, i \in S_{\text{cases},j}\}$ denotes the set of individuals selected from stratum $S_{\text{cases},j}$, where the size of $D_{\text{cases},j}$ is $n_{\text{cases},j}$. Similarly, suppose $D_{\text{noncases},j} = \{i : R_i = 1, i \in S_{\text{noncases},j}\}$ denotes the set of individuals selected from stratum $S_{\text{noncases},j}$, where the size of $D_{\text{noncases},j}$ is $n_{\text{noncases},j}$. Then $\bar{D}_{\text{cases},j} =$

$\{i : R_i = 0, i \in S_{\text{cases},j}\}$ is the set of individuals who are not selected from stratum $S_{\text{cases},j}$ and $\bar{D}_{\text{noncases},j} = \{i : R_i = 0, i \in S_{\text{noncases},j}\}$ is the set of individuals who are not selected from stratum $S_{\text{noncases},j}$. Therefore, the full likelihood function is (2.1) with $V = D_{\text{cases}} \cup D_{\text{noncases}}$, where $D_{\text{cases}} = D_{\text{cases},1} \cup D_{\text{cases},2} \cup D_{\text{cases},3}$ and $D_{\text{noncases}} = D_{\text{noncases},1} \cup D_{\text{noncases},2} \cup D_{\text{noncases},3}$, is the set of individuals who are selected at phase two and $\bar{V} = \bar{D}_{\text{cases}} \cup \bar{D}_{\text{noncases}}$, where $\bar{D}_{\text{cases}} = \bar{D}_{\text{cases},1} \cup \bar{D}_{\text{cases},2} \cup \bar{D}_{\text{cases},3}$ and $\bar{D}_{\text{noncases}} = \bar{D}_{\text{noncases},1} \cup \bar{D}_{\text{noncases},2} \cup \bar{D}_{\text{noncases},3}$, is the set of individuals who are not selected at phase two.

After obtaining the most efficient sampling design (n_{cases}, n_{noncases}) for the strata (S_{cases}, S_{noncases}) which is based on T_1 event indicator, we will next stratify the T_1 case stratum S_{cases} based on T_2 event indicator.

2.1.3 Outcome-dependent sampling design based on the event indicators of the two sequential gap times

In the previous two subsections, we were interested in identifying the efficient sampling design minimizing the variance of the coefficient estimate of the expensive covariate for the first gap time T_1. We may also be interested in exploring the efficient sampling design which minimizes the variance of the coefficient estimate of the expensive covariate for the second gap time T_2.

Assume that we obtained the most efficient sampling design (n_{cases}, n_{noncases}) for the strata (S_{cases}, S_{noncases}) which is based on T_1 event indicator, where $n_{\text{cases}} + n_{\text{noncases}} = n$. In this subsection, a subsample of fixed size (n) will be drawn in order to obtain a covariate which is expensive to measure based on both T_1 event indicator and T_2 event indicator. First, a subsample of size n_{noncases} is drawn from the T_1 non-case stratum S_{noncases}. Note that for the individuals in S_{noncases}, the second event cannot be observed since their first event was censored. Then, a subsample of size n_{cases} is drawn from the T_1 case stratum S_{cases} based on T_2 event indicator. Note that under bivariate sequential survival data, a T_1 case could be either a T_2 case or a T_2 non-case. Let us denote $S_{\text{cases,cases}}$ as the subset of S_{cases} which are T_2 cases and $S_{\text{cases,noncases}}$ as the subset of S_{cases} which are T_2 non-cases. In other words, $S_{\text{cases,cases}} = \{i : \delta_{1i} = 1, \delta_{2i} = 1, 1 \leq i \leq N\}$ and $S_{\text{cases,noncases}} = \{i : \delta_{1i} = 1, \delta_{2i} = 0, 1 \leq i \leq N\}$. Suppose the size of $S_{\text{cases,cases}}$ is M_{cases} and the size of $S_{\text{cases,noncases}}$ is M_{noncases}, then $M_{\text{cases}} + M_{\text{noncases}} = N_{\text{cases}}$. Then a subsample of size n_{cases} can be drawn from the T_1 case stratum S_{cases} based on T_2 event indicator by selecting a subsample from the case-case stratum $S_{\text{cases,cases}}$ and a subsample from the case-noncase stratum $S_{\text{cases,noncases}}$. The size of the subsample from the case-case stratum $S_{\text{cases,cases}}$ is denoted by m_{cases} and the size of the subsample from the case-noncase stratum $S_{\text{cases,noncases}}$ is denoted by m_{noncases}, where $n_{\text{cases}} = m_{\text{cases}} + m_{\text{noncases}}$. Given the fixed size n_{cases} of subsample, one may choose how to allocate it among the strata ($S_{\text{cases,cases}}$, $S_{\text{cases,noncases}}$) which is based on T_2 event indicator. Different allocations (m_{cases}, m_{noncases}) together with n_{noncases} define different outcome-dependent sampling designs based on T_1 and T_2 event indicators.

We adopt the full likelihood estimation method to estimate the regression coefficient of the expensive covariate and to obtain the most efficient sampling design (m_{cases}, m_{noncases}) for the strata ($S_{\text{cases,cases}}$, $S_{\text{cases,noncases}}$) which is based on T_2 event indicator. Efficient sampling design minimizes the variance of the coefficient estimate of the expensive covariate for the second gap time T_2.

Let $R_i = I(\text{individual } i \text{ is selected})$ be the indicator function for individual i being selected at phase two. Suppose $E_{\text{cases}} = \{i : R_i = 1, i \in S_{\text{cases}}, \delta_{2i} = 1\}$ denotes the set of individuals selected from stratum $S_{\text{cases,cases}}$, where the size of E_{cases} is m_{cases}. Similarly, suppose $E_{\text{noncases}} = \{i : R_i = 1, i \in S_{\text{cases}}, \delta_{2i} = 0\}$ denotes the set of individuals selected from stratum $S_{\text{cases,noncases}}$, where the size of E_{noncases} is m_{noncases}. Then $\bar{E}_{\text{cases}} = \{i : R_i = 0, i \in S_{\text{cases}}, \delta_{2i} = 1\}$ is the set of individuals who are not selected from stratum

$S_{\text{cases,cases}}$ and $\bar{E}_{\text{noncases}} = \{i : R_i = 0, i \in S_{\text{cases}}, \delta_{2i} = 0\}$ is the set of individuals who are not selected from stratum $S_{\text{cases,noncases}}$. Therefore, the full likelihood function is defined by (2.1) with $V = E_{\text{cases}} \cup E_{\text{noncases}} \cup D_{\text{noncases}}$ which is the set of individuals selected at phase two and $\bar{V} = \bar{E}_{\text{cases}} \cup \bar{E}_{\text{noncases}} \cup \bar{D}_{\text{noncases}}$ which is the set of individuals not selected at phase two. Both D_{noncases} and $\bar{D}_{\text{noncases}}$ were defined in Section 2.1.1.

After obtaining the most efficient sampling design which is based on T_1 and T_2 event indicators, we will next stratify the case-case stratum $S_{\text{cases,cases}}$ based on the second gap time T_2 and stratify the case-noncase stratum $S_{\text{cases,noncases}}$ based on censoring time $C - T_1$.

2.1.4 Outcome-dependent BSS design based on the two sequential gap times and their event indicators

Recall that the phase one cohort can be stratified into the strata (S_{cases}, S_{noncases}) based on the event indicator of the first gap time T_1. For a fixed phase two sample size n, we can obtain the most efficient sampling design (n_{cases}, n_{noncases}) for the strata (S_{cases}, S_{noncases}) based on the full likelihood-based approach, where $n_{\text{cases}} + n_{\text{noncases}} = n$. Here, efficient sampling design minimizes the variance of the coefficient estimate of the expensive covariate for the first gap time T_1. Note that under bivariate sequential survival data, a first event case could be either a second event case or a second event non-case. Therefore, $S_{\text{cases}} = S_{\text{cases,cases}} \cup S_{\text{cases,noncases}}$ and we can obtain the most efficient sampling design (m_{cases}, m_{noncases}) for the strata ($S_{\text{cases,cases}}$, $S_{\text{cases,noncases}}$) based on the full likelihood-based approach, where $m_{\text{cases}} + m_{\text{noncases}} = n_{\text{cases}}$. Here, efficient sampling design minimizes the variance of the coefficient estimate of the expensive covariate for the second gap time T_2. Greater efficiency may be achieved for outcome-dependent sampling design by selecting the more informative subjects for purposes of detailed covariate measurement.

We can stratify all T_2 cases $S_{\text{cases,cases}}$ into strata ($S_{\text{cases,cases},1}$, $S_{\text{cases,cases},2}$, $S_{\text{cases,cases},3}$) based on time-to-event T_2 using two cut-off values $c_{L2} < c_{U2}$ which are defined in Section 2.2.3:

$$\underbrace{T_{2(1)} < \ldots < T_{2(M_{\text{cases},1})}}_{S_{\text{cases,cases},1}} < c_{L2} < \underbrace{T_{2(M_{\text{cases},1}+1)} < \ldots < T_{2(M_{\text{cases},1}+M_{\text{cases},2})}}_{S_{\text{cases,cases},2}}$$
$$< c_{U2} < \underbrace{T_{2(M_{\text{cases},1}+M_{\text{cases},2}+1)} < \ldots < T_{2(M_{\text{cases},1}+M_{\text{cases},2}+M_{\text{cases},3})}}_{S_{\text{cases,cases},3}}, \tag{2.6}$$

where $M_{\text{cases},j}$ is the size of stratum $S_{\text{cases,cases},j}$, $j = 1, 2, 3$, and $\sum_{j=1}^{3} M_{\text{cases},j} = M_{\text{cases}}$.

Similarly, we can stratify T_2 non-cases $S_{\text{cases,noncases}}$ into strata ($S_{\text{cases,noncases},1}$, $S_{\text{cases,noncases},2}$, $S_{\text{cases,noncases},3}$) based on censoring time $C - T_1$ using two cut-off values $c^*_{L2} < c^*_{U2}$ which are defined in Section 2.2.3:

$$\underbrace{C_{(1)} - T_{1(1)} < \ldots < C_{(M_{\text{noncases},1})} - T_{1(M_{\text{noncases},1})}}_{S_{\text{cases,noncases},1}}$$
$$< c^*_{L2} < \underbrace{C_{(M_{\text{noncases},1}+1)} - T_{1(M_{\text{noncases},1}+1)} < \ldots < C_{(M_{\text{noncases},1}+M_{\text{noncases},2})} - T_{1(M_{\text{noncases},1}+M_{\text{noncases},2})}}_{S_{\text{cases,noncases},2}}$$
$$< c^*_{U2} < \underbrace{C_{(M_{\text{noncases},1}+M_{\text{noncases},2}+1)} - T_{1(M_{\text{noncases},1}+M_{\text{noncases},2}+1)} < \ldots < C_{(M_{\text{noncases}})} - T_{1(M_{\text{noncases}})}}_{S_{\text{cases,noncases},3}}, \tag{2.7}$$

where $M_{\text{noncases},j}$ is the size of stratum $S_{\text{cases,noncases},j}$, $j = 1, 2, 3$, and $\sum_{j=1}^{3} M_{\text{noncases},j} = M_{\text{noncases}}$.

Section 2.2.3 gives details on finding two cut-off values $c_{L2} < c_{U2}$ for T_2 cases $S_{\text{cases,cases}}$ and $c^*_{L2} < c^*_{U2}$ for T_2 non-cases $S_{\text{cases,noncases}}$. We consider a small c_{L2} and c^*_{L2}

values and a high c_{U2} and c^*_{U2} values so that there are less number of individuals in the extreme strata. The data in the extreme strata might be more informative, and one of the main aims of this study is to investigate this as described in Section 2.1.2.

After obtaining the most efficient sampling design (m_{cases}, m_{noncases}) for the strata ($S_{\text{cases,cases}}$, $S_{\text{cases,noncases}}$), we do outcome-dependent BSS on the strata ($S_{\text{cases,cases},1}$, $S_{\text{cases,cases},2}$, $S_{\text{cases,cases},3}$) and ($S_{\text{cases,noncases},1}$, $S_{\text{cases,noncases},2}$, $S_{\text{cases,noncases},3}$). A subsample of fixed size m_{cases} is drawn from the case-case stratum $S_{\text{cases,cases}}$ at phase two in order to obtain the covariate X which is costly or difficult to measure. From the stratum $S_{\text{cases,cases},j}$, $m_{\text{cases},j}$ ($j = 1, 2, 3$) individuals are selected as shown below:

$$\underbrace{T_{2(1)} < \ldots < T_{2(M_{\text{cases},1})}}_{m_{\text{cases},1}} < c_{L2} < \underbrace{T_{2(M_{\text{cases},1}+1)} < \ldots < T_{2(M_{\text{cases},1}+M_{\text{cases},2})}}_{m_{\text{cases},2}}$$
$$< c_{U2} < \underbrace{T_{2(M_{\text{cases},1}+M_{\text{cases},2}+1)} < \ldots < T_{2(M_{\text{cases},1}+M_{\text{cases},2}+M_{\text{cases},3})}}_{m_{\text{cases},3}},$$

where $\sum_{j=1}^{3} m_{\text{cases},j} = m_{\text{cases}}$. Similarly, a subsample of fixed size m_{noncases} is drawn from the case-noncase stratum $S_{\text{cases,noncases}}$. From the stratum $S_{\text{cases,noncases},j}$, $m_{\text{noncases},j}$ ($j = 1, 2, 3$) individuals are selected as shown below:

$$\underbrace{C_{(1)} - T_{1(1)} < \ldots < C_{(M_{\text{noncases},1})} - T_{1(M_{\text{noncases},1})}}_{m_{\text{noncases},1}}$$
$$< c^*_{L2} < \underbrace{C_{(M_{\text{noncases},1}+1)} - T_{1(M_{\text{noncases},1}+1)} < \ldots < C_{(M_{\text{noncases},1}+M_{\text{noncases},2})} - T_{1(M_{\text{noncases},1}+M_{\text{noncases},2})}}_{m_{\text{noncases},2}}$$
$$< c^*_{U2} < \underbrace{C_{(M_{\text{noncases},1}+M_{\text{noncases},2}+1)} - T_{1(M_{\text{noncases},1}+M_{\text{noncases},2}+1)} < \ldots < C_{(M_{\text{noncases}})} - T_{1(M_{\text{noncases}})}}_{m_{\text{noncases},3}},$$

where $\sum_{j=1}^{3} m_{\text{noncases},j} = m_{\text{noncases}}$. Given the fixed sizes ($m_{\text{cases}}, m_{\text{noncases}}$) of subsamples, one may choose how to allocate it among the strata (($S_{\text{cases,cases},j} : j = 1, 2, 3$),

$(S_{\text{cases,noncases},j} : j = 1,2,3))$. Different allocations $((m_{\text{cases},j} : j = 1,2,3), (m_{\text{noncases},j} : j = 1,2,3))$ define different outcome-dependent BSS designs based on the second gap time T_2 and its event indicator.

Given the fixed sizes $(m_{\text{cases}}, m_{\text{noncases}})$ of cases and non-cases to be selected, there is an allocation $((m_{\text{cases},j} : j = 1,2,3), (m_{\text{noncases},j} : j = 1,2,3))$ among the strata $((S_{\text{cases,cases},j} : j = 1,2,3), (S_{\text{cases,noncases},j} : j = 1,2,3))$ satisfying

$$\frac{m_{\text{cases},1}}{M_{\text{cases},1}} = \frac{m_{\text{cases},2}}{M_{\text{cases},2}} = \frac{m_{\text{cases},3}}{M_{\text{cases},3}} \tag{2.8}$$

and

$$\frac{m_{\text{noncases},1}}{M_{\text{noncases},1}} = \frac{m_{\text{noncases},2}}{M_{\text{noncases},2}} = \frac{m_{\text{noncases},3}}{M_{\text{noncases},3}}. \tag{2.9}$$

Here, (2.8) implies that the sampling probability is the same for all T_2 cases in $S_{\text{cases,cases}}$ and (2.9) implies that the sampling probability is the same for all T_2 non-cases in $S_{\text{cases,noncases}}$. Therefore, the outcome-dependent BSS design defined by the allocation $(m_{\text{cases},j}, m_{\text{noncases},j})$ satisfying (2.8) and (2.9) is a SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$, respectively. It is actually a generalized case-cohort design defined by the allocation $(m_{\text{cases}}, m_{\text{noncases}})$ among the strata $(S_{\text{cases,cases}}, S_{\text{cases,noncases}})$.

We use the full likelihood estimation method to estimate the regression coefficient of the expensive covariate and to obtain the most efficient sampling design $((m_{\text{cases},j} : j = 1,2,3), (m_{\text{noncases},j} : j = 1,2,3))$ for the strata $((S_{\text{cases,cases},j} : j = 1,2,3), (S_{\text{cases,noncases},j} : j = 1,2,3))$ which is based on the second gap time T_2 and its event indicator. Efficient sampling design minimizes the variance of the coefficient estimate of the expensive covariate for the second gap time T_2.

Let $R_i = I(\text{individual } i \text{ is selected})$ be the indicator function for individual i being selected at phase two. Suppose $E_{\text{cases},j} = \{i : R_i = 1, i \in S_{\text{cases,cases},j}\}$ denotes the set of individuals selected from stratum $S_{\text{cases,cases},j}$, where the size of $E_{\text{cases},j}$ is $m_{\text{cases},j}$. Similarly, suppose $E_{\text{noncases},j} = \{i : R_i = 1, i \in S_{\text{cases,noncases},j}\}$ denotes the set of individuals selected from stratum $S_{\text{cases,noncases},j}$, where the size of $E_{\text{noncases},j}$ is $m_{\text{noncases},j}$. Then $\bar{E}_{\text{cases},j} = \{i : R_i = 0, i \in S_{\text{cases,cases},j}\}$ is the set of individuals not selected from stratum $S_{\text{cases,cases},j}$ and $\bar{E}_{\text{noncases},j} = \{i : R_i = 0, i \in S_{\text{cases,noncases},j}\}$ is the set of individuals not selected from stratum $S_{\text{cases,noncases},j}$. Therefore, the full likelihood function is defined by (2.1) with $V = E_{\text{cases}} \cup E_{\text{noncases}} \cup D_{\text{noncases}}$, where $E_{\text{cases}} = E_{\text{cases},1} \cup E_{\text{cases},2} \cup E_{\text{cases},3}$ and $E_{\text{noncases}} = E_{\text{noncases},1} \cup E_{\text{noncases},2} \cup E_{\text{noncases},3}$, is the set of individuals selected at phase two and $\bar{V} = \bar{E}_{\text{cases}} \cup \bar{E}_{\text{noncases}} \cup \bar{D}_{\text{noncases}}$, where $\bar{E}_{\text{cases}} = \bar{E}_{\text{cases},1} \cup \bar{E}_{\text{cases},2} \cup \bar{E}_{\text{cases},3}$ and $\bar{E}_{\text{noncases}} = \bar{E}_{\text{noncases},1} \cup \bar{E}_{\text{noncases},2} \cup \bar{E}_{\text{noncases},3}$, is the set of individuals not selected at phase two. Both $D_{\text{noncases}} = D_{\text{noncases},1} \cup D_{\text{noncases},2} \cup D_{\text{noncases},3}$ and $\bar{D}_{\text{noncases}} = \bar{D}_{\text{noncases},1} \cup \bar{D}_{\text{noncases},2} \cup \bar{D}_{\text{noncases},3}$ were defined in Section 2.1.2.

2.2 Simulation study

2.2.1 Data generation

We generate a large random bivariate survival time sample with size $N = 50,000$ from the joint conditional distribution of T_1 and T_2 given $X = x$,

$$F(t_1, t_2|x) = C_\phi(F_1(t_1|x), F_2(t_2|x)) = (F_1(t_1|x)^{-\phi} + F_2(t_2|x)^{-\phi} - 1)^{-1/\phi}, \quad \phi > 0, \quad (2.10)$$

with the Clayton copula in (1.15). Moderate and high dependence levels were considered for T_1 and T_2. The copula parameter values $\phi = \frac{4}{3}$ and $\phi = 8$ were considered corresponding to the Kendall's tau value of $\tau = 0.4$ or $\tau = 0.8$, respectively. Note that the Kendall's tau value is a one-to-one function of ϕ, namely $\tau = \phi/(\phi+2)$. The covariate X follows a Bernoulli distribution with probability of success $p = P(X = 1) = 0.25$.

The marginal distribution of T_1 is assumed to be the Weibull distribution with survival function

$$S_1(t_1|x) = \exp[-e^{\alpha_{10}+\alpha_{11}x}t_1^{\gamma_1}] \tag{2.11}$$

where $\alpha_{10} = 0.6$, $\alpha_{11} = 0.0$ or 1.0, and $\gamma_1 = 0.5$, 1.0 or 1.5. The marginal distribution of T_2 is assumed to be the Weibull distribution with survival function

$$S_2(t_2|x) = \exp[-e^{\alpha_{20}+\alpha_{21}x}t_2^{\gamma_2}] \tag{2.12}$$

where $\alpha_{20} = 0.4$, $\alpha_{21} = 0.0$ or 1.0, and $\gamma_2 = 0.5$. Each set of three parameters $(\alpha_{11}, \alpha_{21}, \gamma_1)$ specifies one scenario.

By virtue of Sklar's theorem, we need to generate a pair (u_1, u_2) of observations of Uniform$(0,1)$ random variables (U_1, U_2) whose joint distribution function is C_ϕ, the Clayton copula of U_1 and U_2, and then transform those uniform variates via the inverse distribution function method.

One procedure for generating such a pair (u_1, u_2) of Uniform$(0,1)$ random variates is the conditional distribution method. For this method, we need the conditional distribution function for U_2 given $U_1 = u_1$, which we denote $c_{u_1}(u_2)$ and is given by

$$c_{u_1}(u_2) = P[U_2 \le u_2 | U_1 = u_1]$$

which can be written in terms of a copula function C_ϕ as

$$\begin{aligned}
c_{u_1}(u_2) &= \lim_{\Delta u_1 \to 0} \frac{P[U_2 \le u_2, u_1 \le U_1 \le u_1 + \Delta u_1]}{P[u_1 \le U_1 \le u_1 + \Delta u_1]} \\
&= \lim_{\Delta u_1 \to 0} \frac{P[U_2 \le u_2, U_1 \le u_1 + \Delta u_1] - P[U_2 \le u_2, U_1 \le u_1]}{P[U_2 \le 1, U_1 \le u_1 + \Delta u_1] - P[U_2 \le 1, U_1 \le u_1]} \\
&= \lim_{\Delta u_1 \to 0} \frac{C_\phi(u_1 + \Delta u_1, u_2) - C_\phi(u_1, u_2)}{C_\phi(u_1 + \Delta u_1, 1) - C_\phi(u_1, 1)} \\
&= \lim_{\Delta u_1 \to 0} \frac{C_\phi(u_1 + \Delta u_1, u_2) - C_\phi(u_1, u_2)}{(u_1 + \Delta u_1) - u_1} \\
&= \lim_{\Delta u_1 \to 0} \frac{C_\phi(u_1 + \Delta u_1, u_2) - C_\phi(u_1, u_2)}{\Delta u_1} \\
&= \frac{\partial C_\phi(u_1, u_2)}{\partial u_1}.
\end{aligned}$$

The conditional distribution method to generate (u_1, u_2) from $C_\phi(u_1, u_2)$ is as follows:

1. Generate a pair (u_1, v_2) of values of two independent Uniform$(0,1)$ random variables U_1 and V_2.

2. Set $u_2 = c_{u_1}^{(-1)}(v_2)$, where $c_{u_1}^{(-1)}$ denotes a quasi-inverse of c_{u_1} (Nelson, 2006).

3. The desired pair is (u_1, u_2).

We then transform such a pair (u_1, u_2) of Uniform$(0,1)$ random variates via the inverse distribution function method to obtain a pair (T_1, T_2) of observations. The pair (T_1, T_2) of observations is obtained by $T_1 = F_1^{(-1)}(u_1)$ and $T_2 = F_2^{(-1)}(u_2)$, where $F_1^{(-1)}$ is any quasi-inverse of $F_1(\cdot|x)$ and $F_2^{(-1)}$ is any quasi-inverse of $F_2(\cdot|x)$.

The censoring time C is generated from Uniform$(0,b)$ such that about 40% of T_1 survival times are censored. When T_1 is censored, T_2 is unobserved. Notice that the upper bound b in the domain $(0,b)$ of Uniform$(0,b)$ is uniquely determined by the model parameters of T_1 and the T_1 censoring rate. For a given model, the censoring rate is a monotone decreasing function of the upper bound b. As an iterative root-finding procedure, bisection method can be used to find the upper bound b to obtain 40% T_1 censoring rate for each given model. Table 2.1 shows values of upper bound b and T_2 censoring rate with 40% T_1 censoring for different scenarios of data generation.

Table 2.1: Percentages of censored second gap time with censoring time generated from Uniform$(0, b)$ to make 40% censored first gap time

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Upper bound b of Uniform$(0, b)$	Percentage of censored T_2 (Kendall's $\tau = 0.4$)	Percentage of censored T_2 (Kendall's $\tau = 0.8$)
$(0, 0, 0.5)$	0.654834	62.43%	56.17%
$(0, 1, 0.5)$	0.654834	58.86%	52.09%
$(1, 0, 0.5)$	0.419287	66.58%	61.20%
$(1, 1, 0.5)$	0.419287	60.44%	55.23%
$(0, 0, 1.0)$	1.235522	59.68%	54.96%
$(0, 1, 1.0)$	1.235522	56.53%	52.37%
$(1, 0, 1.0)$	0.994482	61.45%	56.92%
$(1, 1, 1.0)$	0.994482	56.29%	52.32%
$(0, 0, 1.5)$	1.489063	60.01%	56.36%
$(0, 1, 1.5)$	1.489063	56.74%	53.74%
$(1, 0, 1.5)$	1.300244	60.93%	57.47%
$(1, 1, 1.5)$	1.300244	56.22%	53.22%

We assume that the observed data at phase one is $\{(t_{1i}, \delta_{1i}, t_{2i}, \delta_{2i}) : i = 1, ..., N\}$ where $(t_{1i}, t_{2i}) = (\min(T_{1i}, C_i), \min(T_{2i}, C_i - t_{1i}))$ and $(\delta_{1i}, \delta_{2i}) = (I[T_{1i} = t_{1i}], I[T_{2i} = t_{2i}])$, $i = 1, ..., N$, are the observed gap times and their event indicators, respectively.

2.2.2 Stratification based on the first gap time and its event indicator

Recall that the phase one cohort can be stratified into the strata (S_{cases}, S_{noncases}) based on the event indicator of the first gap time T_1. We can stratify all T_1 cases S_{cases} into strata ($S_{\text{cases},1}$, $S_{\text{cases},2}$, $S_{\text{cases},3}$) based on the first gap time T_1 using two cut-off values $c_{L1} < c_{U1}$ as in (2.2). Similarly, we can stratify all T_1 non-cases S_{noncases}

Table 2.2: Stratification based on the first gap time and its event indicator

Stratum	T_1 cases ($\delta_1 = 1$)
$S_{\text{cases},1}(t_1 \le c_{L1})$	$N_{\text{cases},1} = 5,000$
$S_{\text{cases},2}(c_{L1} < t_1 \le c_{U1})$	$N_{\text{cases},2} = 20,000$
$S_{\text{cases},3}(c_{U1} < t_1)$	$N_{\text{cases},3} = 5,000$
All T_1 cases	$N_{\text{cases}} = 30,000$
	T_1 non-cases ($\delta_1 = 0$)
$S_{\text{noncases},1}(t_1 \le c^*_{L1})$	$N_{\text{noncases},1} = 5,000$
$S_{\text{noncases},2}(c^*_{L1} < t_1 \le c^*_{U1})$	$N_{\text{noncases},2} = 10,000$
$S_{\text{noncases},3}(c^*_{U1} < t_1)$	$N_{\text{noncases},3} = 5,000$
All T_1 non-cases	$N_{\text{noncases}} = 20,000$

into strata ($S_{\text{noncases},1}$, $S_{\text{noncases},2}$, $S_{\text{noncases},3}$) based on censoring time C using two cut-off values $c^*_{L1} < c^*_{U1}$ as in (2.3).

We generated a large sample of size $N = 50,000$ in order to show the asymptotic results. With 40% T_1 censoring, there are about $N_{\text{cases}} = 30,000$ individuals in the case stratum S_{cases} and about $N_{\text{noncases}} = 20,000$ individuals in the non-case stratum S_{noncases}. We set the two cut-off values $c_{L1} < c_{U1}$ and $c^*_{L1} < c^*_{U1}$ in (2.2) and (2.3) as in Table 2.2.

We consider a small c_{L1} and c^*_{L1} value and a high c_{U1} and c^*_{U1} value so that there are less number of individuals in the extreme strata since the data in the extreme strata might be more informative, and our aim is to understand the importance of sampling from the extreme strata.

By ordering the t_{1i} values of $N_{\text{cases}} = 30,000$ first event cases, the two cut-off values $c_{L1} < c_{U1}$ are set to satisfy the conditions in Table 2.2. Using these two case cut-off values $c_{L1} < c_{U1}$, all T_1 cases S_{cases} can be stratified into three groups $S_{\text{cases},j}$, $j = 1, 2, 3$, based on survival time T_1. The first stratum $S_{\text{cases},1}$ consists of T_1 cases with short time-to-first event. The second stratum $S_{\text{cases},2}$ consists of T_1 cases with midrange time-to-first event. The third stratum $S_{\text{cases},3}$ consists of T_1 cases with long time-to-first event.

Similarly, by ordering the t_{1i} values of $N_{\text{noncases}} = 20,000$ first event non-cases, the two cut-off values $c^*_{L1} < c^*_{U1}$ are set to satisfy the conditions in Table 2.2. Using these two non-case cut-off values $c^*_{L1} < c^*_{U1}$, all T_1 non-cases S_{noncases} can be stratified into three groups $S_{\text{noncases},j}$, $j = 1, 2, 3$, based on censoring time C. The first stratum $S_{\text{noncases},1}$ consists of T_1 non-cases with short censoring time. The second stratum $S_{\text{noncases},2}$ consists of T_1 non-cases with midrange censoring time. The third stratum $S_{\text{noncases},3}$ consists of T_1 non-cases with long censoring time.

The case cut-off values $c_{L1} < c_{U1}$ and non-cases cut-off values $c^*_{L1} < c^*_{U1}$ for each model scenario are listed in Table 2.3.

Table 2.3: Cut-off values for stratification based on the first gap time and its event indicator

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	c_{L1}	c_{U1}	c^*_{L1}	c^*_{U1}
$(0, 0, 0.5)$	0.003434457	0.1857223	0.09220276	0.4108654
$(0, 1, 0.5)$	0.003434457	0.1857223	0.09220276	0.4108654
$(1, 0, 0.5)$	0.001742273	0.1098055	0.06043769	0.2671622
$(1, 1, 0.5)$	0.001742273	0.1098055	0.06043769	0.2671622
$(0, 0, 1.0)$	0.06021523	0.5208437	0.1332584	0.6105464
$(0, 1, 1.0)$	0.06021523	0.5208437	0.1332584	0.6105464
$(1, 0, 1.0)$	0.04264397	0.4011779	0.1097162	0.5131692
$(1, 1, 1.0)$	0.04264397	0.4011779	0.1097162	0.5131692
$(0, 0, 1.5)$	0.1576681	0.7135059	0.1498029	0.6125531
$(0, 1, 1.5)$	0.1576681	0.7135059	0.1498029	0.6125531
$(1, 0, 1.5)$	0.1255860	0.6043437	0.1315327	0.5544567
$(1, 1, 1.5)$	0.1255860	0.6043437	0.1315327	0.5544567

2.2.3 Stratification based on the second gap time and its event indicator

Recall that the phase one cohort can be stratified into the strata (S_{cases}, S_{noncases}) based on the event indicator of the first gap time T_1. Note that under bivariate sequential survival data, a first event case could be either a second event case or a second event non-case. Therefore, $S_{\text{cases}} = S_{\text{cases,cases}} \cup S_{\text{cases,noncases}}$ where $S_{\text{cases,cases}}$ is the subset of S_{cases} which are T_2 cases and $S_{\text{cases,noncases}}$ is the subset of S_{cases} which are T_2 non-cases. We can stratify all T_2 cases $S_{\text{cases,cases}}$ into strata ($S_{\text{cases,cases},1}$, $S_{\text{cases,cases},2}$, $S_{\text{cases,cases},3}$) based on time-to-event T_2 using two cut-off values $c_{L2} < c_{U2}$ as in (2.6). Similarly, we can stratify T_2 non-cases $S_{\text{cases,noncases}}$ into strata ($S_{\text{cases,noncases},1}$, $S_{\text{cases,noncases},2}$, $S_{\text{cases,noncases},3}$) based on censoring time $C - T_1$ using two cut-off values $c^*_{L2} < c^*_{U2}$ as in (2.7).

With 40% censoring rate for the first event, there are about $N_{\text{cases}} = 30,000$ individuals in the case stratum S_{cases} and about $N_{\text{noncases}} = 20,000$ individuals in the non-case stratum S_{noncases} based on being T_1 case or T_1 non-case. If we denote M_{cases} as the number of T_2 cases and M_{noncases} as the number of T_2 non-cases, then the total number of T_1 cases is $M_{\text{cases}} + M_{\text{noncases}} = N_{\text{cases}} = 30,000$. The number M_{cases} of T_2 cases and the number M_{noncases} of T_2 non-cases for each model scenario are listed in Table 2.5 and Table 2.6. We set the two cut-off values $c_{L2} < c_{U2}$ and $c^*_{L2} < c^*_{U2}$ in (2.6) and (2.7) as in Table 2.4.

We consider small c_{L2} and c^*_{L2} values and high c_{U2} and c^*_{U2} values so that there are less number of individuals in the extreme strata.

