Two-Phase Response-Dependent Sampling Designs For Time-to-Event Analysis

by

© **Patricia Emily Judd**

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

September 2016

St. John's, Newfoundland and Labrador, Canada

Abstract

Measuring expensive covariates for all subjects within a cohort may not be a feasible option due to a study′s budgetary or logistical constraints. As a result of such limitations, we need to consider sampling designs that account for subjects that have missing data. To design a study allowing incomplete covariate data for some subjects, it is better to employ a cost-efficient sampling design, which balances the efficiency of parameter estimates and power of association tests with the sample size. Response-dependent sampling is a cost-efficient sampling design. In this design a subset of subjects is selected from a cohort, based on the response variable (and inexpensive covariates), which has already been gathered for all subjects in the cohort. In our study, we focus on response-dependent two-phase sampling designs. During phase I of the sampling design, all members in a cohort are measured for the response variable and the inexpensive covariates. In phase II, a subset of the cohort is selected, based on the response variable obtained in phase I, and the expensive covariate(s) are measured only for those selected. In our study, the response variable that determines which individuals are selected for phase II is a continuous time-to-event variable; wherein this type of the response variable maybe subject to censoring. The most common response-dependent sampling design for time-to-event data is the case-cohort sampling design. We explore variations of the case-cohort design which give more efficient association estimates for a given sample size. We stratify cases and non-cases based on the observed time-to-event values and apply basic stratified sampling. Modifying the proportion of observations selected from each strata changes the efficiency of association estimates.

To the memory of my dad, Paul Douglas Judd.

Acknowledgements

I would like to express my greatest appreciation to my supervisor Dr. Yildiz Yilmaz, for giving me the opportunity to work with her over the past few years. Her support, encouragement, and guidance have been invaluable throughout my research.

I express my deep appreciation for the financial support I received from Memorial University of Newfoundland's School of Graduate Studies, the Department of Mathematics & Statistics, and my supervisor in the form of teaching assistantships and graduate fellowships.

To my partner, Matthew Baker, I thank for all of the unconditional love and support during my studies.

Finally, I want to extend my heartfelt gratitude to my parents; Camille & Ron Sutton, Helen McIntosh & Paul Judd, and to my sister, Sarah Iacobaccio for lifelong encouragement and love.

Statement of contribution

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

Table of contents

List of tables

List of figures

List of symbols

$N(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
δ_i	Censoring indicator for individual i
C_i	Censoring time for individual i
$C_{(i)}$	i^{th} ordered censoring time
L_C	Lower Cut-off for Cases
L_{CN}	Lower Cut-off for Non-cases
N	Cohort size
N_{cases}	Number of cases within cohort
N_{cases_j}	Number of cases within case stratum j $(j = 1, 2, 3)$
n	Sample size
n_{cases}	Number of cases initially selected in phase II
n_{cases_j}	Number of cases initially sampled from case stratum j $(j = 1, 2, 3)$ in phase II
n^*_{cases}	Total number of cases sampled in phase II
$n^*_{cases_j}$	Total number of cases sampled from case stratum j $(j = 1, 2, 3)$ in phase II
n_{cohort}	Number sampled from cohort in phase II
$N_{noncases}$	Number of non-cases within cohort
$N_{noncases_h}$	Number of non-cases within stratum h $(h = 1, 2, 3)$
$n_{noncases}$	Total number of non-cases sampled in phase II
$n_{noncases_h}$	Total number of non-cases sampled from non-case stratum h $(h = 1, 2, 3)$ in phase II
$p(x)$	Probability of being cured given covariate(s) x
R_i	Phase II sampling indicator for individual i
T_i	Event time for individual i
$T_{(i)}$	i^{th} ordered event time
t_i	Observed time-to-event for individual i
U_C	Upper Cut-off for Cases
U_{CN}	Upper Cut-off for Non-cases

List of abbreviations

AFT	Accelerated Failure Time
BSS	Basic Stratified Sampling
CDF	Cumulative Distribution Function
EM	Expectation-Maximization
HT	Horvitz-Thompson
IPW	Inverse Probability Weighted
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
pdf	probability density function
PH	Proportional Hazards
SRS	Simple Random Sampling

Chapter 1

Introduction

Measuring expensive covariates for all subjects within a cohort may not be a feasible option due to a study′s budgetary or logistical constraints. As a result of such limitations, we need to consider sampling designs that account for subjects that have missing data. To design a study allowing incomplete covariate data for some subjects, it is better to employ a cost-efficient sampling design, which balances the efficiency of parameter estimates and power of association tests with the sample size. The objective is to explore designs that lead to efficient association estimates for a given sample size. Response-dependent sampling is a cost-efficient sampling design, in which a subset of subjects is selected from a cohort, based on the response variable, which has already been gathered for all subjects in the cohort. In our study, we focus on two phase response-dependent sampling designs. During the first phase of the sampling design, all members in a cohort are measured for the response variable and the inexpensive covariates. In phase II, a subset of the cohort is selected, based on the response variable obtained in phase I, and the expensive covariate(s) is measured only for the selected subjects (Neyman, 1938; Zhao and Lipsitz, 1992). In our study, the response variable that determines which individuals will be selected for phase II is a

continuous time-to-event variable; wherein this type of the response variable is subject to censoring. Difficulties arise when dealing with data that is censored. For example, an individual with a right-censored event time may not have an observed event time; instead it may be censored due to a random process or censored at the end of the study follow-up period. A response variable that may be censored complicates the response-dependent sampling and the method of estimation. There are two commonly used response-dependent sampling designs for time-to-event data: case-cohort sampling designs and nested case-control sampling designs. Our focus, the generalized case-cohort sampling, is performed by selecting a random sample from the case set and a random sample from the cohort, the cohort includes both case and non-case observations in their follow-up period (Chen, 2001).

The aim of our study is to explore efficient case-cohort sampling designs when the response is a continuous time-to-event variable that is subject to censoring. Different variations of case-cohort sampling design will be considered under different survival models. A basic introduction to time-to-event data analysis is given in Section 1.1. To begin, the notation will quickly be outlined and basic definitions will be presented. Following this, regression models used in time-to-event data analysis will be summarized. Parametric, non-parametric and semi-parametric estimation methods will briefly be outlined. Lastly in Section 1.1 the mixture-cure model is presented. Two-phase response-dependent sampling designs will be outlined in Section 1.2, specifically when the response variable is a continuous variable which is not subject to censoring. Section 1.3 contains information regarding common sampling designs for time-to-event data, specifically nested case-control designs and case-cohort sampling designs. Estimation methods under two-phase response-dependent sampling, in particular pseudo-likelihood and likelihood-based approaches under the case-cohort sampling design, are described within Section 1.4. Finally in Section 1.5 the objectives of the study and

the outline of the thesis is presented.

1.1 Introduction to Time-To-Event Data Analysis

Time-to-event, a non-negative continuous variable measures the time from a defined origin to an event occurrence, denoted T_i for individual i. There are three components requiring consideration when defining the time-to-event: measurement scale, point of origin, and the event. The goal is for observations to be as comparable as possible, thus the point of origin must be strictly defined for each observation. Observations may have a staggered entry into the study; they do not require the same entry date. Similarly the event of interest must be exactly defined for all individuals within the study. The measurement scale, which will always be non-negative, must be the same for all observations. Often the measurement scale is based on real time; i.e. hours, days, months, etc. As an example consider death due to cancer as the event of interest to be studied. The point of origin for this study will be defined as date of cancer diagnosis, and the measurement scale will be the number of days between cancer diagnosis and death due to cancer.

One of the difficulties that arises when working with survival data is unobserved time-to-event values, this predicament is termed censoring. The most common type of censoring is right censoring and may occur for any number of reasons. Relating to our example above, a few causes of right censoring may include: the individual may not die due to cancer within the follow up time period, a death unrelated to the cancer, the individual is lost in follow up, and the individual drops out of the study. There are multiple forms of right censoring however our focus is on type I censoring and random censoring. In type I censoring, the censoring time (C_i) is fixed for all individuals ($i = 1, .., N$) at the beginning of the study, thus the total

number of observed events is a random variable. Type I censoring occurs because individuals do not experience the event of interest within the follow-up time period. When there is random censoring, the censoring time (C_i) occurs randomly for each individual within the study. In this case the C_i values are independent of the event time, T_i. In contrast to type I censoring, the censoring time values are considered random variables. Random censoring could occur if an individual is lost in follow up, individual drops out of the study, etc. Often right censored data contains a combination of type I censoring and random censoring.

Under right censored data the observed time for the i^{th} individual, t_i, is equal to the minimum time between the event time and the censoring time, $t_i = min(T_i, C_i)$. An individual has experienced the event if the event time is less than the censoring time. Finally, a censoring indicator δ_i, must be defined to indicate whether the i^{th} individual's event time has been censored or uncensored. The censoring indicator is equal to one when the event time is observed and zero when the censoring time is observed, i.e. $\delta_i = I(T_i \leq C_i)$ where $I(\cdot)$ is an indicator function.

A vital concept in time-to-event data analysis is the hazard function. It is a function of time t, and it is the instantaneous event rate given the individual has survived without the event of interest until time t. The hazard function is defined as:

$$h(t) = \lim_{\Delta \to 0^+} \frac{Pr(t \leq T < t + \Delta | T \geq t)}{\Delta}. \tag{1.1}$$

Another important expression in time-to-event data analysis is the survival function, denoted $S(t)$. This function describes the probability that an individual survives until time point t, $S(t) = P(T \geq t)$. This is a left continuous monotone non-increasing function with $S(0) = 1$ and $S(\infty) = 0$. The probability density function (pdf), denoted $f(t)$, is:

$$f(t) = -\frac{dS(t)}{dt}. \tag{1.2}$$

The expression for hazard function may be simplified using the survival function and the pdf: $h(t) = f(t)/S(t)$. Finally, the cumulative hazard function is defined as:

$$H(t) = \int_0^t h(s)ds. \tag{1.3}$$

There are various parametric families used to model time-to-event data, the most widely used are: Exponential, Weibull, Log-logistic, and Log-normal. The survival function for the Weibull distribution is $S(t) = \exp(-(\rho t)^\alpha)$ and for the Log-logistic distribution is $S(t) = 1/(1 + (\rho t)^\alpha)$, where $\rho > 0$ is the scale parameter and $\alpha > 0$ is the shape parameter in both distributions.

1.1.1 Regression Models for Time-To-Event Data Analysis

In addition to t_i and δ_i, observations may have explanatory information, i.e. covariate information (denoted by x). For example, covariates may indicate the treatment type, exogenous variables (e.g. environmental factors), or intrinsic characteristics (e.g. age, gender, etc). Regression models are used to determine the relationship between the covariates and the response variable. Commonly covariate information is integrated into the survival model using the proportional hazards (PH) regression model or the accelerated failure time (AFT) model. In the PH regression model the expression for the hazard function when considering a single covariate is

$$h(t|x) = h_o(t)e^{\gamma_1 x}, \tag{1.4}$$

where $h_o(t)$ is the baseline hazard function; the hazard function for an individual who

has $x = 0$. Assuming it is constant over time, the hazard ratio, e^{γ_1}, is the measure of change in the hazard function when x is increased by a single unit. When $\gamma_1 > 0$ the length of survival will decrease as x increases; conversely, when $\gamma_1 < 0$ the length of survival will increase as x increases. When considering a single covariate, the survival function becomes:

$$S(t|x) = S_o(t)^{\gamma_1 x}, \tag{1.5}$$

where $S_o(t)$ is the baseline survival function; the survival function for an individual with $x = 0$.

Unlike the PH regression model, the most common form of the AFT model assumes there is a log-linear relationship between x and T. Considering a single covariate the AFT model is:

$$Y = \log(T) = \gamma_0 + \gamma_1 x + \epsilon, \tag{1.6}$$

where ϵ is the error term having a specified distribution. As the effects are additive in the model, the regression function will either increase or decrease the time to event occurrence, it will decelerate or accelerate.

For example the AFT model when ϵ has the extreme value distribution, thus T has the Weibull distribution with the survival function

$$S(t|x) = \exp(-\exp(\gamma_0 + \gamma_1 x)t^{\alpha}), \tag{1.7}$$

where $\alpha > 0$ is the shape parameter. The corresponding hazard function is $h(t|x) = \alpha t^{\alpha} e^{\gamma_0 + \gamma_1 x}$ which reduces to the PH model (Eq. 1.4).

1.1.2 Estimation Methods

The goal of survival analysis is to make inference regarding the survival time of individuals based on collected data; that is, we wish to estimate the survival function in such a way that we may draw meaningful conclusions. Non-parametric, parametric, and semi-parametric approaches can be used to obtain an estimated survival function.

Non-parametric methods can be used when there are no covariates requiring consideration. Consider a sample with N independent individuals with no censoring, the survival function can simply be estimated using the empirical survival function; a step function that decreases by $1/N$ at each event time, T_i. The Kaplan-Meier (KM) estimate, a modification of the empirical survival function, will account for censored observations within observed time-to-event data. The KM estimate of $S(t)$ is a left continuous step function, with the steps occurring at uncensored event times only. The KM estimator is:

$$\hat{S}(t) = \prod_{j|t_j<t} \left(1 - \frac{d_j}{r_j}\right) \tag{1.8}$$

where d_j is the number of individuals who experience an event at t_j and r_j is the number of individuals within the risk set at t_j. The step size at t will be influenced by the number of individuals who experience the event at t as well as individuals who are at risk at t^-.

The parametric approach assumes a specified distribution which completely describes the behaviour of event times given additional explanatory information (covariates). Recall the PH regression model (Eq. 1.4); for a fully parametric PH model we must specify the parametric form of the baseline hazard function. Similarly for the fully parametric AFT model (Eq. 1.6) we must specify the distribution of ϵ. The maximum likelihood estimation (MLE) method is employed to obtain estimates for

the specified distribution parameters as well as the regression parameters. In order to proceed with MLE, first the structure of the likelihood function must be understood. Both censored individuals and uncensored individuals will contribute to the likelihood equation. Consider N individuals, where the T_i for all individuals are independent and identically distributed, and censoring is non-informative. Uncensored individuals with $\delta_i = 1$ will contribute to the likelihood function through the $f(t_i|x_i;\theta)$ since they have experienced the event of interest. Individuals that have not experienced the event of interest, and are thus censored ($\delta_i = 0$), contribute to the likelihood equation through $S(t_i|x_i;\theta)$. The likelihood of the N independent observed data $\{(t_i, \delta_i, x_i); i = 1, .., N\}$ is:

$$L(\theta) = \prod_{i=1}^{N} f(t_i|x_i;\theta)^{\delta_i} S(t_i|x_i;\theta)^{1-\delta_i}, \tag{1.9}$$

where θ denotes the vector of unknown parameters. For example, under the Weibull distribution assumption of the survival time in (Eq. 1.7), the parametric vector becomes $\theta = (\alpha, \gamma_0, \gamma_1)$. The resulting ML estimates are consistent and asymptotically normally distributed.

Finally, there are semi-parametric estimation methods which can be employed to estimate the survival function. Consider the PH regression model (Eq. 1.4). The regression portion of each model is considered the parametric portion, and the baseline hazard or survival function ($h_o(t)$ and $S_o(t)$, respectively) are estimated non-parametrically. The partial likelihood method proposed by Cox (1975) allows estimation of γ_1 without explicitly stating $S_o(t)$, while maintaining consistency and asymptotic normality of the estimate of γ_1.

To non-parametrically estimate $S_o(t)$, one may use the relationship: $S_o(t) = \exp[-H_o(t)]$ where $H_o(t)$ is the baseline cumulative hazard function. The estimate for $H_o(t)$ when there is a single covariate is:

$$\hat{H}_o(t) = \sum_{i:t_i \leq t} \frac{\delta_i}{\sum_{l=1}^{N} Y_l(t_i) e^{\hat{\gamma}_1 x_l}}, \tag{1.10}$$

where $Y_i(t) = I(t_i \geq t)$ where $I(\cdot)$ is the indicator function and $\hat{\gamma}_1$ is an estimated value obtained through the Cox partial likelihood method.

1.1.3 Mixture-Cure Model

The situation may arise in which an individual never experiences the event of interest. For example, not all individuals may experience a disease recurrence. The previous methods discussed assume all individuals will eventually experience the event of interest, thus the survival function will continuously decrease until it reaches zero. However, now consider if the survival function decreases until a positive plateau which accounts for individuals who never experience the event, these individuals are referred to as statistically cured.

The mixture-cure model considers that some individuals within the cohort are statistically cured. Thus the survival function needs to consider individuals that are cured as well as those individuals who will experience the event (termed susceptible). Hence, the survival function could be written as

$$\begin{aligned} S(t|x) &= P(T > t|x) \\ &= \underbrace{p(x)}_{cured} + \underbrace{(1 - p(x))S_o(t|x)}_{susceptible}, \end{aligned} \tag{1.11}$$

where $S_o(t|x)$ is the survival function for the susceptible group, all susceptible individuals will eventually experience the event, i.e. $S_o(0|x) = 1$ and $S_o(\infty|x) = 0$. For example, the standard survival function $S_o(t|x)$ could be modelled by the Weibull

distribution in (Eq. 1.7). In Eq. 1.11, $p(x)$ denotes the probability of being cured conditional on the covariate x, $p(x) = P(T = \infty|x)$. The probability of being cured could be modelled using the logistic regression:

$$p(x) = \exp(\alpha_0 + \alpha_1 x)/(1 + \exp(\alpha_0 + \alpha_1 x)). \tag{1.12}$$

The likelihood function of the N independent observed data $\{(t_i, \delta_i, x_i); i = 1, .., N\}$, under the mixture-cure model is:

$$L(\theta, \alpha_0, \alpha_1) = \prod_{i=1}^{N} [(1 - p(x_i)) f_o(t_i|x_i; \theta)]^{\delta_i} [p(x_i) + (1 - p(x_i)) S_o(t_i|x_i; \theta)]^{1-\delta_i} \tag{1.13}$$

where $f_o(t|x; \theta) = \frac{-\partial S_o(t|x;\theta)}{\partial t}$ and θ is the set of parameters in $S_o(t|x; \theta)$. For our study the $S_o(t|x; \theta)$ will be of Weibull form (regression parameters included) in Eq. 1.7, thus $\theta = (\alpha, \gamma_0, \gamma_1)$.

1.2 Two-Phase Response-Dependent Sampling Designs

Measuring expensive covariates for all subjects within a cohort may not be a feasible option due to a study′s budgetary or logistical constraints. As a result of such limitations, we need to consider sampling designs that account for subjects that have missing data. To design a study allowing incomplete covariate data for some subjects, it is better to employ a cost-efficient sampling design, which balances the efficiency of parameter estimates and power of association tests with the sample size.