By ordering the t_{2i} values of M_{cases} second event cases, the two cut-off values $c_{L2} < c_{U2}$ are set to satisfy the conditions in Table 2.4. Using these two case cut-off values $c_{L2} < c_{U2}$, all T_2 cases $S_{\text{cases,cases}}$ can be stratified into three groups $S_{\text{cases,cases},j}$,

Table 2.4: Stratification based on the second gap time and its event indicator

Stratum	T_2 cases ($\delta_2 = 1$)
$S_{\text{cases,cases},1}(t_2 \le c_{L2})$	$M_{\text{cases},1} = 2,500$
$S_{\text{cases,cases},2}(c_{L2} < t_2 \le c_{U2})$	$M_{\text{cases},2} = M_{\text{cases}} - 5,000$
$S_{\text{cases,cases},3}(c_{U2} < t_2)$	$M_{\text{cases},3} = 2,500$
All T_2 cases	M_{cases}
	T_2 non-cases ($\delta_2 = 0$)
$S_{\text{cases,noncases},1}(t_2 \le c^*_{L2})$	$M_{\text{noncases},1} = 2,500$
$S_{\text{cases,noncases},2}(c^*_{L2} < t_2 \le c^*_{U2})$	$M_{\text{noncases},2} = M_{\text{noncases}} - 5,000$
$S_{\text{cases,noncases},3}(c^*_{U2} < t_2)$	$M_{\text{noncases},3} = 2,500$
All T_2 non-cases	M_{noncases}

$j = 1,2,3$, based on survival time T_2. The first stratum $S_{\text{cases,cases},1}$ consists of T_2 cases with short second gap time. The second stratum $S_{\text{cases,cases},2}$ consists of T_2 cases with midrange second gap time. The third stratum $S_{\text{cases,cases},3}$ consists of T_2 cases with long second gap time.

Similarly, by ordering the t_{2i} values of M_{noncases} second event non-cases, the two cut-off values $c^*_{L2} < c^*_{U2}$ are set to satisfy the conditions in Table 2.4. Using these two non-case cut-off values $c^*_{L2} < c^*_{U2}$, all T_2 non-cases $S_{\text{cases,noncases}}$ can be stratified into three groups $S_{\text{cases,noncases},j}$, $j = 1,2,3$, based on censoring time $C - t_1$. The first stratum $S_{\text{cases,noncases},1}$ consists of T_2 non-cases with short censoring time. The second stratum $S_{\text{cases,noncases},2}$ consists of T_2 non-cases with midrange censoring time. The third stratum $S_{\text{cases,noncases},3}$ consists of T_2 non-cases with long censoring time.

The case cut-off values $c_{L2} < c_{U2}$ and the non-case cut-off values $c^*_{L2} < c^*_{U2}$ for each model scenario when the dependence between time-to-events is moderate are listed in Table 2.5.

Table 2.5: Cut-off values for stratification based on the second gap time and its event indicator when the dependence between gap times is moderate

$(\alpha_{11},\alpha_{21},\gamma_1)$	M_{cases}	M_{noncases}	c_{L2}	c_{U2}	c^*_{L2}	c^*_{U2}
$(0,0,0.5)$	18783	11217	0.0011820310	0.15419920	0.05822070	0.3242031

Continued on next page

Table 2.5 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	M_{cases}	M_{noncases}	c_{L2}	c_{U2}	c^*_{L2}	c^*_{U2}
(0, 1, 0.5)	20570	9430	0.0006062500	0.13750000	0.06527344	0.2815625
(1, 0, 0.5)	16709	13291	0.0011911620	0.10099220	0.03581250	0.2423750
(1, 1, 0.5)	19779	10221	0.0005953125	0.09371094	0.04018066	0.1915039
(0, 0, 1.0)	20160	9840	0.0012597660	0.2379687	0.09917188	0.4472500
(0, 1, 1.0)	21773	8227	0.0006671875	0.2075977	0.11285160	0.3800000
(1, 0, 1.0)	19227	10723	0.0012548830	0.2029175	0.07809375	0.4068457
(1, 1, 1.0)	21855	8145	0.0006235413	0.1715625	0.08852539	0.2960156
(0, 0, 1.5)	19995	10005	0.0013428500	0.2614375	0.10552050	0.4935000
(0, 1, 1.5)	21628	8372	0.0007262207	0.2267969	0.12097660	0.4173438
(1, 0, 1.5)	19537	10463	0.0013417970	0.2407812	0.09414062	0.4689062
(1, 1, 1.5)	21889	8111	0.0006890625	0.2012695	0.10739380	0.3466406

The case cut-off values $c_{L2} < c_{U2}$ and the non-case cut-off values $c^*_{L2} < c^*_{U2}$ for each model scenario when the dependence between time-to-events is high are listed in Table 2.6.

Table 2.6: Cut-off values for stratification based on the second gap time and its event indicator when the dependence between gap times is high

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	M_{cases}	M_{noncases}	c_{L2}	c_{U2}	c^*_{L2}	c^*_{U2}
(0, 0, 0.5)	21916	8084	0.0011421870	0.1585742	0.05941016	0.1752930
(0, 1, 0.5)	23957	6043	0.0005804687	0.1353021	0.06809570	0.1416992
(1, 0, 0.5)	19401	10599	0.0011425780	0.1093750	0.03591406	0.242375
(1, 1, 0.5)	22387	7613	0.0005800781	0.0956012	0.04093750	0.1092773
(0, 0, 1.0)	22519	7481	0.0011796870	0.2175781	0.10122070	0.2582031
(0, 1, 1.0)	23816	6184	0.0006031250	0.1865082	0.11630650	0.2056885
(1, 0, 1.0)	21538	8462	0.0011679080	0.1957275	0.07787500	0.2601563
(1, 1, 1.0)	23842	6158	0.0005914062	0.1559570	0.08837891	0.1513281
(0, 0, 1.5)	21819	8181	0.001238770	0.2415039	0.10761720	0.3224609
(0, 1, 1.5)	23128	6872	0.000643750	0.2068954	0.12207030	0.2696289
(1, 0, 1.5)	21264	8736	0.001238281	0.2230957	0.09255859	0.3207031
(1, 1, 1.5)	23386	6614	0.000628125	0.1827148	0.10664060	0.2116211

Chapter 3

Efficiency of two-phase outcome-dependent sampling designs when the dependence between time-to-events is moderate

The objective of this study is to investigate efficient two-phase outcome-dependent sampling designs for bivariate sequential survival data under a predetermined phase two sample size. Four phase two sampling designs were introduced in Chapter 2: (1) generalized case-cohort design based on the event indicator of the first gap time; (2) outcome-dependent BSS design based on the first gap time and its event indicator; (3) generalized case-cohort design based on the event indicators of the two sequential gap times; and (4) outcome-dependent BSS design based on the two sequential gap times and their event indicators.

A simulation study was conducted to study the efficiency of these phase two sampling designs. We generated a large random bivariate survival time sample with size

$N = 50,000$ from the joint conditional distribution of T_1 and T_2 given $X = x$ in (2.10) with the Clayton copula parameter value $\phi = \frac{4}{3}$, and the covariate X follows the Bernoulli distribution with probability of success $p = P(X = 1) = 0.25$. The corresponding Kendall's tau value was $\tau = \phi/(\phi + 2) = 0.4$, and therefore, there was a moderate dependence between the two sequential gap times T_1 and T_2 given $X = x$. The marginal distributions of T_1 and T_2 given $X = x$ were modelled by Weibull regression with survival functions (2.11) and (2.12), respectively. The censoring time C is generated from Uniform$(0, b)$ such that about 40% of T_1 survival times are censored. The upper bound b of Uniform$(0, b)$ and T_2 censoring rate are given in Table 2.1. At phase one, suppose the observed data is $\{(t_1, \delta_1, t_2, \delta_2) : i = 1, ..., N\}$ where $(t_1, t_2) = (\min(T_1, C), \min(T_2, C - t_1))$ and $(\delta_1, \delta_2) = (I[T_1 = t_1], I[T_2 = t_2])$ are the observed survival times and their event indicators, respectively.

A subsample of fixed size n is drawn at phase two in order to obtain a measurement of covariate X which is costly or difficult to measure. We want to investigate generalized case-cohort and outcome-dependent BSS designs that result in more efficient sampling designs with bivariate sequential survival data.

The phase one cohort can be stratified into the strata (S_{cases}, S_{noncases}) based on the event indicator δ_1 of the first gap time T_1. Suppose the size of the subsample from the case stratum S_{cases} is denoted by n_{cases} and the size of the subsample from the non-case stratum S_{noncases} is denoted by n_{noncases}, where $n_{\text{cases}} + n_{\text{noncases}} = n$. The aim of Section 3.1 is to determine the number of first event cases (n_{cases}) versus the number of first event non-cases (n_{noncases}) that should be selected at phase two where $n_{\text{cases}} + n_{\text{noncases}} = n$. Here, the sampling is only based on the event indicator of the first event.

By selecting the more informative subjects for purposes of detailed covariate measurement, a more efficient generalized case-cohort design could be achieved. We can stratify all first event cases S_{cases} into strata ($S_{\text{cases},1}$, $S_{\text{cases},2}$, $S_{\text{cases},3}$) based on the observed time-to-event T_1 values using two cut-off values $c_{L1} < c_{U1}$ as in (2.2). Similarly, we can stratify all first event non-cases S_{noncases} into strata ($S_{\text{noncases},1}$, $S_{\text{noncases},2}$, $S_{\text{noncases},3}$) based on the observed censoring time C values using two cut-off values $c^*_{L1} < c^*_{U1}$ as in (2.3). The aim of Section 3.2 is to determine sampling probability of each defined stratum leading to a more efficient design while using the most efficient design (n_{cases}, n_{noncases}) obtained in Section 3.1. Here, the sampling is based on both the event indicator of the first event and the time-to-first event.

Under bivariate sequential survival data, a first event case could be either a second event case or a second event non-case. Let us denote $S_{\text{cases,cases}}$ as the subset of S_{cases} which are second event cases and $S_{\text{cases,noncases}}$ as the subset of S_{cases} which are second event non-cases. Using the most efficient design (n_{cases}, n_{noncases}) obtained in Section 3.1, the aim of Section 3.3 is to determine the number of second event cases (m_{cases}) versus the number of second event non-cases (m_{noncases}) that should be selected under the generalized case-cohort design during the sampling procedure where $m_{\text{cases}} + m_{\text{noncases}} = n_{\text{cases}}$. Here, the sampling is based on the event indicators of the two sequential events.

Greater efficiency may be achieved for generalized case-cohort design by selecting the more informative subjects for purposes of detailed covariate measurement. We can stratify all second event cases $S_{\text{cases,cases}}$ into strata ($S_{\text{cases,cases},1}$, $S_{\text{cases,cases},2}$, $S_{\text{cases,cases},3}$) based on the observed time-to-event T_2 values using two cut-off values $c_{L2} < c_{U2}$ as in (2.6). Similarly, we can stratify second event non-cases $S_{\text{cases,noncases}}$ into strata ($S_{\text{cases,noncases},1}$, $S_{\text{cases,noncases},2}$, $S_{\text{cases,noncases},3}$) based on the observed censoring time $C - T_1$ values using two cut-off values $c^*_{L2} < c^*_{U2}$ as in (2.7). The aim of Section 3.4 is to determine sampling probability of each defined stratum leading to a more efficient design using the most efficient design obtained in Section 3.2 and the most efficient design (m_{cases}, m_{noncases}) obtained in Section 3.3. Here, the sampling is based on both the two event indicators and the two sequential gap times.

Finally, the lowest standard errors of the coefficient estimate of the expensive covariate X obtained under the four different phase two sampling designs are compared in Section 3.5.

3.1 Efficiency of generalized case-cohort designs based on the first event indicator

Suppose we observed a large cohort of sequential survival data $\{(t_{1i}, \delta_{1i}, t_{2i}, \delta_{2i}) : i = 1, ..., N\}$, where $N = 50,000$ at phase one. This phase one sample is stratified based on the first event indicators of the survival data. We assume that 40% of the first event is censored. Thus, there are about $N_{\text{cases}} = 30,000$ individuals in the case stratum for the first event S_{cases} and about $N_{\text{noncases}} = 20,000$ individuals in the non-case stratum for the first event S_{noncases}.

A subsample of fixed size $n = 10,000$ is drawn at phase two in order to obtain the covariate which is costly or difficult to measure. The size of the subsample from the case stratum S_{cases} is denoted by n_{cases} and the size of the subsample from the non-case stratum S_{noncases} is denoted by n_{noncases}. Each allocation (n_{cases}, n_{noncases}) defines a generalized case-cohort design based on the first event indicator. Given the fixed size $n = 10,000$ of subsample, one may choose how to allocate it among the strata of phase one. The aim is to gain the efficiency when estimating the regression coefficient of the expensive covariate. Hence, we will determine n_{cases} (and therefore n_{noncases}) which leads to an efficient design where $n_{\text{cases}} + n_{\text{noncases}} = n$. For example, Table 3.1 shows the results of estimates and standard errors for model scenario $(\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 0.5)$, a model defined by (2.11) where $\alpha_{10} = 0.6$, $\alpha_{11} = 1.0$, $\gamma_1 = 0.5$ and by (2.12) where

Table 3.1: Coefficient estimates and their estimated standard errors under generalized case-cohort designs based on the first event indicator

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(1, 1, 0.5)$	1	(1000,9000)	1.013	0.0292	1.042	0.0596
	2	(2000,8000)	0.987	0.0261	1.045	0.0464
	3	(3000,7000)	0.994	0.0251	0.954	0.0408
	4	(4000,6000)	1.010	0.0243	1.019	0.0354
	5	(5000,5000)	0.974	0.0246	0.986	0.0333
	6	(6000,4000)	0.990	0.0250	0.993	0.0311
	7	(7000,3000)	0.964	0.0257	1.028	0.0298
	8	(8000,2000)	0.979	0.0268	0.964	0.0289
	9	(9000,1000)	1.029	0.0280	1.027	0.0277
	10	(10000,0)	0.960	0.0317	0.980	0.0288

$\alpha_{20} = 0.4$, $\alpha_{21} = 1.0$ and $\gamma_2 = 0.5$ as described in Section 2.2.1. Among the ten sampling scenarios, scenario 4 with $(n_{\text{cases}} = 4000,\ n_{\text{noncases}} = 6000)$ and scenario 5 with $(n_{\text{cases}} = n_{\text{noncases}} = 5000)$ give the minimum standard error estimates of the coefficient estimate of the expensive covariate for time to first event thus are the most efficient sampling designs. They will be used in both outcome-dependent BSS design based on time to first event and its event indicator and generalized case-cohort design based on first and second event indicators. Notice that these two sampling scenarios do not yield the most efficient designs for the coefficient estimate of the expensive covariate for time to second event. But we will address this when the sampling also depends on the second event outcome data.

Table 3.2: The most efficient sampling scenario under generalized case-cohort designs based on the first event indicator

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$
$(0, 0, 0.5)$	5	(5000,5000)
$(0, 1, 0.5)$	5	(5000,5000)
$(1, 0, 0.5)$	4	(4000,6000)

Continued on next page

Table 3.2 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$
$(1, 1, 0.5)$	4	(4000,6000)
$(0, 0, 1.0)$	4	(4000,6000)
$(0, 1, 1.0)$	4	(4000,6000)
$(1, 0, 1.0)$	5	(5000,5000)
$(1, 1, 1.0)$	5	(5000,5000)
$(0, 0, 1.5)$	4	(4000,6000)
$(0, 1, 1.5)$	4	(4000,6000)
$(1, 0, 1.5)$	9	(9000,1000)
$(1, 1, 1.5)$	8	(8000,2000)

The simulation results for other model scenarios are listed in Table A.1 of Appendix A. Table 3.2 summarizes the sampling scenario $(n_{\text{cases}}, n_{\text{noncases}})$ which minimizes the standard error estimate thus is the most efficient sampling scenario for the stratification based on the first event indicator under different model scenarios. It shows that the most efficient generalized case-cohort design $(n_{\text{cases}}, n_{\text{noncases}})$ based on the first event indicator is when $n_{\text{cases}} \approx n_{\text{noncases}}$. This is true for all model scenarios except two scenarios: $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 1.5)$ and $(\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 1.5)$. For these two model scenarios, when we increase sampling from the case stratum S_{cases}, the efficiency of the coefficient estimate of the expensive covariate for time to first event improves. The same conclusion can also be obtained from Figure 3.1 which provides the trend of the efficiency for both $\hat{\alpha}_{11}$ and $\hat{\alpha}_{21}$ at various sampling scenarios under different model scenarios.

Figure 3.1 shows that the most efficient sampling design for $\hat{\alpha}_{11}$ does not yield the most efficient designs for $\hat{\alpha}_{21}$. This is true for all model scenarios except two scenarios: $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 1.5)$ and $(\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 1.5)$. For these two model scenarios, when we increase sampling from the case stratum S_{cases}, the estimated standard errors of both $\hat{\alpha}_{11}$ and $\hat{\alpha}_{21}$ decrease.

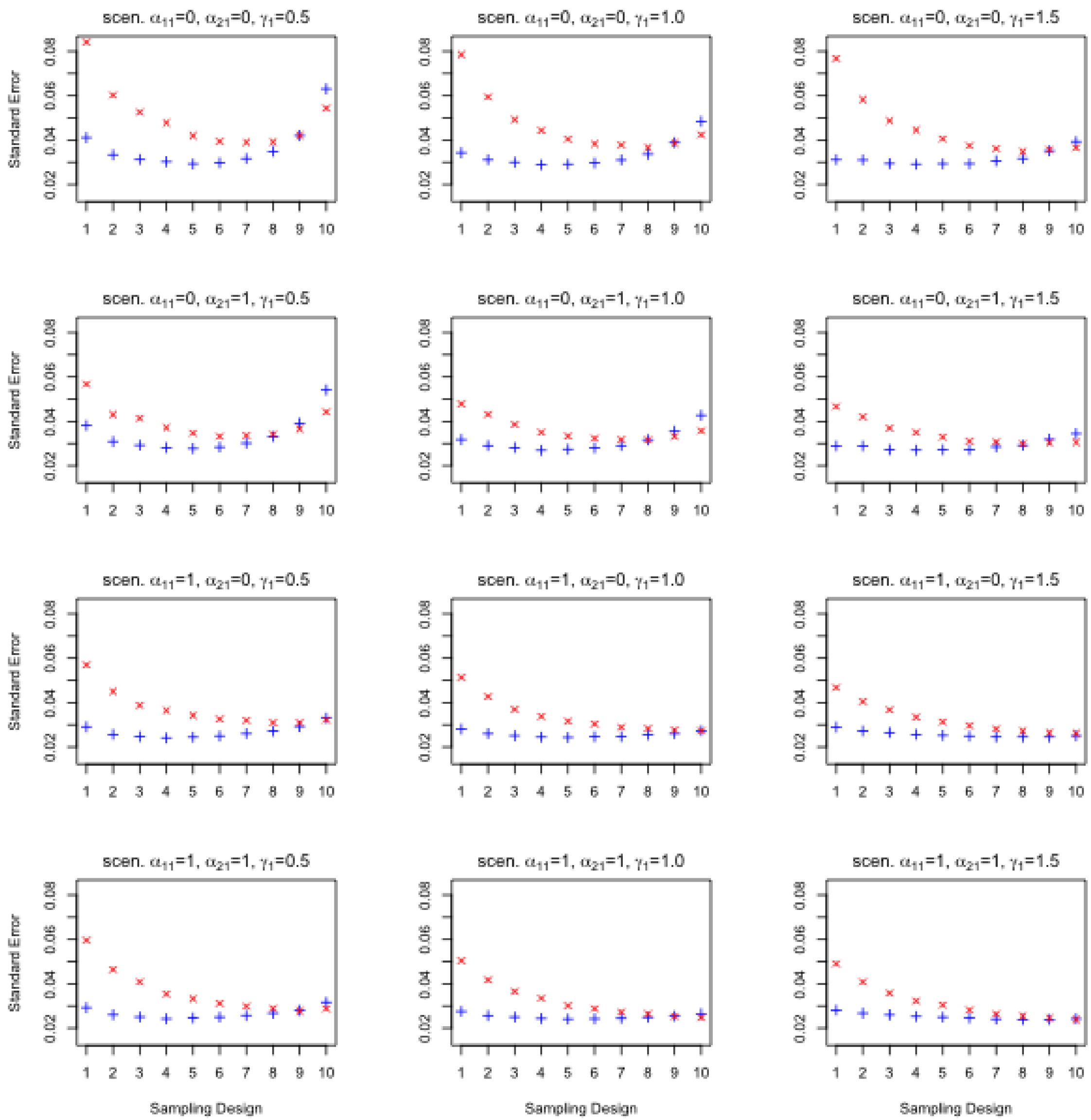

Figure 3.1: Estimated standard errors of the coefficient estimates of the expensive covariate under generalized case-cohort designs based on the first event indicator
$+$ represents standard error of $\hat{\alpha}_{11}$
$\times$ represents standard error of $\hat{\alpha}_{21}$
The sampling scenarios $1, \ldots, 10$ are described in Table 3.1

3.2 Efficiency of outcome-dependent BSS designs based on the first gap time and its event indicator

In order to achieve the possible efficiency gain of generalized case-cohort design, the sampling of subjects could be done such that the sample is enriched with subjects who are especially informative. We can stratify all first event cases S_{cases} into strata $(S_{\text{cases},1}, S_{\text{cases},2}, S_{\text{cases},3})$ based on the observed time-to-event T_1 values using two cut-off values $c_{L1} < c_{U1}$ as in (2.2). Similarly, we can stratify all first event non-cases S_{noncases} into strata ($S_{\text{noncases},1}$, $S_{\text{noncases},2}$, $S_{\text{noncases},3}$) based on the observed censoring time C values using two cut-off values $c^*_{L1} < c^*_{U1}$ as in (2.3).

After obtaining the most efficient sampling design (n_{cases}, n_{noncases}) for the strata (S_{cases}, S_{noncases}) in Section 3.1, we do outcome-dependent BSS on the strata ($S_{\text{cases},1}$, $S_{\text{cases},2}$, $S_{\text{cases},3}$) and $(S_{\text{noncases},1}, S_{\text{noncases},2}, S_{\text{noncases},3})$. Suppose the size of the subsample from the stratum $S_{\text{cases},j}$ is denoted by $n_{\text{cases},j}$, $j = 1, 2, 3$, where $\sum_{j=1}^{3} n_{\text{cases},j} = n_{\text{cases}}$. Similarly, suppose the size of the subsample from the stratum $S_{\text{noncases},j}$ is denoted by $n_{\text{noncases},j}$, $j = 1, 2, 3$, where $\sum_{j=1}^{3} n_{\text{noncases},j} = n_{\text{noncases}}$. Given the fixed sizes (n_{cases}, n_{noncases}) of samples, one may choose how to allocate it among the strata $((S_{\text{cases},j} : j = 1, 2, 3), (S_{\text{noncases},j} : j = 1, 2, 3))$. Different allocations $((n_{\text{cases},j} : j = 1, 2, 3), (n_{\text{noncases},j} : j = 1, 2, 3))$ define different outcome-dependent BSS designs based on the first gap time T_1 and its event indicator δ_1.

The aim is to determine $n_{\text{cases},j}$ and $n_{\text{noncases},j}$, $j = 1, 2, 3$, which lead to an efficient design where $\sum_{j=1}^{3} n_{\text{cases},j} = n_{\text{cases}}$ and $\sum_{j=1}^{3} n_{\text{noncases},j} = n_{\text{noncases}}$. Table 3.3 shows the results of estimates and their standard errors under different allocations (($n_{\text{cases},1}$, $n_{\text{cases},2}$, $n_{\text{cases},3}$), ($n_{\text{noncases},1}$, $n_{\text{noncases},2}$, $n_{\text{noncases},3}$)) for model scenario ($\alpha_{11} = 1, \alpha_{21} =$

$1, \gamma_1 = 0.5)$, a model defined by (2.11) where $\alpha_{10} = 0.6$, $\alpha_{11} = 1.0$, $\gamma_1 = 0.5$ and by (2.12) where $\alpha_{20} = 0.4$, $\alpha_{21} = 1.0$ and $\gamma_2 = 0.5$ as described in Section 2.2.1. We see that sampling scenario 3 with $((n_{\text{cases},j} : j = 1,2,3), (n_{\text{noncases},j} : j = 1,2,3)) = ((4000,0,0),(0,1000,5000))$ minimizes the standard error ($\widehat{\text{SE}}(\hat{\alpha}_{11})$) thus is the most efficient sampling scenario. In scenario 3, there is an increased sampling from the first case stratum $S_{\text{cases},1}$. Selecting individuals with shorter time to first event yields more efficient coefficient estimate. We can see this by looking at sampling scenarios 5, 6, 8, and 9 as well. In addition, in scenario 3, there is an increased sampling from the third non-case stratum $S_{\text{noncases},3}$. When we increase sampling from the stratum with long censoring time, the efficiency improves. We can see this by looking at sampling scenarios 5, 6, 8, and 9 as well. Notice that sampling scenarios 1, 4, and 7 yield larger standard error compared to SRS in S_{cases} and S_{noncases}. These three sampling scenarios with increased sampling from the stratum with short censoring time lead to inefficient designs. Sampling scenario 7 led to largest standard error with increased sampling from both the stratum with large T_1 and the stratum with short censoring time.

The most efficient scenario 3 is used in outcome-dependent BSS design based on the first and second gap times and their event indicators in Section 3.4.

Table 3.3: Coefficient estimates and their estimated standard errors under outcome-dependent BSS designs based on the first gap time and its event indicator

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases},j} : j = 1,2,3)$, $(n_{\text{noncases},j} : j = 1,2,3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(1, 1, 0.5)$	SRS in S_{cases} and S_{noncases}	(666,2668,666),(1500,3000,1500)	1.019	0.0242	1.021	0.0355
	1	(4000,0,0),(5000,1000,0)	0.979	0.0258	0.969	0.0399
	2	(4000,0,0),(0,6000,0)	1.015	0.0205	1.002	0.0373
	3	(4000,0,0),(0,1000,5000)	1.002	0.0189	0.971	0.0369
	4	(0,4000,0),(5000,1000,0)	0.988	0.0321	1.012	0.0367
	5	(3000,1000,0),(0,1000,5000)	1.010	0.0194	0.992	0.0358
	6	(2000,1000,1000),(0,1000,5000)	1.0238	0.0201	0.988	0.0363
	7	(0,0,4000),(5000,1000,0)	0.968	0.0490	0.960	0.0436
	8	(4000,0,0),(1000,1000,4000)	1.009	0.0198	0.976	0.0372
	9	(4000,0,0),(1000,2000,3000)	1.007	0.0201	0.990	0.0372

The simulation results for other model scenarios are listed in Table A.2 of Appendix A. Notice that the first allocation in each model scenario in Table A.2 is a SRS in S_{cases} and S_{noncases} which is defined by (2.4) and (2.5). Thus, it is a generalized case-cohort design.

Table 3.4: The most efficient sampling scenario under outcome-dependent BSS designs based on the first gap time and its event indicator

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases},j} : j = 1, 2, 3)$, $(n_{\text{noncases},j} : j = 1, 2, 3)$
$(0, 0, 0.5)$	3	(5000,0,0),(0,0,5000)
$(0, 1, 0.5)$	3	(5000,0,0),(0,0,5000)
$(1, 0, 0.5)$	3	(4000,0,0),(0,1000,5000)
$(1, 1, 0.5)$	3	(4000,0,0),(0,1000,5000)
$(0, 0, 1.0)$	3	(4000,0,0),(0,1000,5000)
$(0, 1, 1.0)$	3	(4000,0,0),(0,1000,5000)
$(1, 0, 1.0)$	3	(5000,0,0),(0,0,5000)
$(1, 1, 1.0)$	3	(5000,0,0),(0,0,5000)
$(0, 0, 1.5)$	3	(4000,0,0),(0,1000,5000)
$(0, 1, 1.5)$	3	(4000,0,0),(0,1000,5000)
$(1, 0, 1.5)$	2	(1000,4000,4000),(0,0,1000)
$(1, 1, 1.5)$	2	(1000,3000,4000),(0,0,2000)

Table 3.4 summarizes the sampling scenario $((n_{\text{cases},j} : j = 1, 2, 3), (n_{\text{noncases},j} : j = 1, 2, 3))$ which minimizes the standard error thus is the most efficient sampling scenario for stratification based on the first event time and its event indicator under different model scenarios. It shows that the most efficient outcome-dependent BSS

design $((n_{\text{cases},j} : j = 1,2,3), (n_{\text{noncases},j} : j = 1,2,3))$ based on the first event time and its event indicator is the sampling scenario 3 where we increase sampling from the stratum with short first event time (i.e., the first case stratum $S_{\text{cases},1}$) and also increase sampling from the stratum with long censoring time (i.e., the third non-case stratum $S_{\text{noncases},3}$). This is true for all model scenarios except two model scenarios: $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 1.5)$ and $(\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 1.5)$. For these two model scenarios, when we increase sampling from the midrange and long first event time strata (i.e., the second and third case strata $S_{\text{cases},2}$, $S_{\text{cases},3}$) and also increase sampling from the long censoring time stratum (i.e., the third non-case stratum $S_{\text{noncases},3}$), the efficiency of the coefficient estimate of the expensive covariate for time to first event improves. The same conclusion can also be obtained from Figure 3.2 which provides the trend of the efficiency for both $\hat{\alpha}_{11}$ and $\hat{\alpha}_{21}$ at various sampling scenarios under different model scenarios.

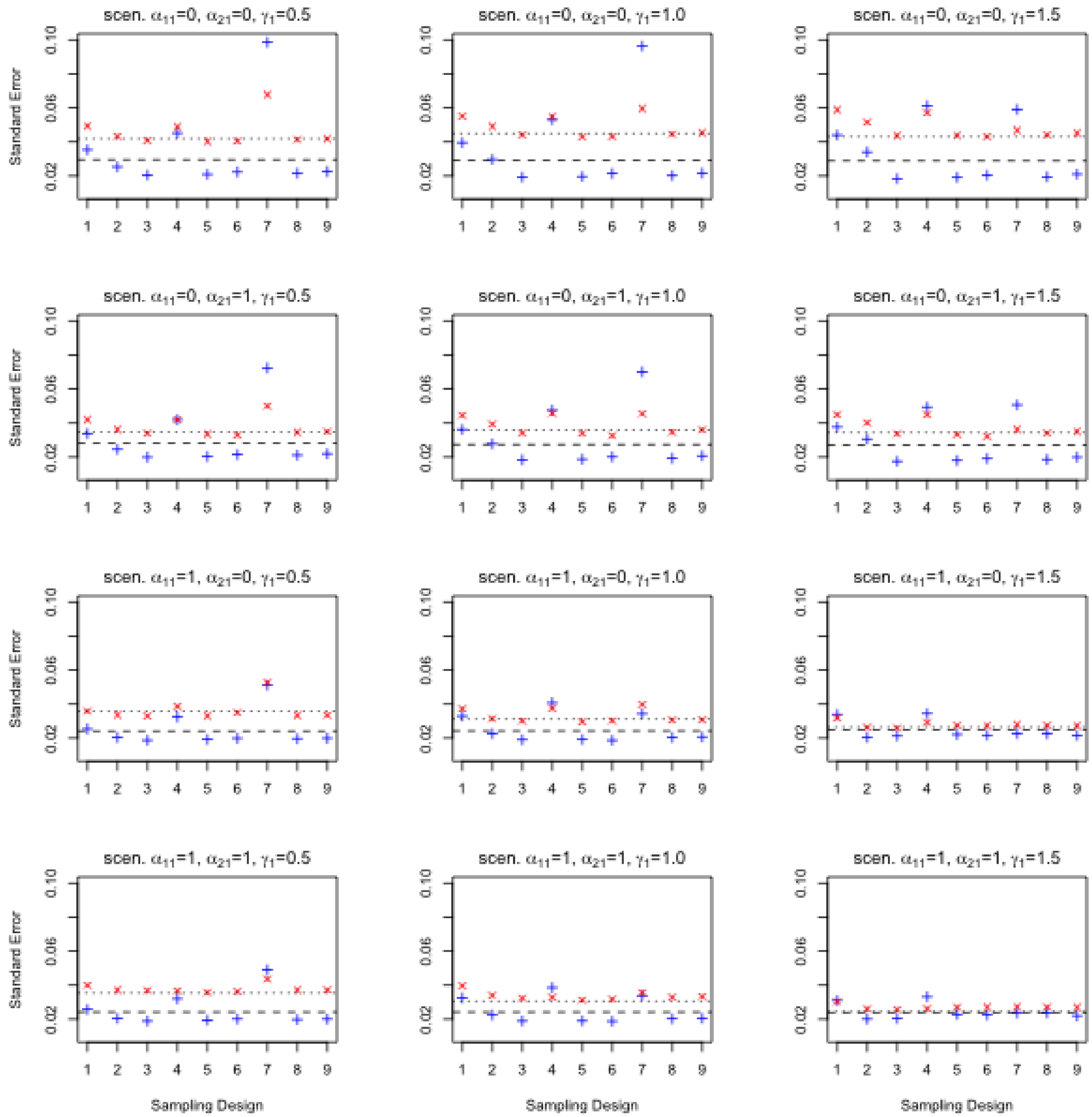

Figure 3.2: Estimated standard errors of the coefficient estimates of the expensive covariate under outcome-dependent BSS designs based on the first gap time and its event indicator
+ represents standard error of $\hat{\alpha}_{11}$
× represents standard error of $\hat{\alpha}_{21}$
dashed line represents standard error of $\hat{\alpha}_{11}$ under SRS in S_{cases} and S_{noncases}
dotted line represents standard error of $\hat{\alpha}_{21}$ under SRS in S_{cases} and S_{noncases}
The sampling scenarios $1, \ldots, 9$ are described in Table A.2

3.3 Efficiency of generalized case-cohort designs based on the event indicators of the two sequential gap times

In Section 3.1, a subsample of fixed size ($n = 10,000$) was drawn in order to obtain a covariate which is expensive to measure based on the first event indicator. Table 3.2 provides us the most efficient sampling scenario for stratification based on the first event indicator under different model scenarios. For example, sampling scenario ($n_{\text{cases}} = 4000$, $n_{\text{noncases}} = 6000$) minimizes the standard error thus is the most efficient sampling scenario for model scenario ($\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 0.5$). The above efficient sampling design minimizes the variance of the coefficient estimate of the expensive covariate for the first gap time. We are also interested in looking for efficient sampling designs which minimize the variance of the coefficient estimate of the expensive covariate for the second gap time.

In this section, a subsample of fixed size ($n = 10,000$) is drawn in order to obtain a covariate which is expensive to measure based on the event indicators of the two sequential gap times. Suppose (n_{cases}, n_{noncases}) is the most efficient sampling scenario for stratification based on the first event indicator. First a subsample of size n_{noncases} is drawn from the first event non-case stratum S_{noncases}. Then a subsample of size n_{cases} can be drawn from the first event case stratum S_{cases} based on the second event indicator. Note that under bivariate sequential survival data, a T_1 case could be either a T_2 case or a T_2 non-case. Let us denote $S_{\text{cases,cases}}$ as the subset of S_{cases} which includes T_2 cases and $S_{\text{cases,noncases}}$ as the subset of S_{cases} which includes T_2 non-cases. The size of the subsample from the first and second event case stratum $S_{\text{cases,cases}}$ is denoted by m_{cases} and the size of the subsample from the first event case and second event non-case stratum $S_{\text{cases,noncases}}$ is denoted by m_{noncases}, where $n_{\text{cases}} = m_{\text{cases}} + m_{\text{noncases}}$.

Given the fixed size n_{cases} of subsample, we investigate how to allocate it among the strata ($S_{\text{cases,cases}}$, $S_{\text{cases,noncases}}$) which is based on T_2 event indicator. Different allocations (m_{cases}, m_{noncases}) in addition to selecting n_{noncases} individuals from S_{noncases} define different generalized case-cohort designs based on the event indicators of the two sequential gap times.

We need to determine m_{cases} and m_{noncases} which lead to an efficient design where $m_{\text{cases}} + m_{\text{noncases}} = n_{\text{cases}}$. Efficient sampling design minimizes the variance of the coefficient estimate of the expensive covariate for the second gap time. Table 3.5 shows the results of estimates and standard errors for model scenario ($\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 0.5$), a model defined by (2.11) where $\alpha_{10} = 0.6$, $\alpha_{11} = 1.0$, $\gamma_1 = 0.5$ and by (2.12) where $\alpha_{20} = 0.4$, $\alpha_{21} = 1.0$ and $\gamma_2 = 0.5$ as described in Section 2.2.1. We see that sampling scenario 5 with ($m_{\text{cases}} = 2500$, $m_{\text{noncases}} = 1500$) minimizes the standard error estimate of $\hat{\alpha}_{21}$, thus is the most efficient sampling scenario based on $\widehat{\text{SE}}(\hat{\alpha}_{21})$. It will be used in outcome-dependent BSS design based on the two sequential gap times and their event indicators. On the other hand, sampling scenario 8 with ($m_{\text{cases}} = 4000$, $m_{\text{noncases}} = 0$) minimizes the standard error estimate of $\hat{\alpha}_{11}$ thus is the most efficient sampling scenario based on $\widehat{\text{SE}}(\hat{\alpha}_{11})$.

Table 3.5: Coefficient estimates and their estimated standard errors under generalized case-cohort designs based on the event indicators of the two sequential gap times

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	(m_{cases}, m_{noncases})	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(1, 1, 0.5)$	1	(500,3500)	1.006	0.0288	1.063	0.0473
	2	(1000,3000)	1.000	0.0269	1.008	0.0423
	3	(1500,2500)	0.983	0.0262	0.983	0.0390
	4	(2000,2000)	0.998	0.0252	0.995	0.0371

Continued on next page

Table 3.5 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	(m_{cases}, m_{noncases})	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	5	(2500,1500)	1.009	0.0244	1.028	0.0360
	6	(3000,1000)	1.011	0.0241	0.976	0.0361
	7	(3500,500)	0.993	0.0238	0.959	0.0362
	8	(4000,0)	1.001	0.0233	1.026	0.0370

Table 3.6: The most efficient sampling scenario under generalized case-cohort designs based on the event indicators of the two sequential gap times

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	(m_{cases}, m_{noncases})
$(0, 0, 0.5)$	10	(5000,0)
$(0, 1, 0.5)$	10	(5000,0)
$(1, 0, 0.5)$	8	(4000,0)
$(1, 1, 0.5)$	5	(2500,1500)
$(0, 0, 1.0)$	6	(3000,1000)
$(0, 1, 1.0)$	6	(3000,1000)
$(1, 0, 1.0)$	7	(3500,1500)
$(1, 1, 1.0)$	5	(2500,2500)
$(0, 0, 1.5)$	7	(3500,500)
$(0, 1, 1.5)$	7	(3500,500)
$(1, 0, 1.5)$	10	(5000,4000)
$(1, 1, 1.5)$	10	(5000,3000)

The simulation results for other model scenarios are listed in Table A.3 of Appendix A. Table 3.6 summarizes the sampling scenario (m_{cases}, m_{noncases}) which minimizes the standard error estimate of $\hat{\alpha}_{21}$ thus is the most efficient sampling scenario based on $\widehat{\mathrm{SE}}(\hat{\alpha}_{21})$ for stratification based on the event indicators of the two sequential gap times under different model scenarios. It shows that, when we increase sampling from the stratum $S_{\text{cases,cases}}$, the efficiency of the coefficient estimate of the expensive covariate for time to second event improves. This is true for all model scenarios except the two scenarios ($\alpha_{11} = 1,\ \alpha_{21} = 1,\ \gamma_1 = 1.0$) and ($\alpha_{11} = 1,\ \alpha_{21} = 0,\ \gamma_1 = 1.5$). For these two model scenarios, the estimated standard errors of $\hat{\alpha}_{21}$ minimizes when $m_{\text{cases}} \approx m_{\text{noncases}}$.

3.4 Efficiency of outcome-dependent BSS designs based on the two sequential gap times and their event indicators

As indicated in Section 3.1, the most efficient sampling design for $\hat{\alpha}_{11}$ based on the first event indicator does not necessarily yield the most efficient sampling design for $\hat{\alpha}_{21}$. To address this, in addition to sampling based on the event indicators, now we consider sampling based on the two sequential gap times. We stratify all T_2 cases $S_{\text{cases,cases}}$ into strata ($S_{\text{cases,cases},1}$, $S_{\text{cases,cases},2}$, $S_{\text{cases,cases},3}$) based on the observed time-to-second event using two cut-off values $c_{L2} < c_{U2}$ as in (2.6). Similarly, we can stratify T_2 non-cases $S_{\text{cases,noncases}}$ into strata ($S_{\text{cases,noncases},1}$, $S_{\text{cases,noncases},2}$, $S_{\text{cases,noncases},3}$) based on observed censoring time $C - T_1$ values using two cut-off values $c^*_{L2} < c^*_{U2}$ as in (2.7).

In Section 3.1, a subsample of fixed size ($n = 10,000$) is drawn from a large cohort of sequential survival data of size $N = 50,000$ under generalized case-cohort designs based on the first event indicator. Table 3.2 provides us the most efficient sampling scenarios ($n_{\text{cases}}, n_{\text{noncases}}$) for $\hat{\alpha}_{11}$ under different model scenarios, where $n_{\text{cases}} + n_{\text{noncases}} = n$.

After obtaining the most efficient sampling design (n_{cases}, n_{noncases}) in Section 3.1, we considered outcome-dependent BSS based on the first gap time and its event indicator in Section 3.2. Table 3.4 summarizes the most efficient sampling scenarios $((n_{\text{cases},j} : j = 1,2,3), (n_{\text{noncases},j} : j = 1,2,3))$ for $\hat{\alpha}_{11}$ under different model scenarios, where $\sum_{j=1}^{3} n_{\text{cases},j} = n_{\text{cases}}$ and $\sum_{j=1}^{3} n_{\text{noncases},j} = n_{\text{noncases}}$. These efficient sampling designs minimize the variance of $\hat{\alpha}_{11}$. We are also interested in looking for efficient sampling designs which minimize the variance of $\hat{\alpha}_{21}$. After obtaining the most efficient sampling design (n_{cases}, n_{noncases}) in Section 3.1, a subsample of size n_{cases} was drawn from the first event case stratum S_{cases} under generalized case-cohort designs based on the second event indicator in Section 3.3. Table 3.6 summarizes the most efficient sampling scenarios (m_{cases}, m_{noncases}) for $\hat{\alpha}_{21}$ under different model scenarios, where $n_{\text{cases}} = m_{\text{cases}} + m_{\text{noncases}}$.

After obtaining the most efficient sampling design (m_{cases}, m_{noncases}) for the strata ($S_{\text{cases,cases}}$, $S_{\text{cases,noncases}}$), we do outcome-dependent BSS on the strata ($S_{\text{cases,cases},1}$, $S_{\text{cases,cases},2}$, $S_{\text{cases,cases},3}$) and ($S_{\text{cases,noncases},1}$, $S_{\text{cases,noncases},2}$, $S_{\text{cases,noncases},3}$). Suppose the size of the subsample from the stratum $S_{\text{cases,cases},j}$ is denoted by $m_{\text{cases},j}$, $j = 1,2,3$, where $\sum_{j=1}^{3} m_{\text{cases},j} = m_{\text{cases}}$. Similarly, suppose the size of the subsample from the stratum $S_{\text{cases,noncases},j}$ is denoted by $m_{\text{noncases},j}$, $j = 1,2,3$, where $\sum_{j=1}^{3} m_{\text{noncases},j} = m_{\text{noncases}}$. Given the fixed sizes (m_{cases}, m_{noncases}) of subsamples, one may choose how to allocate it among the strata $((S_{\text{cases,cases},j} : j = 1,2,3), (S_{\text{cases,noncases},j} : j = 1,2,3))$. Different allocations $((m_{\text{cases},j} : j = 1,2,3), (m_{\text{noncases},j} : j = 1,2,3))$ define different outcome-dependent BSS designs based on the second gap time T_2 and its event indicator.

We need to determine $m_{\text{cases},j}$ and $m_{\text{noncases},j}$, $j = 1,2,3$, which lead to an efficient design where $\sum_{j=1}^{3} m_{\text{cases},j} = m_{\text{cases}}$ and $\sum_{j=1}^{3} m_{\text{noncases},j} = m_{\text{noncases}}$. Efficient sampling design minimizes the variance of the coefficient estimate of the expensive covariate for time-to-event T_2. Table 3.7 shows the results of estimates and standard errors for different allocations $((m_{\text{cases},1},\, m_{\text{cases},2},\, m_{\text{cases},3}),\, (m_{\text{noncases},1},\, m_{\text{noncases},2},\, m_{\text{noncases},3}))$ for model scenario $(\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 0.5)$, a model defined by (2.11) where $\alpha_{10} = 0.6$, $\alpha_{11} = 1.0$, $\gamma_1 = 0.5$ and by (2.12) where $\alpha_{20} = 0.4$, $\alpha_{21} = 1.0$ and $\gamma_2 = 0.5$ as described in Section 2.2.1. We see that sampling scenario 3 with $((m_{\text{cases},j} : j = 1,2,3),$ $(m_{\text{noncases},j} : j = 1,2,3)) = ((2500,0,0),(0,0,1500))$ minimizes the standard error estimate of $\hat{\alpha}_{21}$ thus is the most efficient sampling scenario based on $\widehat{\text{SE}}(\hat{\alpha}_{21})$. In the most efficient scenario 3, there is an increased sampling from the first T_2 case stratum $S_{\text{cases,cases},1}$. When we increase sampling from the stratum with short time-to-second event, the efficiency improves. On the other hand, in the most efficient scenario 3, there is an increased sampling from the third T_2 non-case stratum $S_{\text{cases,noncases},3}$. When we increase sampling from the stratum with long censoring times, the efficiency improves. Notice that sampling scenarios 4, 7 and 8 have larger standard error estimates compared to other sampling scenarios. These three sampling scenarios increase sampling from the stratum with long time-to-second event and/or the short censoring time which yield inefficient designs.

In sampling scenario 3 with $((m_{\text{cases},j} : j = 1,2,3),\ (m_{\text{noncases},j} : j = 1,2,3)) = ((2500,0,0),(0,0,1500))$, we allocate $m_{\text{cases},1} = 2500$ to the intersection of the first (short) T_2 case stratum $S_{\text{cases,cases},1}$ and the first (short) T_1 case stratum $S_{\text{cases},1}$. When $m_{\text{cases},1}$ is larger than the number of individuals in the intersection $S_{\text{cases,cases},1} \cap S_{\text{cases},1}$, the remaining could be allocated to either $S_{\text{cases,cases},1} \cap S_{\text{cases},2}$ or $S_{\text{cases,cases},2} \cap S_{\text{cases},1}$. The first approach ensures gain in efficiency for the estimation of the regression coefficient of the expensive covariate for the second event time as shown in Table 3.7 with $\widehat{\text{SE}}(\hat{\alpha}_{11}) = 0.0221$ and $\widehat{\text{SE}}(\hat{\alpha}_{21}) = 0.0253$. On the other hand, the second approach will gain efficiency when estimating the regression coefficient of the expensive covariate for the first event time with $\widehat{\text{SE}}(\hat{\alpha}_{11}) = 0.0209$ and $\widehat{\text{SE}}(\hat{\alpha}_{21}) = 0.0288$.

Table 3.7: Coefficient estimates and their estimated standard errors under outcome-dependent BSS designs based on the two sequential gap times and their event indicators

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases},j} : j = 1,2,3)$, $(m_{\text{noncases},j} : j = 1,2,3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(1, 1, 0.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(316,1868,316),(367,766,367)	1.003	0.0204	1.021	0.0441
	1	(2500,0,0),(1500,0,0)	1.006	0.0211	1.000	0.0310
	2	(2500,0,0),(0,1500,0)	1.004	0.0215	0.993	0.0277
	3	(2500,0,0),(0,0,1500)	0.993	0.0221	1.021	0.0253
	4	(0,2500,0),(1500,0,0)	1.011	0.0206	1.021	0.0618
	5	(0,2500,0),(0,1500,0)	1.003	0.0205	1.003	0.0518
	6	(0,2500,0),(0,0,1500)	1.010	0.0200	1.043	0.0420
	7	(0,0,2500),(1500,0,0)	1.025	0.0260	1.046	0.0757
	8	(0,0,2500),(0,1500,0)	1.030	0.0269	1.032	0.0670
	9	(0,0,2500),(0,0,1500)	1.034	0.0255	1.068	0.0522

The simulation results for other model scenarios are listed in Table A.4 of Appendix A. Notice that the first allocation in each model scenario in Table A.4 is a SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$ which is defined by (2.8) and (2.9). Thus, it is a generalized case-cohort design.

Table 3.8: The most efficient sampling scenario under outcome-dependent BSS designs based on the two sequential gap times and their event indicators

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases},j} : j = 1,2,3)$, $(m_{\text{noncases},j} : j = 1,2,3)$
$(0, 0, 0.5)$	3	(2500,2500,0),(0,0,0)
$(0, 1, 0.5)$	3	(2500,2500,0),(0,0,0)
$(1, 0, 0.5)$	3	(2500,1500,0),(0,0,0)
$(1, 1, 0.5)$	3	(2500,0,0),(0,0,1500)
$(0, 0, 1.0)$	3	(2500,500,0),(0,0,1000)
$(0, 1, 1.0)$	3	(2500,500,0),(0,0,1000)
$(1, 0, 1.0)$	3	(2500,1000,0),(0,0,1500)
$(1, 1, 1.0)$	3	(2500,0,0),(0,0,2500)
$(0, 0, 1.5)$	3	(2500,100,0),(0,0,500)

Continued on next page

Table 3.8 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases},j} : j = 1,2,3)$, $(m_{\text{noncases},j} : j = 1,2,3)$
$(0, 1, 1.5)$	3	(2500,1000,0),(0,0,500)
$(1, 0, 1.5)$	3	(2500,2500,0),(0,1500,2500)
$(1, 1, 1.5)$	3	(2500,2500,0),(0,500,2500)

Table 3.8 summarizes the sampling scenario $((m_{\text{cases},j} : j = 1,2,3), (m_{\text{noncases},j} : j = 1,2,3))$ which minimizes the standard error estimate of $\hat{\alpha}_{21}$ thus is the most efficient sampling scenario based on $\widehat{\text{SE}}(\hat{\alpha}_{21})$ for stratification based on the two sequential gap times and their event indicators under different model scenarios. It shows that the most efficient outcome-dependent BSS design $((m_{\text{cases},j} : j = 1,2,3),$ $(m_{\text{noncases},j} : j = 1,2,3))$ based on the two sequential gap times and their event indicators is the sampling scenario 3 where we increase sampling from the stratum with short second event times (i.e., the first T_2 case stratum $S_{\text{cases,cases,1}}$) and also increase sampling from the stratum with long censoring times (i.e., the third T_2 non-case stratum $S_{\text{cases,noncases,3}}$). This is true for all model scenarios. The same conclusion can also be obtained from Figure 3.3 which provides the trend of the efficiency for both $\hat{\alpha}_{11}$ and $\hat{\alpha}_{21}$ at various sampling scenarios under different model scenarios.

Notice that in Table 3.8, the sum of $m_{\text{cases},j}$, $j = 1,2,3$, is m_{cases} and the sum of $m_{\text{noncases},j}$, $j = 1,2,3$, is m_{noncases}, where $(m_{\text{cases}}, m_{\text{noncases}})$ is selected based on the most efficient design identified in Table 3.6.

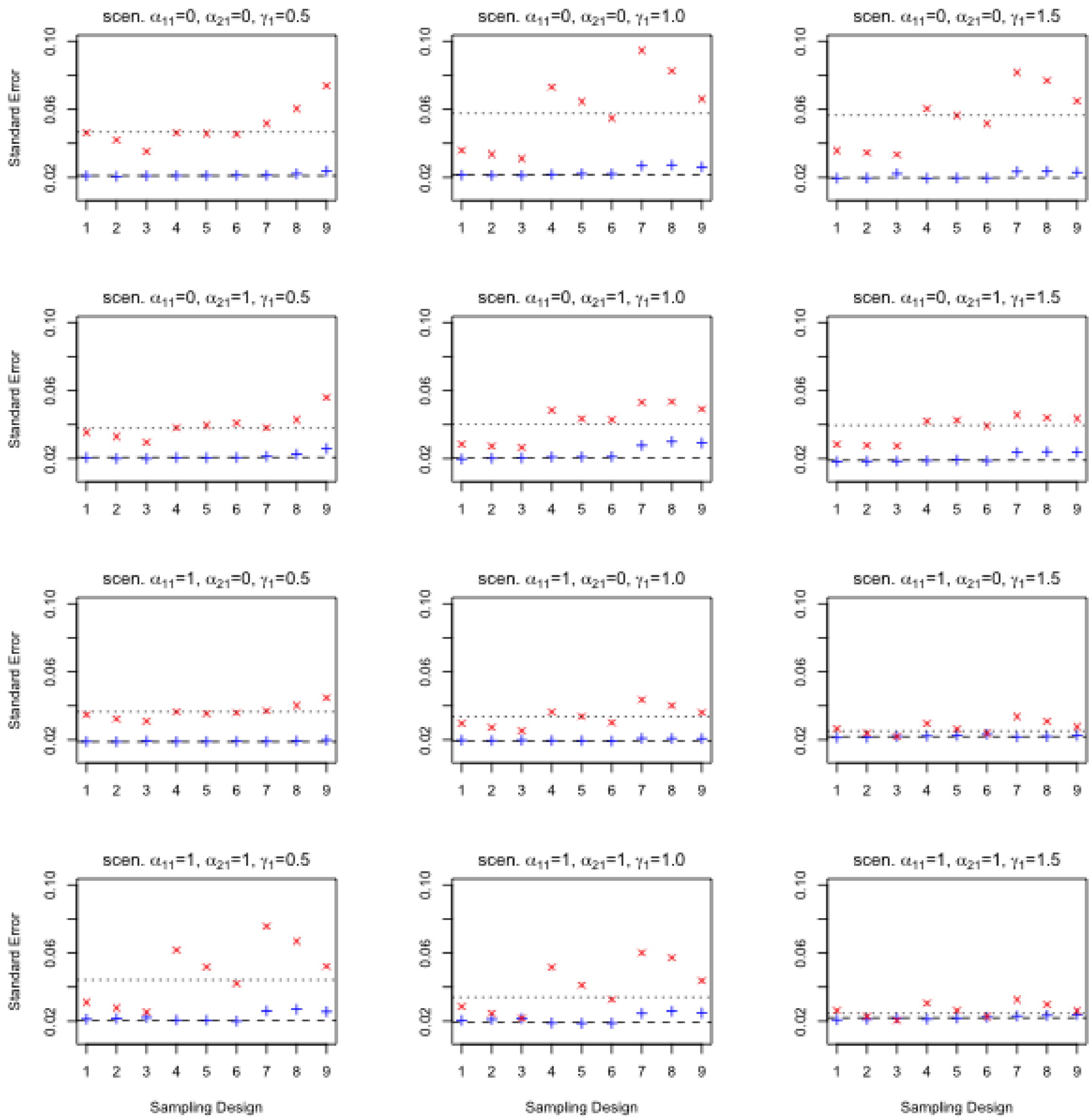

Figure 3.3: Standard errors of the coefficient estimates of the expensive covariate under outcome-dependent BSS designs based on the two sequential gap times and their event indicators

+ represents standard error of $\hat{\alpha}_{11}$

× represents standard error of $\hat{\alpha}_{21}$

dashed line represents standard error of $\hat{\alpha}_{11}$ under SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$

dotted line represents standard error of $\hat{\alpha}_{21}$ under SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$

The sampling scenarios $1, \ldots, 9$ are given in Table A.4

3.5 Summary

Table 3.9 and Figure 3.4 summarize standard errors of $\hat{\alpha}_{11}$ and $\hat{\alpha}_{21}$ for the most efficient sampling scenarios under two-phase outcome-dependent sampling designs for different model scenarios when the dependence between the two sequential gap times is moderate. Design 1 represents a generalized case-cohort design based on the first event indicator. Design 2 represents an outcome-dependent BSS design based on the first gap time and its event indicator. Design 3 represents a generalized case-cohort design based on the event indicators of the two sequential gap times. Design 4 represents an outcome-dependent BSS design based on the two sequential gap times and their event indicators. Recall that the most efficient sampling scenarios for design 1 and design 2 are based on $\widehat{\text{SE}}(\hat{\alpha}_{11})$. On the other hand, the most efficient sampling scenarios for design 3 and design 4 are based on $\widehat{\text{SE}}(\hat{\alpha}_{21})$.

Under design 2, there is a gain on efficiency when estimating the regression coefficient of the expensive covariate for time to first event compared with design 1. Also, under design 4, there is a gain on efficiency when estimating the regression coefficient of the expensive covariate for time to second event compared with design 2. Moreover, under design 4, the difference between standard errors of $\hat{\alpha}_{11}$ and $\hat{\alpha}_{21}$ for the most efficient sampling scenario is reduced. Therefore, design 4 (i.e., outcome-dependent BSS design based on the two sequential gap times and their event indicators) is recommended.

Table 3.9: Lowest standard errors of the coefficient estimates under two-phase outcome-dependent sampling designs

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	standard errors	design 1	design 2	design 3	design 4
$(0, 0, 0.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0293	0.0205	0.0239	0.0209
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0419	0.0408	0.0396	0.0352
$(0, 1, 0.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0278	0.0199	0.0262	0.0200
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0347	0.0339	0.0349	0.0296
$(1, 0, 0.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0241	0.0187	0.0232	0.0193
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0364	0.0330	0.0335	0.0310
$(1, 1, 0.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0243	0.0189	0.0244	0.0221
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0354	0.0369	0.0360	0.0253
$(0, 0, 1.0)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0289	0.0192	0.0285	0.0211
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0444	0.0441	0.0426	0.0311
$(0, 1, 1.0)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0270	0.0183	0.0273	0.0204
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0350	0.0341	0.0355	0.0265
$(1, 0, 1.0)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0243	0.0191	0.0244	0.0198
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0316	0.0301	0.0313	0.0253
$(1, 1, 1.0)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0241	0.0189	0.0246	0.0215
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0301	0.0324	0.0305	0.0217
$(0, 0, 1.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0291	0.0184	0.0275	0.0222
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0445	0.0437	0.0418	0.0334
$(0, 1, 1.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0270	0.0175	0.0260	0.0185
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0350	0.0336	0.0341	0.0274
$(1, 0, 1.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0245	0.0205	0.0236	0.0220
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0265	0.0263	0.0259	0.0219
$(1, 1, 1.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0239	0.0201	0.0238	0.0218
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0255	0.0261	0.0245	0.0204

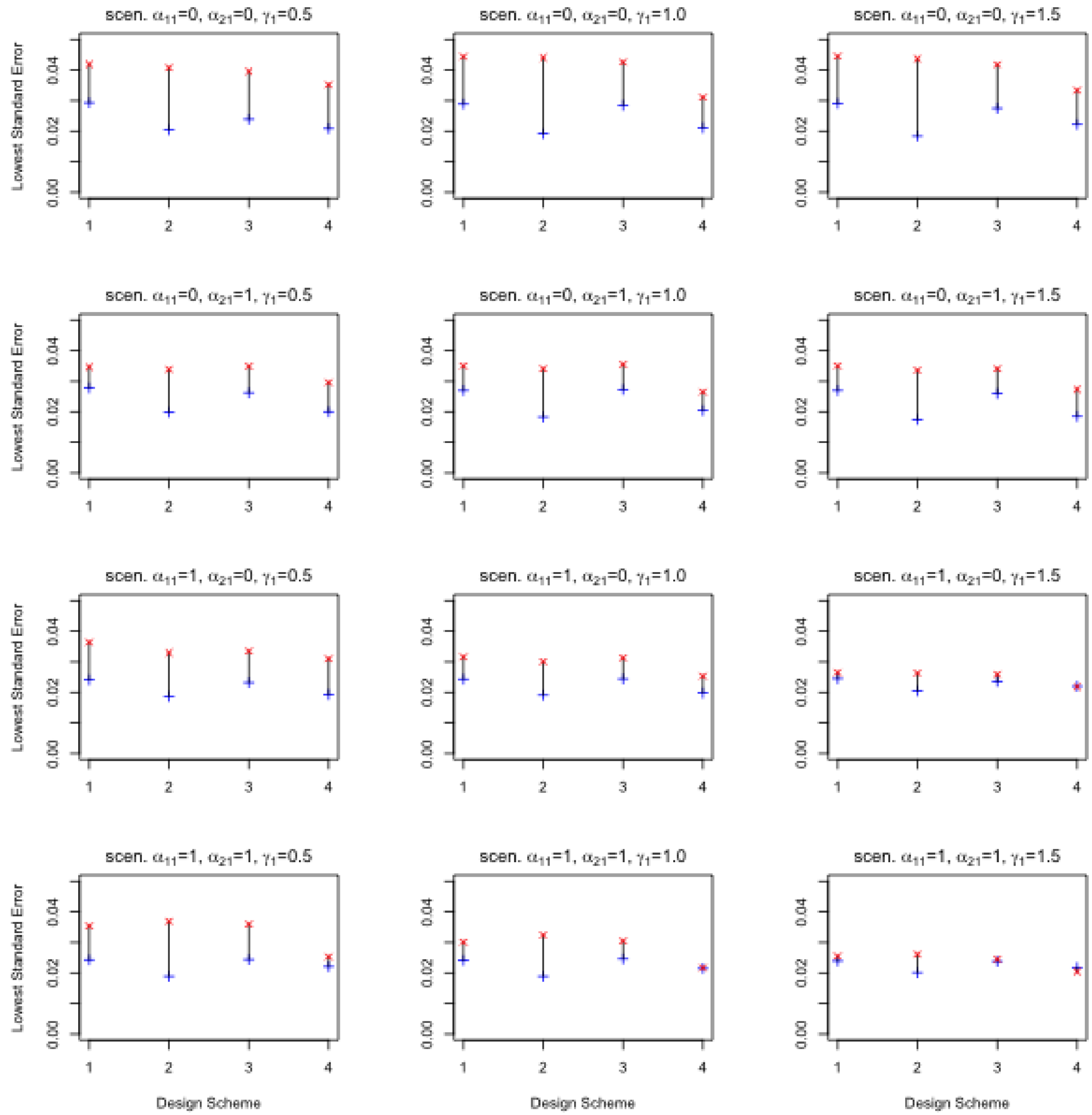

Figure 3.4: Lowest standard errors of the coefficient estimates under two-phase outcome-dependent sampling designs
+ represents standard error of $\hat{\alpha}_{11}$ for the most efficient sampling scenario.
× represents standard error of $\hat{\alpha}_{21}$ for the most efficient sampling scenario.
The design scheme 1 is generalized case-cohort design based on the first event indicator.
The design scheme 2 is outcome-dependent BSS design based on the first gap time and its event indicator.
The design scheme 3 is generalized case-cohort design based on the event indicators of the two sequential gap times.
The design scheme 4 is outcome-dependent BSS design based on the two sequential gap times and their event indicators.

Chapter 4

Efficiency of two-phase outcome-dependent sampling designs when the dependence between time-to-events is high

The objective of this study is to investigate efficient two-phase outcome-dependent sampling designs for bivariate sequential survival data under a predetermined phase two sample size. Four phase two sampling designs were introduced in Chapter 2: (1) generalized case-cohort design based on the event indicator of the first gap time; (2) outcome-dependent BSS design based on the first gap time and its event indicator; (3) generalized case-cohort design based on the event indicators of the two sequential gap times; and (4) outcome-dependent BSS design based on the two sequential gap times and their event indicators.

In Chapter 3, efficiency of outcome-dependent sampling designs were investigated when the dependence between sequential gap times is moderate. In this chapter, a simulation study was conducted to study the efficiency of the phase two sampling designs under strong dependence between sequential gap times. We generated a large random bivariate survival time sample with size $N = 50,000$ from the joint conditional distribution of T_1 and T_2 given $X = x$ in (2.10) with the Clayton copula parameter value $\phi = 8$, and the covariate X follows the Bernoulli distribution with probability of success $p = P(X = 1) = 0.25$. The corresponding Kendall's tau value is $\tau = \phi/(\phi + 2) = 0.8$, and therefore, there is a high dependence between the two sequential gap times T_1 and T_2 given $X = x$. The marginal distributions of T_1 and T_2 given $X = x$ are modelled by Weibull regression with survival functions (2.11) and (2.12), respectively. The censoring time C is generated from Uniform$(0, b)$ such that about 40% of T_1 survival times are censored. The upper bound b of Uniform$(0, b)$ and T_2 censoring rate are given in Table 2.1. At phase one, suppose the observed data is $\{(t_1, \delta_1, t_2, \delta_2) : i = 1, ..., N\}$ where $(t_1, t_2) = (\min(T_1, C), \min(T_2, C - t_1))$ and $(\delta_1, \delta_2) = (I[T_1 = t_1], I[T_2 = t_2])$ are the observed survival times and their event indicators, respectively.

A subsample of fixed size n is drawn at phase two in order to obtain a measurement of covariate X which is costly or difficult to measure. We want to investigate generalized case-cohort and outcome-dependent BSS designs that result in more efficient sampling designs with bivariate sequential survival data.

The phase one cohort can be stratified into the strata (S_{cases}, S_{noncases}) based on the event indicator δ_1 of the first gap time T_1. Suppose the size of the subsample from the case stratum S_{cases} is denoted by n_{cases} and the size of the subsample from the non-case stratum S_{noncases} is denoted by n_{noncases}, where $n_{\text{cases}} + n_{\text{noncases}} = n$. The aim of Section 4.1 is to determine the number of first event cases (n_{cases}) versus the number of first event non-cases (n_{noncases}) that should be selected at phase two where $n_{\text{cases}} + n_{\text{noncases}} = n$. Here, the sampling is only based on the event indicator of the first event.

A more efficient generalized case-cohort design could be achieved by selecting a more informative sample. We can stratify all first event cases S_{cases} into strata ($S_{\text{cases},1}$, $S_{\text{cases},2}$, $S_{\text{cases},3}$) based on the observed time-to-event T_1 values using two cut-off values $c_{L1} < c_{U1}$ as in (2.2). Similarly, we can stratify all first event non-cases S_{noncases} into strata ($S_{\text{noncases},1}$, $S_{\text{noncases},2}$, $S_{\text{noncases},3}$) based on the observed censoring time C values using two cut-off values $c^*_{L1} < c^*_{U1}$ as in (2.3). The aim of Section 4.2 is to determine sampling probability of each defined stratum leading to a more efficient design while using the most efficient design (n_{cases}, n_{noncases}) obtained in Section 4.1. Here, the sampling is based on both the event indicator of the first event and the time-to-first event.

Under bivariate sequential survival data, a first event case could be either a second event case or a second event non-case. Let us denote $S_{\text{cases,cases}}$ as the subset of S_{cases} which are second event cases and $S_{\text{cases,noncases}}$ as the subset of S_{cases} which are second event non-cases. Using the most efficient design (n_{cases}, n_{noncases}) obtained in Section 4.1, the aim of Section 4.3 is to determine the number of second event cases (m_{cases}) versus the number of second event non-cases (m_{noncases}) that should be selected under the generalized case-cohort design during the sampling procedure where $m_{\text{cases}} + m_{\text{noncases}} = n_{\text{cases}}$. Here, the sampling is based on the event indicators of the two sequential events.

By selecting the more informative subjects for purposes of detailed covariate measurement, a more efficient generalized case-cohort design could be achieved. We can stratify all second event cases $S_{\text{cases,cases}}$ into strata ($S_{\text{cases,cases},1}$, $S_{\text{cases,cases},2}$, $S_{\text{cases,cases},3}$) based on the observed time-to-event T_2 values using two cut-off values $c_{L2} < c_{U2}$ as in (2.6). Similarly, we can stratify second event non-cases $S_{\text{cases,noncases}}$ into strata ($S_{\text{cases,noncases},1}$, $S_{\text{cases,noncases},2}$, $S_{\text{cases,noncases},3}$) based on the observed censoring time $C - T_1$ values using two cut-off values $c^*_{L2} < c^*_{U2}$ as in (2.7). The aim of Section 4.4 is to determine sampling probability of each defined stratum leading to a more efficient design using the most efficient design obtained in Section 4.2 and the most efficient design (m_{cases}, m_{noncases}) obtained in Section 4.3. Here, the sampling is based on both the two event indicators and the two sequential gap times.

Finally, the lowest standard errors of the coefficient estimate of the expensive covariate X obtained under the four different phase two sampling designs are compared in Section 4.5.