Response-dependent sampling is a cost-efficient sampling design, in which a subset

of subjects is selected from a cohort, based on the response variable, which has already been gathered for all subjects in the cohort (Lawless et al., 1999). The motivation behind response-dependent sampling is the ability to direct the study resources to the most informative subjects within the sample (Ding et al., 2015). In comparison to using the entire cohort, response-dependent sampling does incur some loss of efficiency to detect the association between the response variable and the expensive covariate; however, it is far more efficient than selecting a SRS of the same size (Yilmaz and Bull, 2011). In our study, we focus on two-phase response-dependent sampling designs. During the first phase of the sampling design, all members in a cohort are measured for the response variable and the inexpensive covariates. In phase II, a subset of the cohort is selected, based on the response variable obtained in phase I, and the expensive covariate(s) is measured only for the selected subjects.

Case-control sampling is the most well known response-dependent sampling design, and methods for two-phase sampling design have been developed (Breslow and Cain, 1988; Flanders and Greenland, 1991; Wacholder and Weinberg, 1994; Breslow and Holubkov, 1997; Breslow and Chatterjee, 1999). In case-control sampling, a random sample is selected from the set of cases and another random sample is selected from the controls. Estimation methods under the two phase case-control sampling designs have been extensively studied (Robins et al., 1994; Carroll et al., 1995; Lawless et al., 1999;, Breslow et al., 2000, 2003; Scott and Wild, 2001; Chatterjee et al., 2003).

1.2.1 Two-Phase Response-Dependent Sampling Designs for Continuous Response Variable

Suppose that the response variable that determines which observations will be selected in phase II is continuous, and is collected during phase I for all observations (along with inexpensive covariate information). Basic stratified sampling (BSS) could be

used for phase II sampling. Assume that there are N independent observations with the response variable and inexpensive covariates available. Suppose K strata are defined by allocating the observations based on $K-1$ cut-off points for response variable. Within the j^{th} strata (S_j) there are N_j observations, where $N = \sum_{j=1}^{K} N_j$. Hence, phase I observations will be allocated to a particular strata depending on the response value. In phase II, assume that n individuals are selected. From the strata S_j a sample of size n_j where $n = \sum_{j=1}^{K} n_j$ is randomly selected, thus the selection probability for strata j is $p_j = n_j/N_j$ (Lawless et al., 1999). The phase II sample with size n is selected based on the stratum selection probabilities, p_j, $j = 1, .., K$.

1.3 Two-Phase Response-Dependent Sampling Designs for Time-To-Event Analysis

In our study, the response variable that determines which individuals will be selected for phase II is a continuous time-to-event variable; wherein this type of the response variable is subject to censoring, the event time may not be fully observed. Difficulties arise when dealing with data that is censored. For example, an individual with a right-censored event time does not have an event time; instead it is censored due to a random process or censored at the end of the study follow-up period. Within our study cases are individuals who have experienced the event of interest, and non-cases are the individuals that have been censored (have not experienced the event of interest until the censoring time).

A response variable that may be censored complicates the response-dependent sampling and the method of estimation; however, there are two commonly used response-dependent sampling designs for time-to-event data: case-cohort sampling designs (Prentice, 1986) and nested case-control sampling designs (Thomas, 1977).

In the nested case-control sampling design, which is a variation of case-control sampling, a small subset of controls (1-5 observations) is selected from those at-risk at the event time of each case (Langholz and Thomas, 1990; Wacholder, 1991). For the standard case-cohort design all case observations are selected and a subset of the cohort, termed the subcohort, is randomly selected (Prentice, 1986). This standard case-cohort design approach is used when the event of interest is rare. When the event is not considered rare and there is an expensive covariate to measure in the second phase, the generalized case-cohort design could be performed. In the generalized case-cohort design, a random subset of the case observations is selected along with a subcohort that is randomly selected. It is important to note the cohort is made up of the non-case (i.e., censored) and case observations.

While advances have been made in the development of methods of analysis for response-dependent sampling designs, the study into identifying efficient response-dependent sampling designs with multiple phases is lacking. Only one study, carried out by Morara et al. (2007), focuses on multi-phase response-dependent sampling designs where the response variable is continuous not subject to censoring. The approach finds the efficient sampling design by minimizing the asymptotic variance of the regression parameter estimate that was obtained through MLE. There have been a few studies analysing the efficiency for the two-phase case-control sampling designs (Palmgren, 1987; Greenland, 1988; McNamee, 2005; Cai and Zeng, 2007) investigated power and sample size calculations specifically for case-cohort designs.

In this study, we consider the response variable as a time-to-event variable subject to censoring and consider extensions of case-cohort design. To improve the efficiency of the sampling designs, we select a more informative sample. The efficiency of the parameter estimates will be the basis for comparison among the sampling scenarios; the aim is to minimize the variance of the coefficient estimate of the expensive covariate

to maximize efficiency. The case-cohort designs considered expand on previous work in which the sampling probabilities depend on the censoring status as well as time-to-event (Ding et al., 2014; Ding et al., 2015; Lawless, 2016).

For this study, the generalized case-cohort sampling design will be modified in such a way that BSS will be applied to the cases only, or applied to both cases and non-cases. Stratification will be based on censoring status and time-to-event. First, the case set will be stratified into three groups based on survival time. Short survival times will be allocated to stratum 1, long survival times will be allocated to stratum 3, and all other survival times will fall into stratum 2 (medium survival times). BSS will be employed on the stratified case set and a simple random sample (SRS) is selected from the cohort.

Secondly, BSS will be applied to the cases and the non-cases. The stratified case sampling design that resulted in the most efficient parameter estimate will be used throughout all sampling scenarios, thus the investigation focuses on BSS of the non-cases. All non-cases are stratified into three groups based on censoring time. Short and long censoring times will be allocated to stratum 1 and 3, respectively. All medium length censoring times are designated to stratum 2. A detailed explanation of the sampling design settings are given in Chapter 2.

1.4 Estimation Methods under Two-Phase Response-Dependent Sampling

Two-phase response-dependent sampling is employed. Assume that N individuals are being generated from $f(y|x,z;\theta)g(x|z)h(z)$ where $f(y|x,z;\theta)$ denotes the conditional pdf of the response variable Y given the expensive covariate X and inexpensive covariate Z, $g(x|z)$ denotes the conditional pdf of X given Z and $h(z)$ denotes the pdf

of Z. In the first phase, the response variable (Y_i) and inexpensive covariates (Z_i) are collected for all individuals ($i = 1, .., N$) in the cohort. In phase II, response-dependent BSS is used to select n individuals that are measured for the expensive covariate X_i. Thus, define $R_i = \text{I[individual } i \text{ is selected in the second phase]}$. The probability of being fully observed depends on Y_i and Z_i, and the sampling probability is $\pi_i = Pr(R_i = 1|x_i, z_i, y_i) = Pr(R_i = 1|z_i, y_i)$. Thus X_i are missing at random (Rubin, 1976).

There are two main methods of estimation for continuous response variables that are not subject to censoring: pseudo-likelihood based and likelihood-based approaches. Lawless et al. (1999) gives a summary of different estimation methods. The most well-known pseudo-likelihood based approach is the Horvitz Thompson (HT) estimation. The estimators obtained through the HT estimation approach provide the solution to the inverse probability weighted (IPW) equation

$$U(\theta) = \sum_{i=1}^{N} \frac{R_i}{\pi_i} \frac{\partial}{\partial \theta} log\{f(y_i|x_i, z_i; \theta)\} = 0, \tag{1.14}$$

which is determined through the sampling probabilities in phase II of sampling. The only information used is from those units that were completely observed. The limitation with this method occurs when the sampling probability π_i for an individual is close to zero.

The likelihood-based approaches that have been considered thus far are: the conditional likelihood method (conditional on being selected into the phase II sample) and the full likelihood method. As in the HT estimation method, in the conditional likelihood approach, only the information from units that were completely observed are used. The conditional likelihood is

$$L_C(\theta) = \prod_{i:R_i=1} Pr(y_i, x_i, z_i | R_i = 1). \tag{1.15}$$

The estimate of θ is obtained by maximizing Eq. 1.15.

The full likelihood uses all of the observed data; which is $\{(y_i, x_i, z_i), i \in V\} \cup \{(y_i, z_i), i \in \overline{V}\}$ where $V = \{i : R_i = 1, i = 1, ..., N\}$ and $\overline{V} = \{i : R_i = 0, i = 1, ..., N\}$. Hence, the full likelihood is

$$L(\theta, G) = \prod_{i \in V} f(y_i | z_i, x_i; \theta) g(x_i | z_i) \prod_{i \in \overline{V}} \int_x f(y_i | z_i, x; \theta) dG(x | z_i). \tag{1.16}$$

Lawless et al. (1999) discussed the semi-parametric estimation method when $g(x|z)$ is treated non-parametrically. When BSS is performed in phase II as described in Section 1.2.1, the stratum identifier is retained for all units, and it is assumed that the fully observed units within each stratum are a random sample. Then, the full likelihood function (Eq. 1.16) could be written under these conditions as

$$L(\theta, G) = \prod_{j=1}^{K} \{ \prod_{i \in S_j} f(y_i | x_i, z_i; \theta) g(x_i | z_i) \} Q_j(\theta, G)^{N_j - n_j} \tag{1.17}$$

where $Q_j(\theta, G) = Pr\{(y, x, z) \in S_j\}$ and $g(x|z)$ is estimated non-parametrically. Semi-parametric ML estimation is performed to estimate the unknown parameters θ and the distribution of x given z.

For missing data and response-selective problems Lawless et al. (1999) presented the semi-parametric ML, conditional likelihood and compared them with the pseudo-likelihood method. They concluded that in most situations the full likelihood method was more efficient over the pseudo-likelihood approach. In particular the pseudo-likelihood approach struggled greatly when the correlation between the response and covariates was strong; conversely when the regression relationship was weak and the

sample size was large, the pseudo-likelihood approach could be more efficient than the ML approach. In general they also found including information from the incomplete observations improves the efficiency.

1.4.1 Estimation Methods Under Case-Cohort Designs

There are two main methods of analysis for case-cohort sampling designs: the pseudo-likelihood based approach and likelihood-based approaches. The pseudo-likelihood based approach or the estimating equation approach discussed by Breslow et al. (2009) is an extension of the IPW estimating equation method. The proposed method reduces the variance (improves efficiency) due to the IPW estimation through calibration or by using auxiliary variable information to estimate the sampling weights. The limitation with this method occurs when at least one strata sampling probability is close to zero. The analysis of case-cohort designs through the Cox partial likelihood method, in particular weighting methods, were discussed by Barlow et al. (1999) and Onland-Moret et al. (2007).

Focusing on two-phase case-cohort designs Zeng and Lin (2014) proposed semi-parametric transformation models that take into consideration correlation between inexpensive phase I covariates and expensive phase II covariates; this is a likelihood-based approach. Considering pdfs conditional on continuous phase I covariates has an added difficulty, thus kernal smoothing was included in the likelihood model. The covariate distribution is modelled non-parametrically. To maximize the proposed likelihood function, they also recommended a new semi-parametric EM algorithm that results in estimators that are consistent, asymptotically efficient and normally distributed.

Our focus will be on the likelihood-based approach. Introducing covariates into the standard likelihood (Eq. 1.9) is simple if all individuals are measured for each

covariate. In case-cohort sampling, however, we have the added complexity that only individuals that are sampled in phase II are measured for an expensive covariate X_i. Suppose a cohort of N individuals is selected in phase I; for those individuals y_i, δ_i and the inexpensive covariate information z_i is collected. In phase II, a sample from the cohort is selected based on the data collected in phase I and the expensive covariate X_i is measured for the selected individuals. The probability of the i^{th} individual being selected in phase II does not depend on the expensive covariate, thus X_i is missing at random. In other words, the sampling probability for the i^{th} individual is $\pi_i = Pr(R_i = 1|\delta_i, Y_i, Z_i, X_i) = Pr(R_i = 1|\delta_i, Y_i, Z_i)$ where R_i = I[individual i is selected in phase II].

A conditional likelihood function, conditional on $R_i = 1$, $i = 1, ..N$, can be applied to estimate the model parameters. This method does not use information pertaining to individuals not sampled in phase II. The likelihood-based approach used within this study takes into account all individuals, both those with complete covariate data and those with incomplete covariate data. The likelihood function based on the data set: $\{(\delta_i, y_i, z_i, x_i), i \in V\} \cup \{(\delta_i, y_i, z_i), i \in \overline{V}\}$ where $V = \{i : R_i = 1, i = 1, ..., N\}$ and $\overline{V} = \{i : R_i = 0, i = 1, ..., N\}$ is:

$$L(\theta, g) = \prod_{i \in V} f(y_i|z_i, x_i; \theta)^{\delta_i} S(y_i|z_i, x_i; \theta)^{1-\delta_i} g(x_i|z_i) \times \prod_{i \in \overline{V}} \int_x f(y_i|z_i, x; \theta)^{\delta_i} S(y_i|z_i, x; \theta)^{1-\delta_i} dG(x|z_i), \quad (1.18)$$

where the conditional distribution function of X given Z is $G(x|z)$ and the corresponding conditional pdf is $g(x|z)$. Semi-parametric ML estimates of θ and $g(x|z)$ are obtained by maximizing (Eq. 1.18); the corresponding estimates are asymptotically normally distributed and consistent.

1.5 Aim and the Outline of the Study

In this study, the aim is to explore the efficiency of the various types of case-cohort sampling designs based on likelihood-based method under different survival models; AFT model, PH regression model, and the mixture-cure model. We want to find the sampling designs that lead to more efficient estimates, in particular we consider sampling based on the observed time (t_i) in addition to the censoring status (δ_i).

In the generalized case-cohort design a random sample is selected from each of the case set and from the cohort. Firstly, in the generalized case-cohort design, we want to determine the proportion of cases versus cohort observations selected into the sample during phase II that lead to efficient estimates; sampling is based only on the censoring indicator. Sampling based on censoring indicator and observed time will then be explored using an extended version of the generalized case-cohort design, in particular we want to determine which BSS design setting with response-dependent strata leads to more efficient parameter estimates. BSS for both the case and the non-case observations will be considered.

In Chapter 2, we outline the investigated sampling design settings and consider their relationship to the study objectives. Section 2.1 considers the efficiency of the generalized case-cohort sampling design, which focus on determining the proportion of cases to cohort observations selected during phase II leading to the most efficient design. The sampling design settings that investigate BSS of the cases and non-cases, stratified by observed time, that result in efficient parameter estimates are outlined in Section 2.2 and 2.3, respectively. The sampling scenarios investigated within each sampling design setting will be compared based on the efficiency of the coefficient estimate of the expensive covariate.

In Chapter 3 the simulation study results under the standard survival model are

presented. The outline of the simulation study performed is found within Section 3.1, including the data generation details. Section 3.2 and 3.3 contain the simulation results for the simulation study based on a large cohort of data generation when the censoring mechanism is generated from the uniform and exponential distribution, respectively. And finally, Section 3.4 assesses the sampling design efficiency when the simulation study is replicated 1000 times for a smaller cohort size.

An investigation regarding the efficient case-cohort sampling designs when the survival times are from the mixture-cure model is found within Chapter 4. The efficiency of the sampling designs are based on the standard error of the expensive covariate coefficient as well as the logistic regression coefficient (in relation to the probability of being cured). A summary of the study results and concluding remarks will be discussed in Chapter 5.

Chapter 2

Case-Cohort Designs

The case-cohort design is generally used when the event of interest is rare or there is an expensive covariate to measure. The standard case-cohort sampling design is used when the event of interest is rare. In this design all cases are selected into the phase II sample, and a simple random sample of size $n - \sum_{i=1}^{N} \delta_i$ is selected from the cohort of size N. The sampling probability of the i^{th} individual, denoted π_i, depends on the censoring indicator. If $\delta_i = 1$ then $\pi_i = 1$; if $\delta_i = 0$ then $\pi_i = (n - \sum_{i=1}^{N} \delta_i) / \sum_{i=1}^{N} (1 - \delta_i)$. An extension of the standard case-cohort design, generally used when there is an expensive covariate and the event is not rare, is the generalized case-cohort design. In this generalized design a simple random sample of size n_{cases} is selected from the case set of size N_{cases}, and a sample sized $n_{cohort} = n - n_{cases}$ is selected from the cohort without replacement. For this generalized design, the sampling probability of individual i still only depends on the censoring indicator, $\pi_i = Pr(R_i = 1 | \delta_i)$.

In our study, using likelihood-based approaches, we want to determine the sampling designs that lead to increased efficiency of the coefficient estimate of the expensive covariate given the phase II sample size n. The efficiency of the parameter

estimate is maximized when the corresponding standard error is minimized. The investigated sampling designs depend on the censoring indicator, or the censoring indicator and the time-to-event variable. In particular, we have the following objectives:

Objective 1: *Given the sample size n, in generalized case-cohort designs determine the proportion of case versus cohort observations that should be selected during phase II that maximizes the efficiency of the coefficient estimate.*

Selecting a more informative case sample could improve the efficiency of the sampling designs. Thus we consider an extension of the generalized case-cohort design in which the selection probabilities depend on the censoring indicator and the case selection probabilities depend on the time-to-event variable. Thus we investigate the following objective:

Objective 2: *Given the sample size n, determine which observations within the case set should have a higher selection probability during phase II to obtain more efficient sampling design. This objective includes*

i) Comparison of the design efficiency of the SRS with the BSS of the case set, and

ii) Determination of the case BSS design which improves the efficiency of the coefficient estimate of the expensive covariate.

Finally, our aim is to determine if the design efficiency improves when the sampling probabilities of the non-case observations depend on the censoring time, C_i. For this final investigation, BSS designs are considered for both the case set and non-case set.

Objective 3: *Using the most efficient design identified under objective 2, determine which observations within the non-case set in phase II should have a higher selection probability to obtain more efficient sampling designs. This objective includes*

i) Comparison of the design efficiency of the SRS with the BSS of the non-cases, and

ii) Determination of the non-case BSS design which improves the efficiency of the coefficient estimate of the expensive covariate.

The following sampling design settings are considered to investigate the objectives. In Section 2.1, we will investigate objective 1 through the sampling design setting 1. Section 2.2 contains the details of sampling design setting 2 which explores objective 2. Finally, the third objective is investigated through sampling design setting 3 details found in Section 2.3. In the simulation study in Chapter 3, we consider these design settings under the Weibull AFT model/PH model when the censoring time is from the uniform distribution and the exponential distribution.