4.1 Efficiency of generalized case-cohort designs based on the first event indicator

Suppose we observed a large cohort of sequential survival data $\{(t_{1i}, \delta_{1i}, t_{2i}, \delta_{2i}) : i = 1, ..., N\}$, where $N = 50,000$ at phase one. This phase one sample is stratified based on the first event indicators of the survival data. We assume that 40% of the first event is censored. Thus, there are about $N_{\text{cases}} = 30,000$ individuals in the case stratum S_{cases} and about $N_{\text{noncases}} = 20,000$ individuals in the non-case stratum S_{noncases}.

A subsample of fixed size $n = 10,000$ is drawn at phase two in order to obtain the covariate which is costly or difficult to measure. The size of the subsample from the case stratum S_{cases} is denoted by n_{cases} and the size of the subsample from the non-case stratum S_{noncases} is denoted by n_{noncases}. Each allocation (n_{cases}, n_{noncases}) defines a generalized case-cohort design based on the first event indicator. Given the fixed size $n = 10,000$ of subsample, one may choose how to allocate it among the strata of phase one. The aim is to gain the efficiency when estimating the regression coefficient of the expensive covariate. Hence, we will determine n_{cases} (and therefore n_{noncases}) which leads to an efficient design where $n_{\text{cases}}+n_{\text{noncases}} = n$. For example, Table 4.1 shows the results of estimates and standard errors for model scenario ($\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 0.5$),

Table 4.1: Coefficient estimates and their estimated standard errors under generalized case-cohort designs based on the first event indicator

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(1, 1, 0.5)$	1	(1000,9000)	0.999	0.0282	1.013	0.0310
	2	(2000,8000)	0.982	0.0253	0.997	0.0271
	3	(3000,7000)	0.979	0.0244	0.969	0.0261
	4	(4000,6000)	1.005	0.0235	1.012	0.0247
	5	(5000,5000)	0.972	0.0238	0.979	0.0248
	6	(6000,4000)	0.982	0.0242	0.985	0.0249
	7	(7000,3000)	0.937	0.0250	0.950	0.0256
	8	(8000,2000)	0.982	0.0258	0.976	0.0263
	9	(9000,1000)	0.994	0.0269	0.995	0.0271
	10	(10000,0)	0.976	0.0294	0.981	0.0294

a model defined by (2.11) where $\alpha_{10} = 0.6$, $\alpha_{11} = 1.0$, $\gamma_1 = 0.5$ and by (2.12) where $\alpha_{20} = 0.4$, $\alpha_{21} = 1.0$ and $\gamma_2 = 0.5$ as described in Section 2.2.1. Among the ten sampling scenarios, scenario 4 with ($n_{\text{cases}} = 4000$, $n_{\text{noncases}} = 6000$) and scenario 5 with ($n_{\text{cases}} = n_{\text{noncases}} = 5000$) give the minimum standard error estimates of the coefficient estimate of the expensive covariate for time to first event thus are the most efficient sampling designs. They will be used in both outcome-dependent BSS design based on time to first event and its event indicator and generalized case-cohort design based on first and second event indicators. Notice that these two sampling scenarios also yield the most efficient designs for the coefficient estimate of the expensive covariate for time to second event.

Table 4.2: The most efficient sampling scenario under generalized case-cohort designs based on the first event indicator

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$
$(0, 0, 0.5)$	5	(5000,5000)
$(0, 1, 0.5)$	1	(1000,9000)
$(1, 0, 0.5)$	1	(1000,9000)

Continued on next page

Table 4.2 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$
$(1, 1, 0.5)$	4	(4000,6000)
$(0, 0, 1.0)$	4	(4000,6000)
$(0, 1, 1.0)$	1	(1000,9000)
$(1, 0, 1.0)$	1	(1000,9000)
$(1, 1, 1.0)$	5	(5000,5000)
$(0, 0, 1.5)$	4	(4000,6000)
$(0, 1, 1.5)$	1	(1000,9000)
$(1, 0, 1.5)$	1	(1000,9000)
$(1, 1, 1.5)$	10	(10000,0)

The simulation results for other model scenarios are listed in Table B.1 of Appendix B. Table 4.2 summarizes the sampling scenario $(n_{\text{cases}}, n_{\text{noncases}})$ which minimizes the standard error estimate thus is the most efficient sampling scenario for the stratification based on the first event indicator under different model scenarios. It shows that, for the following five model scenarios: $(\alpha_{11} = 0, \alpha_{21} = 0, \gamma_1 = 0.5)$, $(\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 0.5)$, $(\alpha_{11} = 0, \alpha_{21} = 0, \gamma_1 = 1.0)$, $(\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 1.0)$, and $(\alpha_{11} = 0, \alpha_{21} = 0, \gamma_1 = 1.5)$, the most efficient generalized case-cohort design $(n_{\text{cases}}, n_{\text{noncases}})$ based on the first event indicator is when $n_{\text{cases}} \approx n_{\text{noncases}}$. For the following six model scenarios: $(\alpha_{11} = 0, \alpha_{21} = 1, \gamma_1 = 0.5)$, $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 0.5)$, $(\alpha_{11} = 0, \alpha_{21} = 1, \gamma_1 = 1.0)$, $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 1.0)$, $(\alpha_{11} = 0, \alpha_{21} = 1, \gamma_1 = 1.5)$, and $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 1.5)$, when we increase sampling from the non-case stratum S_{noncases} for the first event, the efficiency of the coefficient estimate of the expensive covariate for time to first event improves. It requires further study to understand why this design is more efficient for these model scenarios when the dependence between sequential gap times is high.

In Section 3.1, when the dependence between gap times is moderate, it is found that the most efficient design is obtained when $n_{\text{cases}} \approx n_{\text{noncases}}$. Figure C.1 and Table C.1 of Appendix C describe the estimated standard errors of the coefficient estimates of the expensive covariate under generalized case-cohort designs based on the first event indicator for model scenario $(\alpha_{11} = 0, \alpha_{21} = 1, \gamma_1 = 0.5)$ when the dependence between time-to-events is changed from moderate to high. Similarly, Figure C.2 and Table C.2 of Appendix C describe the estimated standard errors of the coefficient estimates of the expensive covariate under generalized case-cohort designs based on the first event indicator for model scenario $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 1.5)$ when the dependence between time-to-events is changed from moderate to high.

For the model scenario $(\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 1.5)$, when we increase sampling from the case stratum S_{cases}, the efficiency of the coefficient estimate of the expensive covariate for time to first event improves. The same conclusion can also be obtained from Figure 4.1 which provides the trend of the efficiency for both $\hat{\alpha}_{11}$ and $\hat{\alpha}_{21}$ at various sampling scenarios under different model scenarios.

Figure 4.1 shows that the most efficient sampling design for $\hat{\alpha}_{11}$ yields the most efficient designs for $\hat{\alpha}_{21}$. This is true for all model scenarios.

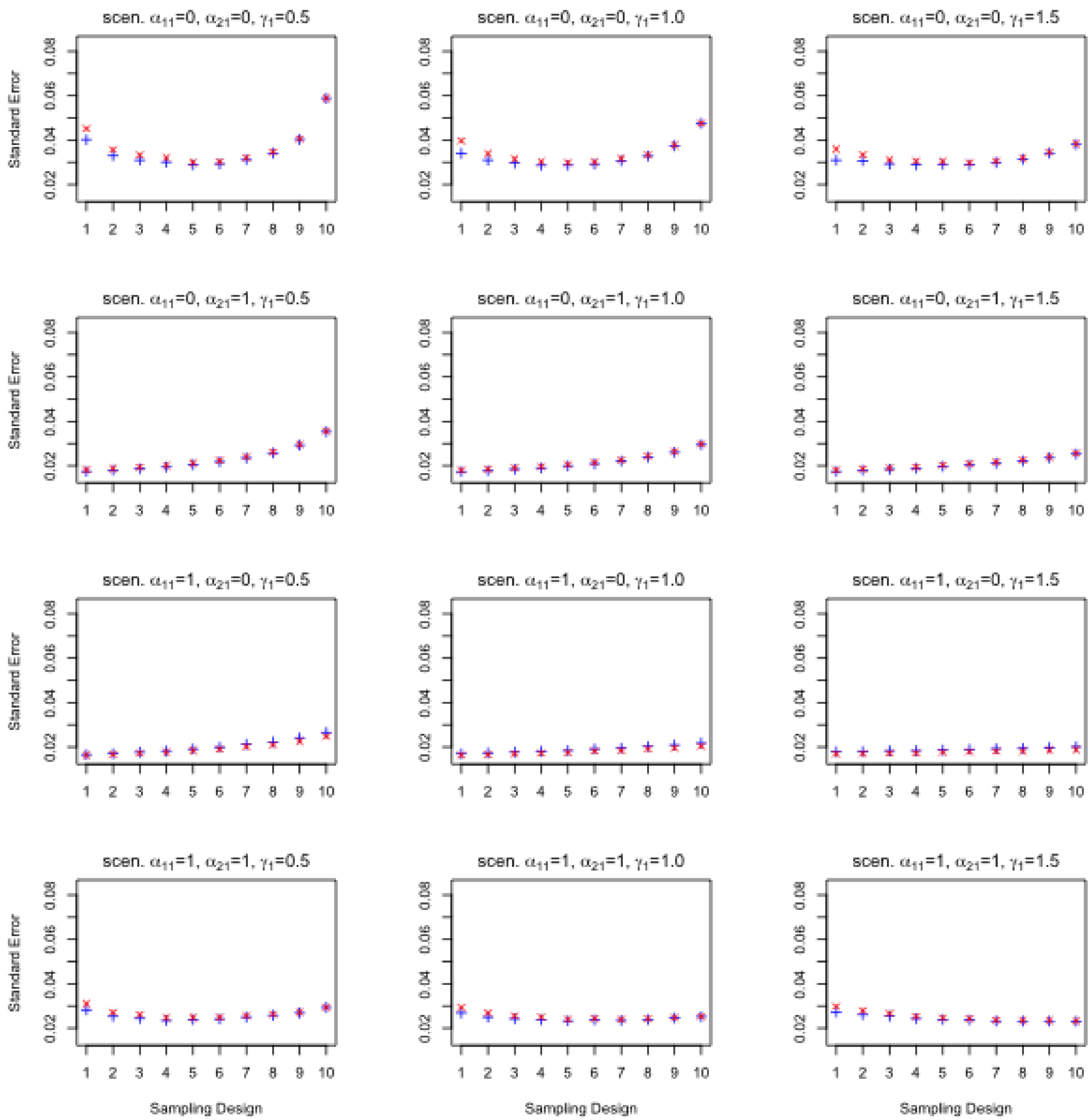

Figure 4.1: Estimated standard errors of the coefficient estimates of the expensive covariate under generalized case-cohort designs based on the first event indicator
+ represents standard error of $\hat{\alpha}_{11}$
$\times$ represents standard error of $\hat{\alpha}_{21}$
The sampling scenarios $1, ..., 10$ are described in Table 4.1

4.2 Efficiency of outcome-dependent BSS designs based on the first gap time and its event indicator

Greater efficiency may be achieved for generalized case-cohort design by selecting the more informative subjects for purposes of detailed covariate measurement. We can stratify all first event cases S_{cases} into strata $(S_{\text{cases},1}, S_{\text{cases},2}, S_{\text{cases},3})$ based on the observed time-to-event T_1 values using two cut-off values $c_{L1} < c_{U1}$ as in (2.2). Similarly, we can stratify all first event non-cases S_{noncases} into strata $(S_{\text{noncases},1}$, $S_{\text{noncases},2}$, $S_{\text{noncases},3})$ based on the observed censoring time C values using two cut-off values $c^*_{L1} < c^*_{U1}$ as in (2.3).

After obtaining the most efficient sampling design (n_{cases}, n_{noncases}) for the strata (S_{cases}, S_{noncases}) in Section 4.1, we do outcome-dependent BSS on the strata ($S_{\text{cases},1}$, $S_{\text{cases},2}$, $S_{\text{cases},3}$) and $(S_{\text{noncases},1}, S_{\text{noncases},2}, S_{\text{noncases},3})$. Suppose the size of the subsample from the stratum $S_{\text{cases},j}$ is denoted by $n_{\text{cases},j}$, $j = 1, 2, 3$, where $\sum_{j=1}^{3} n_{\text{cases},j} = n_{\text{cases}}$. Similarly, suppose the size of the subsample from the stratum $S_{\text{noncases},j}$ is denoted by $n_{\text{noncases},j}$, $j = 1, 2, 3$, where $\sum_{j=1}^{3} n_{\text{noncases},j} = n_{\text{noncases}}$. Given the fixed sizes (n_{cases}, n_{noncases}) of samples, one may choose how to allocate it among the strata $((S_{\text{cases},j} : j = 1, 2, 3), (S_{\text{noncases},j} : j = 1, 2, 3))$. Different allocations $((n_{\text{cases},j} : j = 1, 2, 3), (n_{\text{noncases},j} : j = 1, 2, 3))$ define different outcome-dependent BSS designs based on the first gap time T_1 and its event indicator δ_1.

The aim is to determine $n_{\text{cases},j}$ and $n_{\text{noncases},j}$, $j = 1, 2, 3$, which lead to an efficient design where $\sum_{j=1}^{3} n_{\text{cases},j} = n_{\text{cases}}$ and $\sum_{j=1}^{3} n_{\text{noncases},j} = n_{\text{noncases}}$. Table 4.3 shows the results of estimates and their standard errors under different allocations (($n_{\text{cases},1}$, $n_{\text{cases},2}$, $n_{\text{cases},3}$), ($n_{\text{noncases},1}$, $n_{\text{noncases},2}$, $n_{\text{noncases},3}$)) for model scenario ($\alpha_{11} = 1, \alpha_{21} =$

$1, \gamma_1 = 0.5)$, a model defined by (2.11) where $\alpha_{10} = 0.6$, $\alpha_{11} = 1.0$, $\gamma_1 = 0.5$ and by (2.12) where $\alpha_{20} = 0.4$, $\alpha_{21} = 1.0$ and $\gamma_2 = 0.5$ as described in Section 2.2.1. We see that sampling scenario 3 with $((n_{\text{cases},j} : j = 1,2,3), (n_{\text{noncases},j} : j = 1,2,3)) = ((4000,0,0),(0,1000,5000))$ minimizes the standard error ($\widehat{\text{SE}}(\hat{\alpha}_{11})$) thus is the most efficient sampling scenario. In scenario 3, there is an increased sampling from the first case stratum $S_{\text{cases},1}$. Selecting individuals with shorter time to first event yields more efficient coefficient estimate. In addition, in scenario 3, there is an increased sampling from the third non-case stratum $S_{\text{noncases},3}$. When we increase sampling from the stratum with long censoring time, the efficiency improves. Notice that in this chapter, there are six sampling scenarios 1, 4, 5, 7, 8, and 9 which yield larger standard error compare to SRS in S_{cases} and S_{noncases} while only three sampling scenarios 1, 4, and 7 yield larger standard error compare to SRS in S_{cases} and S_{noncases} in Section 3.2. Sampling scenarios 1, 4, and 7 with increased sampling from the stratum with short censoring time yield inefficient designs. Sampling scenarios 8 and 9 with increased sampling from the stratum with long time to first event yield inefficient designs. Sampling scenario 5 also yields a larger standard error with increased sampling from both the stratum with midrange time to first event and the stratum with midrange censoring time.

The most efficient scenario 3 is used in outcome-dependent BSS design based on the first and second gap times and their event indicators in Section 4.4.

Table 4.3: Coefficient estimates and their estimated standard errors under outcome-dependent BSS designs based on the first gap time and its event indicator

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases},j} : j = 1,2,3)$, $(n_{\text{noncases},j} : j = 1,2,3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(1, 1, 0.5)$	SRS in S_{cases} and S_{noncases}	(666,2668,666),(1500,3000,1500)	1.011	0.0234	1.017	0.0246
	1	(4000,0,0),(5000,1000,0)	0.976	0.0252	0.975	0.0261
	2	(4000,0,0),(0,6000,0)	0.990	0.0204	0.986	0.0215
	3	(4000,0,0),(0,1000,5000)	1.003	0.0187	1.001	0.0198
	4	(0,4000,0),(5000,1000,0)	0.956	0.0305	0.963	0.0315
	5	(0,4000,0),(0,6000,0)	0.961	0.0245	0.962	0.0259

Continued on next page

Table 4.3 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases},j} : j = 1,2,3)$, $(n_{\text{noncases},j} : j = 1,2,3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	6	(0,4000,0),(0,1000,5000)	0.976	0.0217	0.984	0.0232
	7	(0,0,4000),(5000,1000,0)	0.970	0.0406	0.979	0.0399
	8	(0,0,4000),(0,6000,0)	0.953	0.0370	0.961	0.0368
	9	(0,0,4000),(0,1000,5000)	0.976	0.0311	0.986	0.0322

The simulation results for other model scenarios are listed in Table B.2 of Appendix B. Notice that the first allocation in each model scenario in Table B.2 is a SRS in S_{cases} and S_{noncases} which is defined by (2.4) and (2.5). Thus, it is a generalized case-cohort design.

Table 4.4: The most efficient sampling scenario under outcome-dependent BSS designs based on the first gap time and its event indicator

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases},j} : j = 1,2,3)$, $(n_{\text{noncases},j} : j = 1,2,3)$
$(0, 0, 0.5)$	3	(5000,0,0),(0,0,5000)
$(0, 1, 0.5)$	3	(1000,0,0),(0,4000,5000)
$(1, 0, 0.5)$	9	(0,0,1000),(0,4000,5000)
$(1, 1, 0.5)$	3	(4000,0,0),(0,1000,5000)
$(0, 0, 1.0)$	3	(4000,0,0),(0,1000,5000)
$(0, 1, 1.0)$	9	(0,0,1000),(0,4000,5000)
$(1, 0, 1.0)$	9	(0,0,1000),(0,4000,5000)
$(1, 1, 1.0)$	3	(5000,0,0),(0,0,5000)
$(0, 0, 1.5)$	3	(4000,0,0),(0,1000,5000)
$(0, 1, 1.5)$	9	(0,0,1000),(0,4000,5000)
$(1, 0, 1.5)$	9	(0,0,1000),(0,4000,5000)
$(1, 1, 1.5)$	8	(1000,5000,4000),(0,0,0)

Table 4.4 summarizes the sampling scenario $((n_{\text{cases},j} : j = 1, 2, 3), (n_{\text{noncases},j} : j = 1, 2, 3))$ which minimizes the standard error thus is the most efficient sampling scenario for stratification based on the first event time and its event indicator under different model scenarios. It shows that, for the following six model scenarios: $(\alpha_{11} = 0, \alpha_{21} = 0, \gamma_1 = 0.5)$, $(\alpha_{11} = 0, \alpha_{21} = 1, \gamma_1 = 0.5)$, $(\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 0.5)$, $(\alpha_{11} = 0, \alpha_{21} = 0, \gamma_1 = 1.0)$, $(\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 1.0)$, and $(\alpha_{11} = 0, \alpha_{21} = 0, \gamma_1 = 1.5)$, the most efficient outcome-dependent BSS design $((n_{\text{cases},j} : j = 1, 2, 3), (n_{\text{noncases},j} : j = 1, 2, 3))$ based on the first event time and its event indicator is the sampling scenario 3 where we increase sampling from the stratum with short first event time (i.e., the first case stratum $S_{\text{cases},1}$) and also increase sampling from the stratum with long censoring time (i.e., the third non-case stratum $S_{\text{noncases},3}$). For the following five model scenarios: $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 0.5)$, $(\alpha_{11} = 0, \alpha_{21} = 1, \gamma_1 = 1.0)$, $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 1.0)$, $(\alpha_{11} = 0, \alpha_{21} = 1, \gamma_1 = 1.5)$, and $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 1.5)$, the most efficient design is the sampling scenario 9 where we increase sampling from the stratum with long first event time (i.e., the third case stratum $S_{\text{cases},3}$) and also increase sampling from the stratum with long censoring time (i.e., the third non-case stratum $S_{\text{noncases},3}$). Among these five model scenarios with sampling scenario 9 as the most efficient design, four of them with $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 0.5)$, $(\alpha_{11} = 0, \alpha_{21} = 1, \gamma_1 = 1.0)$, $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 1.0)$, and $(\alpha_{11} = 0, \alpha_{21} = 1, \gamma_1 = 1.5)$ yield the sampling scenario 3 as the next efficient design with the standard error $(\widehat{\text{SE}}(\hat{\alpha}_{11}))$ very close to that of the sampling scenario 9 and can be thought as another most efficient design. Hence, as in Section 3.2, the most efficient outcome-dependent BSS design $((n_{\text{cases},j} : j = 1, 2, 3), (n_{\text{noncases},j} : j = 1, 2, 3))$ based on the first event time and its event indicator is considered as the sampling scenario 3. This is true for all model scenarios except two model scenarios: $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 1.5)$ and $(\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 1.5)$. Due to the high dependence between the two sequential gap times, the most efficient sampling design for $\hat{\alpha}_{11}$ yields the most efficient designs for $\hat{\alpha}_{21}$. This is true for all model scenarios as seen in Table B.2 of Appendix B. The same conclusion can also be obtained from Figure 4.2 which provides the trend of the efficiency for both $\hat{\alpha}_{11}$ and $\hat{\alpha}_{21}$ at various sampling scenarios under different model scenarios.

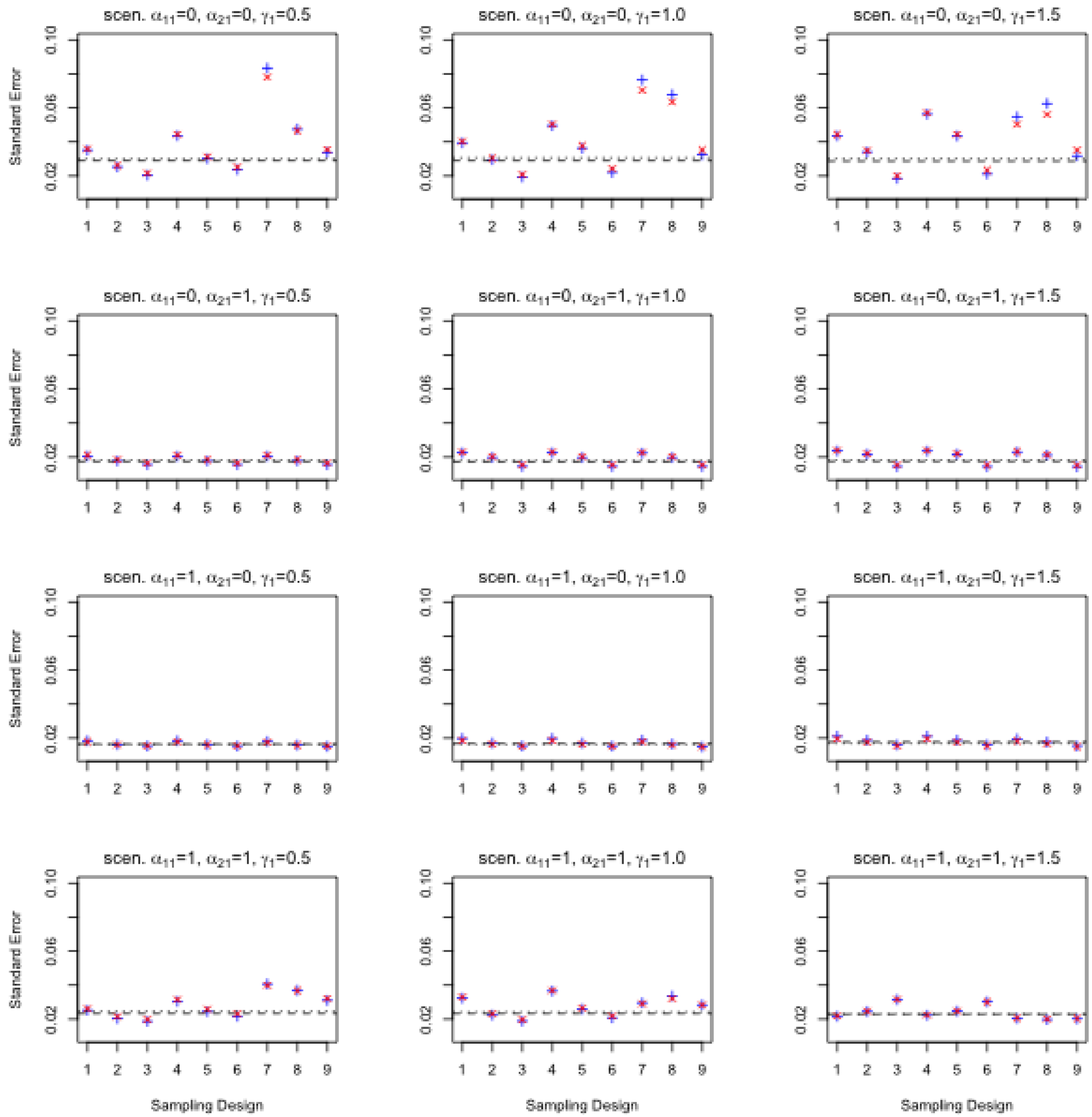

Figure 4.2: Estimated standard errors of the coefficient estimates of the expensive covariate under outcome-dependent BSS designs based on the first gap time and its event indicator
+ represents standard error of $\hat{\alpha}_{11}$
× represents standard error of $\hat{\alpha}_{21}$
dashed line represents standard error of $\hat{\alpha}_{11}$ under SRS in S_{cases} and S_{noncases}
dotted line represents standard error of $\hat{\alpha}_{21}$ under SRS in S_{cases} and S_{noncases}
The sampling scenarios $1, \ldots, 9$ are described in Table B.2

4.3 Efficiency of generalized case-cohort designs based on the event indicators of the two sequential gap times

In Section 4.1, a subsample of fixed size ($n = 10,000$) was drawn in order to obtain a covariate which is expensive to measure based on the first event indicator. Table 4.2 provides us the most efficient sampling scenario for stratification based on the first event indicator under different model scenarios. For example, sampling scenario ($n_{\text{cases}} = 4000$, $n_{\text{noncases}} = 6000$) minimizes the standard error thus is the most efficient sampling scenario for model scenario ($\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 0.5$). It minimizes the variance of the coefficient estimate of the expensive covariate for the first gap time. In addition, due to the high dependence between the two sequential gap times, it also minimizes the variance of the coefficient estimate of the expensive covariate for the second gap time. We are interested in exploring efficient sampling designs considering stratification based on the event indicators of the two sequential gap times so that the efficiency can be improved further.

In this section, a subsample of fixed size ($n = 10,000$) is drawn in order to obtain a covariate which is expensive to measure based on the event indicators of the two sequential gap times. Suppose (n_{cases}, n_{noncases}) is the most efficient sampling scenario for stratification based on the first event indicator. First, a subsample of size n_{noncases} is drawn from the first event non-case stratum S_{noncases}. Then, a subsample of size n_{cases} is drawn from the first event case stratum S_{cases} based on the second event indicator. Note that under bivariate sequential survival data, a T_1 case could be either a T_2 case or a T_2 non-case. Let us denote $S_{\text{cases,cases}}$ as the subset of S_{cases} which includes T_2 cases and $S_{\text{cases,noncases}}$ as the subset of S_{cases} which includes T_2 non-cases. The size of the subsample from the first and second event case stratum $S_{\text{cases,cases}}$ is denoted by m_{cases} and the size of the subsample from the first event case and second event non-case stratum $S_{\text{cases,noncases}}$ is denoted by m_{noncases}, where $n_{\text{cases}} = m_{\text{cases}} + m_{\text{noncases}}$.

Given the fixed size n_{cases} of subsample, we investigate how to allocate it among the strata ($S_{\text{cases,cases}}$, $S_{\text{cases,noncases}}$) which is based on T_2 event indicator. Different allocations (m_{cases}, m_{noncases}) in addition to selecting n_{noncases} individuals from S_{noncases} define different generalized case-cohort designs based on the event indicators of the two sequential gap times.

We need to determine m_{cases} and m_{noncases} which lead to an efficient design where $m_{\text{cases}} + m_{\text{noncases}} = n_{\text{cases}}$. Efficient sampling design minimizes the variance of the coefficient estimate of the expensive covariate for the second gap time. Table 4.5 shows the results of estimates and standard errors for model scenario ($\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 0.5$), a model defined by (2.11) where $\alpha_{10} = 0.6$, $\alpha_{11} = 1.0$, $\gamma_1 = 0.5$ and by (2.12) where $\alpha_{20} = 0.4$, $\alpha_{21} = 1.0$ and $\gamma_2 = 0.5$ as described in Section 2.2.1. We see that sampling scenario 8 with ($m_{\text{cases}} = 4000$, $m_{\text{noncases}} = 0$) minimizes the standard error estimate of $\hat{\alpha}_{21}$, thus is the most efficient sampling scenario based on $\widehat{\text{SE}}(\hat{\alpha}_{21})$. Notice that sampling scenario 8 with ($m_{\text{cases}} = 4000$, $m_{\text{noncases}} = 0$) also minimizes the standard error estimate of $\hat{\alpha}_{11}$, thus is the most efficient sampling scenario based on $\widehat{\text{SE}}(\hat{\alpha}_{11})$. Moreover, both $\widehat{\text{SE}}(\hat{\alpha}_{11})$ and $\widehat{\text{SE}}(\hat{\alpha}_{21})$ are smaller compared to sampling scenario 4 of Table 4.1. Thus, the efficiency of generalized case-cohort designs based on the first event indicator can be improved by generalized case-cohort designs based on the event indicators of the two sequential gap times. When we increase sampling from the first and second event case stratum $S_{\text{cases,cases}}$, the efficiency of the coefficient estimate of the expensive covariate for times to first and second event improves. It will be used in outcome-dependent BSS design based on the two sequential gap times and their event indicators.

Table 4.5: Coefficient estimates and their estimated standard errors under generalized case-cohort designs based on the event indicators of the two sequential gap times

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases}},\, m_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(1, 1, 0.5)$	1	(500,3500)	1.010	0.0276	1.024	0.0323
	2	(1000,3000)	1.014	0.0260	1.028	0.0290
	3	(1500,2500)	0.969	0.0257	0.981	0.0280
	4	(2000,2000)	0.992	0.0247	0.987	0.0266
	5	(2500,1500)	0.953	0.0246	0.949	0.0262
	6	(3000,1000)	0.971	0.0238	0.969	0.0250
	7	(3500,500)	0.966	0.0235	0.963	0.0246
	8	(4000,0)	0.996	0.0229	1.000	0.0238

Table 4.6: The most efficient sampling scenario under generalized case-cohort designs based on the event indicators of the two sequential gap times

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases}},\, m_{\text{noncases}})$
$(0, 0, 0.5)$	10	(5000,0)
$(0, 1, 0.5)$	2	(1000,0)
$(1, 0, 0.5)$	1	(500,500)
$(1, 1, 0.5)$	8	(4000,0)
$(0, 0, 1.0)$	8	(4000,0)
$(0, 1, 1.0)$	1	(500,500)
$(1, 0, 1.0)$	1	(500,500)
$(1, 1, 1.0)$	7	(3500,1500)
$(0, 0, 1.5)$	8	(4000,0)
$(0, 1, 1.5)$	1	(500,500)

Continued on next page

Table 4.6 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases}},\ m_{\text{noncases}})$
$(1, 0, 1.5)$	1	(500,500)
$(1, 1, 1.5)$	7	(3500,6500)

The simulation results for other model scenarios are listed in Table B.3 of Appendix B. Table 4.6 summarizes the sampling scenario $(m_{\text{cases}},\ m_{\text{noncases}})$ which minimizes the standard error estimate of $\hat{\alpha}_{21}$ thus is the most efficient sampling scenario based on $\widehat{\text{SE}}(\hat{\alpha}_{21})$ for stratification based on the event indicators of the two sequential gap times under different model scenarios. As in Section 3.3, when we increase sampling from the stratum $S_{\text{cases,cases}}$, the efficiency of the coefficient estimate of the expensive covariate for time to second event improves. This is true for all model scenarios except the six scenarios $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 0.5)$, $(\alpha_{11} = 0, \alpha_{21} = 1, \gamma_1 = 1.0)$, $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 1.0)$, $(\alpha_{11} = 0, \alpha_{21} = 1, \gamma_1 = 1.5)$, $(\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 1.5)$, and $(\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 1.5)$. For these six model scenarios, the estimated standard error of $\hat{\alpha}_{21}$ minimizes when $m_{\text{cases}} \approx m_{\text{noncases}}$ or when we increase sampling from the stratum $S_{\text{cases,noncases}}$.

Due to the high dependence between the two sequential gap times, the sampling scenario $(m_{\text{cases}},\ m_{\text{noncases}})$ which minimizes the standard error estimate of $\hat{\alpha}_{21}$ also minimizes the standard error estimate of $\hat{\alpha}_{11}$. Thus, the most efficient sampling scenario based on $\widehat{\text{SE}}(\hat{\alpha}_{21})$ is also the most efficient sampling scenario based on $\widehat{\text{SE}}(\hat{\alpha}_{11})$ for stratification based on the event indicators of the two sequential gap times. This is true for all model scenarios as seen in Table B.3 of Appendix B.

4.4 Efficiency of outcome-dependent BSS designs based on the two sequential gap times and their event indicators

In Section 4.3, a subsample of fixed size ($n = 10,000$) was drawn in order to obtain a covariate which is expensive to measure based on the event indicators of the two sequential gap times. In order to achieve the possible efficiency gain of generalized case-cohort design, the sampling of subjects could be done such that the sample is enriched with subjects who are especially informative. In addition to sampling based on the event indicators, now we consider sampling based on the two sequential gap times. We stratify all T_2 cases $S_{\text{cases,cases}}$ into strata ($S_{\text{cases,cases},1}$, $S_{\text{cases,cases},2}$, $S_{\text{cases,cases},3}$) based on the observed time-to-second event using two cut-off values $c_{L2} < c_{U2}$ as in (2.6). Similarly, we can stratify T_2 non-cases $S_{\text{cases,noncases}}$ into strata ($S_{\text{cases,noncases},1}$, $S_{\text{cases,noncases},2}$, $S_{\text{cases,noncases},3}$) based on observed censoring time $C - T_1$ values using two cut-off values $c^*_{L2} < c^*_{U2}$ as in (2.7).

In Section 4.1, a subsample of fixed size ($n = 10,000$) is drawn from a large cohort of sequential survival data of size $N = 50,000$ under generalized case-cohort designs based on the first event indicator. Table 4.2 provides us the most efficient sampling scenarios ($n_{\text{cases}}, n_{\text{noncases}}$) for $\hat{\alpha}_{11}$ under different model scenarios, where $n_{\text{cases}} + n_{\text{noncases}} = n$. After obtaining the most efficient sampling design (n_{cases}, n_{noncases}) in Section 4.1, we do outcome-dependent BSS based on the first gap time and its event indicator in Section 4.2. Table 4.4 summarizes the most efficient sampling scenarios (($n_{\text{cases},j} : j = 1, 2, 3$), ($n_{\text{noncases},j} : j = 1, 2, 3$)) for $\hat{\alpha}_{11}$ under different model scenarios, where $\sum_{j=1}^{3} n_{\text{cases},j} = n_{\text{cases}}$ and $\sum_{j=1}^{3} n_{\text{noncases},j} = n_{\text{noncases}}$. The above efficient sampling designs minimize the variance of $\hat{\alpha}_{11}$. We are also interested in looking for efficient sampling designs which minimize the variance of $\hat{\alpha}_{21}$. After obtaining the most efficient sampling design (n_{cases}, n_{noncases}) in Section 4.1, a subsample of size n_{cases} was drawn from the first event case stratum S_{cases} under generalized case-cohort designs based on the second event indicator in Section 4.3. Table 4.6 summarizes the most efficient sampling scenarios (m_{cases}, m_{noncases}) for $\hat{\alpha}_{21}$ under different model scenarios, where $n_{\text{cases}} = m_{\text{cases}} + m_{\text{noncases}}$.

After obtaining the most efficient sampling design (m_{cases}, m_{noncases}) for the strata ($S_{\text{cases,cases}}$, $S_{\text{cases,noncases}}$), we do outcome-dependent BSS on the strata ($S_{\text{cases,cases,1}}$, $S_{\text{cases,cases,2}}$, $S_{\text{cases,cases,3}}$) and ($S_{\text{cases,noncases,1}}$, $S_{\text{cases,noncases,2}}$, $S_{\text{cases,noncases,3}}$). Suppose the size of the subsample from the stratum $S_{\text{cases,cases},j}$ is denoted by $m_{\text{cases},j}$, $j = 1, 2, 3$, where $\sum_{j=1}^{3} m_{\text{cases},j} = m_{\text{cases}}$. Similarly, suppose the size of the subsample from the stratum $S_{\text{cases,cases},j}$ is denoted by $m_{\text{cases},j}$, $j = 1, 2, 3$, where $\sum_{j=1}^{3} m_{\text{noncases},j} = m_{\text{noncases}}$. Given the fixed sizes (m_{cases}, m_{noncases}) of subsamples, one may choose how to allocate it among the strata $((S_{\text{cases,cases},j} : j = 1, 2, 3), (S_{\text{cases,noncases},j} : j = 1, 2, 3))$. Different allocations $((m_{\text{cases},j} : j = 1, 2, 3), (m_{\text{noncases},j} : j = 1, 2, 3))$ define different outcome-dependent BSS designs based on the second gap time T_2 and its event indicator.

Our objective is to determine $m_{\text{cases},j}$ and $m_{\text{noncases},j}$, $j = 1, 2, 3$, which lead to an efficient design where $\sum_{j=1}^{3} m_{\text{cases},j} = m_{\text{cases}}$ and $\sum_{j=1}^{3} m_{\text{noncases},j} = m_{\text{noncases}}$. Efficient sampling design minimizes the variance of the coefficient estimate of the expensive covariate for time-to-event T_2. Table 4.7 shows the results of estimates and standard errors for different allocations ($m_{\text{cases},1}$, $m_{\text{cases},2}$, $m_{\text{cases},3}$), ($m_{\text{noncases},1}$, $m_{\text{noncases},2}$, $m_{\text{noncases},3}$) for model scenario $(\alpha_{11} = 1, \alpha_{21} = 1, \gamma_1 = 0.5)$, a model defined by (2.11) where $\alpha_{10} = 0.6$, $\alpha_{11} = 1.0$, $\gamma_1 = 0.5$ and by (2.12) where $\alpha_{20} = 0.4$, $\alpha_{21} = 1.0$ and $\gamma_2 = 0.5$ as described in Section 2.2.1. We see that sampling scenario 3 with $((m_{\text{cases},j} : j = 1, 2, 3), (m_{\text{noncases},j} : j = 1, 2, 3)) = ((2500, 1500, 0), (0, 0, 0))$ minimizes the standard error estimate of $\hat{\alpha}_{21}$ thus are the most efficient sampling scenarios based on $\widehat{\text{SE}}(\hat{\alpha}_{21})$. In scenario 3, there is an increased sampling from the first T_2 case stratum $S_{\text{cases,cases},1}$. When we increase sampling from the stratum with short time-to-second event, the efficiency improves. Notice that sampling scenarios 7, 8 and 9 have larger standard error estimates compared to other sampling scenarios. These three sampling scenarios increase sampling from the stratum with long time-to-second event which yield inefficient designs.

Table 4.7: Coefficient estimates and their estimated standard errors under outcome-dependent BSS designs based on the two sequential gap times and their event indicators

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases},j} : j = 1,2,3)$, $(m_{\text{noncases},j} : j = 1,2,3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(1, 1, 0.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(447,3106,447),(0,0,0)	1.001	0.0194	1.002	0.0213
	1	(1500,1500,1000),(0,0,0)	0.994	0.0194	0.990	0.0208
	2	(2000,1500,500),(0,0,0)	1.005	0.0190	1.001	0.0202
	3	(2500,1500,0),(0,0,0)	1.004	0.0187	0.999	0.0196
	4	(500,3000,500),(0,0,0)	0.987	0.0195	0.990	0.0214
	5	(250,3500,250),(0,0,0)	1.005	0.0195	1.010	0.0212
	6	(0,4000,0),(0,0,0)	0.988	0.0196	0.990	0.0213
	7	(1000,1500,1500),(0,0,0)	1.009	0.0198	1.017	0.0217
	8	(500,1500,2000),(0,0,0)	0.997	0.0207	0.993	0.0236
	9	(0,1500,2500),(0,0,0)	1.008	0.0217	1.012	0.0252

The simulation results for other model scenarios are listed in Table B.4 of Appendix B. Notice that the first allocation in each model scenario in Table B.4 is a SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$ which is defined by (2.8) and (2.9). Thus, it is a generalized case-cohort design.

Table 4.8: The most efficient sampling scenario under outcome-dependent BSS designs based on the two sequential gap times and their event indicators

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases},j} : j = 1,2,3)$, $(m_{\text{noncases},j} : j = 1,2,3)$
$(0, 0, 0.5)$	3	(2500,2500,0),(0,0,0)
$(0, 1, 0.5)$	3	(1000,0,0),(0,0,0)

Continued on next page

Table 4.8 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases},j} : j = 1,2,3)$, $(m_{\text{noncases},j} : j = 1,2,3)$
$(1, 0, 0.5)$	3	(500,0,0),(0,0,500)
$(1, 1, 0.5)$	3	(2500,1500,0),(0,0,0)
$(0, 0, 1.0)$	3	(2500,1500,0),(0,0,0)
$(0, 1, 1.0)$	3	(500,0,0),(0,0,500)
$(1, 0, 1.0)$	3	(500,0,0),(0,0,500)
$(1, 1, 1.0)$	3	(2500,1000,0),(0,0,1500)
$(0, 0, 1.5)$	3	(2500,1500,0),(0,0,0)
$(0, 1, 1.5)$	3	(500,0,0),(0,0,500)
$(1, 0, 1.5)$	3	(500,0,0),(0,0,500)
$(1, 1, 1.5)$	3	(2500,1000,0),(2386,1614,2500)

Table 4.8 summarizes the sampling scenario ($(m_{\text{cases},j} : j = 1,2,3)$, $(m_{\text{noncases},j} : j = 1,2,3)$) which minimizes the standard error estimate of $\hat{\alpha}_{21}$ thus is the most efficient sampling scenario based on $\widehat{\text{SE}}(\hat{\alpha}_{21})$ for stratification based on the two sequential gap times and their event indicators under different model scenarios. It shows that the most efficient outcome-dependent BSS design ($(m_{\text{cases},j} : j = 1,2,3)$, $(m_{\text{noncases},j} : j = 1,2,3)$) based on the two sequential gap times and their event indicators is the sampling scenario 3 where we increase sampling from the stratum with short second event times (i.e., the first T_2 case stratum $S_{\text{cases,cases},1}$) and also increase sampling from the stratum with long censoring times (i.e., the third T_2 non-case stratum $S_{\text{cases,noncases},3}$). This is true for all model scenarios. The same conclusion can also be obtained from Figure 4.3 which provides the trend of the efficiency for both $\hat{\alpha}_{11}$ and $\hat{\alpha}_{21}$ at various sampling scenarios under different model scenarios.

Notice that in Table 4.8, the sum of $m_{\text{cases},j}$, $j = 1,2,3$, is m_{cases} and the sum of

$m_{\text{noncases},j}$, $j = 1, 2, 3$, is m_{noncases}, where (m_{cases}, m_{noncases}) is selected based on the most efficient design identified in Table 4.6.

Due to the high dependence between the two sequential gap times in this chapter, the sampling scenario (m_{cases}, m_{noncases}) which minimizes the standard error estimate of $\hat{\alpha}_{21}$ also minimizes the standard error estimate of $\hat{\alpha}_{11}$. Thus, the most efficient sampling scenario based on $\widehat{\text{SE}}(\hat{\alpha}_{21})$ is also the most efficient sampling scenario based on $\widehat{\text{SE}}(\hat{\alpha}_{11})$ for stratification based on the event indicators of the two sequential gap times. This is true for all model scenarios as seen in Table B.4 of Appendix B.

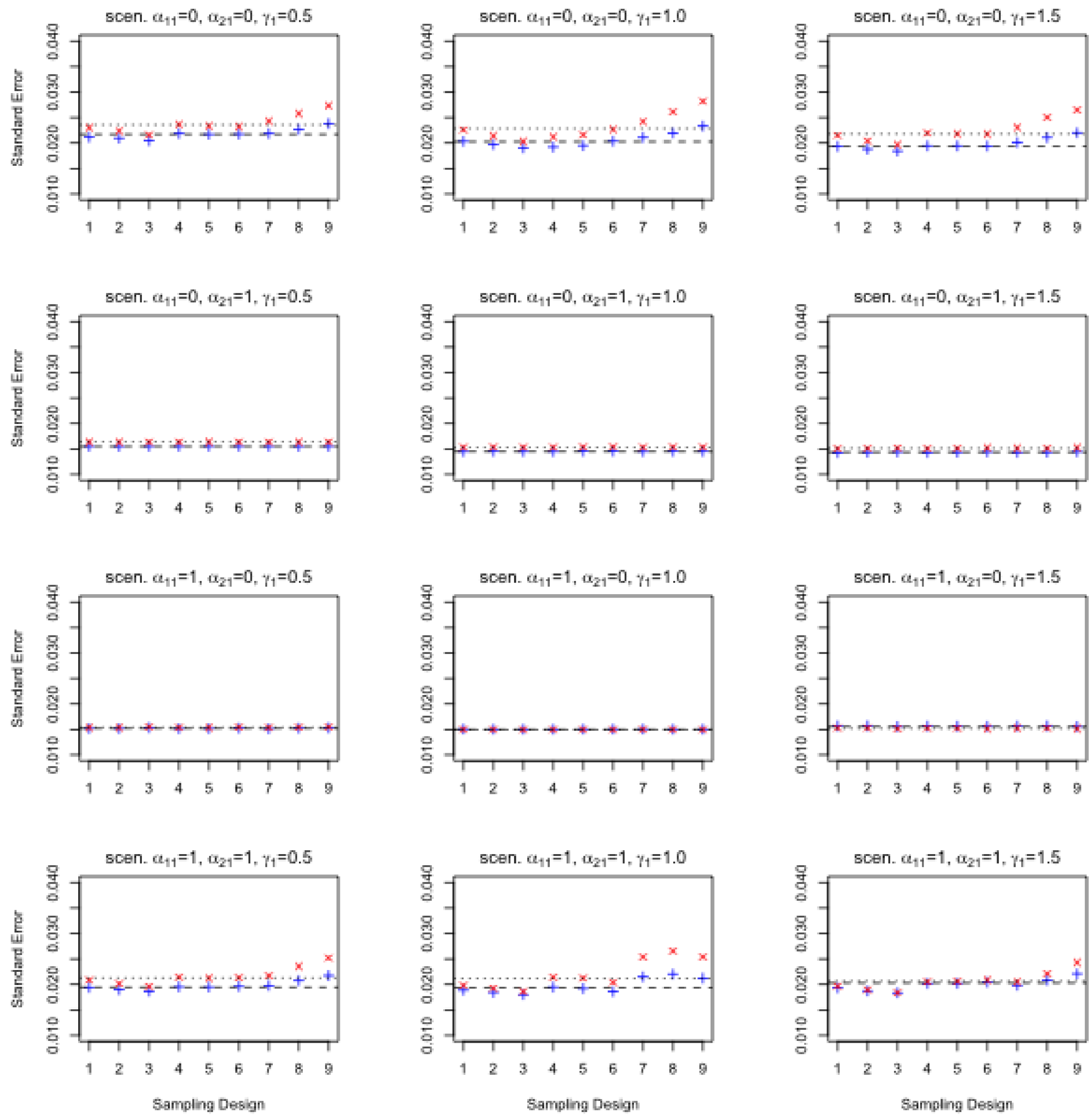

Figure 4.3: Estimated standard errors of the coefficient estimates of the expensive covariate under outcome-dependent BSS designs based on the two sequential gap times and their event indicators

+ represents standard error of $\hat{\alpha}_{11}$

× represents standard error of $\hat{\alpha}_{21}$

dashed line represents standard error of $\hat{\alpha}_{11}$ under SRS

dotted line represents standard error of $\hat{\alpha}_{21}$ under SRS

The sampling scenarios $1, \ldots, 9$ are given in Table B.4

4.5 Summary

Table 4.9 and Figure 4.4 summarize standard errors of $\hat{\alpha}_{11}$ and $\hat{\alpha}_{21}$ for the most efficient sampling scenarios under two-phase outcome-dependent sampling designs for different model scenarios when the dependence between the two sequential gap times is high. Design 1 represents a generalized case-cohort design based on the first event indicator. Design 2 represents an outcome-dependent BSS design based on the first gap time and its event indicator. Design 3 represents a generalized case-cohort design based on the event indicators of the two sequential gap times. Design 4 represents an outcome-dependent BSS design based on the two sequential gap times and their event indicators. Recall that the most efficient sampling scenarios for design 1 and design 2 are based on $\widehat{\text{SE}}(\hat{\alpha}_{11})$. On the other hand, the most efficient sampling scenarios for design 3 and design 4 are based on $\widehat{\text{SE}}(\hat{\alpha}_{21})$.

Under design 2, there is a gain on efficiency when estimating the regression coefficient of the expensive covariate for time to first event compared with design 1. Due to the high dependence between the two sequential gap times, standard errors of $\hat{\alpha}_{11}$ and $\hat{\alpha}_{21}$ for the most efficient sampling scenario are close to each other. Moreover, under design 4, there is no gain or only gain a little on efficiency when estimating the regression coefficient of the expensive covariate for time to second event compared with design 2. Therefore, it is suffice to use design 2 (i.e., outcome-dependent BSS design based on the first gap time and its event indicator) when there is high dependence between the two sequential gap times.

Under design 3, there is a gain on efficiency when estimating the regression coefficient of the expensive covariate for time to first event compared with design 1. This is true for all but one model scenario. Due to the high dependence between the two sequential gap times, design 3 also has a gain on efficiency when estimating the regression coefficient of the expensive covariate for time to second event compared with design 1. This is true for all but one model scenario. Therefore, design 3 (i.e., generalized case-cohort design based on the event indicators of the two sequential gap times) is better than design 1 (i.e., generalized case-cohort design based on the first event indicator).

Table 4.9: Lowest standard errors of the coefficient estimates under two-phase outcome-dependent sampling designs

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	standard errors	design 1	design 2	design 3	design 4
$(0, 0, 0.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0289	0.0205	0.0280	0.0204
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0299	0.0216	0.0287	0.0216
$(0, 1, 0.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0174	0.0155	0.0174	0.0155
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0181	0.0164	0.0181	0.0164
$(1, 0, 0.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0165	0.0151	0.0165	0.0151
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0163	0.0153	0.0163	0.0154
$(1, 1, 0.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0235	0.0187	0.0229	0.0187
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0247	0.0198	0.0238	0.0196
$(0, 0, 1.0)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0287	0.0193	0.0277	0.0190
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0303	0.0208	0.0285	0.0202
$(0, 1, 1.0)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0173	0.0136	0.0172	0.0133
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0180	0.0139	0.0179	0.0135
$(1, 0, 1.0)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0169	0.0149	0.0167	0.0150
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0163	0.0149	0.0163	0.0149
$(1, 1, 1.0)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0233	0.0188	0.0234	0.0180
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0240	0.0197	0.0243	0.0186
$(0, 0, 1.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0289	0.0185	0.0277	0.0184
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0306	0.0201	0.0285	0.0197
$(0, 1, 1.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0174	0.0143	0.0173	0.0143
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0180	0.0151	0.0179	0.0151
$(1, 0, 1.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0178	0.0154	0.0176	0.0156
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0169	0.0150	0.0168	0.0152
$(1, 1, 1.5)$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	0.0231	0.0200	0.0195	0.0183
	$\widehat{\text{SE}}(\hat{\alpha}_{21})$	0.0233	0.0203	0.0202	0.0184

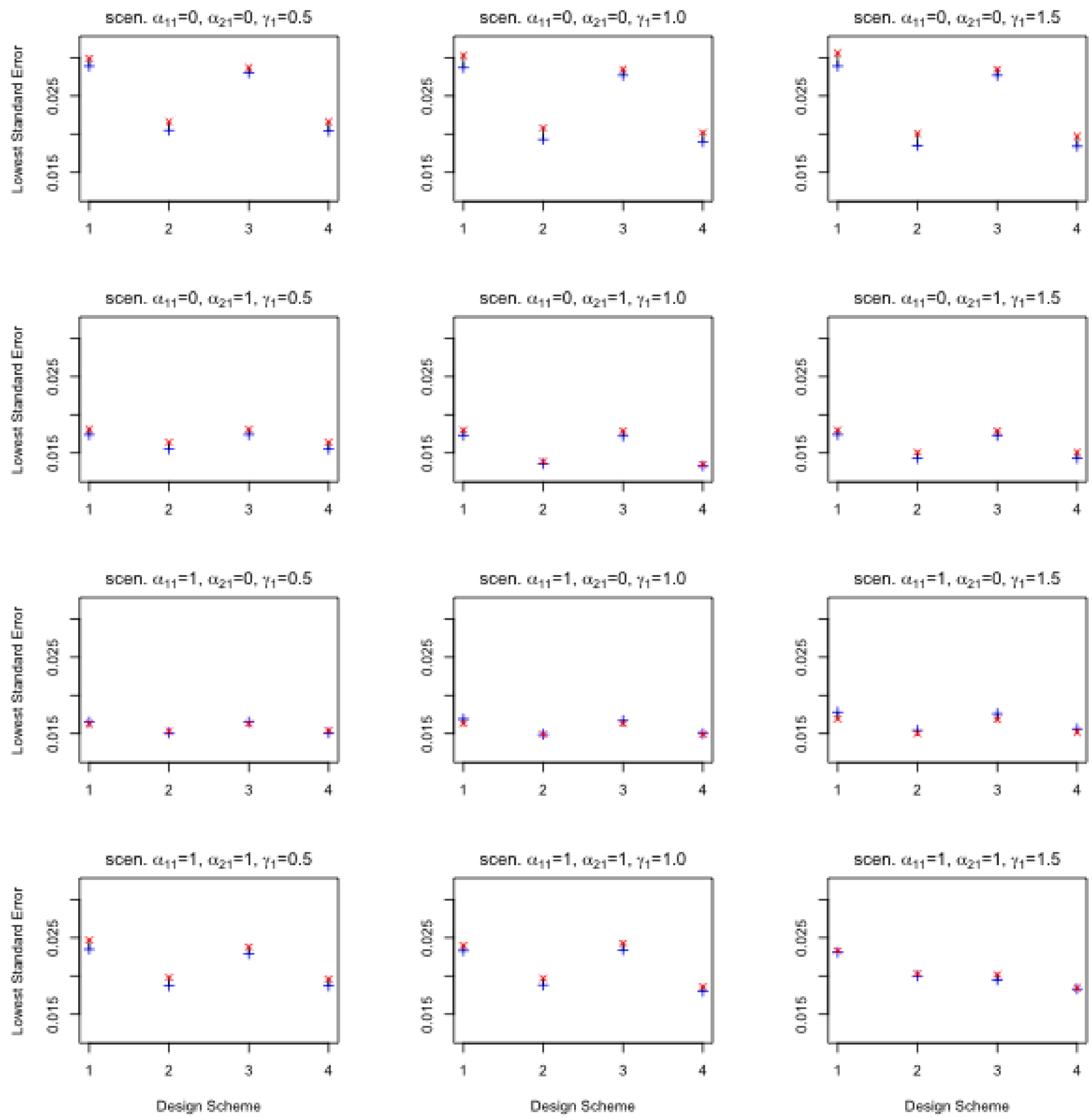

Figure 4.4: Lowest standard errors of the coefficient estimates under two-phase outcome-dependent sampling designs
+ represents standard error of $\hat{\alpha}_{11}$ for the most efficient sampling scenario.
× represents standard error of $\hat{\alpha}_{21}$ for the most efficient sampling scenario.
The design scheme 1 is generalized case-cohort design based on the first event indicator.
The design scheme 2 is outcome-dependent BSS design based on the first gap time and its event indicator.
The design scheme 3 is generalized case-cohort design based on the event indicators of the two sequential gap times.
The design scheme 4 is outcome-dependent BSS design based on the two sequential gap times and their event indicators.

Chapter 5

Conclusion

In some observational studies, the covariates of interest might be expensive to measure although the outcome variable could easily be obtained. In this situation, a cost-efficient two-phase outcome-dependent sampling design could be employed to measure the expensive covariate for more informative subjects. In phase one, all members of a random sample from a population or a cohort are measured for the outcome variable and inexpensive covariates. In phase two, a subset of the cohort is selected based on the outcome variable, and the expensive covariate is measured only for the selected individuals.

In this study, we investigated efficient two-phase outcome-dependent sampling designs with bivariate sequential time-to-event data for a predetermined phase two sample size under the likelihood-based approach. We considered sampling designs depending on the event indicators and gap times. A likelihood-based method was used to estimate the associations between the expensive covariate and the two gap times. We showed that when the selection probability at phase two depends on the two observed gap times and censoring times in addition to their event indicators, the efficiency of the design might improve compared to a generalized case-cohort design.

Bivariate sequential time-to-event data consists of two gap times T_1 and T_2 observed in sequence, and a right censoring time (total followup time) C. Let X be the expensive covariate. As the phase one data, in Section 2.2.1 we generated a $N = 50,000$ random bivariate sequential time-to-event sample from the joint conditional distribution of T_1 and T_2 given $X = x$ in (2.10) modelled by the Clayton copula (1.15). Moderate and high dependence levels were considered between the first and second event times. The covariate X follows the Bernoulli distribution. The marginal distributions of T_1 and T_2 given $X = x$ are modelled with Weibull regression with survival functions (2.11) and (2.12), respectively. The censoring time C is generated from Uniform$(0, b)$ such that about 40% of T_1 survival times are censored. When T_1 is censored, T_2 is unobserved.

The generated phase one data can be stratified based on the event indicators and the survival times. A phase two sample of fixed size ($n = 10,000$) was drawn based on the strata of phase one in order to obtain the covariate which is costly or difficult to measure. In Section 2.1, we described four phase two sampling designs: (1) generalized case-cohort design based on the event indicator of the first gap time; (2) outcome-dependent BSS design based on the first gap time and its event indicator; (3) generalized case-cohort design based on the event indicators of the two sequential gap times; and (4) outcome-dependent BSS design based on the two sequential gap times and their event indicators.

We adopted the full likelihood-based approach to estimate the regression coefficients of the expensive covariate for the first and second gap times. A simulation study was conducted to study the efficiency of these phase two sampling designs. The simulation results in Chapter 3 and Chapter 4 showed that when the selection probability at phase two depends on the two observed gap times and censoring times in addition to their event indicators, the efficiency of the design might improve compared to a generalized case-cohort design. When the dependence between time-to-events is moderate, the outcome-dependent BSS design based on both of the two sequential gap times and their event indicators is recommended. When the dependence between time-to-events is high, the outcome-dependent BSS design based on the first gap time and its event indicator is recommended.

Our results of phase two sampling designs for efficiency improvement are implicitly conditional on knowing the true distributions of all random variables of interest. As a further work, we would like to explore the efficiency of the sampling designs with bivariate sequential time-to-event data when the underlying model is misspecified before phase two sampling occurs. In this study, we also assume that there is only one expensive covariate and no other covariates. As a further work, we would like to investigate the efficiency of the sampling designs with bivariate sequential time-to-event data when there are other inexpensive covariates.

Bibliography

[1] N. E. Breslow, T. Lumley, C. M. Ballantyne, L. E. Chambless, and M. Kulich. Improved horvitz–thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Stat Biosci*, 1:32–49, 2009.

[2] D. R. Cox. Regression models and life tables (with discussion). *J. Roy. Stat. Soc. B*, 34:187–220, 1972.

[3] D. R. Cox and D. Oakes. *Analysis of Survival Data*. Chapman and Hall, London, 1984.

[4] J. Ding, H. Zhou, L. Liu, K. Cai, and M. Longnecker. Estimating effect of environmental contaminants on women's subfecundity for the moba study data with an outcome-dependent sampling scheme. *Biostatistics*, 15(4):636–650, 2014.

[5] T. R. Fleming and D. P. Harrington. *Counting Processes and Survival Analysis*. Wiley, New York, 1991.

[6] D. A. Hsieh, C. F. Manski, and D. McFadden. Estimation of response probabilities from augmented retrospective observations. *J. Amer. Statist. Assoc.*, 80:651–662, 1985.

[7] B. E. Huang and D. Y. Lin. Efficient association mapping of quantitative trait loci with selective genotyping. *The American Journal of Human Genetics*, 80:567–576, 2007.

[8] G. Imbens and T. Lancaster. Case-control studies with contaminated controls. *Journal of Econometrics*, 71(1):145–160, 1996.

[9] H. Joe. *Multivariate Models and Dependence Concepts*. Chapman and Hall, London, 1997.

[10] P. Judd. *Two-phase response-dependent sampling designs for time-to-event analysis*. M.Sc. thesis, Memorial University of Newfoundland, 2016.

[11] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley and Sons, New York, 2nd edition, 2002.

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

[13] R. H. Keogh and D. R. Cox. *Case-Control Studies*. Cambridge University Press, 2014.

[14] M. Kulich and D. Y. Lin. Improving the efficiency of relative-risk estimation in case-cohort studies. *J Am Stat Assoc*, 99:832–844, 2004.

[15] J. F. Lawless. *Statistical Models and Methods for Lifetime Data*. Wiley, Hoboken, 2nd edition, 2003.

[16] J. F. Lawless. Two-phase outcome-dependent studies for failure times and testing for effects of expensive covariates. *Lifetime Data Anal*, 24:28–44, 2018.

[17] J. F. Lawless, C. J. Wild, and J. D. Kalbfleisch. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society Series B*, 61(413 - 438), 1999.

[18] F. D. K. Liddell, J. C. McDonald, D. C. Thomas, and S. V. Cuniffe. Methods of cohort analysis: Appraisal by application to asbestos mining. *Journal of the Royal Statistical Society, Series A*, 140:469–491, 1977.

[19] D. Y. Lin, D. Zeng, and Z. Z. Tang. Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proceedings of the National Academy of Sciences*, 110(30):12247–12252, 2013.

[20] R. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer-Verlag New York, Inc, Secaucus, NJ, USA, 2nd edition, 2006.

[21] J. Neyman. Contribution to the theory of sampling from human populations. *Journal of the American Statistical Association*, 33:101–116, 1938.

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

[23] R. L. Prentice and N. E. Breslow. Retrospective studies and failure time models. *Biometrika*, 65:153–158, 1978.

[24] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*, 89:846–866, 1994.

[25] D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[26] A. J. Scott and C. J. Wild. Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84:57–71, 1997.

[27] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

[28] D. C. Thomas. Addendum to: Liddell et al. (1977). *Journal of the Royal Statistical Society, Series A*, 140:483–485, 1977.

[29] C. J. Wild. Fitting prospective regression models to case-control data. *Biometrika*, 78:705–717, 1991.

[30] Y. E. Yilmaz and S. B. Bull. Are quantitative trait-dependent sampling designs cost-effective for analysis of rare and common variants? *BMC Proceedings*, 5(Suppl 9):S111, 2011.

[31] Y. E. Yilmaz and J. F. Lawless. Semiparametric estimation in copula models for bivariate sequential survival times. *Biometrical Journal*, 5:779–796, 2011.

[32] D. Zeng and D. Y. Lin. Efficient estimation of semiparametric transformation models for two-phase cohort studies. *Journal of the American Statistical Association*, 109:371–383, 2014.

[33] Z. Zhang and H. E. Rockette. On maximum likelihood estimation in parametric regression with missing covariates. *J Stat Plan Inference*, 134:206–223, 2005.

[34] L. P. Zhao and S. Lipsitz. Designs and analysis of two-stage studies. *Statistics in Medicine*, 11:769–782, 1992.

[35] Y. Zhao, J. F. Lawless, and D. L. McLeish. Likelihood methods for regression models with expensive variables missing by design. *Biom J*, 51:123–136, 2009.

[36] H. Zhou, J. Chen, T. H. Rissanen, S. A. Korrick, H. Hu, J. T. Salonen, and M. P. Longnecker. Outcome-dependent sampling: An efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology*, 18:461–468, 2007.

[37] H. Zhou, M. A. Weaver, J. Qin, M. Longnecker, and M. C. Wang. A semiparametric empirical likelihood method for data from an outcome dependent sampling scheme with a continuous outcome. *Biometrics*, 58:413–421, 2002.