2.1 Sampling Design Setting 1

The first sampling method studied is the generalized case-cohort sampling design in which sampling depends only on the censoring indicator.

A single large phase I cohort is used throughout all sampling scenarios. The cohort size is $N = 50000$ with approximately 10000 cases within the cohort ($N_{cases} \approx 10000$). A sample is selected from the cohort, the phase II sample, and is measured for the expensive covariate. For all sampling scenarios the phase II overall sample size n is 10000, but different case sample size (n_{cases}) and cohort sample size (n_{cohort}) values are considered. The purpose of this design setting is to determine the proportion of case versus cohort observations to be selected in phase II which lead to more efficient estimates, objective 1.

The six sampling scenarios outlined in Table 2.1 are performed on the same generated phase I data set. The percentage of cases initially selected into the sample decreases by 10% as the sampling scenario number increases. The total number of cases selected in the phase II is a combination of the n_{cases} sampled and cases selected

Table 2.1: Sampling Design Setting 1

Sampling Scenario	n_{cases}	n_{cohort}
1	7000	3000
2	6000	4000
3	5000	5000
4	4000	6000
5	3000	7000
6	2000	8000

during cohort sampling; this is given by

$$n^*_{cases} = \sum_{i=1}^{N} R_i \delta_i, \tag{2.1}$$

where $R_i = \text{I[individual } i \text{ is selected in the second phase]}$. The total number of cases within the sample n^*_{cases} is greater than or equal to the cases initially selected into the sample n_{cases}.

2.2 Sampling Design Setting 2

The second objective, how to select an informative case sample which leads to increased efficiency of the coefficient estimate, is explored in sampling design setting 2. To determine the cases that should have a higher selection probability we employ the generalized case-cohort sampling design and a modified generalized case-cohort sampling design with BSS for case set only. The sample collected under the generalized case-cohort sampling design is a random sample selected from the case set and a random sample selected from the cohort. This was completed in sampling design setting 1. Sampling scenario 2 from sampling design setting 1 ($n_{cases} = 6000$ and $n_{cohort} = 4000$) is used as the SRS of the case set to compare with the BSS sampling

scenarios outlined in sampling design setting 2A. Sampling design setting 2B investigates case BSS when $n_{cases} = 4000$ and $n_{cohort} = 6000$, thus sampling scenario 4 from sampling design setting 1 will be used to compare SRS with these case BSS scenarios.

For the modified case-cohort sampling design, BSS is applied to the cases and a random sample is selected from the cohort. For case BSS, the case observations are ordered based on length of survival time, shortest to longest. Cases will be allocated to one of three possible case strata using fixed cut-off survival time values; lower cut-off (L_C) and upper cut-off (U_C). There will be N_{cases_1} cases below L_C thus are placed the first, short survival time stratum. The second stratum will have N_{cases_2} mid-range survival time cases that fall between L_C and U_C. Finally, N_{cases_3} cases are allocated to the stratum 3, which contains the survival times that are longer than U_C. The defined stratum cut-offs allocate approximately 3000 observations to stratum 1 and to stratum 3, and $N_{cases_2} = N_{cases} - N_{cases_1} - N_{cases_3}$. The cut-off values were chosen in such a way that we tried to not allocate too many individuals to the first and third strata to see the importance of selecting individuals from these extreme strata.

$$\underbrace{T_{(1)} < ... < T_{(N_{cases_1})}}_{n_{cases_1}} < L_C < \underbrace{T_{(N_{cases_1}+1)} < ... < T_{(N_{cases_1}+N_{cases_2})}}_{n_{cases_2}} < U_C < \underbrace{T_{(N_{cases_1}+N_{cases_2}+1)} < ... < T_{(N_{cases_1}+N_{cases_2}+N_{cases_3})}}_{n_{cases_3}}, \quad (2.2)$$

where $T_{(i)}$ denotes the i^{th} smallest survival time in the case set.

A random sample is selected from each stratum independently with sizes n_{cases_1}, n_{cases_2} and n_{cases_3} as shown in Eq. 2.2. The sampling probability assigned to each stratum varies, however the total sample size remains fixed, $n = 10000$. In Table 2.2, the first five sampling scenarios outlined select 6000 case observations, and 4000 cohort

observations. Firstly, we select the predetermined number from each case strata to make up the total n_{cases}. Following this a random sample is collected from the cohort without replacement. Recall the cohort is made up of cases as well as non-cases, this implies the sample size from each case strata will be larger than those presented in Table 2.2. In the first sampling scenario in Table 2.2, case observations with short or long survival times (stratum 1 and 3) will be sampled heavily and no mid-range survival time observations will be selected during the case sampling. For sampling scenario 2, an equal number of cases will be selected from each of the three defined case strata. In sampling scenario 3 the case observations with long survival times, case stratum 3, are sampled more than the other case strata. In contrast sampling from stratum 1, the short survival times, is increased in sampling scenario 4. Finally sampling scenario 5, like sampling scenario 1, focuses case sampling mostly on the observations from stratum 1 or 3, however a sixth of the initial case sample will be mid-range survival time observations (stratum 2). The proportion of cases to cohort observations selected into the sample is modified in sampling scenario 6, $n_{cases} = 4000$ and $n_{cohort} = 6000$, which is found to be the most efficient sampling design in sampling design setting 1 in Chapter 3. The percentage of observations sampled from each case strata mimics that of sampling scenario 4 in Table 2.2 which is found to be the most efficient design among the first five scenarios in Chapter 3.

Table 2.2: Sampling Design Setting 2A

Sampling Scenario	n_{cases_1}	n_{cases_2}	n_{cases_3}	n_{cohort}
1	3000	0	3000	4000
2	2000	2000	2000	4000
3	1000	2000	3000	4000
4	3000	2000	1000	4000
5	3000	1000	2000	4000
6	2500	1000	500	6000

The goal of sampling design setting 2B in Table 2.3 is to obtain the most efficient estimates when the case strata selection probabilities are modified between sampling scenarios, as in sampling design setting 2A in Table 2.2, combined with the knowledge gained from sampling design setting 1. All six scenarios begin by selecting a predefined number of cases from each case strata totalling $n_{cases} = 4000$ cases; then the cohort sample, where $n_{cohort} = 6000$ observations (a mixture of cases and non-cases) are randomly selected without replacement. For sampling scenarios 1 and 6 in sampling design setting 2B, most or all of the case sample is selected from the first and third strata. In sampling scenario 2, a random sample will be selected from each case strata of approximately equal size. In the third sampling scenario cases with long survival times, cases within stratum 3, will be largely sampled. Finally the short survival times, cases within the first stratum, are heavily sampled in sampling scenarios 4, 5 and 6 with different stratum 2 and 3 sample size scenarios.

Table 2.3: Sampling Design Setting 2B

Sampling Scenario	n_{cases_1}	n_{cases_2}	n_{cases_3}	n_{cohort}
1	2000	0	2000	6000
2	1333	1334	1333	6000
3	500	1000	2500	6000
4	2500	1000	500	6000
5	2666	667	667	6000
6	2500	500	1000	6000

The total number of cases sampled from the j^{th} strata is the combination of the cases initially selected into the sample from strata j and cases selected during cohort sampling. This is given by

$$n^*_{cases_j} = \sum_{i \in S_j} R_i \delta_i, \tag{2.3}$$

where S_j denotes the j^{th} stratum ($j = 1, 2, 3$). The total number of cases sampled from the j^{th} strata $n^*_{cases_j}$ is greater than or equal to the number of cases initially sampled from the j^{th} strata n_{cases_j}.

2.3 Sampling Design Setting 3

Objective 3 is explored in sampling design setting 3; we want to obtain the most efficient design by selecting the non-case observations in an informative manner. For this analysis we will be selecting an informative sample from the case set and the non-case set; these sets are mutually exclusive.

The cases are selected based on knowledge gained from sampling design setting 2; we consider the most efficient design obtained in Section 2.2. We set $n_{cases_j} = n^*_{cases_j}$ and perform sampling on the case set and non-case set separately instead of sampling from the case set or the cohort. The non-case observations are ordered by censoring time, and stratified into three groups. Suppose the first non-case stratum includes $N_{noncases_1}$ non-cases with a censoring time shorter than the lower censoring time threshold (L_{CN}), and the third non-case stratum includes $N_{noncases_3}$ non-cases with a censoring time longer than the upper censoring time threshold (U_{CN}). The remaining $N_{noncases_2} = N_{noncases} - N_{noncases_1} - N_{noncases_3}$ non-cases with intermediate censoring times fall in non-case stratum 2. The defined stratum cutoffs, L_{CN} and U_{CN}, result in: $N_{noncases_1} \approx 7000$, $N_{noncases_2} \approx 26000$ and $N_{noncases_3} \approx 7000$.

$$\underbrace{C_{(1)} < ... < C_{(N_{noncases_1})}}_{n_{noncases_1}} < L_{CN} < \underbrace{C_{(N_{noncases_1}+1)} < ... < C_{(N_{noncases_1}+N_{noncases_2})}}_{n_{noncases_2}}$$
$$< U_{CN} < \underbrace{C_{(N_{noncases_1}+N_{noncases_2}+1)} < ... < C_{(N_{noncases_1}+N_{noncases_2}+N_{noncases_3})}}_{n_{noncases_3}}, \quad (2.4)$$

where $C_{(i)}$ denotes the i^{th} smallest censoring time. Firstly, a random sample of size $n^*_{cases_j}$, will be randomly selected from the j^{th} case strata ($j = 1, 2, 3$); then a random sample is selected from each non-case stratum independently; a sample of size $n_{noncases_h}$ is selected from the h^{th} non-case stratum ($h = 1, 2, 3$).

Table 2.4 outlines the sampling scenarios investigated for sampling design setting 3. Throughout all of the simulations n, $n^*_{cases_1}$, $n^*_{cases_2}$ and $n^*_{cases_3}$ remain fixed. There will be $n_{noncases} = 5000$ throughout all of the sampling scenarios. The sampling probabilities of the non-case strata change between different sampling scenarios. In sampling scenarios 1 and 6 non-cases will be sampled mostly from stratum 1 and 3, short and long censoring times respectively. The non-case sampling performed in sampling scenario 2 is selected mostly from the mid-range censoring times, non-case strata 2. Each non-case strata will be sampled from in approximately equal numbers in sampling scenario 3. Sampling scenarios 4 and 7 will sample non-case observations mostly from non-case stratum 1, the short censoring times. Finally, the long censoring time observations within stratum 3 will be heavily sampled in sampling scenario 5.

Table 2.4: Sampling Design Setting 3

Sampling Scenario	$n^*_{cases_1}$	$n^*_{cases_2}$	$n^*_{cases_3}$	$n_{noncases_1}$	$n_{noncases_2}$	$n_{noncases_3}$
1	2700	1400	900	2500	0	2500
2	2700	1400	900	875	3250	875
3	2700	1400	900	1667	1666	1667
4	2700	1400	900	2700	1400	900
5	2700	1400	900	900	1400	2700
6	2700	1400	900	2700	900	1400
7	2700	1400	900	3334	834	834

Chapter 3

Simulation Study Under the Standard Survival Model

For a given phase II sample size n, we explore extensions of the generalized case-cohort sampling design that result in more efficient sampling designs. Efficient sampling designs minimize the variance of the coefficient estimate of the expensive covariate X. In Chapter 2, different sampling design settings are outlined relating to the three main study objectives. These defined sampling design settings will explore phase II sampling where the sampling probabilities depend on censoring indicator only, or censoring indicator and the observed survival time.

In two-phase sampling designs, for all N individuals within the cohort (i.e. phase I sample) the observed time t_i and censoring indicator δ_i is obtained. No inexpensive covariates are considered in the simulation study. In phase II, a sample is selected from the cohort, based on δ_i and/or t_i generated in phase I, and the expensive covariate is obtained. In the simulation study it is assumed that there is an expensive covariate X which is a binary variable. Individuals within the cohort, both those with complete (δ_i, t_i, x_i) and incomplete (δ_i, t_i) data will contribute to the likelihood function. For

the simulation study performed the likelihood function (Eq. 1.16) can be simplified to:

$$L(\theta; q) = \prod_{i \in V} f(t_i|x_i;\theta)^{\delta_i} S(t_i|x_i;\theta)^{1-\delta_i} g(x_i; q) \prod_{i \in \overline{V}} \sum_{x=0}^{1} f(t_i|x;\theta)^{\delta_i} S(t_i|x;\theta)^{1-\delta_i} g(x; q), \tag{3.1}$$

where the notation used was described in Section 1.4. The ML estimate of θ and q is obtained by maximizing Eq. 3.1.

3.1 Simulation Procedure

A large cohort of individuals, $N = 50000$, was generated for each simulation study. First the expensive covariate, X_i, was generated from the Bernoulli distribution with the probability of success $p = 0.25$. The time-to-event value for individual i, T_i was generated from the Weibull distribution with survival function $S(t|x) = \exp(-e^{0.5+\gamma_1 x} t^{\alpha})$. Throughout the simulation study different values of the shape parameter, α and regression parameter, γ_1, were considered: $\alpha = 0.5, 1.0, 1.5$ and $\gamma_1 = 0, 0.25, 1.0$. Two censoring mechanisms were considered during the simulation study. Firstly, the censoring time, C_i, was generated from the Uniform distribution $(0,\ b)$. The results to be found in Section 3.2. In Section 3.3, the censoring time was generated from the Exponential distribution with rate λ. The values of the parameters b and λ in the censoring distributions were determined in such a way that approximately 20% of the observations are not censored; in other words, there will be approximately $N_{cases} \approx 10000$ cases within the cohort, with size $N = 50000$.

In phase I, the response variable, $(\delta_i,\ t_i)$ is collected for all individuals N within the cohort. A sample will be selected in phase II, based on the response variable, and will

be measured for the expensive covariate. The expensive covariate x_i is assumed to be unknown for unselected individuals in the second phase. Thus individuals selected in phase II will be completely observed (δ_i, t_i, x_i), and individuals not selected in phase II will be incomplete (δ_i, t_i). The case-cohort sampling designs outlined in Chapter 2 were performed on the simulated dataset and the ML estimates were obtained for the model parameters. The nlm function that is built into the R software environment was used to maximize the likelihood function (Eq. 3.1) and obtain the ML estimates of the parameters and their corresponding standard errors. The sampling designs were compared based on the efficiency of coefficient of the expensive covariate (γ_1 in Eq. 1.7), thus we want to minimize the standard error of the ML estimate of γ_1, $\hat{\gamma_1}$.

3.2 Uniform Censoring

Although the sampling procedure, as described in Section 3.1, is a response-dependent sampling design the sampling designs investigated in Chapter 2 under uniform censoring result in ML estimates that are unbiased. The figures in Appendix A.1 illustrate the ML estimates of γ_1 obtained through maximizing Eq. 3.1 are close to the true value.

3.2.1 Results under Sampling Design Setting 1

We consider the efficiency of sampling designs that are extensions of the generalized case-cohort design. In particular in sampling design setting 1 (Section 2.1), we aim to understand the efficiency of the generalized case-cohort designs based on the number of case and cohort observations selected in phase II (Objective 1) given the phase II sample size n. The sampling scenarios considered are in Table 2.1.

For each value of γ_1 a plot of the standard error of $\hat{\gamma_1}$ is presented in Figure 3.1

for all sampling scenarios in Table 2.1 explored and all values of the shape parameter that were considered. All of the plots indicate that scenario 1, when selecting 7000 case observations into the sample, produces the highest standard error estimate. The standard error estimate continues to decrease as the number of selected cases observations decreases until scenario 4. For scenarios 5 and 6 the standard error of $\hat{\gamma}_1$ begins to increase. This trend is seen across all values of γ_1 and all values of the shape parameter.

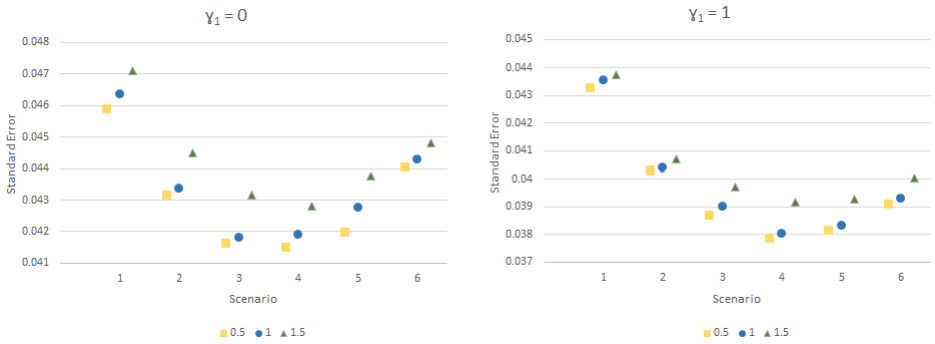

Figure 3.1: Standard error estimate of $\hat{\gamma}_1$ under sampling design setting 1 and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Uniform distribution.

Our results indicate that sampling scenario 4 produces the minimum standard error for all γ_1 and α values investigated. Sampling scenario 4 selects 4000 cases and 6000 subcohort observations however this does not reflect the true number of cases found within the total sample in the second phase. Recall the cohort sampled in the second phase includes non-case observations as well as cases that have not been sampled. Thus the true number of cases, n^*_{cases} (Eq. 2.1) is generally larger than n_{cases}. Table 3.1 gives n^*_{cases} and the number of non-cases sampled, $n_{noncases}$, in each sampling scenario. For sampling scenario 4, the number of cases within the sample was 4802. Comparing the total number of cases in the other sampling scenarios with scenario 4 (Table 3.1), scenario 4 achieves near balance between the number of cases

and non-cases selected into the sample.

As in the case-control sampling design, when the phase II sample size n is fixed, a balanced number of cases versus non-cases selected into the sample minimizes the variance. Therefore the parameter estimates that are produced are the most efficient.

Table 3.1: Count of Cases versus Non-cases under the Sampling Design Setting 1 and Standard Survival Model

Sampling Scenario	n^*_{cases}	$n_{noncases}$
1	7211	2789
2	6383	3617
3	5573	4427
4	4802	5198
5	4063	5938
6	3360	6640

3.2.2 Results under Sampling Design Setting 2

In this sampling design setting, phase II sampling is dependent on the censoring indicator as well as the observed survival time. In particular for Objective 2, defined in Chapter 2, focuses on the selection probability assigned to the case observations. This objective is explored through sampling design setting 2, Section 2.2, and considers SRS and BSS of the case set. Case observations will be stratified into three groups based on their survival time, and each stratum will be sampled from independently. Table 2.2 and 2.3 contains the sampling scenarios considered in this design setting.