Appendix A

Tables for Chapter 3

A.1 Generalized case-cohort designs based on the event indicator of the first gap time

Table A.1: Coefficient estimates and their estimated standard errors under generalized case-cohort designs based on the first event indicator

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(0, 0, 0.5)$	1	(1000,9000)	0.023	0.0411	-0.029	0.0842
	2	(2000,8000)	0.068	0.0334	0.137	0.0602
	3	(3000,7000)	-0.009	0.0313	-0.056	0.0526
	4	(4000,6000)	-0.052	0.0303	-0.025	0.0478
	5	(5000,5000)	-0.008	0.0293	-0.006	0.0419
	6	(6000,4000)	0.021	0.0298	0.042	0.0394
	7	(7000,3000)	0.013	0.0317	0.002	0.0390
	8	(8000,2000)	-0.039	0.0349	-0.042	0.0391
	9	(9000,1000)	-0.047	0.0421	-0.058	0.0421

Continued on next page

Table A.1 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	10	(10000,0)	0.044	0.0631	0.004	0.0544
$(0, 1, 0.5)$	1	(1000,9000)	0.015	0.0381	0.938	0.0567
	2	(2000,8000)	0.044	0.0309	1.065	0.0430
	3	(3000,7000)	0.003	0.0292	0.957	0.0414
	4	(4000,6000)	-0.028	0.0281	1.008	0.0371
	5	(5000,5000)	-0.023	0.0278	1.000	0.0347
	6	(6000,4000)	0.016	0.0284	1.039	0.0332
	7	(7000,3000)	0.010	0.0302	0.995	0.0336
	8	(8000,2000)	-0.041	0.0332	0.963	0.0341
	9	(9000,1000)	-0.048	0.0391	0.947	0.0366
	10	(10000,0)	0.040	0.0541	0.988	0.0442
$(1, 0, 0.5)$	1	(1000,9000)	0.998	0.0291	0.087	0.0570
	2	(2000,8000)	0.983	0.0256	0.055	0.0450
	3	(3000,7000)	0.985	0.0247	-0.047	0.0388
	4	(4000,6000)	1.006	0.0241	0.071	0.0364
	5	(5000,5000)	0.966	0.0244	-0.011	0.0342
	6	(6000,4000)	0.990	0.0249	0.008	0.0326
	7	(7000,3000)	0.942	0.0261	0.007	0.0320
	8	(8000,2000)	0.991	0.0271	-0.024	0.0309
	9	(9000,1000)	0.979	0.0292	-0.001	0.0310
	10	(10000,0)	0.947	0.0330	-0.027	0.0322
$(1, 1, 0.5)$	1	(1000,9000)	1.013	0.0292	1.042	0.0596
	2	(2000,8000)	0.987	0.0261	1.045	0.0464
	3	(3000,7000)	0.994	0.0251	0.954	0.0408
	4	(4000,6000)	1.010	0.0243	1.019	0.0354

Continued on next page

Table A.1 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}},\, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	5	(5000,5000)	0.974	0.0246	0.986	0.0333
	6	(6000,4000)	0.990	0.0250	0.993	0.0311
	7	(7000,3000)	0.964	0.0257	1.028	0.0298
	8	(8000,2000)	0.979	0.0268	0.964	0.0289
	9	(9000,1000)	1.029	0.0280	1.027	0.0277
	10	(10000,0)	0.960	0.0317	0.980	0.0288
$(0, 0, 1.0)$	1	(1000,9000)	0.012	0.0343	0.084	0.0785
	2	(2000,8000)	0.018	0.0312	-0.017	0.0594
	3	(3000,7000)	0.020	0.0300	0.003	0.0492
	4	(4000,6000)	0.027	0.0289	0.034	0.0444
	5	(5000,5000)	-0.008	0.0291	-0.007	0.0404
	6	(6000,4000)	-0.038	0.0297	-0.021	0.0383
	7	(7000,3000)	-0.049	0.0311	-0.037	0.0378
	8	(8000,2000)	0.00479	0.0336	-0.00032	0.0367
	9	(9000,1000)	-0.020	0.0389	-0.061	0.0387
	10	(10000,0)	0.075	0.0484	0.058	0.0424
$(0, 1, 1.0)$	1	(1000,9000)	0.013	0.0318	1.023	0.0478
	2	(2000,8000)	0.019	0.0291	0.990	0.0431
	3	(3000,7000)	0.018	0.0282	0.988	0.0386
	4	(4000,6000)	0.031	0.0270	1.045	0.0350
	5	(5000,5000)	-0.011	0.0274	0.994	0.0333
	6	(6000,4000)	-0.021	0.0280	0.978	0.0323
	7	(7000,3000)	-0.018	0.0291	1.000	0.0318
	8	(8000,2000)	0.038	0.0317	1.026	0.0316
	9	(9000,1000)	0.000	0.0355	0.992	0.0332

Continued on next page

Table A.1 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	10	(10000,0)	0.051	0.0427	1.029	0.0357
$(1, 0, 1.0)$	1	(1000,9000)	0.998	0.0281	0.096	0.0513
	2	(2000,8000)	0.986	0.0261	0.070	0.0427
	3	(3000,7000)	0.999	0.0251	0.040	0.0369
	4	(4000,6000)	0.994	0.0245	-0.004	0.0337
	5	(5000,5000)	1.040	0.0243	0.081	0.0316
	6	(6000,4000)	0.972	0.0246	0.020	0.0304
	7	(7000,3000)	0.998	0.0247	0.035	0.0289
	8	(8000,2000)	0.947	0.0255	-0.029	0.0283
	9	(9000,1000)	0.994	0.0260	0.012	0.0277
	10	(10000,0)	0.980	0.0273	0.014	0.0273
$(1, 1, 1.0)$	1	(1000,9000)	1.011	0.0276	1.060	0.0504
	2	(2000,8000)	1.002	0.0257	1.041	0.0417
	3	(3000,7000)	0.999	0.0250	1.012	0.0366
	4	(4000,6000)	1.002	0.0244	0.997	0.0335
	5	(5000,5000)	1.044	0.0241	1.079	0.0301
	6	(6000,4000)	0.990	0.0243	1.036	0.0288
	7	(7000,3000)	0.992	0.0245	1.013	0.0273
	8	(8000,2000)	0.955	0.0249	0.990	0.0265
	9	(9000,1000)	0.992	0.0255	0.999	0.0254
	10	(10000,0)	0.977	0.0265	1.004	0.0248
$(0, 0, 1.5)$	1	(1000,9000)	0.082	0.0312	0.031	0.0768
	2	(2000,8000)	-0.064	0.0311	-0.078	0.0582
	3	(3000,7000)	0.018	0.0294	0.040	0.0487
	4	(4000,6000)	-0.033	0.0291	-0.048	0.0445

Continued on next page

Table A.1 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	5	(5000,5000)	-0.007	0.0293	-0.038	0.0405
	6	(6000,4000)	0.043	0.0293	0.021	0.0375
	7	(7000,3000)	0.011	0.0306	-0.021	0.0362
	8	(8000,2000)	0.057	0.0317	0.051	0.0348
	9	(9000,1000)	-0.013	0.0352	-0.025	0.0357
	10	(10000,0)	-0.011	0.0393	-0.014	0.0365
$(0, 1, 1.5)$	1	(1000,9000)	0.069	0.0289	1.045	0.0466
	2	(2000,8000)	-0.064	0.0289	0.936	0.0420
	3	(3000,7000)	0.014	0.0273	1.037	0.0369
	4	(4000,6000)	-0.019	0.0270	0.983	0.0350
	5	(5000,5000)	0.009	0.0272	1.001	0.0329
	6	(6000,4000)	0.046	0.0273	1.032	0.0310
	7	(7000,3000)	0.024	0.0285	0.990	0.0308
	8	(8000,2000)	0.027	0.0292	1.026	0.0299
	9	(9000,1000)	-0.041	0.0319	0.959	0.0304
	10	(10000,0)	-0.001	0.0344	1.004	0.0306
$(1, 0, 1.5)$	1	(1000,9000)	1.007	0.0290	0.044	0.0467
	2	(2000,8000)	0.973	0.0272	-0.001	0.0404
	3	(3000,7000)	0.993	0.0265	0.038	0.0367
	4	(4000,6000)	1.029	0.0256	0.062	0.0334
	5	(5000,5000)	0.983	0.0252	-0.012	0.0313
	6	(6000,4000)	0.981	0.0249	0.002	0.0296
	7	(7000,3000)	1.005	0.0245	0.028	0.0282
	8	(8000,2000)	0.978	0.0245	-0.012	0.0273
	9	(9000,1000)	1.003	0.0245	-0.003	0.0265

Continued on next page

Table A.1 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	(n_{cases}, n_{noncases})	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	10	(10000,0)	0.964	0.0251	-0.013	0.0261
$(1, 1, 1.5)$	1	(1000,9000)	1.018	0.0282	0.999	0.0490
	2	(2000,8000)	0.971	0.0268	0.989	0.0408
	3	(3000,7000)	1.003	0.0260	1.019	0.0358
	4	(4000,6000)	1.029	0.0253	1.055	0.0323
	5	(5000,5000)	0.980	0.0250	0.966	0.0304
	6	(6000,4000)	0.981	0.0246	0.999	0.0282
	7	(7000,3000)	0.992	0.0241	1.014	0.0264
	8	(8000,2000)	0.985	0.0239	0.997	0.0255
	9	(9000,1000)	1.005	0.0239	0.991	0.0245
	10	(10000,0)	0.963	0.0243	1.005	0.0239

A.2 Outcome-dependent BSS designs based on the first gap time and its event indicator

Table A.2: Coefficient estimates and their estimated standard errors under outcome-dependent BSS designs based on the first gap time and its event indicator

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	($n_{\text{cases},j} : j = 1, 2, 3$), ($n_{\text{noncases},j} : j = 1, 2, 3$)	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(0, 0, 0.5)$	SRS in S_{cases} and S_{noncases}	(833,3334,833),(1250,2500,1250)	0.012	0.0295	0.020	0.0419
	1	(5000,0,0),(5000,0,0)	-0.031	0.0354	-0.015	0.0494
	2	(5000,0,0),(0,5000,0)	-0.018	0.0254	-0.002	0.0433
	3	(5000,0,0),(0,0,5000)	0.004	0.0205	0.019	0.0408
	4	(0,5000,0),(5000,0,0)	-0.061	0.0451	0.012	0.0489
	5	(4000,1000,0),(0,0,5000)	0.009	0.0209	0.034	0.0403
	6	(3000,1000,1000),(0,0,5000)	-0.010	0.0223	0.029	0.0405
	7	(0,0,5000),(5000,0,0)	-0.092	0.0986	-0.066	0.0678
	8	(5000,0,0),(1000,0,4000)	0.005	0.0216	0.020	0.0414

Continued on next page

Table A.2 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases},j} : j = 1,2,3)$, $(n_{\text{noncases},j} : j = 1,2,3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	9	(5000,0,0),(1000,1000,3000)	0.005	0.0226	0.019	0.0418
$(0, 1, 0.5)$	SRS in S_{cases} and S_{noncases}	(833,3334,833),(1250,2500,1250)	0.005	0.0281	0.998	0.0348
	1	(5000,0,0),(5000,0,0)	-0.047	0.0338	0.943	0.0419
	2	(5000,0,0),(0,5000,0)	-0.029	0.0245	0.960	0.0362
	3	(5000,0,0),(0,0,5000)	-0.006	0.0199	0.983	0.0339
	4	(0,5000,0),(5000,0,0)	-0.066	0.0420	0.988	0.0419
	5	(4000,1000,0),(0,0,5000)	-0.000	0.0203	0.997	0.0334
	6	(3000,1000,1000),(0,0,5000)	-0.015	0.0214	0.989	0.0329
	7	(0,0,5000),(5000,0,0)	-0.081	0.0724	0.971	0.0499
	8	(5000,0,0),(1000,0,4000)	-0.005	0.0210	0.983	0.0345
	9	(5000,0,0),(1000,1000,3000)	-0.006	0.0219	0.983	0.0349
$(1, 0, 0.5)$	SRS in S_{cases} and S_{noncases}	(666,2668,666),(1500,3000,1500)	1.011	0.0240	0.029	0.0358
	1	(4000,0,0),(5000,1000,0)	0.977	0.0254	-0.012	0.0359
	2	(4000,0,0),(0,6000,0)	1.010	0.0204	0.023	0.0337
	3	(4000,0,0),(0,1000,5000)	0.996	0.0187	-0.015	0.0330
	4	(0,4000,0),(5000,1000,0)	0.963	0.0324	-0.011	0.0387
	5	(3000,1000,0),(0,1000,5000)	1.007	0.0192	0.011	0.0331
	6	(2000,1000,1000),(0,1000,5000)	1.013	0.0199	0.033	0.0349
	7	(0,0,4000),(5000,1000,0)	0.921	0.0511	-0.028	0.0527
	8	(4000,0,0),(1000,1000,4000)	1.002	0.0196	-0.009	0.0334
	9	(4000,0,0),(1000,2000,3000)	1.003	0.0199	0.004	0.0334
$(1, 1, 0.5)$	SRS in S_{cases} and S_{noncases}	(666,2668,666),(1500,3000,1500)	1.019	0.0242	1.021	0.0355
	1	(4000,0,0),(5000,1000,0)	0.979	0.0258	0.969	0.0399
	2	(4000,0,0),(0,6000,0)	1.015	0.0205	1.002	0.0373
	3	(4000,0,0),(0,1000,5000)	1.002	0.0189	0.971	0.0369
	4	(0,4000,0),(5000,1000,0)	0.988	0.0321	1.012	0.0367
	5	(3000,1000,0),(0,1000,5000)	1.010	0.0194	0.992	0.0358
	6	(2000,1000,1000),(0,1000,5000)	1.0238	0.0201	0.988	0.0363
	7	(0,0,4000),(5000,1000,0)	0.968	0.0490	0.960	0.0436
	8	(4000,0,0),(1000,1000,4000)	1.009	0.0198	0.976	0.0372
	9	(4000,0,0),(1000,2000,3000)	1.007	0.0201	0.990	0.0372
$(0, 0, 1.0)$	SRS in S_{cases} and S_{noncases}	(666,2668,666),(1500,3000,1500)	0.022	0.0292	-0.021	0.0449
	1	(4000,0,0),(5000,1000,0)	-0.017	0.0395	-0.024	0.0551
	2	(4000,0,0),(0,6000,0)	-0.017	0.0297	-0.028	0.0492
	3	(4000,0,0),(0,1000,5000)	-0.002	0.0192	0.024	0.0441
	4	(0,4000,0),(5000,1000,0)	-0.065	0.0531	-0.036	0.0549
	5	(3000,1000,0),(0,1000,5000)	0.012	0.0196	0.003	0.0430
	6	(2000,1000,1000),(0,1000,5000)	-0.007	0.0215	0.031	0.0431
	7	(0,0,4000),(5000,1000,0)	-0.000	0.0963	0.032	0.0596
	8	(4000,0,0),(1000,1000,4000)	-0.014	0.0202	0.012	0.0445
	9	(4000,0,0),(1000,2000,3000)	-0.011	0.0217	-0.025	0.0454
$(0, 1, 1.0)$	SRS in S_{cases} and S_{noncases}	(666,2668,666),(1500,3000,1500)	0.028	0.0273	1.016	0.0359
	1	(4000,0,0),(5000,1000,0)	-0.016	0.0362	0.960	0.0445
	2	(4000,0,0),(0,6000,0)	-0.016	0.0278	0.965	0.0394
	3	(4000,0,0),(0,1000,5000)	-0.007	0.0183	1.006	0.0341
	4	(0,4000,0),(5000,1000,0)	-0.029	0.0475	1.000	0.0456

Continued on next page

Table A.2 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases},j} : j = 1, 2, 3)$, $(n_{\text{noncases},j} : j = 1, 2, 3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	5	(3000,1000,0),(0,1000,5000)	0.012	0.0187	1.002	0.0339
	6	(2000,1000,1000),(0,1000,5000)	0.001	0.0201	1.017	0.0326
	7	(0,0,4000),(5000,1000,0)	-0.023	0.0701	1.013	0.0454
	8	(4000,0,0),(1000,1000,4000)	-0.019	0.0193	0.994	0.0345
	9	(4000,0,0),(1000,2000,3000)	-0.011	0.0207	0.973	0.0360
$(1, 0, 1.0)$	SRS in S_{cases} and S_{noncases}	(833,3334,833),(1250,2500,1250)	1.002	0.0242	0.021	0.0313
	1	(5000,0,0),(5000,0,0)	0.965	0.0330	-0.024	0.0373
	2	(5000,0,0),(0,5000,0)	0.998	0.0226	0.003	0.0314
	3	(5000,0,0),(0,0,5000)	1.000	0.0191	0.005	0.0301
	4	(0,5000,0),(5000,0,0)	0.985	0.0408	0.016	0.0375
	5	(4000,1000,0),(0,0,5000)	1.001	0.0192	-0.003	0.0296
	6	(3000,1000,1000),(0,0,5000)	0.991	0.0187	-0.018	0.0302
	7	(0,0,5000),(5000,0,0)	0.999	0.0343	0.033	0.0398
	8	(5000,0,0),(1000,0,4000)	0.999	0.0204	0.004	0.0306
	9	(5000,0,0),(1000,1000,3000)	0.971	0.0207	-0.020	0.0308
$(1, 1, 1.0)$	SRS in S_{cases} and S_{noncases}	(833,3334,833),(1250,2500,1250)	1.010	0.0242	1.000	0.0305
	1	(5000,0,0),(5000,0,0)	0.970	0.0326	0.963	0.0397
	2	(5000,0,0),(0,5000,0)	0.999	0.0225	0.986	0.0340
	3	(5000,0,0),(0,0,5000)	1.001	0.0189	0.988	0.0324
	4	(0,5000,0),(5000,0,0)	1.012	0.0386	1.017	0.0329
	5	(4000,1000,0),(0,0,5000)	1.003	0.0190	0.975	0.0312
	6	(3000,1000,1000),(0,0,5000)	0.994	0.0187	0.954	0.0318
	7	(0,0,5000),(5000,0,0)	1.014	0.0339	0.964	0.0357
	8	(5000,0,0),(1000,0,4000)	1.000	0.0202	0.988	0.0329
	9	(5000,0,0),(1000,1000,3000)	0.973	0.0206	0.965	0.0333
$(0, 0, 1.5)$	SRS in S_{cases} and S_{noncases}	(666,2668,666),(1500,3000,1500)	-0.008	0.0290	-0.021	0.0432
	1	(4000,0,0),(5000,1000,0)	0.012	0.0440	0.005	0.0589
	2	(4000,0,0),(0,6000,0)	0.007	0.0340	0.004	0.0515
	3	(4000,0,0),(0,1000,5000)	0.004	0.0184	-0.012	0.0437
	4	(0,4000,0),(5000,1000,0)	0.015	0.0612	0.002	0.0571
	5	(3000,1000,0),(0,1000,5000)	-0.010	0.0191	-0.029	0.0439
	6	(2000,1000,1000),(0,1000,5000)	0.004	0.0204	0.001	0.0431
	7	(0,0,4000),(5000,1000,0)	0.022	0.0591	0.018	0.0468
	8	(4000,0,0),(1000,1000,4000)	0.007	0.0195	-0.009	0.0441
	9	(4000,0,0),(1000,2000,3000)	0.006	0.0210	-0.022	0.0452
$(0, 1, 1.5)$	SRS in S_{cases} and S_{noncases}	(666,2668,666),(1500,3000,1500)	0.002	0.0270	0.998	0.0345
	1	(4000,0,0),(5000,1000,0)	0.017	0.0376	1.020	0.0449
	2	(4000,0,0),(0,6000,0)	0.004	0.0306	0.997	0.0401
	3	(4000,0,0),(0,1000,5000)	0.001	0.0175	0.988	0.0336
	4	(0,4000,0),(5000,1000,0)	0.031	0.0493	1.027	0.0449
	5	(3000,1000,0),(0,1000,5000)	-0.003	0.0180	1.014	0.0332
	6	(2000,1000,1000),(0,1000,5000)	0.008	0.0191	1.024	0.0321
	7	(0,0,4000),(5000,1000,0)	0.013	0.0507	1.005	0.0363
	8	(4000,0,0),(1000,1000,4000)	0.003	0.0185	0.989	0.0341
	9	(4000,0,0),(1000,2000,3000)	0.007	0.0199	0.993	0.0351
$(1, 0, 1.5)$	SRS in S_{cases} and S_{noncases}	(1500,6000,1500),(250,500,250)	1.004	0.0248	0.002	0.0265

Continued on next page

Table A.2 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases},j} : j = 1,2,3)$, $(n_{\text{noncases},j} : j = 1,2,3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	1	(5000,4000,0),(1000,0,0)	0.976	0.0339	-0.027	0.0319
	2	(1000,4000,4000),(0,0,1000)	1.010	0.0205	0.025	0.0263
	3	(1000,5000,3000),(0,0,1000)	0.996	0.0213	-0.003	0.0259
	4	(0,9000,0),(500,0,0)	1.014	0.0344	0.024	0.0292
	5	(0,4000,5000),(0,0,500)	1.005	0.0221	0.017	0.0275
	6	(0,4000,5000),(0,500,500)	1.020	0.0217	0.036	0.0274
	7	(0,4000,5000),(1000,0,0)	0.991	0.0226	-0.013	0.0279
	8	(0,4000,5000),(0,1000,0)	0.997	0.0225	-0.009	0.0275
	9	(0,4000,5000),(0,0,1000)	1.015	0.0216	0.008	0.0273
$(1, 1, 1.5)$	SRS in S_{cases} and S_{noncases}	(1333,5334,1333),(500,1000,500)	0.981	0.0239	1.005	0.0245
	1	(5000,3000,0),(2000,0,0)	1.026	0.0311	1.009	0.0303
	2	(1000,3000,4000),(0,0,2000)	1.012	0.0201	1.029	0.0261
	3	(1000,4000,3000),(0,0,2000)	1.018	0.0205	0.985	0.0256
	4	(0,8000,0),(2000,0,0)	1.012	0.0332	1.021	0.0262
	5	(0,3000,5000),(1000,0,1000)	0.991	0.0228	0.993	0.0270
	6	(0,3000,5000),(0,1000,1000)	1.011	0.0225	1.034	0.0271
	7	(0,3000,5000),(2000,0,0)	0.994	0.0236	0.985	0.0274
	8	(0,3000,5000),(0,2000,0)	0.996	0.0237	0.978	0.0271
	9	(0,3000,5000),(0,0,2000)	1.021	0.0219	0.994	0.0269

A.3 Generalized case-cohort designs based on the event indicators of the two sequential gap times

Table A.3: Coefficient estimates and their estimated standard errors under generalized case-cohort designs based on the event indicators of the two sequential gap times

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases}}, m_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(0, 0, 0.5)$	1	(500,4500)	-0.042	0.0407	-0.145	0.0782
	2	(1000,4000)	0.036	0.0345	0.051	0.0587
	3	(1500,3500)	-0.049	0.0333	-0.071	0.0542
	4	(2000,3000)	-0.056	0.0319	-0.075	0.0489
	5	(2500,2500)	-0.011	0.0302	-0.052	0.0453
	6	(3000,2000)	-0.042	0.0301	-0.095	0.0435

Continued on next page

Table A.3 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases}}, m_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	7	(3500,1500)	-0.011	0.0290	-0.024	0.0410
	8	(4000,1000)	0.009	0.0286	-0.002	0.0399
	9	(4500,500)	0.018	0.0281	-0.016	0.0392
	10	(5000,0)	0.027	0.0279	-0.062	0.0389
$(0, 1, 0.5)$	1	(500,4500)	-0.126	0.0413	0.935	0.0507
	2	(1000,4000)	0.022	0.0339	1.050	0.0412
	3	(1500,3500)	-0.043	0.0327	0.981	0.0404
	4	(2000,3000)	-0.040	0.0312	0.997	0.0380
	5	(2500,2500)	-0.032	0.0299	0.979	0.0372
	6	(3000,2000)	0.016	0.0284	0.985	0.0358
	7	(3500,1500)	0.018	0.0276	1.016	0.0351
	8	(4000,1000)	-0.015	0.0272	0.981	0.0349
	9	(4500,500)	0.005	0.0265	0.973	0.0350
	10	(5000,0)	-0.028	0.0262	0.98105	0.0349
$(1, 0, 0.5)$	1	(500,3500)	0.980	0.0275	-0.021	0.0510
	2	(1000,3000)	1.000	0.0256	0.057	0.0435
	3	(1500,2500)	0.984	0.0249	0.002	0.0395
	4	(2000,2000)	1.012	0.0242	-0.026	0.0361
	5	(2500,1500)	0.967	0.0242	0.006	0.0358
	6	(3000,1000)	0.990	0.0238	-0.018	0.0342
	7	(3500,500)	0.980	0.0236	-0.017	0.0336
	8	(4000,0)	0.995	0.0232	-0.002	0.0335
$(1, 1, 0.5)$	1	(500,3500)	1.00623	0.0288	1.06334	0.0473
	2	(1000,3000)	1.00021	0.0269	1.00849	0.0423
	3	(1500,2500)	0.98290	0.0262	0.98336	0.0390

Continued on next page

Table A.3 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases}}, m_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	4	(2000,2000)	0.99760	0.0252	0.99513	0.0371
	5	(2500,1500)	1.00878	0.0244	1.02751	0.0360
	6	(3000,1000)	1.01085	0.0241	0.97606	0.0361
	7	(3500,500)	0.99341	0.0238	0.95936	0.0362
	8	(4000,0)	1.00102	0.0233	1.02640	0.0370
$(0, 0, 1.0)$	1	(500,3500)	-0.008	0.0359	0.073	0.0610
	2	(1000,3000)	0.017	0.0327	0.036	0.0534
	3	(1500,2500)	-0.050	0.0320	-0.072	0.0511
	4	(2000,2000)	-0.023	0.0304	-0.011	0.0467
	5	(2500,1500)	-0.009	0.0292	-0.040	0.0448
	6	(3000,1000)	0.018	0.0285	0.089	0.0426
	7	(3500,500)	-0.017	0.0282	-0.033	0.0429
	8	(4000,0)	-0.001	0.0278	-0.002	0.0432
$(0, 1, 1.0)$	1	(500,3500)	0.000	0.0348	1.042	0.0408
	2	(1000,3000)	0.006	0.0319	1.028	0.0385
	3	(1500,2500)	-0.016	0.0306	0.99	0.0378
	4	(2000,2000)	-0.040	0.0298	0.933	0.0375
	5	(2500,1500)	0.012	0.0278	1.022	0.0355
	6	(3000,1000)	0.006	0.0273	0.983	0.0355
	7	(3500,500)	0.011	0.0265	0.999	0.0357
	8	(4000,0)	0.007	0.0260	0.974	0.0362
$(1, 0, 1.0)$	1	(500,4500)	0.952	0.0265	-0.003	0.0423
	2	(1000,4000)	0.952	0.0260	-0.003	0.0389
	3	(1500,3500)	0.954	0.0253	-0.012	0.0354
	4	(2000,3000)	0.949	0.0248	-0.025	0.0339

Continued on next page

Table A.3 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	(m_{cases}, m_{noncases})	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	5	(2500,2500)	0.978	0.0246	0.008	0.0328
	6	(3000,2000)	0.964	0.0244	-0.014	0.0318
	7	(3500,1500)	0.959	0.0244	-0.048	0.0313
	8	(4000,1000)	0.956	0.0247	0.020	0.0324
	9	(4500,500)	0.948	0.0249	-0.006	0.0330
	10	(5000,0)	0.974	0.0246	-0.070	0.0323
$(1, 1, 1.0)$	1	(500,4500)	0.946	0.0278	1.017	0.0365
	2	(1000,4000)	0.945	0.0266	0.995	0.0335
	3	(1500,3500)	0.969	0.0256	0.984	0.0325
	4	(2000,3000)	0.956	0.0253	0.970	0.0315
	5	(2500,2500)	0.979	0.0246	1.023	0.0305
	6	(3000,2000)	0.950	0.0248	0.931	0.0311
	7	(3500,1500)	0.972	0.0243	0.962	0.0312
	8	(4000,1000)	0.974	0.0243	0.951	0.0317
	9	(4500,500)	0.975	0.0243	1.017	0.0319
	10	(5000,0)	0.978	0.0244	0.940	0.0344
$(0, 0, 1.5)$	1	(500,3500)	0.027	0.0324	0.073	0.0545
	2	(1000,3000)	0.001	0.0310	0.103	0.0493
	3	(1500,2500)	0.023	0.0300	0.111	0.0472
	4	(2000,2000)	-0.010	0.0297	-0.020	0.0454
	5	(2500,1500)	-0.027	0.0295	-0.061	0.0450
	6	(3000,1000)	-0.008	0.0286	-0.029	0.0429
	7	(3500,500)	0.078	0.0275	-0.003	0.0418
	8	(4000,0)	-0.002	0.0281	-0.044	0.0440
$(0, 1, 1.5)$	1	(500,3500)	-0.019	0.0317	1.016	0.0376

Continued on next page

Table A.3 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	(m_{cases}, m_{noncases})	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	2	(1000,3000)	-0.005	0.0299	1.010	0.0356
	3	(1500,2500)	-0.023	0.0293	0.970	0.0360
	4	(2000,2000)	0.013	0.0278	1.020	0.0344
	5	(2500,1500)	-0.021	0.0277	0.982	0.0352
	6	(3000,1000)	-0.003	0.0268	0.987	0.0347
	7	(3500,500)	0.013	0.0260	1.038	0.0341
	8	(4000,0)	-0.002	0.0260	0.970	0.0357
$(1, 0, 1.5)$	1	(500,8500)	0.983	0.0233	-0.010	0.0374
	2	(1000,8000)	0.986	0.0227	0.024	0.0334
	3	(1500,7500)	0.979	0.0226	0.009	0.0312
	4	(2000,7000)	0.983	0.0225	-0.020	0.0295
	5	(2500,6500)	0.983	0.0225	-0.003	0.0281
	6	(3000,6000)	0.959	0.0228	-0.016	0.0275
	7	(3500,5500)	1.018	0.0226	0.013	0.0261
	8	(4000,5000)	1.013	0.0228	0.017	0.0262
	9	(4500,4500)	0.998	0.0234	-0.005	0.0260
	10	(5000,4000)	1.015	0.0236	-0.005	0.0259
	11	(5500,3500)	1.012	0.0243	0.012	0.0261
	12	(6000,3000)	0.995	0.0247	0.015	0.0265
	13	(6500,2500)	1.014	0.0254	0.002	0.0272
	14	(7000,2000)	1.011	0.0264	0.010	0.0280
	15	(7500,1500)	0.966	0.0277	-0.029	0.0295
	16	(8000,1000)	0.997	0.0286	0.008	0.0307
	17	(8500,500)	1.022	0.0313	0.017	0.0340

Continued on next page

Table A.3 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases}}, m_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	18	(9000,0)	1.007	0.0347	-0.016	0.0384
$(1, 1, 1.5)$	1	(500,7500)	0.999	0.0256	1.024	0.0305
	2	(1000,7000)	0.968	0.0253	0.974	0.0290
	3	(1500,6500)	0.958	0.0253	0.970	0.0275
	4	(2000,6000)	0.974	0.0250	0.969	0.0263
	5	(2500,5500)	0.987	0.0243	0.992	0.0257
	6	(3000,5000)	0.986	0.0242	1.021	0.0249
	7	(3500,4500)	1.021	0.0238	1.017	0.0245
	8	(4000,4000)	1.003	0.0242	0.985	0.0246
	9	(4500,3500)	1.004	0.0240	1.009	0.0246
	10	(5000,3000)	1.004	0.0238	1.032	0.0245
	11	(5500,2500)	0.988	0.0241	0.980	0.0252
	12	(6000,2000)	1.037	0.0238	1.043	0.0255
	13	(6500,1500)	0.991	0.0244	0.960	0.0265
	14	(7000,1000)	1.005	0.0245	1.016	0.0270
	15	(7500,500)	1.002	0.0251	1.010	0.0283
	16	(8000,0)	1.024	0.0254	1.043	0.0291

A.4 Outcome-dependent BSS designs based on the two sequential gap times and their event indicators

Table A.4: Coefficient estimates and their estimated standard errors under outcome-dependent BSS designs based on the two sequential gap times and their event indicators

$(\alpha_{11},\alpha_{21},\gamma_1)$	Sampling scenario	$(m_{\text{cases},j}: j=1,2,3)$, $(m_{\text{noncases},j}: j=1,2,3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(0,0,0.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(666,3669,665),(0,0,0)	0.001	0.0210	-0.030	0.0467
	1	(1500,2500,1000),(0,0,0)	-0.004	0.0209	0.002	0.0462
	2	(2000,2500,500),(0,0,0)	0.007	0.0206	0.029	0.0420
	3	(2500,2500,0),(0,0,0)	-0.001	0.0209	-0.005	0.0352
	4	(500,4000,500),(0,0,0)	-0.005	0.0210	-0.041	0.0462
	5	(250,4500,250),(0,0,0)	-0.005	0.0211	-0.044	0.0457
	6	(0,5000,0),(0,0,0)	-0.010	0.0212	-0.075	0.0454
	7	(1000,2500,1500),(0,0,0)	0.002	0.0212	0.008	0.0517
	8	(500,2500,2000),(0,0,0)	0.000	0.0222	0.0202	0.0605
	9	(0,2500,2500),(0,0,0)	0.004	0.0236	-0.010	0.0738
$(0,1,0.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(608,3785,607),(0,0,0)	-0.008	0.0207	0.992	0.0379
	1	(1500,2500,1000),(0,0,0)	-0.012	0.0206	0.980	0.0354
	2	(2000,2500,500),(0,0,0)	-0.006	0.0201	0.980	0.0330
	3	(2500,2500,0),(0,0,0)	-0.000	0.0200	0.973	0.0296
	4	(500,4000,500),(0,0,0)	-0.001	0.0206	1.003	0.0382
	5	(250,4500,250),(0,0,0)	-0.001	0.0205	0.979	0.0396
	6	(0,5000,0),(0,0,0)	-0.011	0.0206	0.964	0.0409
	7	(1000,2500,1500),(0,0,0)	-0.011	0.0213	0.976	0.0382
	8	(500,2500,2000),(0,0,0)	-0.022	0.0226	0.997	0.0429
	9	(0,2500,2500),(0,0,0)	-0.008	0.0259	0.972	0.0559
$(1,0,0.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(599,2803,598),(0,0,0)	1.006	0.0189	0.012	0.0366
	1	(1500,1500,1000),(0,0,0)	1.009	0.0190	0.045	0.0348
	2	(2000,1500,500),(0,0,0)	1.025	0.0189	0.018	0.0322
	3	(2500,1500,0),(0,0,0)	0.996	0.0193	0.010	0.0310
	4	(500,3000,500),(0,0,0)	0.999	0.0190	0.010	0.0365
	5	(250,3500,250),(0,0,0)	0.998	0.0190	-0.030	0.0354
	6	(0,4000,0),(0,0,0)	0.993	0.0191	-0.001	0.0359
	7	(1000,1500,1500),(0,0,0)	1.011	0.0190	0.038	0.0371
	8	(500,1500,2000),(0,0,0)	1.010	0.0194	0.034	0.0402
	9	(0,1500,2500),(0,0,0)	1.007	0.0198	0.042	0.0447
$(1,1,0.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(316,1868,316),(367,766,367)	1.003	0.0204	1.021	0.0441
	1	(2500,0,0),(1500,0,0)	1.006	0.0211	1.000	0.0310
	2	(2500,0,0),(0,1500,0)	1.004	0.0215	0.993	0.0277
	3	(2500,0,0),(0,0,1500)	0.993	0.0221	1.021	0.0253
	4	(0,2500,0),(1500,0,0)	1.011	0.0206	1.021	0.0618
	5	(0,2500,0),(0,1500,0)	1.003	0.0205	1.003	0.0518
	6	(0,2500,0),(0,0,1500)	1.010	0.0200	1.043	0.0420
	7	(0,0,2500),(1500,0,0)	1.025	0.0260	1.046	0.0757
	8	(0,0,2500),(0,1500,0)	1.030	0.0269	1.032	0.0670
	9	(0,0,2500),(0,0,1500)	1.034	0.0255	1.068	0.0522
$(0,0,1.0)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(372,2256,372),(254,492,254)	0.004	0.0216	-0.034	0.0577
	1	(2500,500,0),(1000,0,0)	-0.006	0.0214	0.009	0.0359
	2	(2500,500,0),(0,1000,0)	0.012	0.0211	-0.015	0.0336

Continued on next page

Table A.4 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases},j} : j = 1, 2, 3)$, $(m_{\text{noncases},j} : j = 1, 2, 3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	3	(2500,500,0),(0,0,1000)	0.013	0.0211	-0.009	0.0311
	4	(0,3000,0),(1000,0,0)	-0.005	0.0218	-0.018	0.0730
	5	(0,3000,0),(0,1000,0)	-0.011	0.0221	-0.046	0.0646
	6	(0,3000,0),(0,0,1000)	-0.005	0.0220	-0.039	0.0548
	7	(0,500,2500),(1000,0,0)	0.017	0.0268	0.072	0.0946
	8	(0,500,2500),(0,1000,0)	0.016	0.0273	0.070	0.0825
	9	(0,500,2500),(0,0,1000)	0.023	0.0261	0.058	0.0661
$(0, 1, 1.0)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(345,2311,344),(304,392,304)	0.016	0.0205	1.006	0.0402
	1	(2500,500,0),(1000,0,0)	0.010	0.0197	0.979	0.0285
	2	(2500,500,0),(0,1000,0)	-0.002	0.0202	0.983	0.0273
	3	(2500,500,0),(0,0,1000)	0.007	0.0204	0.972	0.0265
	4	(0,3000,0),(1000,0,0)	0.008	0.0210	1.030	0.0484
	5	(0,3000,0),(0,1000,0)	0.007	0.0210	0.984	0.0435
	6	(0,3000,0),(0,0,1000)	0.001	0.0212	1.044	0.0429
	7	(0,500,2500),(1000,0,0)	0.018	0.0278	1.041	0.0531
	8	(0,500,2500),(0,1000,0)	-0.006	0.0300	1.050	0.0534
	9	(0,500,2500),(0,0,1000)	-0.006	0.0292	1.052	0.0491
$(1, 0, 1.0)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(455,2590,455),(348,804,348)	0.995	0.0192	-0.003	0.0336
	1	(2500,1000,0),(1500,0,0)	0.995	0.0197	0.009	0.0297
	2	(2500,1000,0),(0,1500,0)	0.999	0.0196	-0.003	0.0273
	3	(2500,1000,0),(0,0,1500)	0.995	0.0198	0.010	0.0253
	4	(0,3500,0),(1500,0,0)	0.990	0.0196	-0.035	0.0364
	5	(0,3500,0),(0,1500,0)	0.997	0.0195	-0.026	0.0337
	6	(0,3500,0),(0,0,1500)	1.001	0.0193	-0.013	0.0302
	7	(0,1000,2500),(1500,0,0)	1.016	0.0209	-0.017	0.0436
	8	(0,1000,2500),(0,1500,0)	1.011	0.0208	-0.013	0.0402
	9	(0,1000,2500),(0,0,1500)	0.999	0.0207	-0.026	0.0361
$(1, 1, 1.0)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(286,1928,286),(768,965,767)	1.015	0.0193	1.052	0.0340
	1	(2500,0,0),(2500,0,0)	0.992	0.0204	1.009	0.0285
	2	(2500,0,0),(0,2500,0)	0.993	0.0210	1.007	0.0245
	3	(2500,0,0),(0,0,2500)	1.000	0.0215	0.999	0.0217
	4	(0,2500,0),(2500,0,0)	0.998	0.0189	1.012	0.0516
	5	(0,2500,0),(0,2500,0)	0.995	0.0187	1.015	0.0411
	6	(0,2500,0),(0,0,2500)	0.996	0.0188	0.988	0.0327
	7	(0,0,2500),(2500,0,0)	0.993	0.0246	1.013	0.0601
	8	(0,0,2500),(0,2500,0)	0.989	0.0258	0.992	0.0573
	9	(0,0,2500),(0,0,2500)	0.994	0.0247	0.987	0.0437
$(0, 0, 1.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(438,2625,437),(125,250,125)	0.004	0.0198	0.011	0.0566
	1	(2500,1000,0),(500,0,0)	0.008	0.0196	-0.029	0.0357
	2	(2500,1000,0),(0,500,0)	0.004	0.0197	-0.015	0.0345
	3	(2500,100,0),(0,0,500)	0.007	0.0222	-0.027	0.0334
	4	(0,3500,0),(500,0,0)	-0.014	0.0195	-0.055	0.0604
	5	(0,3500,0),(0,500,0)	-0.011	0.0196	-0.071	0.0564
	6	(0,3500,0),(0,0,500)	-0.002	0.0196	-0.004	0.0516
	7	(0,1000,2500),(500,0,0)	0.019	0.0234	0.060	0.0816
	8	(0,1000,2500),(0,500,0)	0.040	0.0235	0.080	0.0769

Continued on next page

Table A.4 – *Continued from previous page*

$(\alpha_{11},\alpha_{21},\gamma_1)$	Sampling scenario	$(m_{\text{cases},j}: j=1,2,3)$, $(m_{\text{noncases},j}: j=1,2,3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	9	(0,1000,2500),(0,0,500)	0.031	0.0229	0.092	0.0649
$(0,1,1.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(405,2691,404),(150,201,149)	-0.005	0.0191	1.006	0.0395
	1	(2500,1000,0),(500,0,0)	0.004	0.0183	0.979	0.0284
	2	(2500,1000,0),(0,500,0)	-0.001	0.0185	0.983	0.0278
	3	(2500,1000,0),(0,0,500)	0.004	0.0185	0.972	0.0274
	4	(0,3500,0),(500,0,0)	0.006	0.0188	1.030	0.0421
	5	(0,3500,0),(0,500,0)	-0.008	0.0192	0.984	0.0426
	6	(0,3500,0),(0,0,500)	0.008	0.0188	1.044	0.0392
	7	(0,1000,2500),(500,0,0)	0.010	0.0236	1.041	0.0457
	8	(0,1000,2500),(0,500,0)	-0.001	0.0240	1.050	0.0441
	9	(0,1000,2500),(0,0,500)	0.015	0.0238	1.052	0.0436
$(1,0,1.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(640,3720,640),(956,2089,955)	1.016	0.0218	0.016	0.0251
	1	(2500,2500,0),(2500,1500,0)	0.995	0.0213	0.033	0.0266
	2	(2500,2500,0),(0,4000,0)	0.989	0.0214	0.005	0.0239
	3	(2500,2500,0),(0,1500,2500)	1.009	0.0220	0.020	0.0219
	4	(0,5000,0),(2500,1500,0)	0.991	0.0223	0.036	0.0296
	5	(0,5000,0),(0,4000,0)	0.988	0.0227	0.019	0.0264
	6	(0,5000,0),(0,1500,2500)	0.996	0.0233	0.017	0.0239
	7	(0,2500,2500),(2500,1500,0)	1.025	0.0218	0.009	0.0336
	8	(0,2500,2500),(0,4000,0)	1.019	0.0220	-0.001	0.0309
	9	(0,2500,2500),(0,1500,2500)	1.035	0.0226	0.034	0.0277
$(1,1,1.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(571,3858,571),(925,1151,924)	1.001	0.0219	1.007	0.0247
	1	(2500,2500,0),(2500,500,0)	1.000	0.0207	0.999	0.0263
	2	(2500,2500,0),(0,3000,0)	1.018	0.0211	1.024	0.0228
	3	(2500,2500,0),(0,500,2500)	1.013	0.0218	0.997	0.0204
	4	(0,5000,0),(2500,500,0)	1.022	0.0212	1.022	0.0305
	5	(0,5000,0),(0,3000,0)	1.002	0.0218	1.007	0.0263
	6	(0,5000,0),(0,500,2500)	1.014	0.0223	1.002	0.0227
	7	(0,2500,2500),(2500,500,0)	0.959	0.0229	0.965	0.0326
	8	(0,2500,2500),(0,3000,0)	0.975	0.0234	0.989	0.0298
	9	(0,2500,2500),(0,500,2500)	0.976	0.0239	0.979	0.0261

Appendix B

Tables for Chapter 4

B.1 Generalized case-cohort design based on the event indicator of the first gap time

Table B.1: Coefficient estimates and their estimated standard errors under generalized case-cohort designs based on the first event indicator

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(0, 0, 0.5)$	1	(1000,9000)	0.019	0.0402	-0.007	0.0451
	2	(2000,8000)	0.062	0.0329	0.080	0.0356
	3	(3000,7000)	-0.013	0.0309	-0.030	0.0333
	4	(4000,6000)	-0.058	0.0300	-0.057	0.0320
	5	(5000,5000)	-0.011	0.0289	-0.017	0.0299
	6	(6000,4000)	0.014	0.0293	0.022	0.0302
	7	(7000,3000)	0.006	0.0312	-0.000	0.0319
	8	(8000,2000)	-0.045	0.0341	-0.042	0.0346
	9	(9000,1000)	-0.018	0.0403	-0.012	0.0405

Continued on next page

Table B.1 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	10	(10000,0)	-0.002	0.0588	-0.009	0.0588
$(0, 1, 0.5)$	1	(1000,9000)	-0.021	0.0174	0.973	0.0181
	2	(2000,8000)	-0.007	0.0180	0.987	0.0187
	3	(3000,7000)	-0.020	0.0186	0.974	0.0193
	4	(4000,6000)	-0.016	0.0194	0.978	0.0201
	5	(5000,5000)	-0.028	0.0205	0.967	0.0211
	6	(6000,4000)	-0.006	0.0218	0.988	0.0224
	7	(7000,3000)	-0.022	0.0235	0.973	0.0240
	8	(8000,2000)	-0.020	0.0257	0.974	0.0261
	9	(9000,1000)	-0.030	0.0292	0.965	0.0295
	10	(10000,0)	-0.024	0.0354	0.971	0.0355
$(1, 0, 0.5)$	1	(1000,9000)	0.979	0.0165	-0.021	0.0163
	2	(2000,8000)	0.953	0.0170	-0.044	0.0167
	3	(3000,7000)	0.980	0.0176	-0.021	0.0172
	4	(4000,6000)	0.981	0.0182	-0.019	0.0177
	5	(5000,5000)	0.970	0.0190	-0.028	0.0184
	6	(6000,4000)	0.990	0.0198	-0.010	0.0190
	7	(7000,3000)	0.944	0.0212	-0.052	0.0202
	8	(8000,2000)	0.987	0.0222	-0.013	0.0210
	9	(9000,1000)	0.962	0.0240	-0.036	0.0226
	10	(10000,0)	0.959	0.0265	-0.037	0.0247
$(1, 1, 0.5)$	1	(1000,9000)	1.000	0.0282	1.013	0.0310
	2	(2000,8000)	0.982	0.0253	0.997	0.0271
	3	(3000,7000)	0.979	0.0244	0.969	0.0261
	4	(4000,6000)	1.005	0.0235	1.012	0.0247

Continued on next page

Table B.1 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	5	(5000,5000)	0.972	0.0238	0.979	0.0248
	6	(6000,4000)	0.982	0.0242	0.985	0.0249
	7	(7000,3000)	0.937	0.0250	0.950	0.0256
	8	(8000,2000)	0.982	0.0258	0.976	0.0263
	9	(9000,1000)	0.994	0.0269	0.995	0.0271
	10	(10000,0)	0.976	0.0294	0.981	0.0294
$(0, 0, 1.0)$	1	(1000,9000)	0.038	0.0340	0.054	0.0396
	2	(2000,8000)	0.003	0.0309	-0.045	0.0339
	3	(3000,7000)	0.020	0.0297	0.012	0.0315
	4	(4000,6000)	0.018	0.0287	0.046	0.0303
	5	(5000,5000)	-0.022	0.0287	0.025	0.0299
	6	(6000,4000)	-0.050	0.0292	-0.005	0.0301
	7	(7000,3000)	-0.058	0.0307	-0.007	0.0317
	8	(8000,2000)	0.005	0.0328	-0.012	0.0334
	9	(9000,1000)	-0.031	0.0375	-0.009	0.0377
	10	(10000,0)	-0.041	0.0476	-0.010	0.0476
$(0, 1, 1.0)$	1	(1000,9000)	-0.014	0.0173	0.978	0.0180
	2	(2000,8000)	-0.011	0.0178	0.982	0.0184
	3	(3000,7000)	-0.001	0.0184	0.992	0.0190
	4	(4000,6000)	0.002	0.0190	0.995	0.0196
	5	(5000,5000)	-0.008	0.0199	0.984	0.0205
	6	(6000,4000)	-0.027	0.0209	0.967	0.0214
	7	(7000,3000)	-0.021	0.0221	0.973	0.0225
	8	(8000,2000)	-0.008	0.0239	0.985	0.0243
	9	(9000,1000)	-0.016	0.0261	0.979	0.0264

Continued on next page

Table B.1 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	10	(10000,0)	-0.009	0.0295	0.985	0.0298
$(1, 0, 1.0)$	1	(1000,9000)	0.995	0.0169	-0.005	0.0163
	2	(2000,8000)	0.973	0.0172	-0.026	0.0166
	3	(3000,7000)	0.980	0.0177	-0.019	0.0170
	4	(4000,6000)	0.976	0.0180	-0.023	0.0173
	5	(5000,5000)	1.004	0.0184	0.004	0.0175
	6	(6000,4000)	0.961	0.0191	-0.036	0.0181
	7	(7000,3000)	0.979	0.0194	-0.018	0.0184
	8	(8000,2000)	0.945	0.0202	-0.049	0.0191
	9	(9000,1000)	0.979	0.0208	-0.020	0.0195
	10	(10000,0)	0.965	0.0217	-0.033	0.0203
$(1, 1, 1.0)$	1	(1000,9000)	1.002	0.0269	1.019	0.0294
	2	(2000,8000)	0.995	0.0250	1.012	0.0269
	3	(3000,7000)	0.991	0.0243	1.001	0.0256
	4	(4000,6000)	0.988	0.0238	0.989	0.0250
	5	(5000,5000)	1.029	0.0233	1.046	0.0240
	6	(6000,4000)	0.971	0.0237	0.981	0.0244
	7	(7000,3000)	0.991	0.0235	0.996	0.0239
	8	(8000,2000)	0.955	0.0238	0.964	0.0243
	9	(9000,1000)	0.979	0.0246	0.981	0.0248
	10	(10000,0)	0.970	0.0252	0.970	0.0253
$(0, 0, 1.5)$	1	(1000,9000)	0.078	0.0309	0.059	0.0360
	2	(2000,8000)	-0.062	0.0307	-0.061	0.0334
	3	(3000,7000)	0.006	0.0292	0.006	0.0311
	4	(4000,6000)	-0.045	0.0289	-0.059	0.0306

Continued on next page

Table B.1 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	(n_{cases}, n_{noncases})	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	5	(5000,5000)	-0.015	0.0291	-0.026	0.0303
	6	(6000,4000)	0.038	0.0289	0.032	0.0298
	7	(7000,3000)	-0.009	0.0299	-0.019	0.0306
	8	(8000,2000)	-0.002	0.0313	-0.005	0.0319
	9	(9000,1000)	-0.055	0.0341	-0.051	0.0344
	10	(10000,0)	-0.005	0.0380	-0.008	0.0382
$(0, 1, 1.5)$	1	(1000,9000)	0.008	0.0174	1.001	0.0180
	2	(2000,8000)	-0.030	0.0180	0.964	0.0185
	3	(3000,7000)	0.001	0.0184	0.995	0.0190
	4	(4000,6000)	-0.018	0.0188	0.976	0.0194
	5	(5000,5000)	-0.001	0.0197	0.992	0.0203
	6	(6000,4000)	0.008	0.0203	1.002	0.0209
	7	(7000,3000)	-0.000	0.0212	0.995	0.0217
	8	(8000,2000)	-0.002	0.0221	0.992	0.0225
	9	(9000,1000)	-0.029	0.0236	0.965	0.0239
	10	(10000,0)	-0.010	0.0253	0.983	0.0256
$(1, 0, 1.5)$	1	(1000,9000)	0.989	0.0178	-0.011	0.0169
	2	(2000,8000)	0.972	0.0179	-0.026	0.0170
	3	(3000,7000)	0.984	0.0182	-0.015	0.0173
	4	(4000,6000)	1.006	0.0183	0.004	0.0173
	5	(5000,5000)	0.977	0.0186	-0.022	0.0176
	6	(6000,4000)	0.973	0.0189	-0.025	0.0178
	7	(7000,3000)	0.983	0.0192	-0.017	0.0180
	8	(8000,2000)	0.977	0.0194	-0.023	0.0182
	9	(9000,1000)	0.967	0.0197	-0.030	0.0185

Continued on next page

Table B.1 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	10	(10000,0)	0.971	0.0201	-0.028	0.0188
$(1, 1, 1.5)$	1	(1000,9000)	1.010	0.0273	1.019	0.0298
	2	(2000,8000)	0.966	0.0261	0.970	0.0279
	3	(3000,7000)	0.982	0.0254	0.986	0.0266
	4	(4000,6000)	1.024	0.0243	1.034	0.0252
	5	(5000,5000)	0.988	0.0239	0.989	0.0246
	6	(6000,4000)	0.985	0.0237	0.987	0.0243
	7	(7000,3000)	0.992	0.0232	1.002	0.0236
	8	(8000,2000)	0.982	0.0232	0.992	0.0235
	9	(9000,1000)	0.971	0.0232	0.973	0.0234
	10	(10000,0)	0.971	0.0231	0.973	0.0233

B.2 Outcome-dependent BSS designs based on the first gap time and its event indicator

Table B.2: Coefficient estimates and their estimated standard errors under outcome-dependent BSS designs based on the first gap time and its event indicator

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases},j} : j = 1, 2, 3)$, $(n_{\text{noncases},j} : j = 1, 2, 3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(0, 0, 0.5)$	SRS in S_{cases} and S_{noncases}	(833,3334,833),(1250,2500,1250)	0.009	0.0291	0.013	0.0301
	1	(5000,0,0),(5000,0,0)	-0.031	0.0351	-0.029	0.0358
	2	(5000,0,0),(0,5000,0)	-0.021	0.0253	-0.019	0.0262
	3	(5000,0,0),(0,0,5000)	0.004	0.0205	0.005	0.0216
	4	(0,5000,0),(5000,0,0)	-0.058	0.0436	-0.071	0.0444
	5	(0,5000,0),(0,5000,0)	-0.032	0.0302	-0.034	0.0313
	6	(0,5000,0),(0,0,5000)	-0.021	0.0237	-0.030	0.0252
	7	(0,0,5000),(5000,0,0)	-0.100	0.0834	-0.109	0.0781
	8	(0,0,5000),(0,5000,0)	-0.023	0.0475	-0.039	0.0466

Continued on next page

Table B.2 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases},j} : j = 1,2,3)$, $(n_{\text{noncases},j} : j = 1,2,3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	9	(0,0,5000),(0,0,5000)	-0.013	0.0338	-0.029	0.0355
$(0, 1, 0.5)$	SRS in S_{cases} and S_{noncases}	(166,668,166),(2250,4500,2250)	-0.023	0.0174	0.971	0.0181
	1	(1000,0,0),(5000,4000,0)	-0.020	0.0205	0.974	0.0210
	2	(1000,0,0),(0,9000,0)	-0.017	0.0177	0.977	0.0184
	3	(1000,0,0),(0,4000,5000)	-0.005	0.0155	0.989	0.0164
	4	(0,1000,0),(5000,4000,0)	-0.033	0.0204	0.961	0.0209
	5	(0,1000,0),(0,9000,0)	-0.029	0.0177	0.965	0.0184
	6	(0,1000,0),(0,4000,5000)	-0.012	0.0155	0.982	0.0164
	7	(0,0,1000),(5000,4000,0)	-0.049	0.0205	0.946	0.0210
	8	(0,0,1000),(0,9000,0)	-0.030	0.0178	0.965	0.0185
	9	(0,0,1000),(0,4000,5000)	-0.011	0.0155	0.983	0.0164
$(1, 0, 0.5)$	SRS in S_{cases} and S_{noncases}	(166,668,166),(2250,4500,2250)	0.978	0.0165	-0.023	0.0163
	1	(1000,0,0),(5000,4000,0)	0.968	0.0184	-0.030	0.0178
	2	(1000,0,0),(0,9000,0)	0.979	0.0162	-0.021	0.0160
	3	(1000,0,0),(0,4000,5000)	0.988	0.0153	-0.013	0.0155
	4	(0,1000,0),(5000,4000,0)	0.960	0.0184	-0.038	0.0178
	5	(0,1000,0),(0,9000,0)	0.980	0.0162	-0.020	0.0160
	6	(0,1000,0),(0,4000,5000)	0.978	0.0153	-0.022	0.0155
	7	(0,0,1000),(5000,4000,0)	0.956	0.0180	-0.042	0.0175
	8	(0,0,1000),(0,9000,0)	0.982	0.0159	-0.018	0.0158
	9	(0,0,1000),(0,4000,5000)	0.989	0.0151	-0.013	0.0153
$(1, 1, 0.5)$	SRS in S_{cases} and S_{noncases}	(666,2668,666),(1500,3000,1500)	1.011	0.0234	1.017	0.0246
	1	(4000,0,0),(5000,1000,0)	0.976	0.0252	0.975	0.0261
	2	(4000,0,0),(0,6000,0)	0.990	0.0204	0.986	0.0215
	3	(4000,0,0),(0,1000,5000)	1.003	0.0187	1.001	0.0198
	4	(0,4000,0),(5000,1000,0)	0.956	0.0305	0.963	0.0315
	5	(0,4000,0),(0,6000,0)	0.961	0.0245	0.962	0.0259
	6	(0,4000,0),(0,1000,5000)	0.976	0.0217	0.984	0.0232
	7	(0,0,4000),(5000,1000,0)	0.970	0.0406	0.979	0.0399
	8	(0,0,4000),(0,6000,0)	0.953	0.0370	0.961	0.0368
	9	(0,0,4000),(0,1000,5000)	0.976	0.0311	0.986	0.0322
$(0, 0, 1.0)$	SRS in S_{cases} and S_{noncases}	(666,2668,666),(1500,3000,1500)	0.006	0.0291	-0.009	0.0307
	1	(4000,0,0),(5000,1000,0)	-0.050	0.0394	-0.054	0.0403
	2	(4000,0,0),(0,6000,0)	-0.018	0.0295	-0.015	0.0305
	3	(4000,0,0),(0,1000,5000)	-0.008	0.0193	-0.012	0.0208
	4	(0,4000,0),(5000,1000,0)	-0.082	0.0495	-0.097	0.0504
	5	(0,4000,0),(0,6000,0)	-0.005	0.0363	-0.004	0.0376
	6	(0,4000,0),(0,1000,5000)	-0.000	0.0221	-0.013	0.0243
	7	(0,0,4000),(5000,1000,0)	0.009	0.0764	0.023	0.0704
	8	(0,0,4000),(0,6000,0)	-0.048	0.0678	-0.012	0.0634
	9	(0,0,4000),(0,1000,5000)	0.040	0.0324	0.043	0.0353
$(0, 1, 1.0)$	SRS in S_{cases} and S_{noncases}	(166,668,166),(2250,4500,2250)	-0.000	0.0175	0.992	0.0181
	1	(1000,0,0),(5000,4000,0)	-0.015	0.0236	0.977	0.0239
	2	(1000,0,0),(0,9000,0)	-0.013	0.0216	0.980	0.0219
	3	(1000,0,0),(0,4000,5000)	0.005	0.0144	0.998	0.0152
	4	(0,1000,0),(5000,4000,0)	-0.020	0.0235	0.972	0.0238

Continued on next page

Table B.2 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases},j} : j = 1,2,3)$, $(n_{\text{noncases},j} : j = 1,2,3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	5	(0,1000,0),(0,9000,0)	-0.017	0.0216	0.976	0.0219
	6	(0,1000,0),(0,4000,5000)	-0.001	0.0144	0.991	0.0152
	7	(0,0,1000),(5000,4000,0)	-0.024	0.0229	0.969	0.0231
	8	(0,0,1000),(0,9000,0)	-0.013	0.0210	0.981	0.0213
	9	(0,0,1000),(0,4000,5000)	0.001	0.0143	0.994	0.0151
$(1, 0, 1.0)$	SRS in S_{cases} and S_{noncases}	(166,668,166),(2250,4500,2250)	0.973	0.0169	-0.025	0.0163
	1	(1000,0,0),(5000,4000,0)	0.966	0.0197	-0.031	0.0186
	2	(1000,0,0),(0,9000,0)	0.973	0.0171	-0.026	0.0164
	3	(1000,0,0),(0,4000,5000)	0.996	0.0153	-0.005	0.0152
	4	(0,1000,0),(5000,4000,0)	0.966	0.0197	-0.031	0.0186
	5	(0,1000,0),(0,9000,0)	0.970	0.0171	-0.028	0.0164
	6	(0,1000,0),(0,4000,5000)	0.976	0.0153	-0.023	0.0151
	7	(0,0,1000),(5000,4000,0)	0.971	0.0188	-0.028	0.0179
	8	(0,0,1000),(0,9000,0)	0.975	0.0165	-0.024	0.0160
	9	(0,0,1000),(0,4000,5000)	0.987	0.0149	-0.013	0.0149
$(1, 1, 1.0)$	SRS in S_{cases} and S_{noncases}	(833,3334,833),(1250,2500,1250)	1.000	0.0234	1.003	0.0242
	1	(5000,0,0),(5000,0,0)	0.959	0.0324	0.957	0.0329
	2	(5000,0,0),(0,5000,0)	0.983	0.0224	0.982	0.0232
	3	(5000,0,0),(0,0,5000)	0.997	0.0188	0.995	0.0197
	4	(0,5000,0),(5000,0,0)	0.958	0.0366	0.959	0.0369
	5	(0,5000,0),(0,5000,0)	0.986	0.0259	1.001	0.0265
	6	(0,5000,0),(0,0,5000)	0.991	0.0208	0.993	0.0218
	7	(0,0,5000),(5000,0,0)	0.992	0.0296	0.995	0.0291
	8	(0,0,5000),(0,5000,0)	0.961	0.0336	0.968	0.0321
	9	(0,0,5000),(0,0,5000)	0.970	0.0284	0.975	0.0283
$(0, 0, 1.5)$	SRS in S_{cases} and S_{noncases}	(666,2668,666),(1500,3000,1500)	-0.025	0.0287	-0.033	0.0302
	1	(4000,0,0),(5000,1000,0)	-0.012	0.0436	-0.014	0.0442
	2	(4000,0,0),(0,6000,0)	-0.022	0.0341	-0.026	0.0349
	3	(4000,0,0),(0,1000,5000)	-0.002	0.0185	-0.009	0.0201
	4	(0,4000,0),(5000,1000,0)	-0.010	0.0563	-0.022	0.0570
	5	(0,4000,0),(0,6000,0)	-0.107	0.0435	-0.106	0.0444
	6	(0,4000,0),(0,1000,5000)	0.005	0.0212	0.010	0.0234
	7	(0,0,4000),(5000,1000,0)	0.029	0.0548	0.024	0.0503
	8	(0,0,4000),(0,6000,0)	0.009	0.0625	-0.006	0.0561
	9	(0,0,4000),(0,1000,5000)	0.012	0.0317	0.021	0.0352
$(0, 1, 1.5)$	SRS in S_{cases} and S_{noncases}	(166,668,166),(2250,4500,2250)	0.006	0.0175	0.998	0.0181
	1	(1000,0,0),(5000,4000,0)	-0.011	0.0236	0.982	0.0239
	2	(1000,0,0),(0,9000,0)	-0.022	0.0216	0.971	0.0219
	3	(1000,0,0),(0,4000,5000)	-0.005	0.0144	0.988	0.0152
	4	(0,1000,0),(5000,4000,0)	-0.007	0.0235	0.986	0.0238
	5	(0,1000,0),(0,9000,0)	-0.030	0.0216	0.963	0.0219
	6	(0,1000,0),(0,4000,5000)	-0.006	0.0144	0.987	0.0152
	7	(0,0,1000),(5000,4000,0)	-0.036	0.0229	0.958	0.0231
	8	(0,0,1000),(0,9000,0)	-0.025	0.0210	0.968	0.0213
	9	(0,0,1000),(0,4000,5000)	0.000	0.0143	0.993	0.0151
$(1, 0, 1.5)$	SRS in S_{cases} and S_{noncases}	(166,668,166),(2250,4500,2250)	0.976	0.0177	-0.023	0.0169

Continued on next page

Table B.2 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(n_{\text{cases},j} : j = 1,2,3)$, $(n_{\text{noncases},j} : j = 1,2,3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	1	(1000,0,0),(5000,4000,0)	0.969	0.0210	-0.029	0.0196
	2	(1000,0,0),(0,9000,0)	0.978	0.0188	-0.021	0.0176
	3	(1000,0,0),(0,4000,5000)	1.005	0.0161	0.003	0.0155
	4	(0,1000,0),(5000,4000,0)	0.964	0.0211	-0.033	0.0196
	5	(0,1000,0),(0,9000,0)	0.972	0.0188	-0.026	0.0176
	6	(0,1000,0),(0,4000,5000)	1.006	0.0160	0.002	0.0155
	7	(0,0,1000),(5000,4000,0)	0.974	0.0194	-0.025	0.0183
	8	(0,0,1000),(0,9000,0)	0.981	0.0176	-0.018	0.0167
	9	(0,0,1000),(0,4000,5000)	1.008	0.0154	0.005	0.0150
$(1, 1, 1.5)$	SRS in S_{cases} and S_{noncases}	(1666,6668,1666):(0,0,0)	0.977	0.0230	0.978	0.0232
	1	(3000,5000,2000):(0,0,0)	0.975	0.0217	0.973	0.0220
	2	(4000,5000,1000):(0,0,0)	0.981	0.0245	0.978	0.0247
	3	(5000,5000,0):(0,0,0)	0.949	0.0315	0.949	0.0316
	4	(2000,6000,2000):(0,0,0)	0.945	0.0223	0.951	0.0225
	5	(1000,8000,1000):(0,0,0)	0.993	0.0247	0.996	0.0248
	6	(0,10000,0):(0,0,0)	0.959	0.0303	0.963	0.0301
	7	(2000,5000,3000):(0,0,0)	0.974	0.0205	0.980	0.0207
	8	(1000,5000,4000):(0,0,0)	1.009	0.0198	0.999	0.0201
	9	(0,5000,5000):(0,0,0)	0.987	0.0202	0.990	0.0204

B.3 Generalized case-cohort designs based on the event indicators of the two sequential gap times

Table B.3: Coefficient estimates and their estimated standard errors under generalized case-cohort designs based on the event indicators of the two sequential gap times

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases}}, m_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(0, 0, 0.5)$	1	(500 , 4500)	-0.081	0.0382	-0.097	0.0475
	2	(1000 , 4000)	-0.077	0.0351	-0.089	0.0413
	3	(1500 , 3500)	-0.031	0.0330	-0.048	0.0368
	4	(2000 , 3000)	-0.020	0.0319	-0.035	0.0351
	5	(2500 , 2500)	-0.026	0.0308	-0.039	0.0328
	6	(3000 , 2000)	0.028	0.0294	0.009	0.0312

Continued on next page

Table B.3 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	(m_{cases}, m_{noncases})	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	7	(3500 , 1500)	-0.008	0.0292	0.003	0.0305
	8	(4000 , 1000)	-0.011	0.0287	-0.007	0.0295
	9	(4500 , 500)	0.001	0.0282	-0.003	0.0291
	10	(5000 , 0)	-0.024	0.0280	-0.019	0.0287
$(0, 1, 0.5)$	1	(500 , 500)	-0.016	0.0174	0.978	0.0182
	2	(1000 , 0)	-0.017	0.0174	0.978	0.0181
$(1, 0, 0.5)$	1	(500 , 500)	0.974	0.0165	-0.026	0.0163
	2	(1000 , 0)	0.973	0.0165	-0.027	0.0164
$(1, 1, 0.5)$	1	(500 , 3500)	1.010	0.0276	1.024	0.0323
	2	(1000 , 3000)	1.014	0.0260	1.028	0.0290
	3	(1500 , 2500)	0.969	0.0257	0.981	0.0280
	4	(2000 , 2000)	0.992	0.0247	0.987	0.0266
	5	(2500 , 1500)	0.953	0.0246	0.949	0.0262
	6	(3000 , 1000)	0.971	0.0238	0.969	0.0250
	7	(3500 , 500)	0.966	0.0235	0.963	0.0246
	8	(4000 , 0)	0.996	0.0229	1.000	0.0238
$(0, 0, 1.0)$	1	(500 , 3500)	-0.045	0.0360	-0.055	0.0453
	2	(1000 , 3000)	-0.019	0.0335	-0.040	0.0391
	3	(1500 , 2500)	-0.013	0.0318	0.004	0.0360
	4	(2000 , 2000)	0.028	0.0300	0.010	0.0329
	5	(2500 , 1500)	-0.011	0.0295	-0.011	0.0317
	6	(3000 , 1000)	-0.025	0.0289	-0.026	0.0304
	7	(3500 , 500)	-0.005	0.0281	-0.016	0.0295
	8	(4000 , 0)	-0.016	0.0277	-0.031	0.0285
$(0, 1, 1.0)$	1	(500 , 500)	0.004	0.0172	0.996	0.0179

Continued on next page

Table B.3 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases}},\, m_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	2	(1000 , 0)	0.003	0.0172	0.995	0.0179
$(1, 0, 1.0)$	1	(500 , 500)	0.958	0.0167	-0.038	0.0163
	2	(1000 , 0)	0.959	0.0169	-0.038	0.0164
$(1, 1, 1.0)$	1	(500 , 4500)	0.989	0.0256	0.998	0.0290
	2	(1000 , 4000)	0.977	0.0249	0.978	0.0276
	3	(1500 , 3500)	0.970	0.0243	0.965	0.0265
	4	(2000 , 3000)	0.962	0.0241	0.960	0.0257
	5	(2500 , 2500)	0.969	0.0236	0.970	0.0251
	6	(3000 , 2000)	0.945	0.0236	0.941	0.0248
	7	(3500 , 1500)	0.978	0.0234	0.987	0.0243
	8	(4000 , 1000)	0.936	0.0234	0.944	0.0243
	9	(4500 , 500)	0.945	0.0241	0.948	0.0247
	10	(5000 , 0)	0.949	0.0243	0.940	0.0251
$(0, 0, 1.5)$	1	(500 , 3500)	-0.017	0.0338	-0.028	0.0425
	2	(1000 , 3000)	-0.006	0.0319	0.002	0.0374
	3	(1500 , 2500)	-0.016	0.0304	-0.025	0.0342
	4	(2000 , 2000)	-0.016	0.0297	-0.024	0.0323
	5	(2500 , 1500)	-0.002	0.0290	-0.007	0.0308
	6	(3000 , 1000)	0.001	0.0285	0.010	0.0298
	7	(3500 , 500)	0.016	0.0279	0.013	0.0290
	8	(4000 , 0)	-0.013	0.0277	-0.020	0.0285
$(0, 1, 1.5)$	1	(500 , 500)	-0.016	0.0173	0.977	0.0179
	2	(1000 , 0)	-0.012	0.0174	0.981	0.0180
$(1, 0, 1.5)$	1	(500 , 500)	0.979	0.0176	-0.020	0.0168

Continued on next page

Table B.3 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases}}, m_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	2	(1000 , 0)	0.983	0.0179	-0.017	0.0170
$(1, 1, 1.5)$	7	(3500 , 6500)	1.006	0.0195	1.015	0.0202
	8	(4000 , 6000)	1.002	0.0196	1.006	0.0202
	9	(4500 , 5500)	0.990	0.0199	0.994	0.0204
	10	(5000 , 5000)	0.980	0.0202	0.973	0.0207
	11	(5500 , 4500)	0.974	0.0205	0.977	0.0208
	12	(6000 , 4000)	0.994	0.0207	0.999	0.0210
	13	(6500 , 3500)	0.997	0.0211	0.994	0.0213
	14	(7000 , 3000)	1.007	0.0215	1.010	0.0217
	15	(7500 , 2500)	0.993	0.0226	0.994	0.0227
	16	(8000 , 2000)	0.972	0.0234	0.973	0.0236
	17	(8500 , 1500)	0.992	0.0240	0.990	0.0243
	18	(9000 , 1000)	0.965	0.0259	0.968	0.0260
	19	(9500 , 500)	0.970	0.0279	0.965	0.0280
	20	(10000 , 0)	0.990	0.0299	0.991	0.0301