Using sampling scenario 2 of sampling design setting 1 (Table 2.1), we can compare SRS with BSS of the case set (sampling scenarios 1 through 5 of sampling design setting 2A). The efficiency of $\hat{\gamma}_1$ is affected when employing BSS versus SRS. In particular, when the BSS design selects more cases from the left side of the distribution the efficiency of $\hat{\gamma}_1$ is improved over the SRS design.

For sampling scenarios one to five in Table 2.2 the proportion of cases to non-cases was consistent, in general 6400 to 3600. Therefore the change in standard error between sampling scenarios is not due to the number of cases versus non-cases within the sample.

Figure 3.2 depicting standard error of the γ_1 estimate, by sampling scenario and shape parameter, indicate selecting more cases from the first and second strata will result in a lower standard error. In particular, sampling scenarios four and five show lower standard errors in comparison to the other sampling scenarios with $n_{cases} = 6000$. In these scenarios increased sampling from the left side of the distribution and results in more efficient estimates. Reflecting on the shape of the true underlying distribution of the simulated data, the Weibull distribution, this conclusion seems logical. The Weibull distribution is skewed to the right thus selecting more from the left hand side would provide more information on the shape of the distribution. The pdfs for $\gamma_1 = 1$ for the three shape parameters explored can be found in the Appendix B.1.

Sampling scenario three selects more from case stratum 3, thus selecting more cases from the right side of the distribution. The standard error of $\hat{\gamma_1}$ for this scenario is larger than all other scenarios. Again this is a logical conclusion, the sampling scenario is selecting more cases from the right tail of the distribution and such observations are less frequent in the cohort.

The lowest standard error is achieved when selecting more observations from the cohort ($n_{cohort} = 6000$) the the case set ($n_{cases} = 4000$) in addition to increased sampling from stratum 1 and 2, denoted sampling scenario six. This is a reflection of a conclusion made in sampling design 1, balance between the number of cases and non-cases in the sample produce more efficient estimates.

The results discussed hold for all values of γ_1 and all defined values of the shape

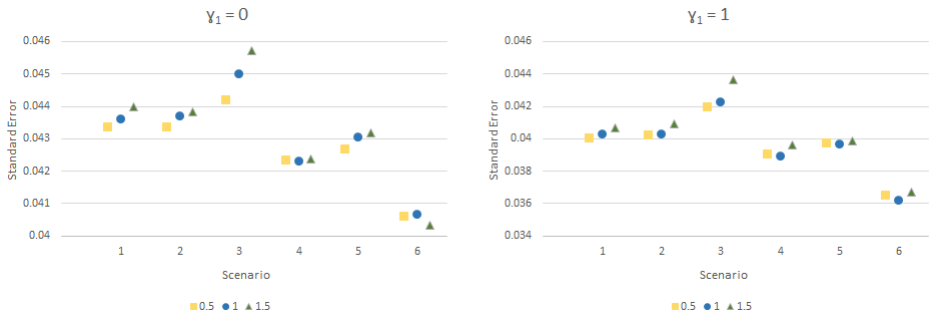

Figure 3.2: Standard error estimate of $\hat{\gamma}_1$ under sampling design setting 2A and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Uniform distribution.

parameter used throughout the analysis.

In sampling design setting 2B, we aim to determine the most efficient BSS case sampling scenario when there is approximate balance between cases and non-cases within the sample, the most efficient sampling scenario from sampling design setting 1. Using sampling scenario 4 of sampling design setting 1 (Table 2.1) and sampling design setting 2B (Table 2.3), we can compare SRS with BSS of the case set. Recall in sampling design setting 1 that the generalized case-cohort sampling design was employed; a SRS is selected from the case set and then from the cohort. Thus sampling scenario 4 from sampling design setting 1 is used as the SRS of the case set in the required comparison. Further, sampling scenario 4 is used for comparison since the proportion of cases initially selected into the sample is the same as the scenarios outlined in Table 2.3, $n_{cases} = 4000$. The sampling design efficiency may improve or degrade based on the BSS design employed, when compared to the SRS design.

For all of the sampling scenarios in sampling design setting 2B the proportion of cases to non-cases was consistent, approximately 4800 to 5200. This indicates change in standard error between sampling scenarios is not due to the proportion of cases to non-cases within the sample, but instead a result of the case strata sampling. Figure 3.3 displays the standard error of $\hat{\gamma}_1$ for all sampling scenarios in Table 2.3 and shape

parameter values. Sampling scenarios four, five and six result in the lowest standard error estimate. These three sampling scenarios correspond to selecting more cases from stratum 1 and 2 versus stratum 3, in other words selecting more from shorter time-to-event cases. A more efficient sampling design can be obtained by selecting more individuals from the high frequency areas of the distribution.

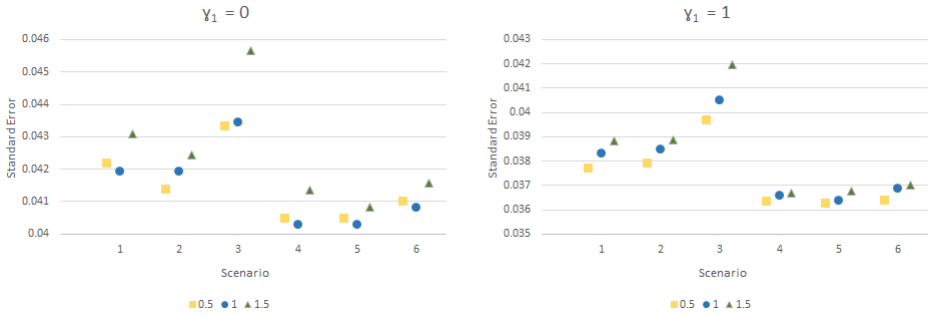

Figure 3.3: Standard error estimate of $\hat{\gamma}_1$ under sampling design setting 2B and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Uniform distribution.

The largest standard error for $\hat{\gamma}_1$ is produced when more cases are being sampled from stratum 3. Increased sampling from stratum 3 translates to increased sampling from the right tail of the Weibull distribution, a rather uninformative portion of the distribution.

3.2.3 Results under Sampling Design Setting 3

Employing the most efficient BSS design of the case set from sampling design setting 2, we aim to determine which non-cases within the cohort should be assigned a higher selection probability during phase II sampling. In Chapter 2, this is discussed as Objective 3 and will be explored using sampling design setting 3 as outlined in Section 2.3.

Firstly, to compare SRS with BSS of the non-cases, we compare sampling scenario 4 from sampling design setting 2B with the sampling design setting 3 in Table 2.4.

Sampling scenario 4 from sampling design setting 2B was the most efficient case BSS sampling scenario, in which a BSS is performed on the case sample and a SRS is collected from the cohort. The efficiency of the γ_1 estimate is affected when employing BSS versus SRS of the non-case set, either it will degrade or improve depending on the BSS design employed.

To determine the efficiency of the tested non-case BSS scenarios in Table 2.4, the standard error of $\hat{\gamma_1}$ is plotted in Figure 3.4. The least efficient sampling scenarios are 4 and 7, which have increased sampling from non-case strata 1 and 2. In general, as the number of observations sampled from the first stratum decreases, the design efficiency improves.

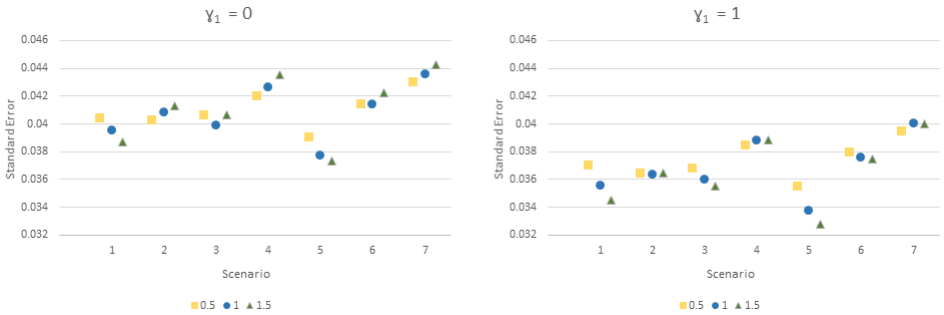

Figure 3.4: Standard error estimate of $\hat{\gamma_1}$ under sampling design setting 3 and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Uniform distribution.

For all γ_1 and shape parameter values explored, the minimum standard error occurs in sampling scenario 5. In sampling scenario 5, sampling from the third stratum is highest among the scenarios investigated and then is coupled with little sampling from stratum 1. Thus the design efficiency is improved when the non-cases within stratum 3 have been assigned a higher selection probability.

3.3 Exponential Censoring

Although the sampling procedure, as described in Section 3.1, is a response-dependent sampling design the sampling designs investigated in Chapter 2 under exponential censoring result in ML estimates that are unbiased. The figures in Appendix A.2 illustrate the ML estimates of γ_1, under exponential censoring, obtained through maximizing Eq. 3.1 are close to the true value.

3.3.1 Results under Sampling Design Setting 1

To determine the most efficient generalized case-cohort design under exponential censoring, the standard errors are plotted in Figure 3.5 for all Weibull shape parameter α values and sampling scenarios in Table 2.1. When $\gamma_1 = 0$, the efficiency of $\hat{\gamma_1}$ increases as the number of selected cases decreases, until sampling scenario 3 for all shape parameters investigated. When $\alpha = 0.5$ the efficiency increases again as the number of selected cases decreases to 4000. Thus the most efficient sampling design when $\gamma_1 = 0$ and $\alpha = 0.5$ is sampling scenario 4. When $\alpha = 1$ or $\alpha = 1.5$ the efficiency of the γ_1 estimate decreases as the number of cases selected decreases, after sampling scenario 3. Therefore sampling scenario 3 produces the most efficient γ_1 estimates when the shape parameter is equal to 1 or 1.5 and the true value of $\gamma_1 = 0$.

For $\gamma_1 = 1$, the standard error when $\alpha = 0.5$ or $\alpha = 1.0$ behaves similarly to what was observed when the censoring indicator was generated from a uniform distribution. The minimum standard error was observed under sampling scenario 4, thus the most efficient parameter estimate was obtained by this sampling design. The standard errors calculated for the other sampling scenarios steadily decrease until sampling scenario 4, then begin to increase moving from fourth to the fifth sampling scenario. The behaviour of the standard error when $\gamma_1 = 1$ and $\alpha = 1.5$ is slightly different than

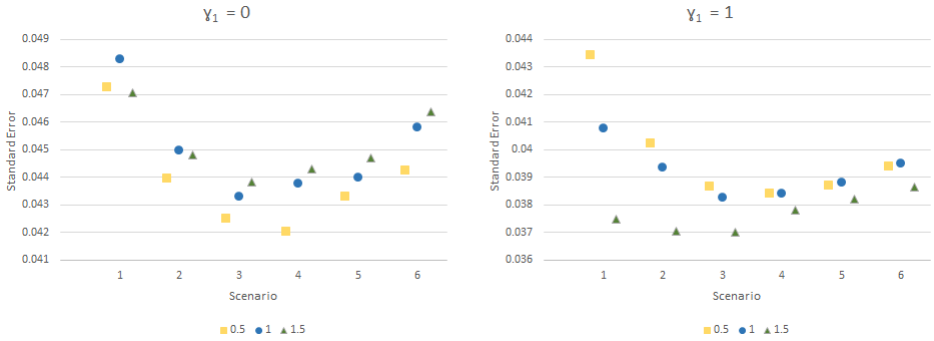

Figure 3.5: Standard error estimate of $\hat{\gamma}_1$ under sampling design setting 1 and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Exponential distribution.

observed under uniform censoring. The standard error for sampling scenarios 1, 2 and 3 are smaller then the standard errors for sampling scenarios 4, 5 and 6. Previously, under uniform censoring, sampling scenarios 1 and 2 were the least efficient parameter estimates, the standard errors were the highest among all scenarios. In Appendix C, the empirical conditional pdfs of the cases in the cohort and of the cases selected in phase II are shown for each shape parameter value investigated. In Appendix C.3, when $\alpha = 1.5$, it shows that the sample from the case set must be large enough to sufficiently sample cases from the extremes of the distribution.

3.3.2 Results under Sampling Design Setting 2

Comparing sampling scenario 2 from sampling design setting 1 with sampling scenarios 1 to 5 of sampling design setting 2A in Table 2.2, we conclude BSS of the case set does change the efficiency of the γ_1 estimate when compared with a SRS from the same case set. To determine which stratified sampling design is the most efficient, the standard error of $\hat{\gamma}_1$ for each sampling scenario and α value is presented in the Figure 3.6. When $\gamma_1 = 0$, the standard error patterns follow what was observed under uniform censoring. Within the first 5 sampling scenarios, the standard error is

minimized when we decrease the case sampling from stratum 3, few observations with long-term survival sampled. Combining this conclusion with the knowledge gained during sampling design setting 1, balance between case and non-case observations, sampling scenario 6 is presented. Sampling scenario 6 does in fact produce the most efficient γ_1 estimate.

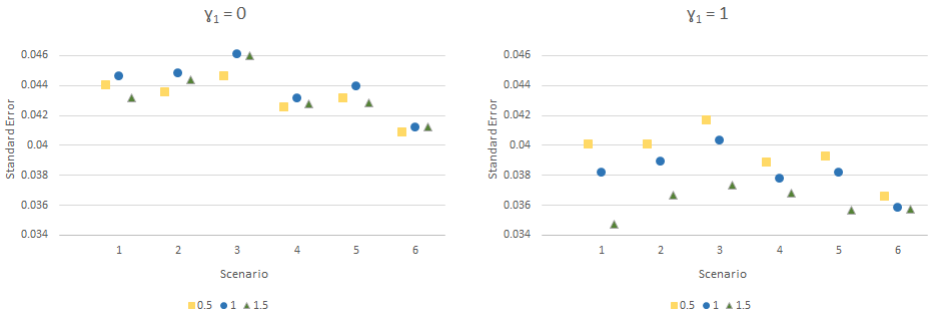

Figure 3.6: Standard error estimate of $\hat{\gamma}_1$ under sampling design setting 2A and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Exponential distribution.

For $\gamma_1 = 1$ when the α is equal to 0.5 or 1.0 our conclusion does not change. Increased sampling from the first two stratum produce more efficient estimates, and sampling scenario 6 obtains the most efficient estimates among the sampling scenarios given. When $\gamma_1 = 1$ and $\alpha = 1.5$, however, the standard error pattern is drastically different then what was observed under uniform censoring, and what was observed for the other shape parameters under exponential censoring. The standard error is now minimized when there is increased sampling from stratum 1 and stratum 3, therefore sampling scenario 1 results in the most efficient estimates. Decreasing the sampling from either stratum 1 or 3 increases the standard error, thus produces less efficient estimates. When the censoring mechanism is from the exponential distribution, sampling from the stratum 1 and 3 (the tails of the distribution) becomes increasingly important as the empirical conditional distribution for the cases becomes more symmetric when the shape parameter α is 1.5.

Sampling design setting 2B is used to compare SRS with BSS of the case sample and to determine the most efficient case BSS sampling scenario under the most efficient sampling scenario obtained in sampling design setting 1; cases and non-cases should be approximately balanced within the sample. The efficiency of $\hat{\gamma}_1$ is affected when BSS is employed instead of SRS of the case set, in particular an increase or decrease in efficiency occurs depending on the BSS performed. This conclusion was drawn by comparing sampling scenario 4 from sampling design setting 1 and all sampling scenarios in Figure 3.7 from sampling design setting 2B (Table 2.3).

When $\gamma_1 = 0$, the standard error of $\hat{\gamma}_1$ is minimized when there is increased sampling on stratum 1 and 2, sampling scenarios 4, 5, and 6 (Figure 3.7). When selecting heavily from stratum 3 there is an increase in the standard error, the estimates are less efficient. This agrees with the conclusions made in sampling design setting 2B under exponential censoring and when the censoring times were generated from a uniform distribution.

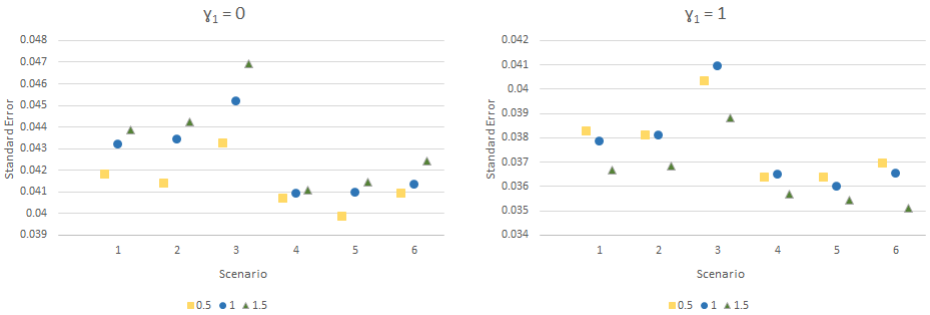

Figure 3.7: Standard error estimate of $\hat{\gamma}_1$ under sampling design setting 2B and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Exponential distribution.

The standard error of $\hat{\gamma}_1$ when $\gamma_1 = 1$ and the shape parameter α is 0.5 or 1.0 increases as sampling from stratum 1 decreases. For these two shape parameters increased sampling from stratum 1 and stratum 2 produces the most efficient estimates.

When the shape parameter is 1.5 however, the standard error decreases when sampling from stratum 1 and stratum 3 is increased, thus sampling scenario 6 is the most efficient sampling design.

3.3.3 Results under Sampling Design Setting 3

The standard error of $\hat{\gamma}_1$ for each sampling scenario in Table 2.4 and shape parameter investigated is plotted in Figure 3.8. When BSS of the non-case set is employed there is a change in $\hat{\gamma}_1$ efficiency when compared with SRS. The change in efficiency can be seen by comparing sampling scenario 4 in Figure 3.7 from sampling design setting 2B with the sampling design setting 3 sampling scenarios (Table 2.4) depicted in Figure 3.8.

The standard error is maximized in sampling scenario 7 when there is increased sampling from non-case stratum 1. Thus increased sampling of the short censoring times leads to a less efficient parameter estimate. Sampling scenario 5 is the most efficient design; the standard error is low when sampling from non-case stratum 3 is increased.