B.4 Outcome-dependent BSS designs based on the two sequential gap times and their event indicators

Table B.4: Coefficient estimates and their estimated standard errors under outcome-dependent BSS designs based on the two sequential gap times and their event indicators

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases},j} : j = 1, 2, 3)$, $(m_{\text{noncases},j} : j = 1, 2, 3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$(0, 0, 0.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(570,3860,570),(0,0,0)	0.003	0.0217	0.010	0.0236
	1	(1500,2500,1000),(0,0,0)	0.014	0.0211	0.020	0.0230
	2	(2000,2500,500),(0,0,0)	-0.005	0.0209	-0.008	0.0224
	3	(2500,2500,0),(0,0,0)	0.006	0.0204	0.006	0.0216
	4	(500,4000,500),(0,0,0)	-0.018	0.0219	-0.012	0.0236
	5	(250,4500,250),(0,0,0)	-0.003	0.0217	-0.009	0.0233
	6	(0,5000,0),(0,0,0)	-0.001	0.0217	0.001	0.0232
	7	(1000,2500,1500),(0,0,0)	0.009	0.0219	0.016	0.0243
	8	(500,2500,2000),(0,0,0)	0.004	0.0227	0.008	0.0258
	9	(0,2500,2500),(0,0,0)	-0.009	0.0238	-0.008	0.0273
$(0, 1, 0.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(104,792,104),(0,0,0)	-0.012	0.0155	0.982	0.0164
	1	(600,200,200),(0,0,0)	-0.010	0.0155	0.985	0.0164
	2	(800,100,100),(0,0,0)	-0.008	0.0155	0.986	0.0164
	3	(1000,0,0),(0,0,0)	-0.004	0.0155	0.990	0.0164
	4	(100,800,100),(0,0,0)	-0.016	0.0155	0.978	0.0164
	5	(50,900,50),(0,0,0)	-0.012	0.0155	0.982	0.0164
	6	(0,1000,0),(0,0,0)	-0.010	0.0155	0.984	0.0164
	7	(200,200,600),(0,0,0)	-0.013	0.0155	0.981	0.0164
	8	(100,100,800),(0,0,0)	-0.005	0.0155	0.989	0.0164
	9	(0,0,1000),(0,0,0)	-0.012	0.0155	0.982	0.0164
$(1, 0, 0.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(64,372,64),(118,264,118)	0.986	0.0152	-0.015	0.0154
	1	(500,0,0),(500,0,0)	0.991	0.0153	-0.010	0.0155
	2	(500,0,0),(0,500,0)	0.986	0.0152	-0.015	0.0154
	3	(500,0,0),(0,0,500)	0.992	0.0151	-0.010	0.0154
	4	(0,500,0),(500,0,0)	0.982	0.0152	-0.019	0.0154
	5	(0,500,0),(0,500,0)	0.982	0.0152	-0.019	0.0154
	6	(0,500,0),(0,0,500)	0.972	0.0153	-0.028	0.0154
	7	(0,0,500),(500,0,0)	0.977	0.0152	-0.023	0.0154
	8	(0,0,500),(0,500,0)	0.999	0.0152	-0.004	0.0154
	9	(0,0,500),(0,0,500)	0.987	0.0153	-0.015	0.0155
$(1, 1, 0.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(447,3106,447),(0,0,0)	1.001	0.0194	1.002	0.0213
	1	(1500,1500,1000),(0,0,0)	0.994	0.0194	0.990	0.0208
	2	(2000,1500,500),(0,0,0)	1.005	0.0190	1.001	0.0202
	3	(2500,1500,0),(0,0,0)	1.004	0.0187	0.999	0.0196
	4	(500,3000,500),(0,0,0)	0.987	0.0195	0.990	0.0214
	5	(250,3500,250),(0,0,0)	1.005	0.0195	1.010	0.0212
	6	(0,4000,0),(0,0,0)	0.988	0.0196	0.990	0.0213
	7	(1000,1500,1500),(0,0,0)	1.009	0.0198	1.017	0.0217
	8	(500,1500,2000),(0,0,0)	0.997	0.0207	0.993	0.0236
	9	(0,1500,2500),(0,0,0)	1.008	0.0217	1.012	0.0252
$(0, 0, 1.0)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(444,3112,444),(0,0,0)	-0.005	0.0203	-0.001	0.0228
	1	(1500,1500,1000),(0,0,0)	0.006	0.0201	0.013	0.0225
	2	(2000,1500,500),(0,0,0)	0.019	0.0194	0.012	0.0210

Continued on next page

Table B.4 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases},j} : j = 1,2,3)$, $(m_{\text{noncases},j} : j = 1,2,3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	3	(2500,1500,0),(0,0,0)	0.009	0.0190	0.005	0.0202
	4	(500,4000,500),(0,0,0)	-0.002	0.0192	-0.003	0.0211
	5	(250,4000,250),(0,0,0)	-0.015	0.0197	-0.024	0.0218
	6	(0,4000,0),(0,0,0)	0.014	0.0199	0.013	0.0222
	7	(1000,1500,1500),(0,0,0)	0.006	0.0209	0.005	0.0240
	8	(500,1500,2000),(0,0,0)	-0.018	0.0223	-0.015	0.0265
	9	(0,1500,2500),(0,0,0)	0.002	0.0234	0.013	0.0283
$(0, 1, 1.0)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(52,396,52),(202,96,202)	0.002	0.0144	0.996	0.0152
	1	(500,0,0),(500,0,0)	-0.006	0.0144	0.987	0.0152
	2	(500,0,0),(0,500,0)	-0.002	0.0143	0.991	0.0151
	3	(500,0,0),(0,0,500)	-0.006	0.0143	0.987	0.0151
	4	(0,500,0),(500,0,0)	-0.006	0.0143	0.987	0.0151
	5	(0,500,0),(0,500,0)	-0.003	0.0143	0.990	0.0151
	6	(0,500,0),(0,0,500)	-0.003	0.0144	0.990	0.0152
	7	(0,0,500),(500,0,0)	-0.005	0.0143	0.988	0.0151
	8	(0,0,500),(0,500,0)	-0.003	0.0143	0.990	0.0151
	9	(0,0,500),(0,0,500)	0.000	0.0144	0.994	0.0152
$(1, 0, 1.0)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(58,384,58),(148,204,148)	0.981	0.0150	-0.018	0.0149
	1	(500,0,0),(500,0,0)	0.990	0.0151	-0.010	0.0150
	2	(500,0,0),(0,500,0)	0.979	0.0150	-0.020	0.0150
	3	(500,0,0),(0,0,500)	0.986	0.0150	-0.014	0.0149
	4	(0,500,0),(500,0,0)	0.982	0.0150	-0.018	0.0150
	5	(0,500,0),(0,500,0)	0.991	0.0150	-0.009	0.0150
	6	(0,500,0),(0,0,500)	0.995	0.0150	-0.006	0.0150
	7	(0,0,500),(500,0,0)	0.983	0.0151	-0.017	0.0150
	8	(0,0,500),(0,500,0)	0.982	0.0150	-0.018	0.0150
	9	(0,0,500),(0,0,500)	0.980	0.0150	-0.020	0.0149
$(1, 1, 1.0)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(367,2766,367),(609,282,609)	0.995	0.0193	0.993	0.0212
	1	(2500,1000,0),(1500,0,0)	0.998	0.0189	0.997	0.0198
	2	(2500,1000,0),(0,1500,0)	1.001	0.0184	1.002	0.0193
	3	(2500,1000,0),(0,0,1500)	1.006	0.0180	1.003	0.0186
	4	(0,3500,0),(1500,0,0)	0.995	0.0195	1.001	0.0214
	5	(0,3500,0),(0,1500,0)	0.996	0.0192	0.999	0.0212
	6	(0,3500,0),(0,0,1500)	0.982	0.0186	0.979	0.0204
	7	(0,1000,2500),(1500,0,0)	1.000	0.0215	1.007	0.0254
	8	(0,1000,2500),(0,1500,0)	0.975	0.0220	0.977	0.0266
	9	(0,1000,2500),(0,0,1500)	0.993	0.0212	0.996	0.0254
$(0, 0, 1.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(458,3084,458),(0,0,0)	-0.001	0.0193	-0.009	0.0218
	1	(1500,1500,1000),(0,0,0)	0.005	0.0193	0.001	0.0215
	2	(2000,1500,500),(0,0,0)	0.010	0.0188	0.011	0.0204
	3	(2500,1500,0),(0,0,0)	0.004	0.0184	-0.001	0.0197
	4	(500,3000,500),(0,0,0)	-0.013	0.0194	-0.012	0.0220
	5	(250,3500,250),(0,0,0)	-0.011	0.0194	-0.011	0.0218
	6	(0,4000,0),(0,0,0)	-0.007	0.0194	-0.007	0.0218
	7	(1000,1500,1500),(0,0,0)	0.001	0.0201	0.014	0.0230
	8	(500,1500,2000),(0,0,0)	-0.009	0.0211	-0.008	0.0250

Continued on next page

Table B.4 – *Continued from previous page*

$(\alpha_{11}, \alpha_{21}, \gamma_1)$	Sampling scenario	$(m_{\text{cases},j} : j = 1,2,3)$, $(m_{\text{noncases},j} : j = 1,2,3)$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	9	(0,1500,2500),(0,0,0)	-0.002	0.0220	0.004	0.0265
$(0, 1, 1.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(54,392,54),(182,136,182)	0.002	0.0144	0.996	0.0152
	1	(500,0,0),(500,0,0)	-0.006	0.0144	0.987	0.0152
	2	(500,0,0),(0,500,0)	-0.002	0.0143	0.991	0.0151
	3	(500,0,0),(0,0,500)	-0.006	0.0143	0.987	0.0151
	4	(0,500,0),(500,0,0)	-0.006	0.0143	0.987	0.0151
	5	(0,500,0),(0,500,0)	-0.003	0.0143	0.990	0.0151
	6	(0,500,0),(0,0,500)	-0.003	0.0144	0.990	0.0152
	7	(0,0,500),(500,0,0)	-0.005	0.0143	0.988	0.0151
	8	(0,0,500),(0,500,0)	-0.003	0.0143	0.990	0.0151
	9	(0,0,500),(0,0,500)	0.000	0.0144	0.994	0.0152
$(1, 0, 1.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(59,382,59),(143,214,143)	1.000	0.0156	-0.001	0.0152
	1	(500,0,0),(500,0,0)	0.999	0.0157	-0.003	0.0153
	2	(500,0,0),(0,500,0)	1.002	0.0157	-0.000	0.0153
	3	(500,0,0),(0,0,500)	0.995	0.0156	-0.007	0.0152
	4	(0,500,0),(500,0,0)	0.998	0.0156	-0.004	0.0152
	5	(0,500,0),(0,500,0)	0.992	0.0156	-0.009	0.0152
	6	(0,500,0),(0,0,500)	1.003	0.0156	0.001	0.0152
	7	(0,0,500),(500,0,0)	1.001	0.0156	-0.001	0.0153
	8	(0,0,500),(0,500,0)	1.004	0.0157	0.001	0.0153
	9	(0,0,500),(0,0,500)	0.993	0.0156	-0.009	0.0152
$(1, 1, 1.5)$	SRS in $S_{\text{cases,cases}}$ and $S_{\text{cases,noncases}}$	(374,2752,374),(2457,1586,2457)	0.974	0.0203	0.981	0.0207
	1	(1500,1000,1000),(2500,1614,2386)	0.987	0.0193	0.990	0.0196
	2	(2000,1000,500),(2433,1614,2443)	0.993	0.0187	0.997	0.0189
	3	(2500,1000,0),(2386,1614,2500)	0.997	0.0183	0.993	0.0184
	4	(500,2500,500),(2500,1614,2386)	0.985	0.0202	0.987	0.0205
	5	(250,3000,250),(2433,1614,2443)	0.992	0.0202	1.001	0.0206
	6	(0,3500,0),(2386,1614,2500)	0.974	0.0204	0.985	0.0208
	7	(1000,1000,1500),(2500,1614,2386)	0.993	0.0198	0.991	0.0205
	8	(500,1000,2000),(2433,1614,2443)	0.987	0.0208	0.989	0.0221
	9	(0,1000,2500),(2386,1614,2500)	0.977	0.0221	0.976	0.0243

Appendix C

Tables and figures for Section 4.1

C.1 Table and figure for model scenario ($\alpha_{11} = 0, \alpha_{21} = 1, \gamma_1 = 0.5$)

Table C.1: Coefficient estimates and their estimated standard errors under generalized case-cohort designs based on the first event indicator for model scenario ($\alpha_{11} = 0, \alpha_{21} = 1, \gamma_1 = 0.5$) when the dependence between time-to-events is changed from moderate to high

Dependence	Sampling scenario	(n_{cases}, n_{noncases})	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$\tau = 0.4$	1	(1000,9000)	0.015	0.0381	0.939	0.0567
	2	(2000,8000)	0.044	0.0309	1.065	0.0430
	3	(3000,7000)	0.003	0.0292	0.957	0.0414
	4	(4000,6000)	-0.027	0.0281	1.008	0.0371
	5	(5000,5000)	-0.023	0.0278	1.000	0.0347
	6	(6000,4000)	0.016	0.0284	1.039	0.0332
	7	(7000,3000)	0.010	0.0302	0.995	0.0336
	8	(8000,2000)	-0.040	0.0332	0.963	0.0341
	9	(9000,1000)	-0.048	0.0391	0.947	0.0366

Continued on next page

Table C.1 – *Continued from previous page*

Dependence	Sampling scenario	(n_{cases}, n_{noncases})	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	10	(10000,0)	0.040	0.0541	0.988	0.0442
τ = 0.45	1	(1000,9000)	0.034	0.0356	1.056	0.0431
	2	(2000,8000)	0.024	0.0311	1.002	0.0407
	3	(3000,7000)	-0.022	0.0289	1.001	0.0363
	4	(4000,6000)	0.005	0.0278	1.012	0.0339
	5	(5000,5000)	0.002	0.0273	1.019	0.0322
	6	(6000,4000)	-0.021	0.0283	0.975	0.0327
	7	(7000,3000)	-0.001	0.0300	0.988	0.0326
	8	(8000,2000)	0.005	0.0333	1.002	0.0340
	9	(9000,1000)	-0.009	0.0395	0.991	0.0377
	10	(10000,0)	0.040	0.0542	1.033	0.0474
τ = 0.5	1	(1000,9000)	0.028	0.0373	1.030	0.0373
	2	(2000,8000)	0.025	0.0353	1.001	0.0353
	3	(3000,7000)	-0.022	0.0325	0.988	0.0325
	4	(4000,6000)	0.004	0.0310	1.002	0.0310
	5	(5000,5000)	-0.004	0.0301	1.000	0.0301
	6	(6000,4000)	-0.022	0.0311	0.970	0.0311
	7	(7000,3000)	-0.005	0.0316	0.984	0.0316
	8	(8000,2000)	-0.003	0.0338	0.991	0.0338
	9	(9000,1000)	-0.019	0.0385	0.978	0.0385
	10	(10000,0)	0.031	0.0506	1.023	0.0506
τ = 0.55	1	(1000,9000)	0.017	0.0318	1.012	0.0320
	2	(2000,8000)	0.023	0.0287	1.003	0.0305
	3	(3000,7000)	-0.021	0.0271	0.983	0.0289
	4	(4000,6000)	0.002	0.0266	0.997	0.0284

Continued on next page

Table C.1 – *Continued from previous page*

Dependence	Sampling scenario	$(n_{\text{cases}},\, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	5	(5000,5000)	-0.009	0.0264	0.989	0.0281
	6	(6000,4000)	-0.021	0.0277	0.972	0.0293
	7	(7000,3000)	-0.006	0.0295	0.984	0.0305
	8	(8000,2000)	-0.009	0.0329	0.985	0.0332
	9	(9000,1000)	-0.022	0.0393	0.975	0.0387
	10	(10000,0)	0.037	0.0550	1.030	0.0523
$\tau = 0.6$	1	(1000,9000)	0.007	0.0280	0.995	0.0273
	2	(2000,8000)	0.021	0.0262	1.004	0.0265
	3	(3000,7000)	-0.020	0.0254	0.976	0.0260
	4	(4000,6000)	0.001	0.0253	0.992	0.0261
	5	(5000,5000)	-0.016	0.0255	0.977	0.0263
	6	(6000,4000)	-0.022	0.0268	0.970	0.0277
	7	(7000,3000)	-0.007	0.0288	0.983	0.0293
	8	(8000,2000)	-0.016	0.0321	0.977	0.0324
	9	(9000,1000)	-0.026	0.0385	0.969	0.0382
	10	(10000,0)	0.019	0.0539	1.012	0.0525
$\tau = 0.65$	1	(1000,9000)	0.000	0.0237	0.986	0.0234
	2	(2000,8000)	0.016	0.0232	1.001	0.0233
	3	(3000,7000)	-0.018	0.0232	0.973	0.0235
	4	(4000,6000)	0.000	0.0236	0.989	0.0241
	5	(5000,5000)	-0.020	0.0241	0.970	0.0246
	6	(6000,4000)	-0.023	0.0256	0.969	0.0261
	7	(7000,3000)	-0.006	0.0276	0.983	0.0280
	8	(8000,2000)	-0.024	0.0308	0.969	0.0310
	9	(9000,1000)	-0.038	0.0366	0.956	0.0366

Continued on next page

Table C.1 – *Continued from previous page*

Dependence	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	10	(10000,0)	-0.007	0.0501	0.985	0.0495
τ = 0.7	1	(1000,9000)	-0.003	0.0204	0.984	0.0206
	2	(2000,8000)	0.011	0.0206	0.999	0.0210
	3	(3000,7000)	-0.017	0.0211	0.973	0.0215
	4	(4000,6000)	0.001	0.0218	0.991	0.0223
	5	(5000,5000)	-0.019	0.0227	0.971	0.0231
	6	(6000,4000)	-0.023	0.0241	0.970	0.0246
	7	(7000,3000)	-0.001	0.0261	0.989	0.0265
	8	(8000,2000)	-0.027	0.0291	0.967	0.0294
	9	(9000,1000)	-0.045	0.0341	0.949	0.0342
	10	(10000,0)	-0.024	0.0445	0.968	0.0443
τ = 0.75	1	(1000,9000)	-0.007	0.0185	0.983	0.0190
	2	(2000,8000)	0.004	0.0190	0.995	0.0196
	3	(3000,7000)	-0.019	0.0196	0.973	0.0202
	4	(4000,6000)	0.001	0.0204	0.993	0.0210
	5	(5000,5000)	-0.017	0.0213	0.975	0.0219
	6	(6000,4000)	-0.025	0.0227	0.969	0.0232
	7	(7000,3000)	0.005	0.0245	0.996	0.0250
	8	(8000,2000)	-0.025	0.0273	0.968	0.0276
	9	(9000,1000)	-0.043	0.0314	0.951	0.0316
	10	(10000,0)	-0.031	0.0393	0.963	0.0393
τ = 0.8	1	(1000,9000)	-0.021	0.0174	0.973	0.0181
	2	(2000,8000)	-0.007	0.0180	0.987	0.0187
	3	(3000,7000)	-0.020	0.0186	0.974	0.0193
	4	(4000,6000)	-0.016	0.0194	0.978	0.0201

Continued on next page

Table C.1 – *Continued from previous page*

Dependence	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	5	(5000,5000)	-0.028	0.0205	0.967	0.0211
	6	(6000,4000)	-0.006	0.0218	0.988	0.0224
	7	(7000,3000)	-0.022	0.0235	0.973	0.0240
	8	(8000,2000)	-0.020	0.0257	0.974	0.0261
	9	(9000,1000)	-0.030	0.0292	0.965	0.0295
	10	(10000,0)	-0.024	0.0354	0.971	0.0355

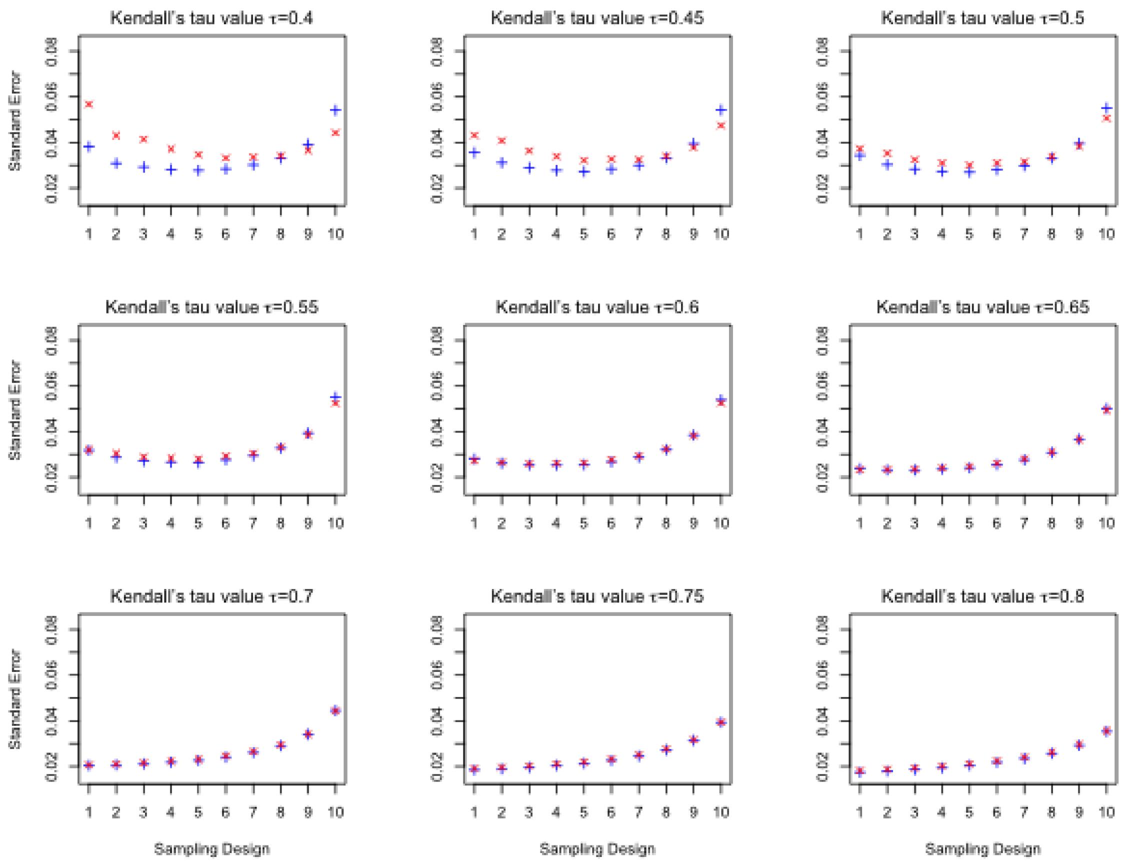

Figure C.1: Estimated standard errors of the coefficient estimates of the expensive covariate under generalized case-cohort designs based on the first event indicator for model scenario ($\alpha_{11} = 0, \alpha_{21} = 1, \gamma_1 = 0.5$) when the dependence between time-to-events is changed from moderate to high

$+$ represents standard error of $\hat{\alpha}_{11}$

$\times$ represents standard error of $\hat{\alpha}_{21}$

The sampling scenarios $1, ..., 10$ are described in Table C.1

C.2 Table and figure for model scenario ($\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 1.5$)

Table C.2: Coefficient estimates and their estimated standard errors under generalized case-cohort designs based on the first event indicator for model scenario ($\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 1.5$) when the dependence between time-to-events is changed from moderate to high

Dependence	Sampling scenario	(n_{cases}, n_{noncases})	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
$\tau = 0.4$	1	(1000,9000)	1.007	0.0290	0.044	0.0467
	2	(2000,8000)	0.97	0.0272	-0.001	0.0404
	3	(3000,7000)	0.993	0.0265	0.038	0.0367
	4	(4000,6000)	1.029	0.0256	0.062	0.0334
	5	(5000,5000)	0.983	0.0252	-0.012	0.0313
	6	(6000,4000)	0.981	0.0249	0.002	0.0296
	7	(7000,3000)	1.005	0.0245	0.028	0.0282
	8	(8000,2000)	0.978	0.0245	-0.012	0.0273
	9	(9000,1000)	1.003	0.0245	-0.003	0.0265
	10	(10000,0)	0.964	0.0251	-0.013	0.0261
$\tau = 0.45$	1	(1000,9000)	1.005	0.0282	0.032	0.0426
	2	(2000,8000)	0.985	0.0272	0.042	0.0384
	3	(3000,7000)	0.997	0.0259	0.034	0.0344
	4	(4000,6000)	0.983	0.0253	-0.019	0.0314
	5	(5000,5000)	1.018	0.0245	0.038	0.0294
	6	(6000,4000)	0.985	0.0243	-0.010	0.0283
	7	(7000,3000)	0.978	0.0243	0.021	0.0276
	8	(8000,2000)	0.984	0.0243	0.007	0.0266
	9	(9000,1000)	0.973	0.0243	-0.007	0.0258
	10	(10000,0)	0.990	0.0245	-0.016	0.0255

Continued on next page

Table C.2 – *Continued from previous page*

Dependence	Sampling scenario	(n_{cases}, n_{noncases})	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
τ = 0.5	1	(1000,9000)	1.003	0.0273	0.031	0.0378
	2	(2000,8000)	0.985	0.0265	0.033	0.0349
	3	(3000,7000)	0.998	0.0252	0.025	0.0316
	4	(4000,6000)	0.982	0.0246	-0.019	0.0294
	5	(5000,5000)	1.016	0.0240	0.035	0.0278
	6	(6000,4000)	0.983	0.0238	-0.011	0.0270
	7	(7000,3000)	0.979	0.0238	0.013	0.0265
	8	(8000,2000)	0.984	0.0239	0.004	0.0258
	9	(9000,1000)	0.972	0.0239	-0.011	0.0251
	10	(10000,0)	0.986	0.0242	-0.021	0.0249
τ = 0.55	1	(1000,9000)	1.003	0.0260	0.017	0.0326
	2	(2000,8000)	0.987	0.0253	0.013	0.0309
	3	(3000,7000)	1.000	0.0243	0.017	0.0287
	4	(4000,6000)	0.987	0.0238	-0.015	0.0271
	5	(5000,5000)	1.018	0.0233	0.030	0.0260
	6	(6000,4000)	0.982	0.0232	-0.019	0.0254
	7	(7000,3000)	0.980	0.0233	-0.000	0.0252
	8	(8000,2000)	0.985	0.0233	-0.002	0.0247
	9	(9000,1000)	0.973	0.0234	-0.018	0.0242
	10	(10000,0)	0.989	0.0236	-0.017	0.0241
τ = 0.6	1	(1000,9000)	0.995	0.0240	-0.004	0.0278
	2	(2000,8000)	0.984	0.0237	-0.006	0.0270
	3	(3000,7000)	0.997	0.0230	0.001	0.0257
	4	(4000,6000)	0.980	0.0227	-0.028	0.0248
	5	(5000,5000)	1.013	0.0224	0.017	0.0241

Continued on next page

Table C.2 – *Continued from previous page*

Dependence	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	6	(6000,4000)	0.977	0.0224	-0.029	0.0237
	7	(7000,3000)	0.976	0.0225	-0.018	0.0236
	8	(8000,2000)	0.981	0.0226	-0.013	0.0234
	9	(9000,1000)	0.968	0.0228	-0.029	0.0232
	10	(10000,0)	0.985	0.0230	-0.022	0.0231
τ = 0.65	1	(1000,9000)	0.990	0.0218	-0.014	0.0236
	2	(2000,8000)	0.984	0.0219	-0.015	0.0235
	3	(3000,7000)	0.994	0.0215	-0.009	0.0228
	4	(4000,6000)	0.982	0.0215	-0.024	0.0224
	5	(5000,5000)	1.008	0.0214	0.008	0.0221
	6	(6000,4000)	0.976	0.0215	-0.029	0.0220
	7	(7000,3000)	0.976	0.0217	-0.024	0.0221
	8	(8000,2000)	0.979	0.0219	-0.019	0.0220
	9	(9000,1000)	0.966	0.0221	-0.034	0.0220
	10	(10000,0)	0.981	0.0223	-0.025	0.0221
τ = 0.7	1	(1000,9000)	0.987	0.0200	-0.018	0.0204
	2	(2000,8000)	0.982	0.0202	-0.019	0.0205
	3	(3000,7000)	0.992	0.0201	-0.011	0.0203
	4	(4000,6000)	0.980	0.0203	-0.026	0.0203
	5	(5000,5000)	1.001	0.0204	-0.002	0.0203
	6	(6000,4000)	0.974	0.0206	-0.028	0.0204
	7	(7000,3000)	0.976	0.0207	-0.025	0.0205
	8	(8000,2000)	0.977	0.0210	-0.023	0.0206
	9	(9000,1000)	0.959	0.0214	-0.041	0.0208

Continued on next page

Table C.2 – *Continued from previous page*

Dependence	Sampling scenario	$(n_{\text{cases}}, n_{\text{noncases}})$	$\hat{\alpha}_{11}$	$\widehat{\text{SE}}(\hat{\alpha}_{11})$	$\hat{\alpha}_{21}$	$\widehat{\text{SE}}(\hat{\alpha}_{21})$
	10	(10000,0)	0.975	0.0216	-0.029	0.0209
τ = 0.75	1	(1000,9000)	0.983	0.0186	-0.020	0.0182
	2	(2000,8000)	0.977	0.0189	-0.024	0.0184
	3	(3000,7000)	0.990	0.0189	-0.013	0.0184
	4	(4000,6000)	0.977	0.0192	-0.026	0.0186
	5	(5000,5000)	0.994	0.0194	-0.009	0.0186
	6	(6000,4000)	0.969	0.0197	-0.032	0.0189
	7	(7000,3000)	0.973	0.0198	-0.028	0.0190
	8	(8000,2000)	0.971	0.0202	-0.028	0.0193
	9	(9000,1000)	0.953	0.0206	-0.047	0.0196
	10	(10000,0)	0.971	0.0208	-0.030	0.0198
τ = 0.8	1	(1000,9000)	0.989	0.0178	-0.011	0.0169
	2	(2000,8000)	0.972	0.0179	-0.026	0.0170
	3	(3000,7000)	0.984	0.0182	-0.015	0.0173
	4	(4000,6000)	1.006	0.0183	0.004	0.0173
	5	(5000,5000)	0.977	0.0186	-0.022	0.0176
	6	(6000,4000)	0.973	0.0189	-0.025	0.0178
	7	(7000,3000)	0.983	0.0192	-0.017	0.0180
	8	(8000,2000)	0.977	0.0194	-0.023	0.0182
	9	(9000,1000)	0.967	0.0197	-0.030	0.0185
	10	(10000,0)	0.971	0.0201	-0.028	0.0188

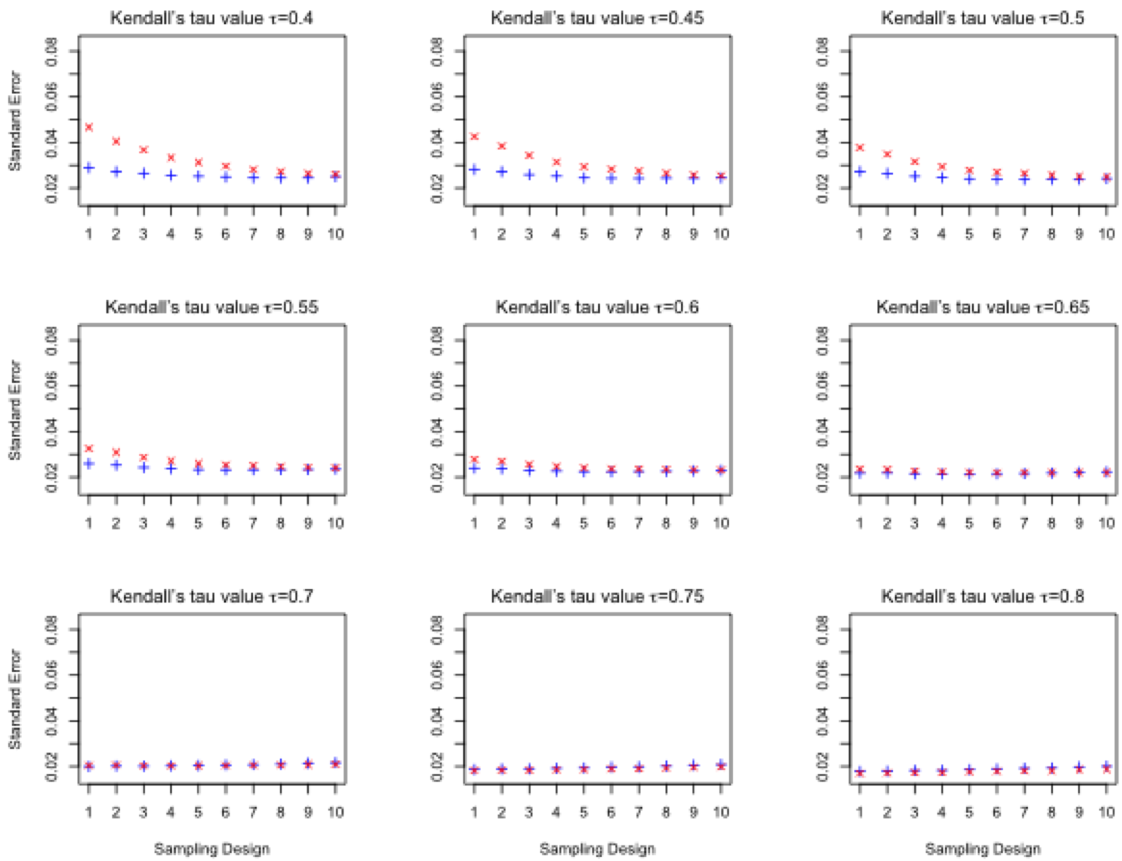

Figure C.2: Estimated standard errors of the coefficient estimates of the expensive covariate under generalized case-cohort designs based on the first event indicator for model scenario ($\alpha_{11} = 1, \alpha_{21} = 0, \gamma_1 = 1.5$) when the dependence between time-to-events is changed from moderate to high

$+$ represents standard error of $\hat{\alpha}_{11}$

$\times$ represents standard error of $\hat{\alpha}_{21}$

The sampling scenarios $1, ..., 10$ are described in Table C.2