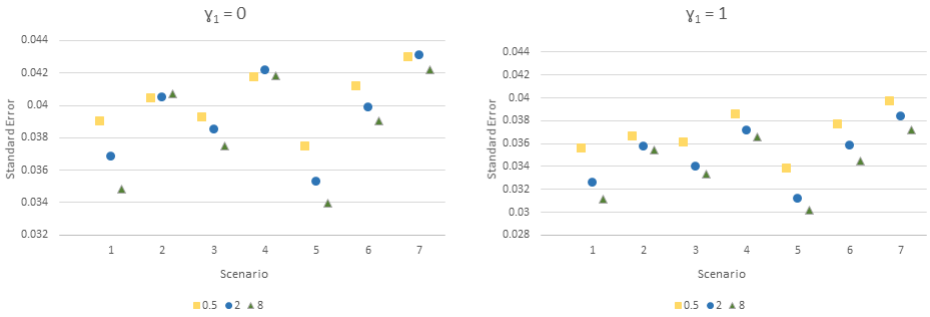

Figure 3.8: Standard error estimate of $\hat{\gamma}_1$ under sampling design setting 3 and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Exponential distribution.

Generally, the design efficiency improves as $n_{noncases_3}$ increases. In other words,

selecting longer censoring times improves the efficiency of γ_1 estimate. This agrees with conclusions for sampling design setting 3 with uniform censoring.

3.4 A Standard Monte Carlo Simulation Study

The simulation study performed is similar to the previously defined study in Section 3.1, with a few modifications. Instead of using a very large cohort size N and sample size n, we consider that the cohort size is $N = 5000$ and the phase II sample size is $n = 1000$. The data generation, sampling and estimation was replicated 1000 times for each set of parameters investigated ($\alpha = 0.5, 1, 1.5$ and $\gamma_1 = 0, 0.25$). This is a typical Monte Carlo simulation study, and it was conducted to see whether the previous simulation study based on a large cohort of data and a large sample size is valid to understand the efficiency of the sampling designs. Since the standard Monte Carlo simulation study with multiple replications is computationally expensive, one would prefer to use a simulation study with a large cohort.

The censoring mechanism used in the simulation study is from the uniform distribution with a censoring rate of 80%, thus there are approximately 1000 cases within the cohort, $N_{cases} \approx 1000$. Depending on the aim of the sampling designs investigated the proportion of cases to cohort observations selected is modified, or the case strata selection probabilities are changed.

The results capture the mean of the γ_1 coefficient estimate and the mean standard error estimates of $\hat{\gamma_1}$ for each shape parameter and sampling scenario considered over 1000 simulation replications. A summary of the mean standard error estimates of $\hat{\gamma_1}$ is provided in the following sections. The mean coefficient estimates were found to be very close to the true value.

3.4.1 Results under Sampling Design Setting 1

Sampling design setting 1 investigates the efficiency of the sampling scenarios when changing the proportion of case to cohort observations selected in phase II. Table 3.2 outlines the sampling scenarios investigated in this simulation study.

Table 3.2: Sampling Design Setting 1 Used for the Standard Monte Carlo Simulation Study

Sampling Scenario	n_{cases}	n_{cohort}
1	250	750
2	400	600
3	500	500
4	600	400
5	750	250

As previously seen, the true number of cases within the sample is made up of two components: the cases initially selected into sample, and the case observations that were in the cohort portion of the sample. Table 3.3 presents the mean number of cases and non-cases within the sample over 1000 replications for each sampling scenario.

Table 3.3: Mean Number of Cases versus Non-cases for Sampling Design Setting 1 over 1000 Replications

Sampling Scenario	Mean Number of Cases	Mean Number of Non-Cases
1	371	629
2	480	520
3	557	443
4	637	363
5	766	234

Figure 3.9 depicts the mean of the standard error estimates of $\hat{\gamma}_1$ over 1000 replications for each sampling scenario and shape parameter investigated. The standard

error is minimized for sampling scenario 2, in which 400 cases and 600 cohort observations were selected, and thus was the most efficient design. Sampling scenario 2 is the sampling design that is closest to balancing the number of cases to non-cases within the sample.

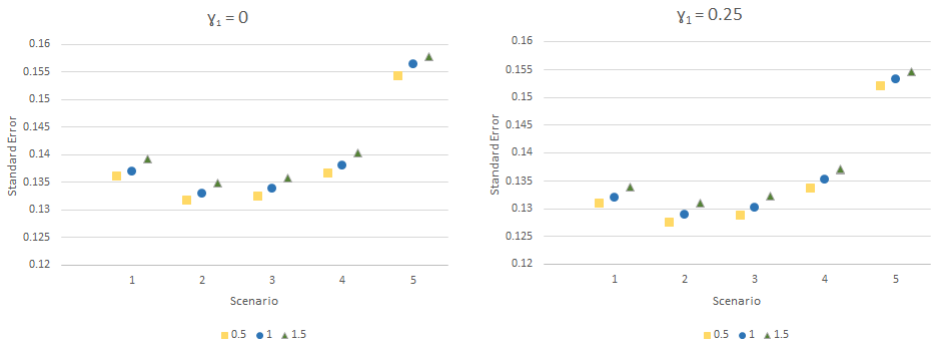

Figure 3.9: Mean standard error estimates of $\hat{\gamma}_1$ over 1000 replications under sampling design setting 1 and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Uniform distribution.

The result of this sampling design setting agrees with the results from the large sample $N = 50000$ simulation study in Section 3.2.1, the sampling design with approximate balance between cases and non-cases within the sample is the most efficient design.

3.4.2 Results under Sampling Design Setting 2

To determine the sampling probabilities for the cases that lead to the most efficient sampling design we stratify the case observations into 3 strata based on observed survival time and sample from each independently. There will be approximately $N_{cases_1} \approx N_{cases_3} \approx 300$ cases allocated to stratum 1 and 3, and $N_{cases_2} \approx 400$ cases to stratum 2. The proportion of cases selected from each of the strata is modified however $n_{cases} = 600$ is not altered. Table 3.4 shows the sampling scenarios performed for sampling design setting 2A.

Table 3.4: Sampling Design Setting 2A Used for the Standard Monte Carlo Simulation Study

Sampling Scenario	n_{cases_1}	n_{cases_2}	n_{cases_3}	n_{cohort}
1	300	0	300	400
2	200	200	200	400
3	300	100	200	400
4	100	200	300	400
5	300	200	100	400

Figure 3.10 depicts the mean standard error estimates of $\hat{\gamma}_1$ over 1000 replications for the scenarios investigated. From the figure notice that increased sampling from stratum 1 and 2 improves the design efficiency. Sampling scenario 4, in which sampling is increased from stratum 3 is, in fact, the least efficient sampling design. Sampling scenario 5 is the most efficient design with the lowest standard error and is the design in which case sampling is mostly done from stratum 1 and 2.

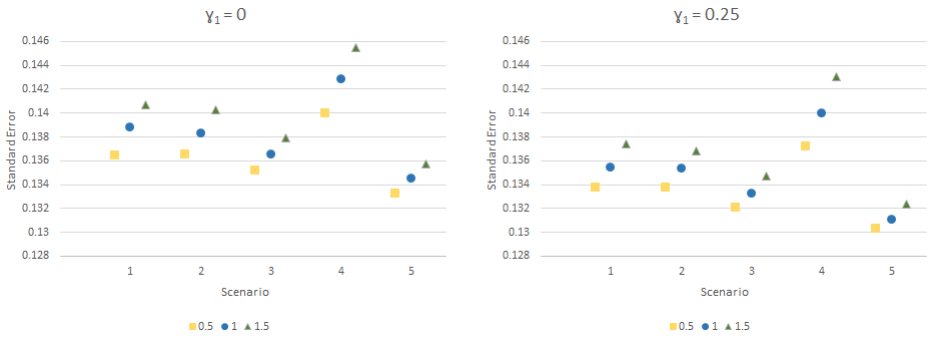

Figure 3.10: Mean standard error estimates of $\hat{\gamma}_1$ over 1000 replications under sampling design setting 2A and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Uniform distribution.

The most efficient design is achieved when sampling from stratum 1 and stratum 2 is increased, this agrees with the results obtained in the large sample $N = 50000$ simulation study in Section 3.2.2.

In sampling design setting 2B, the number of cases initially selected into the phase

II sample is $n_{cases} = 500$ and the proportion of cases sampled from each strata changes as provided in Table 3.5.

Table 3.5: Sampling Design Setting 2B Used for the Standard Monte Carlo Simulation Study

Sampling Scenario	n_{cases_1}	n_{cases_2}	n_{cases_3}	n_{cohort}
1	250	0	250	500
2	175	150	175	500
3	167	166	167	500
4	85	165	250	500
5	250	165	85	500

The mean standard error estimate of $\hat{\gamma}_1$ for each sampling scenario is plotted in Figure 3.11. We conclude that as sampling is increased from stratum 1 and 2, the standard error decreases. The most efficient sampling design is sampling scenario 5, in which sampling is mostly from stratum 1 and 2. The mean standard error estimate under sampling scenario 5 is also less than the mean standard error estimates in Figure 3.10. The least efficient sampling design is when sampling from the right side of the distribution is increased, sampling scenario 4.

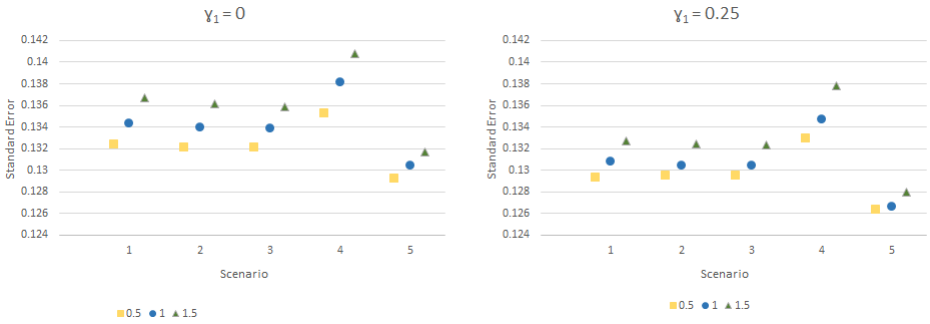

Figure 3.11: Mean of standard error estimates of $\hat{\gamma}_1$ over 1000 replications under sampling design setting 2B and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Uniform distribution.

Again we concluded that increased sampling from case stratum 1 and 2 improves the sampling design efficiency, which agrees with previous conclusions made in the

sampling design setting 2A of the standard Monte Carlo simulation study and the large sample $N = 50000$ simulation study.

The efficiency conclusions obtained in the standard Monte Carlo simulation study agree with those obtained in the large sample $N = 50000$ simulation study. Hence, performing a large sample simulation study is sufficient to understand the efficiency of sampling designs, and the framework applied in Section 3.1 could be safely used.

Chapter 4

Efficiency of Sampling Designs Under Mixture Cure Model

The mixture cure model (Eq. 1.11) considers that some individuals are statistically cured. Within this study, the survival times are from a mixture cure model which has a non-standard survival function. The aim is to determine the most efficient sampling designs when data is from the mixture cure model. In particular, the aim is to determine the proportion of cases versus cohort observations selected and to understand how the sampling probabilities should be assigned to members of the cohort, to obtain the most efficient case-cohort design.

The previously discussed simplified likelihood equation (Eq. 3.1) is modified in such a way to consider data from the mixture cure model:

$$L(\theta;q) = \prod_{i\in V}[p(x_i)f_o(t_i|x_i;\theta)]^{\delta_i}[p(x_i)+(1-p(x_i))S_o(t_i|x_i;\theta)]^{1-\delta_i}g(x_i;q)\times \prod_{i\in\overline{V}}\sum_{x=0}^{1}[p(x)f_o(t_i|x;\theta)]^{\delta_i}[p(x)+(1-p(x))S_o(t_i|x;\theta)]^{1-\delta_i}g(x;q), \quad (4.1)$$

where $p(x_i)$ is the probability of being cured, $f_o(t_i|x_i;\theta)$ is the standard pdf and $S_o(t_i|x_i;\theta)$ is the standard survival function for the susceptible group of individuals.

4.1 Simulation Study Set-up under Mixture Cure Model

To identify efficient two-phase sampling designs, following the aims in Chapter 2 under the mixture cure model (Eq. 1.11), based on the likelihood-based method, a large cohort of data was generated from the mixture cure model.

For each simulation study a large cohort of individuals, $N = 50000$, was generated. First for each individual i the expensive covariate, x_i, which is binary, was generated from the Bernoulli distribution with the probability of success 0.25. For each observation the probability of being cured $p(x_i)$ is calculated using Eq. 1.12. The probability of being cured has introduced two new parameters into the model, α_0 and α_1. These two parameters were defined in such a way that when the expensive covariate is equal to zero the probability of being cured is 0.5, and when the covariate is equal to one the probability of being cured is 0.6. Thus, we set $\alpha_0 = 0$ and $\alpha_1 = 0.405$. Using the probability of being cured, the cure status (Q_i) for each observation is generated using a Bernoulli distribution with probability of being cured, $p(x_i)$. If $Q_i = 1$, the observation is cured, otherwise $Q_i = 0$ and the observation is not cured.

The time-to-event value for a susceptible individual i, T_i, was generated from the Weibull distribution with survival function

$$S_o(t_i|x_i) = \exp(-e^{0.5+\gamma_1 x_i} t_i^{\alpha}). \tag{4.2}$$

Throughout the simulation study three values of the shape parameter (α) and

regression parameter (γ_1) are considered: $\alpha = 0.5, 1.0, 1.5$ and $\gamma_1 = 0, 0.25, 1.0$. The right censoring time (C_i) was generated from Uniform(a,b) where a and b are selected in such a way that the censoring rate is approximately 70% and there is no early censoring.

Each individual has a censoring indicator (δ_i) and observation time (t_i). All cured individuals, $Q_i = 1$, are censored ($\delta_i = 0$) thus $t_i = C_i$. Susceptible individuals with $Q_i = 0$ are censored if $C_i < T_i$, otherwise $\delta_i = 1$ when $C_i \geq T_i$. The observed time for susceptible individuals is obtained by $t_i = min(T_i, C_i)$.

The case-cohort sampling designs were performed on the simulated dataset are outlined in Section 4.2. Parameter estimates were obtained through MLE by maximizing the likelihood function in Eq. 4.1. The nlm function that is built into the R software environment was used to obtain the estimates for the parameters and their corresponding standard errors.

The sampling designs were compared based on the efficiency of the coefficient estimate of the expensive covariate in both the probability of being cured model (Eq. 1.12) and the Weibull survival model for the susceptible individuals (Eq. 4.2). Thus we want to minimize the standard error estimates of the ML estimate of γ_1 and α_1, $\hat{\gamma_1}$ and $\hat{\alpha_1}$.

4.2 Sampling Designs Settings for Mixture Cure Model

For the mixture cure model we will be investigating the efficiency of both regression coefficient estimates $\hat{\gamma_1}$ and $\hat{\alpha_1}$. We would like to achieve optimal efficiency for both parameter estimates, thus the most efficient sampling design described within our investigation may balance the efficiency of the two parameter estimates.

4.2.1 Mixture Cure Model Sampling Design Setting 1

Mixture cure model sampling design setting 1 investigates the proportion of case observations versus cohort observations that should be selected in a generalized case-cohort design where a SRS is performed on the cases and the cohort given a phase II sample size n.

Table 4.1: Mixture Cure Model Sampling Design Setting 1

Sampling Scenario	n_{cases}	n_{cohort}
1	9000	1000
2	8000	2000
3	7000	3000
4	6000	4000
5	5000	5000
6	4000	6000
7	3000	7000
8	2000	8000

There are 8 sampling scenarios considered within this design setting (Table 4.1) when $n = 10000$. Each sampling scenario is applied to the same generated cohort, thus allowing comparison. The percentage of cases selected into the sample decreases by 10% between each sampling scenario. Recall, the total number of cases n^*_{cases} selected in the phase II is a combination of the n_{cases} sampled and cases selected during cohort sampling (Eq. 2.1). The total number of cases within the sample n^*_{cases} will be greater than or equal to the cases initially selected into the sample n_{cases}.

4.2.2 Mixture Cure Model Sampling Design Setting 2

For mixture cure model sampling design setting 2 the focus is to determine the most efficient sampling scenario when the case set is strategically sampled using BSS. As discussed previously, in standard survival model sampling design setting 2 in Chapter

2, the case observations are stratified into three mutually exclusive groups. There are approximately 15000 case observations within the cohort, thus the stratification is a modification of what was previously seen. The allocation to each case stratum is: $N_{cases_1} \approx 3500$, $N_{cases_2} \approx 8000$ and $N_{cases_3} \approx 3500$.

Table 4.2: Mixture Cure Model Sampling Design Setting 2

Sampling Scenario	n_{cases_1}	n_{cases_2}	n_{cases_3}	n_{cohort}
1	3500	0	3500	3000
2	3500	1000	2500	3000
3	2333	2334	2333	3000
4	1000	2500	3500	3000
5	1633	3734	1633	3000
6	3500	2500	1000	3000

Table 4.2 describes the sampling scenarios investigated. The phase II sample size n is set to 10000. In the first two sampling scenarios in Table 4.2, case observations with short or long survival times (stratum 1 and 3) will be sampled heavily during the case sampling. For sampling scenario 3, an equal number of cases will be selected from each of the three defined case strata. In sampling scenario 4 the case observations with long survival times, case stratum 3, are sampled more than the other case strata. In contrast sampling from stratum 1, the short survival times, is increased in sampling scenario 6. Sampling scenario 5 is equivalent to performing a SRS on the case set. The cohort sample will always be collected through SRS and n_{cohort} will remain the same across all scenarios. Recall, the total number of cases sampled from the j^{th} strata $n^*_{cases_j}$ is the combination of the cases initially selected into the sample from strata j and cases selected during cohort sampling (Eq. 2.3).

4.2.3 Mixture Cure Model Sampling Design Setting 3

The focus of this sampling design setting is to determine how to best select the non-case observations in an informative matter, either through SRS or BSS. For this analysis we will be selecting a BSS sample from the case set and a BSS sample from the non-case set; these sets are mutually exclusive. Non-cases are stratified into three strata based on censoring time as described in Section 2.3. The defined stratum cutoffs, L_{CN} and U_{CN}, result in: $N_{noncases_1} \approx 6000$, $N_{noncases_2} \approx 23000$ and $N_{noncases_3} \approx 6000$. A sample of size $n_{noncases_h}$ $(h = 1, 2, 3)$ is selected from the h^{th} non-case stratum where $\sum_{h=1}^{3} n_{noncases_h} = n_{noncases}$.

Within Table 4.3, sampling scenarios 2-8 describe a SRS selected from the case set ($n_{cases} = n^*_{cases} = 7000$) and a BSS selected from the non-case set ($n_{noncases} = 3000$). Efficiency conclusions comparing different non-case stratum sampling probabilities can be made using these 7 scenarios. In sampling scenario 2 all sampled non-cases are selected from stratum 3, observations with long censoring times. For sampling scenarios 2 to 5, as the sampling scenario number increases, the number of non-cases sampled from the third non-case stratum decreases. In sampling scenarios 6 to 8 there is no sampling performed from the non-case stratum 3 and the number of observations selected from non-case stratum 2 decreases as the sampling scenario increases. Sampling scenario 8 corresponds to the design where non-cases are only selected from stratum 1, the short censoring times. Sampling scenario 1 is included to assess the efficiency of γ_1 and α_1 when BSS is employed to the non-cases under the most efficient case BSS sampling scenario found in sampling design setting 2 (Table 4.2).

To further explore BSS of cases and non-cases within the cohort, mixture cure model sampling design setting 3B is presented in Table 4.4. For all 8 sampling scenarios, the most efficient BSS sampling scenario of the case set, as determined in

Table 4.3: Mixture Cure Model Sampling Design Setting 3A

Sampling Scenario	$n^*_{cases_1}$	$n^*_{cases_2}$	$n^*_{cases_3}$	$n_{noncases_1}$	$n_{noncases_2}$	$n_{noncases_3}$
1	3500	0	3500	0	0	3000
2	1633	3734	1633	0	0	3000
3	1633	3734	1633	0	1500	1500
4	1633	3734	1633	1500	0	1500
5	1633	3734	1633	515	1970	515
6	1633	3734	1633	0	3000	0
7	1633	3734	1633	1500	1500	0
8	1633	3734	1633	3000	0	0

sampling design setting 2 (Table 4.2); 3500 cases will be selected from each case stratum 1 and 3, and zero cases are selected from the second case stratum. The proportion of non-cases selected from each of the non-case strata is modified throughout the 8 sampling scenarios. In general, as the sampling scenario increases the number of non-cases selected from the left side of the distribution increases; in other words sampling observations with short censoring times increases as the sampling scenario increases.

Table 4.4: Mixture Cure Model Sampling Design Setting 3B

Sampling Scenario	$n^*_{cases_1}$	$n^*_{cases_2}$	$n^*_{cases_3}$	$n_{noncases_1}$	$n_{noncases_2}$	$n_{noncases_3}$
1	3500	0	3500	0	0	3000
2	3500	0	3500	0	1500	1500
3	3500	0	3500	1500	0	1500
4	3500	0	3500	515	1970	515
5	3500	0	3500	1000	1000	1000
6	3500	0	3500	0	3000	0
7	3500	0	3500	1500	1500	0
8	3500	0	3500	3000	0	0

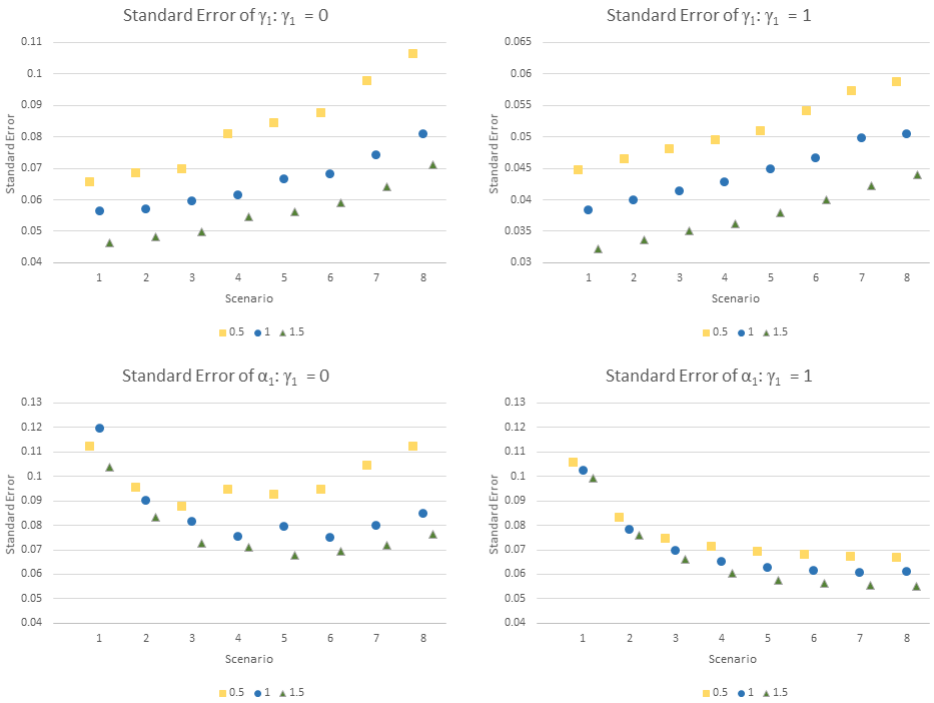

Figure 4.1: Standard error estimate of $\hat{\gamma}_1$ (upper panel) and $\hat{\alpha}_1$ (lower panel) under sampling design setting 1 for the mixture cure model.

4.3 Results of the Simulation Study

4.3.1 Results under Mixture Cure Model Sampling Design Setting 1

In sampling design setting 1 we consider the generalized case-cohort design when data is from the mixture cure model. In particular, we aim to understand the efficiency of $\hat{\gamma}_1$ and $\hat{\alpha}_1$ based on the proportion of case to cohort observations selected into the phase II sample. Table 4.1 contains the sampling scenarios used in this investigation.

In general, for $\hat{\gamma}_1$, the standard error increases as the number of cases initially selected into the sample decreases (Figure 4.1). Thus sampling scenario 1 in Table 4.1 results in the most efficient γ_1 parameter estimate among those scenarios investigated.

As shown in Figure 4.1, the standard error estimate of $\hat{\alpha}_1$ decreases among the

first three sampling scenarios. Following sampling scenario 3, the standard error estimate of $\hat{\alpha_1}$ either increases or stays relatively stable depending on the Weibull shape parameter and the true value of γ_1. Therefore, sampling scenario 3 is selected as the most efficient sampling design to approximately balance the efficiency of the γ_1 parameter with the efficiency of α_1 parameter. Hence, moving forward we will use sampling scenario 3 ($n_{cases} = 7000$) as the base design for more complex designs.

4.3.2 Results under Mixture Cure Model Sampling Design Setting 2

Now consider an extended version of the generalized case-cohort design in which BSS is performed on the case set. We wish to determine, during phase II sampling, which case observations should be assigned a higher selection probability so that more efficient estimates of γ_1 and α_1 are obtained. The sampling scenarios used in this investigation are outlined in Table 4.2.

Firstly, to compare SRS with BSS of the case set, we compare sampling scenario 5 with the remaining sampling scenarios within this sampling design setting 2. Figure 4.2 shows that the efficiency of $\hat{\gamma_1}$ does increase or decrease when BSS is employed, depending on the BSS design setting. Conversely, there is little to no effect on the efficiency of $\hat{\alpha_1}$ when BSS is performed.

Generally, the efficiency of $\hat{\gamma_1}$ decreases as the number of cases sampled from stratum 2 increases (Figure 4.2). An exception to this is sampling scenario 6 has a smaller n_{cases_2} when compared with sampling scenario 5, but the efficiency has decreased. Between sampling scenarios 5 and 6 there is a significant reduction in n_{cases_3} which could lead to the loss in efficiency. Sampling scenario 1 resulted in the most efficient γ_1 estimate. This sampling scenario selects only from the extremes in the case set.

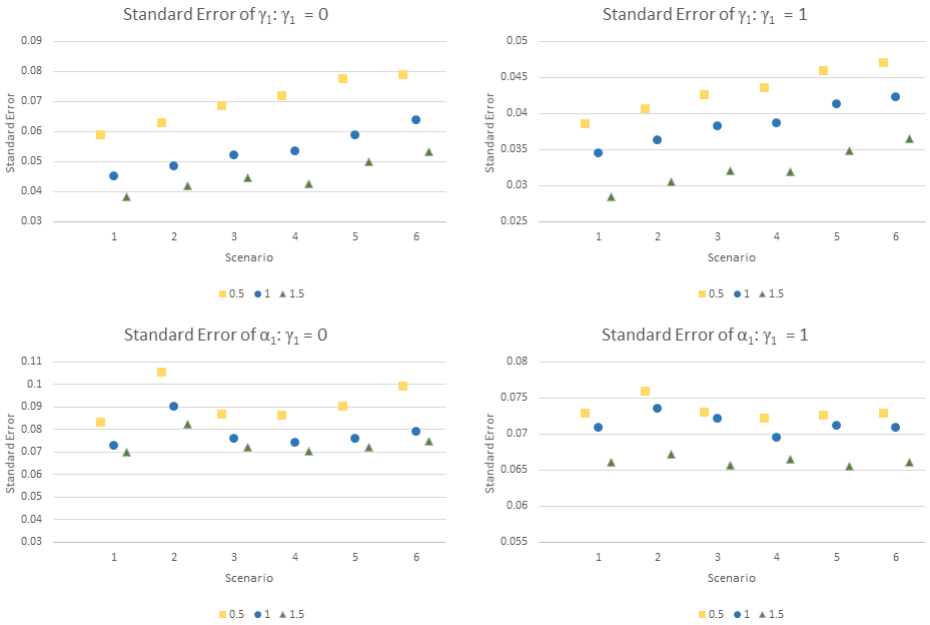

Figure 4.2: Standard error estimate of $\hat{\gamma_1}$ (upper panel) and $\hat{\alpha_1}$ (lower panel) under sampling design setting 2 for the mixture cure model.

The efficiency of $\hat{\alpha_1}$ throughout the sampling scenarios does not drastically change as shown in Figure 4.2. Sampling scenarios 2 and 6 (sampling from case stratum 3 is decreased) are slightly less efficient then the remaining sampling scenarios particularly when the true value of γ_1 is zero. Additionally, when $\gamma_1 = 0$ sampling scenario 1 is slightly more efficient for α_1.

In conclusion, under sampling design setting 2 the most efficient design for both γ_1 and α_1 is sampling scenario 1, in which sampling is performed only on case stratum 1 and 3 ($n_{cases_1} = 3500$, $n_{cases_2} = 0$ & $n_{cases_3} = 3500$). Thus, when completing the case sampling phase the probability of sampling from case stratum 2 will be zero.

4.3.3 Results under Mixture Cure Model Sampling Design Setting 3

Again consider two-phase response-dependent sampling, in which phase II sampling is dependent on the censoring indicator as well as the time-to-event variable. In particular, the design setting aims to determine how non-case sampling probabilities should be assigned to obtain the most efficient parameter estimates. Table 4.3 and 4.4 contain the sampling scenarios considered.

In Figure 4.3, sampling scenarios 2 through 8 given in Table 4.3, outline the changes in efficiency due to modification of the sampling probabilities of the non-cases; the cases are selected through SRS. To compare SRS with BSS of the non-cases, compare sampling scenario 5 (SRS of non-cases) with the remaining sampling scenarios (BSS of non-cases). The efficiency of $\hat{\gamma_1}$ does not change when BSS is performed, however the efficiency of $\hat{\alpha_1}$ improves or degrades when BSS is performed. The standard error of $\hat{\alpha_1}$ increases as the sampling from non-case stratum 3 decreases. Further still, once $n_{noncases_3} = 0$ the standard error continues to increase as sampling from non-case stratum 2 decreases (i.e. sampling non-case stratum 1 is increased).

Increased sampling from the third non-case stratum (long censoring times) improves the efficiency of $\hat{\alpha_1}$ which is the ML estimate of the coefficient of the expensive covariate in the model of the probability of being cured. Stratified sampling of the non-cases does not greatly influence the $\hat{\gamma_1}$ efficiency. From Figure 4.3, sampling scenario 1 is the most efficient design for both of the parameters among those scenarios investigated. Sampling scenario 1 combines the most efficient case stratification design determined in sampling design setting 2, with the most efficient non-case stratification sampling design.

Sampling design setting 3B only considers sampling scenarios in which the most

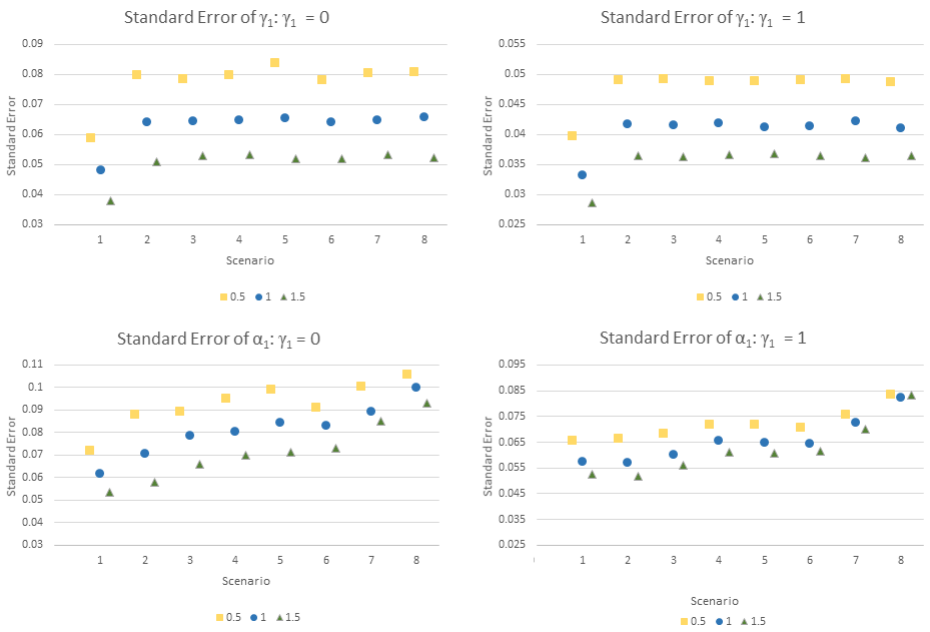

Figure 4.3: Standard error estimate of $\hat{\gamma}_1$ (upper panel) and $\hat{\alpha}_1$ (lower panel) under sampling design setting 3A for the mixture cure model.

efficient case stratified sampling design is employed. The difference between all of the sampling scenarios investigated is the change in non-case stratum sampling probabilities. Again, sampling scenario 5 is the SRS of the non-case set. As before, the efficiency of only $\hat{\alpha}_1$ will either increase or decrease depending on the BSS design performed. The efficiency of $\hat{\gamma}_1$ remains unchanged regardless of the stratified sampling design completed on the non-case set. The $\hat{\alpha}_1$ efficiency improves as sampling from non-case stratum 3 increases (Figure 4.4).

In conclusion, the sampling design which results in the most efficient parameter estimates, samples heavily from case strata 1 and 3, as well as heavy sampling from the longer censoring times (non-case stratum 3).

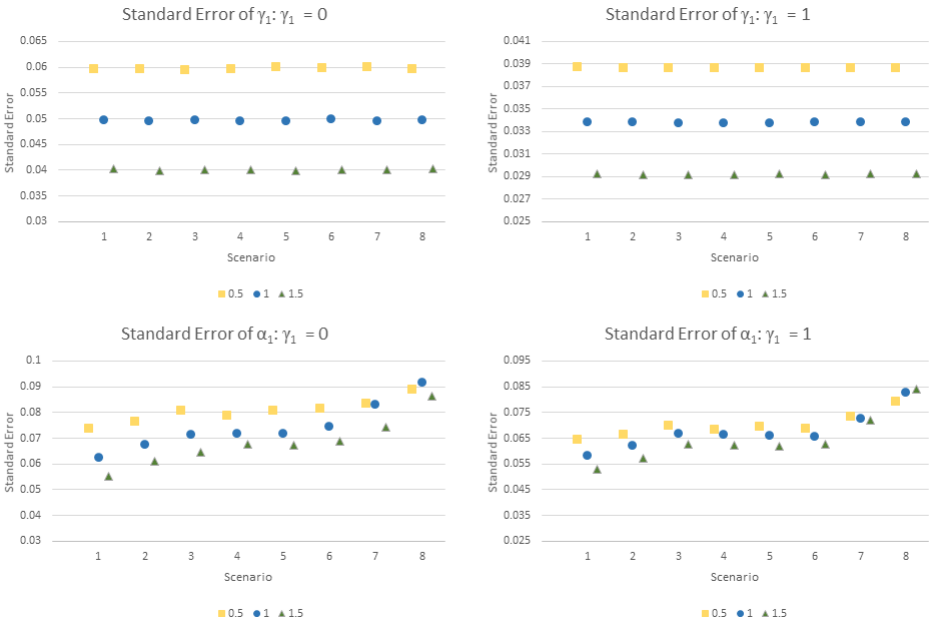

Figure 4.4: Standard error estimate of $\hat{\gamma}_1$ (upper panel) and $\hat{\alpha}_1$ (lower panel) under sampling design setting 3B for the mixture cure model.

Chapter 5

Summary and Conclusions

5.1 Summary

This study was motivated by important problems in biomedicine and genetics, particularly in genetic association studies. For example, in genetic association studies, survival data and inexpensive covariates can be collected for all individuals in a cohort but budgetary constraints prevent genotyping all individuals within the cohort (Huang and Lin, 2007; Lin et al. 2013; Lin, 2014). Two-phase response-dependent sampling designs are cost-efficient and consider observations with incomplete data. During phase I, all members of the cohort are measured for the response variable and the inexpensive covariates. In phase II, observations are sampled based on the response variable collected, and measured for the expensive covariate. In particular, throughout the study we focused on a time-to-event response variable and case-cohort sampling designs; during Phase II a random sample is selected from the case set and the cohort. Firstly, when the phase II sample size n is fixed, the objective was to determine the proportion of case versus cohort observations that should be selected into the Phase II sample, resulting in efficient parameter estimates. The efficiency of

parameter estimates are improved by selecting a more informative sample, thus sampling based on censoring status as well as observed survival time was considered. In particular, strategic sampling of the cases and the non-cases was performed through BSS to assess the design efficiency. Cases and non-cases were each stratified into three groups based on observed survival time. Each of the defined strata were sampled from independently.

There are two main methods of analysis for case-cohort sampling designs: the pseudo-likelihood based approach and likelihood based approaches. The pseudo-likelihood based approaches are generally less efficient than the likelihood-based approaches and cannot be used when the phase II sampling probability for an individual is near to zero. Likelihood based approaches can be divided into two main approaches: the conditional likelihood based approach (conditional on being selected into the sample) and the full likelihood based approach. We employed the full likelihood based approach which considers all individuals within the cohort regardless of whether they were sampled in phase II. The ML estimates obtained through the likelihood based approach are consistent, asymptotically efficient and normally distributed.

The framework, the large sample $N = 50000$ simulation study, used to determine the efficient sampling design within this study was applied under certain assumptions. In a different study with different assumptions this framework can be employed to obtain the efficient sampling designs.

5.2 Conclusions

Within this study, the efficiency of various types of case-cohort sampling designs based on likelihood methods were explored under different survival models: PH model, AFT model and the mixture cure model. In particular we focused on the Weibull survival

model, one of the most common survival models, which is both an AFT and a PH model.

For the standard survival model we concluded, under Uniform and Exponential censoring, selecting a sample that approximately balances the number of cases with the non-cases leads to a more efficient design. Further, applying BSS to the cases and non-cases can improve the efficiency of the coefficient estimate of the expensive covariate. The most efficient sampling design, among those explored in the analysis, selects an increased number of case observations with short survival times and samples more non-cases with long censoring times.

In Chapter 4 we extended the analysis to the mixture cure model which accounts for individuals that are statistically cured. For the mixture cure model, we were interested in balancing the efficiency of two parameters; the coefficients of the expensive covariate γ_1 and α_1 in the survival model for susceptible individuals and the model for the probability of being cured, respectively. The efficiency of the two coefficient estimates was approximately balanced when selecting more from the cases such that $n_{cases} = 7000$ and $n_{cohort} = 3000$ when $n = 10000$ is given. When employing BSS on the case set the efficiency of $\hat{\gamma_1}$ was affected whereas the efficiency of $\hat{\alpha_1}$ was unaffected. Selecting more cases with short- and long-survival times, the extremes of the distribution, improved the design efficiency. The efficiency of $\hat{\alpha_1}$ was affected when BSS was performed on the non-cases. In particular, selecting more non-cases with long censoring times lead to more efficient estimates of α_1. The efficiency of γ_1 was unaffected by BSS of the non-cases.

5.3 Recommendations and Future Work

The general framework purposed within this thesis may be extended to consider alternative situations. The design efficiency conclusions may change if the performed analysis considers other survival models, parameter values, censoring distributions or rates, and choice of the cutoff points when determining the strata in the BSS setting. Further, datasets with multiple response variables may also be considered by extending this framework.

Within this study there was a single covariate considered, the expensive covariate, and the strata were defined through the response variable (censoring status and observed time). The proposed framework could be extended to include inexpensive covariates. The design efficiency can additionally be improved when the stratification is also based on an inexpensive covariate, collected for all individuals within the cohort, that is highly correlated with the expensive covariate (Borgan, et al.,2000; Nan et al., 2006).

Finally, the efficiency of the sampling designs were assessed when likelihood-based approaches were used. Efficient sampling designs under pseudo-likelihood based approaches should be explored and the conclusions compared with those found within this thesis.

Bibliography

[1] Barlow, W. E., Ichikawa, L., Rosner, D., & Izumi, S. (1999). *Analysis of case-cohort designs*. Journal of Clinical Epidemiology, 52(12), 1165-1172.

[2] Borgan, O., Langholz, B., Samuelsen, S. O., Goldstein, L., & Pogoda, J. (2000). *Exposure stratified case-cohort designs*. Lifetime Data Analysis, 6(1), 39-58.

[3] Breslow, N. E., & Cain, K. C. (1988). *Logistic regression for two-stage case-control data*. Biometrika, 75(1), 11-20.

[4] Breslow, N. E., & Chatterjee, N. (1999). *Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis*. Journal of the Royal Statistical Society: Series C 48(4), 457-468.

[5] Breslow, N. E., & Holubkov, R. (1997). *Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data*. Statistics in Medicine, 16(1), 103-116.

[6] Breslow, N., McNeney, B., & Wellner, J. (2003). *Large sample theory for semiparametric regression models with two-phase, outcome-dependent sampling*. Annals of Statistics, 31, 1110-1139.

[7] Breslow, N., Robins, J., & Wellner, J. (2000). *On the semiparametric efficiency of logistic regression under case-control sampling*. Bernoulli 5: 447-455.

[8] Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., & Kulich, M. (2009). *Improved HorvitzThompson estimation of model parameters from two-phase stratified samples: Applications in Epidemiology*. Statistics in Biosciences, 1(1), 32-49.

[9] Breslow, N. E., Amorim, G., Pettinger, M. B., & Rossouw, J. (2013). *Using the whole cohort in the analysis of case-control data*. Statistics in Biosciences, 5(2), 232-249.

[10] Cai, J., & Zeng, D. (2007). *Power calculation for case-cohort studies with nonrare events*. Biometrics, 63(4), 1288-1295.

[11] Carroll, R., Wang, S., & Wang, C. (1995). *Prospective analysis of logistic case-control studies*. Journal of the American Statistical Association, 90, 157-159.

[12] Chatterjee, N., Chen, Y-H, & Breslow, N. (2003). *A pseudoscore estimator for regression problems with two-phase sampling*. Journal of the American Statistical Association, 98, 158-168.

[13] Chen, K. (2001) *Generalized case-cohort sampling*. Journal of Royal Statistical Society: Series B, 63, 791-809.

[14] Cox, D. R. (1975). *Partial likelihood*. Biometrika, 62, 269-76.

[15] Ding, J., Zhou, H., Liu, Y., Cai, J., & Longnecker, M. P. (2014). *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.

[16] Ding, J., Lu, T., Cai, J., & Zhou, H. (2015). *Recent progresses in outcome-dependent sampling with failure time data*. Lifetime Data Analysis, 1-26.

[17] Flanders, W. D., & Greenland, S. (1991). *Analytic methods for two-stage case-control studies and other stratified designs*. Statistics in Medicine,10(5), 739-747.

[18] Greenland, S. (1988). *Statistical uncertainty due to misclassification: Implications for validation substudies*. Journal of Clinical Epidemiology, 41(12), 1167-1174.

[19] Huang, B.E., & Lin, D-Y (2007). *Efficient association mapping of quantitative trait loci with selective genotyping*. American Journal of Human Genetics, 80, 567-576.

[20] Langholz, B., & Thomas, D. C. (1990). *Nested case-control and case-cohort methods of sampling from a cohort: A critical comparison*. American Journal of Epidemiology, 131(1), 169-176.

[21] Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data (2nd ed.)*. New York: Wiley.

[22] Lawless, J.F. (2016). *Two-phase outcome-dependent studies for failure times and testing for effects of expensive covariates*. Manuscript.

[23] Lawless, J. F., Kalbfleisch, J. D., & Wild, C. J. (1999). *Semiparametric methods for response-selective and missing data problems in regression*. Journal of the Royal Statistical Society, 61(2), 413-438.

[24] Lin, D.-Y. (2014). *Survival analysis with incomplete genetic data*. Lifetime Data Analysis, 20, 16-22.

[25] Lin, D.-Y., Zeng D., & Tang Z.-Z. (2013). *Quantitative trait analysis in sequencing studies under trait-dependent sampling*. Proceedings of the National Academy of Science, 110, 12247-12252.

[26] McNamee, R. (2005). *Optimal design and efficiency of two-phase case-control studies with error-prone and error-free exposure measures*. Biostatistics, 6(4), 590-603.

[27] Morara, M., Ryan, L., Houseman, A., & Strauss, W. (2007) *Optimal design for epidemiological studies with error-prone and error-free exposure measures*. Biostatistics, 6, 583-605.

[28] Neyman, J. (1938). *Contribution to the theory of sampling human populations*. Journal of the American Statistical Association, 33, 101-116.

[29] Nan, B., Yu, M., & Kalbfleisch, J. D. (2006). *Censored linear regression for case-cohort studies*. Biometrika, 93(4), 747-762.

[30] Onland-Moret, N. C., van der A, D. L., van der Schouw, Y. T., Buschers, W., Elias, S. G., van Gils, C. H., Koerselman, J., Roest, M., Grobbee, D. E., & Peeters, P. H. M. (2007). *Analysis of case-cohort data: A comparison of different methods*. Journal of Clinical Epidemiology, 60(4), 350-355.

[31] Palmgren, J. (1987). *Precision of double sampling estimators for comparing two probabilities*. Biometrika, 74(4), 687-694.

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

[33] Robins, J.M., Rotnitzky, A., Zhao, L.P., & Liplitz, S. (1994). *Estimation of regression coefficients when some regressors are not always observed*. Journal of the American Statistical Association, 89, 846-866.

[34] Rubin, D. B. (1976). *Inference and missing data*. Biometrika, 63(3), 581-592.

[35] Scott, A. J., & Wild, C. J. (2001a). *Case-control studies with complex sampling*. Journal of Royal Statistical Society C, 50, 389-401.

[36] Scott, A. J., & Wild, C. J. (2001b). *Maximum likelihood for generalized case-control studies*. Journal of Statistical Planning and Inference, 96, 3-27.

[37] Scott, A. J., & Wild, C. J. (2011). *Fitting regression models with response-biased samples*. Canadian Journal of Statistics, 39, 513-536.

[38] Thomas, D.C. (1977). *Addendum to Methods for cohort analysis: appraised by application to asbestos mining by Liddell, McDonald and Thomas*. Journal of Royal Statistical Society A, 140, 469-419.

[39] Wacholder, S. (1991). *Practical considerations in choosing between the case-cohort and nested case-control designs*. Epidemiology, 2(2), 155-158.

[40] Wacholder, S., & Weinberg, C. R. (1994). *Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling*. Biometrics, 50(2), 350.

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

[42] Zeng, D., & Lin, D. Y. (2014). *Efficient estimation of semiparametric transformation models for two-Phase cohort studies*. Journal of the American Statistical Association, 109(505), 371-383.

[43] Zhao, L.P., & Lipsitz, S. (1992). *Designs and analysis of two-stage studies*. Statistics in Medicine, 11, 769-782.

Appendix A

Estimated γ_1 and α_1 Plots

A.1 Uniform Censoring

A.1.1 Sampling Design Setting 1

For each scenario in Table 2.1 and each shape parameter value the estimated γ_1 and the 95% confidence interval for γ_1 was plotted in Figure A.1. There are a couple of instances where the 95% confidence interval does not include the true value of the parameter. For these few instances the 99% confidence interval was calculated and the true parameter value was always included in the interval. Hence, $\hat{\gamma_1}$ under each sampling scenario is close to the true value of γ_1.

A.1.2 Sampling Design Setting 2

For each scenario in Table 2.2 and each shape parameter value the estimated γ_1 was plotted and the 95% confidence interval for γ_1 was displayed (Figure A.2). Most of the 95% confidence intervals contained the true value of the parameter. For instances

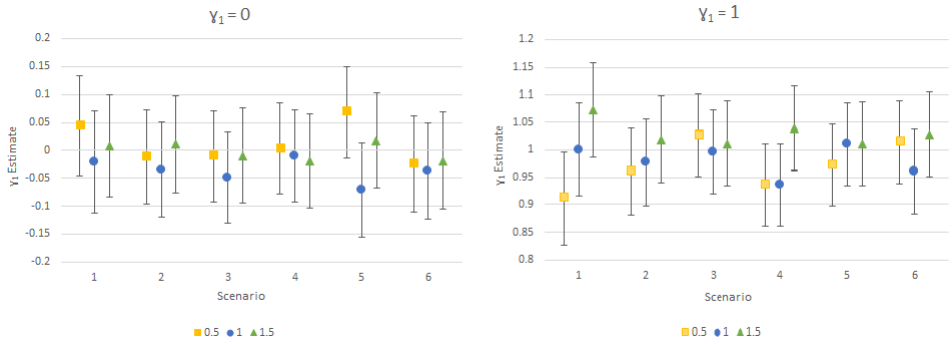

Figure A.1: Estimated γ_1 under sampling design setting 1 and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Uniform distribution.

in which the true value did not fall within the 95% confidence interval, the 99% confidence interval was calculated. The true value always fell within the 99% confidence interval, thus the estimated γ_1 adequately captures the true value.

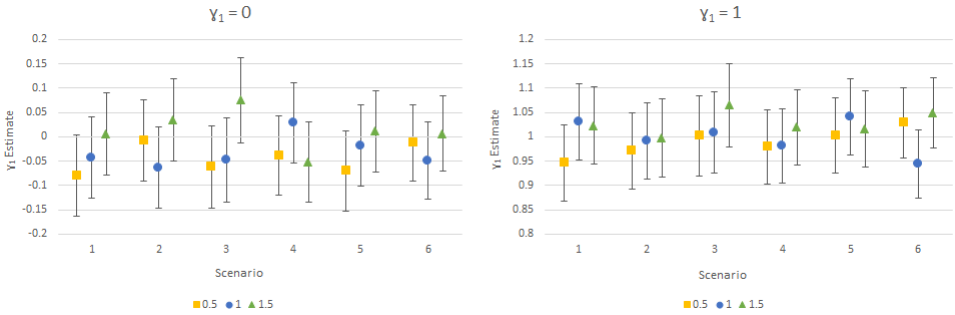

Figure A.2: Estimated γ_1 under sampling design setting 2A and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Uniform distribution.

For sampling design setting 2B, the estimated γ_1 values for all sampling scenarios in Table 2.3, and shape parameter values, were plotted with the corresponding 95% confidence interval in Figure A.3. Most of the 95% confidence intervals contained the true value of γ_1 and all 99% confidence intervals contained the true value of γ_1. Thus the ML estimate of γ_1 is adequately capturing the true γ_1 value.

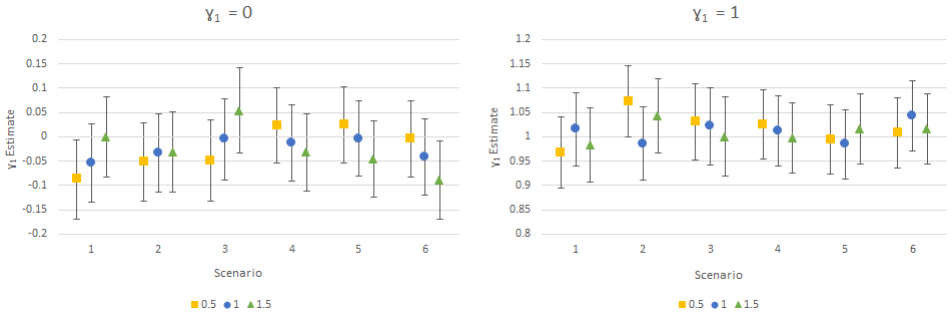

Figure A.3: Estimated γ_1 under sampling design setting 2B and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Uniform distribution.

A.1.3 Sampling Design Setting 3

Figure A.4 contains the estimated γ_1 values under the sampling scenarios outlined in Table 2.4. The most efficient BSS design for the cases is applied to all scenarios, and the sampling probabilities of the non-case strata are being modified. All 99% γ_1 confidence intervals contain the true value of γ_1.

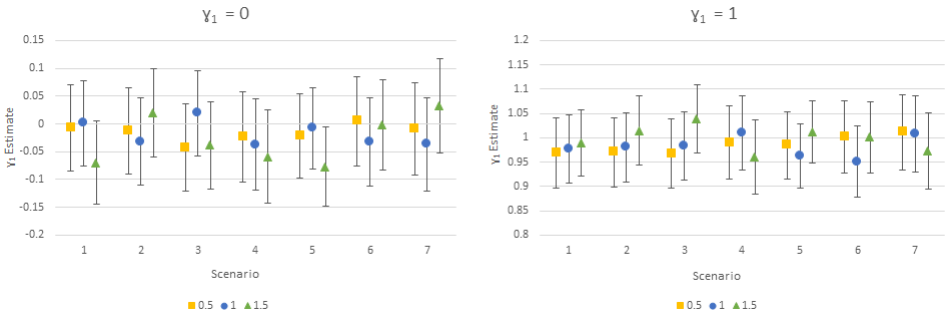

Figure A.4: Estimated γ_1 under sampling design setting 3 and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Uniform distribution.

A.2 Exponential Censoring

A.2.1 Sampling Design Setting 1

The γ_1 estimates are plotted in Figure A.5, along with the corresponding 95% confidence interval for γ_1, for all sampling scenarios in Table 2.1 and shape parameters explored. The estimated 95% confidence intervals contain the true value in almost all scenarios, the exceptions occur when $\gamma_1 = 1$. When $\alpha = 0.5$ for sampling scenarios 2 and 3, the 95% confidence intervals do not contain the true value of γ_1. The 99% confidence intervals for these situations were calculated, and the true value was found to fall within the interval. We conclude these estimates adequately represent the true value of γ_1.

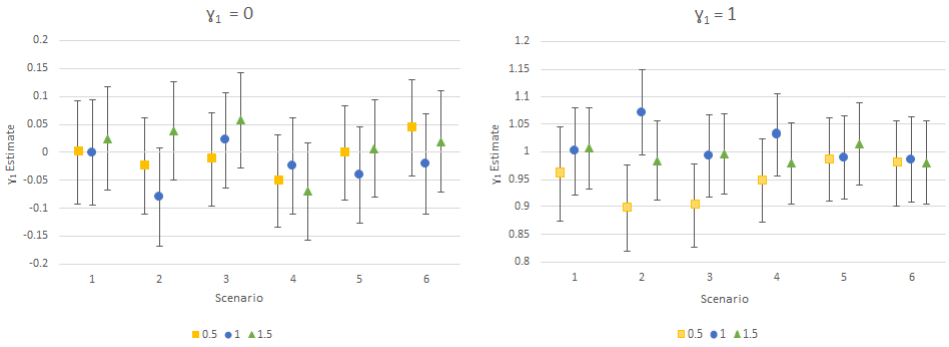

Figure A.5: Estimated γ_1 under sampling design setting 1 and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Exponential distribution.

A.2.2 Sampling Design Setting 2

Firstly, the sampling scenarios in Table 2.2 were considered under Exponential censoring. The 95% confidence intervals for γ_1 generally contain the true value of the parameter, as depicted in the Figure A.6. The few situations where the true value is not contained within the 95% confidence interval, the 99% confidence interval was

calculated. The true value was always within the wider confidence interval.

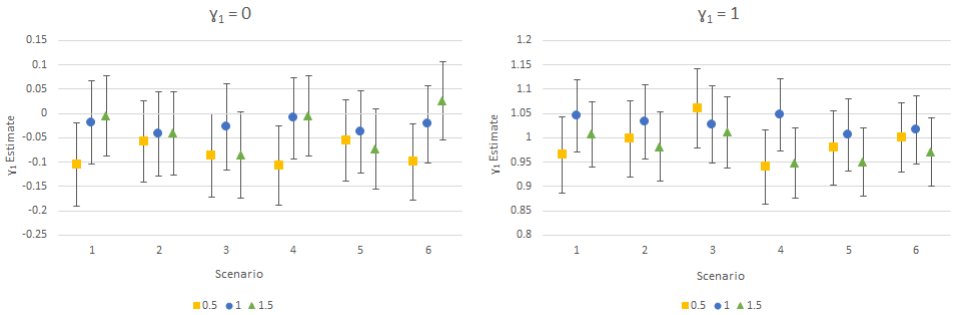

Figure A.6: Estimated γ_1 under sampling design setting 2A and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Exponential distribution.

Sampling design setting 2B explores stratified sampling when the proportion of cases to non-cases within the sample is approximately equal. The sampling scenarios investigated are outlined in Table 2.3. Figure A.7 depicts the 95% confidence intervals for γ_1 under each sampling scenario investigated; the true value is adequately captured. This conclusion was made by recognizing that all of the 95% confidence intervals contain the true value of γ_1.

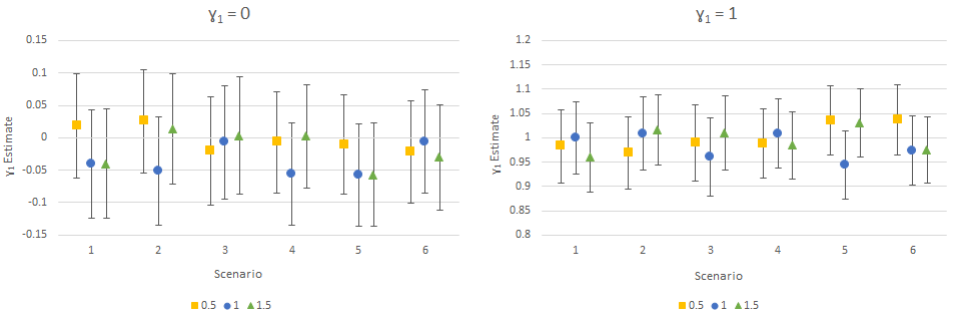

Figure A.7: Estimated γ_1 under sampling design setting 2B and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Exponential distribution.

A.2.3 Sampling Design Setting 3

Figure A.8 depicts the estimated γ_1 and corresponding 95% confidence intervals under each shape parameter and sampling scenario investigated in Table 2.4. The true value falls within the 99% confidence interval of γ_1 in most cases. The 99% confidence intervals for sampling scenarios 4 and 6 when $\gamma_1 = 1$ do not contain the true value of γ_1.

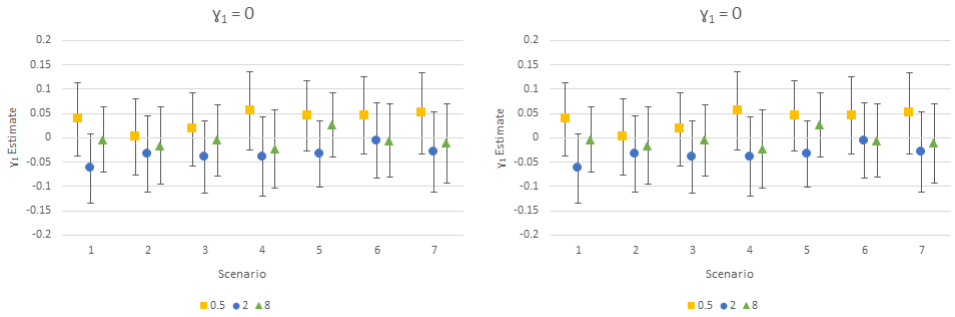

Figure A.8: Estimated γ_1 under sampling design setting 3 and the Weibull survival model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Exponential distribution.

A.3 Mixture Cure Model

A.3.1 Sampling Design Setting 1

The estimated values for γ_1 and α_1 under the sampling scenarios in Table 4.1 are given in Figure A.9, along with the corresponding 95% confidence intervals. In general, for both parameters, the true value falls within the 99% confidence interval for γ_1 and α_1 therefore we can conclude the model is adequately estimating the parameter values.

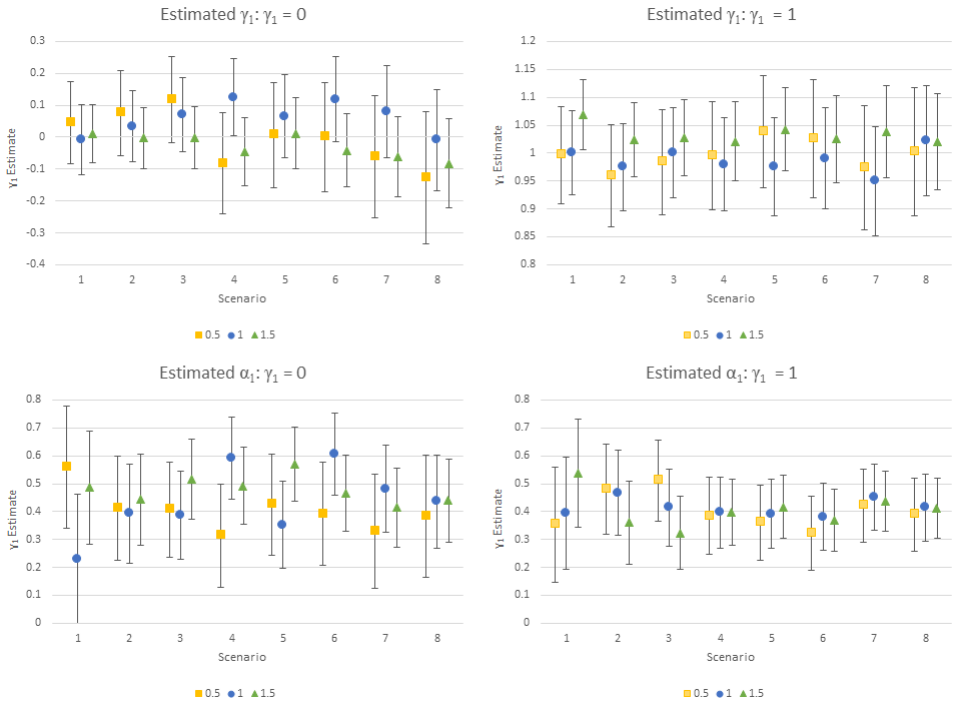

Figure A.9: Estimated γ_1 (upper panel) and α_1 (lower panel) under sampling design setting 1 and the Weibull mixture cure model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Uniform distribution.

A.3.2 Sampling Design Setting 2

Figure A.10 displays the estimated γ_1 and α_1 values under the sampling scenarios in Table 2.4 in which BSS of the cases is investigated. The model appears to be adequately capturing the true value of the parameter as the estimated 95% confidence interval for γ_1 and α_1 always contains the true value for both γ_1 and α_1.

A.3.3 Sampling Design Setting 3

The estimated values for γ_1 and α_1 under the sampling scenarios in Table 4.3 are given in Figure A.11. We observe that when $\gamma_1 = 0$, the 95% confidence intervals for γ_1 always contain the true value. When $\gamma_1 = 1$, however, it appears that the value of γ_1 is regularly under-estimated, particularly when the Weibull shape parameter is

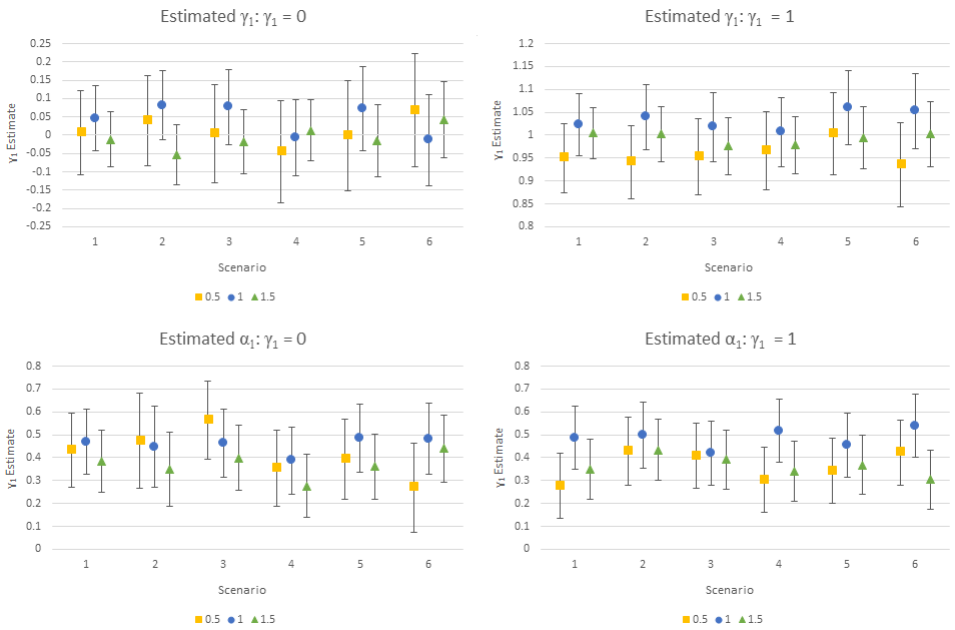

Figure A.10: Estimated γ_1 (upper panel) and α_1 (lower panel) under sampling design setting 2 and the Weibull mixture cure model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Uniform distribution.

1.5. For α_1 the model appears to adequately capture the true value of the parameter, as all of the 95% confidence intervals contain the true value, $\alpha_1 = 0.405$.

Sampling design setting 3B described in Table 4.4 only considers sampling scenarios in which the most efficient case stratified sampling design is employed. The difference between all of the sampling scenarios investigated is the change in non-case stratum sampling probabilities. From Figure A.12, we see that the resulting estimated γ_1 for all of the sampling scenarios appear close to the true value. Due to the method of case stratification and the selection of the case set (essentially all cases from stratum 1 and 3 were selected), the 7000 case observations selected for each sampling scenario are almost identical. This leads to the γ_1 estimates being extremely similar across all of the sampling scenarios. The model appears to capture the true value of α_1 in most cases, as well.

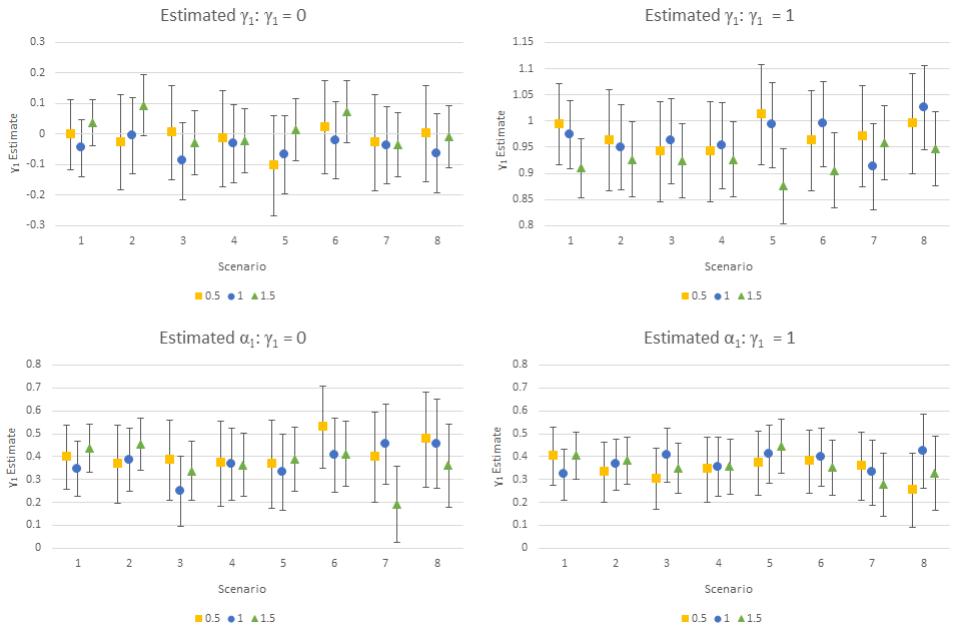

Figure A.11: Estimated γ_1 (upper panel) and α_1 (lower panel) under sampling design setting 3A and the Weibull mixture cure model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Uniform distribution.

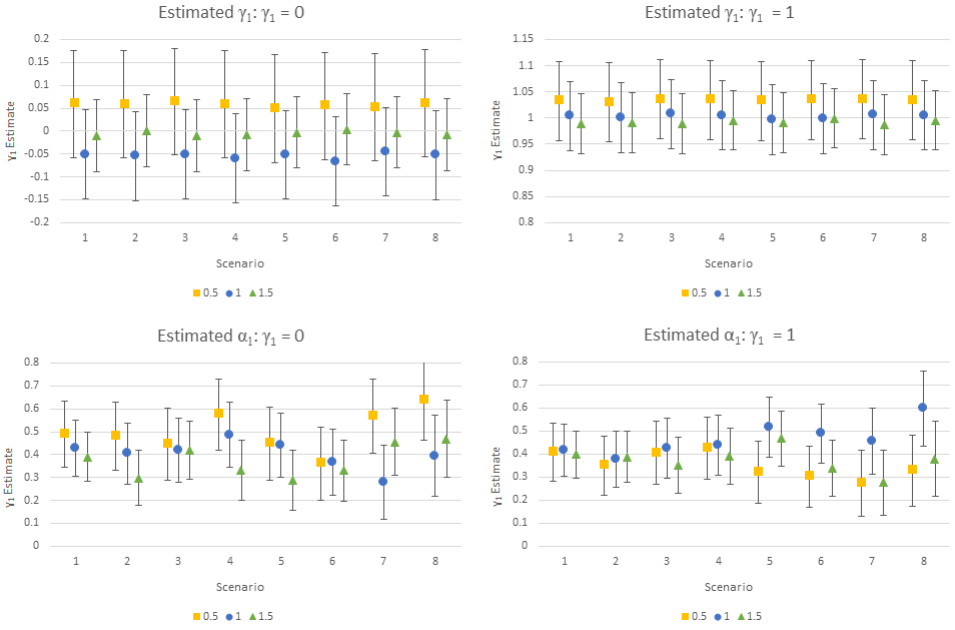

Figure A.12: Estimated γ_1 (upper panel) and α_1 (lower panel) under sampling design setting 3B and the Weibull mixture cure model with shape parameter $\alpha = 0.5, 1, 1.5$. Censoring times were generated from the Uniform distribution.

Appendix B

Empirical Conditional Pdfs of Survival Models

B.1 Empirical Conditional Pdf of the Weibull Survival Model

To understand how the shape parameter influences the shape of the Weibull distribution, Figure B.1 depicts the empirical conditional pdfs of the Weibull distribution with shape parameters $\alpha = 0.5, 1.0, 1.5$.

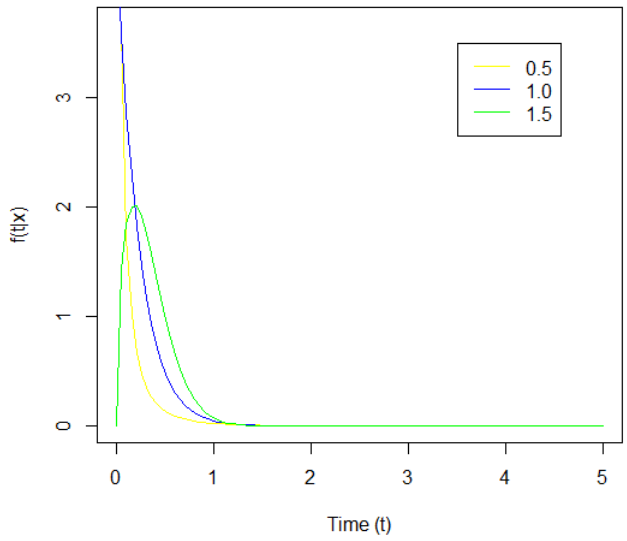

Figure B.1: The empirical conditional pdfs of the Weibull survival model (Eq. 1.7) with $\gamma_1 = 1$ shape parameters $\alpha = 0.5, 1, 1.5$

Appendix C

Sampling from the Weibull Distribution under Exponential Censoring

In Sections 3.2.1 and 3.3.1, we observed a difference between the efficiency conclusions drawn under Uniform censoring and Exponential censoring. In particular, when the Weibull shape parameter $\alpha = 1.5$, the efficiency conclusions changed depending on the censoring mechanism in place. Under Uniform censoring when $\alpha = 1.5$ we concluded approximately balancing the cases and the non-cases within the sample lead to a more efficient design. However, when Exponential censoring was employed and $\alpha = 1.5$ the efficiency of γ_1 improved when n_{cases} was large. To understand these differences the following Appendix contains the empirical conditional pdfs, for all shape parameters investigated $\alpha = 0.5, 1.0, 1.5$ under the Weibull distribution, of the case sample drawn during phase II for each sampling scenario investigated in Table 2.1 (sampling design setting 1).

C.1 Empirical Conditional Pdf when $\alpha = 0.5$

Figure C.1 depicts each empirical conditional pdf of the observed survival times for the case set under Weibull survival model with $\alpha = 0.5$ and Exponential censoring, produced from the sampling scenarios considered in sampling design setting 1 (Table 2.1). The black points represent the cases within the entire cohort. The orange points indicate sampled cases under each sampling scenario. From the plots we can conclude that sampling scenario 4 ($n_{cases} = 4000$ & $n_{cohort} = 6000$) captures enough case observations from the left hand side the of the distribution, a highly informative portion, through SRS of the cases.

C.2 Empirical Conditional Pdf when $\alpha = 1.0$

Figure C.2 depicts each empirical conditional pdf of the observed survival times for the case set under Weibull survival model with $\alpha = 1.0$ and Exponential censoring, produced from the sampling scenarios considered in sampling design setting 1 (Table 2.1). The black points represent the cases within the entire cohort. The orange points indicate sampled cases under each sampling scenario. From the plots we can conclude that sampling scenario 4 randomly selects enough observations from the informative portion of the distribution, cases with short survival times.

C.3 Empirical Conditional Pdf when $\alpha = 1.5$

Figure C.3 depicts each empirical conditional pdf of the observed survival times for the case set under Weibull survival model with $\alpha = 1.5$ and Exponential censoring, produced from the sampling scenarios considered in sampling design setting 1 (Table 2.1). The black points represent the cases within the entire cohort. The orange points

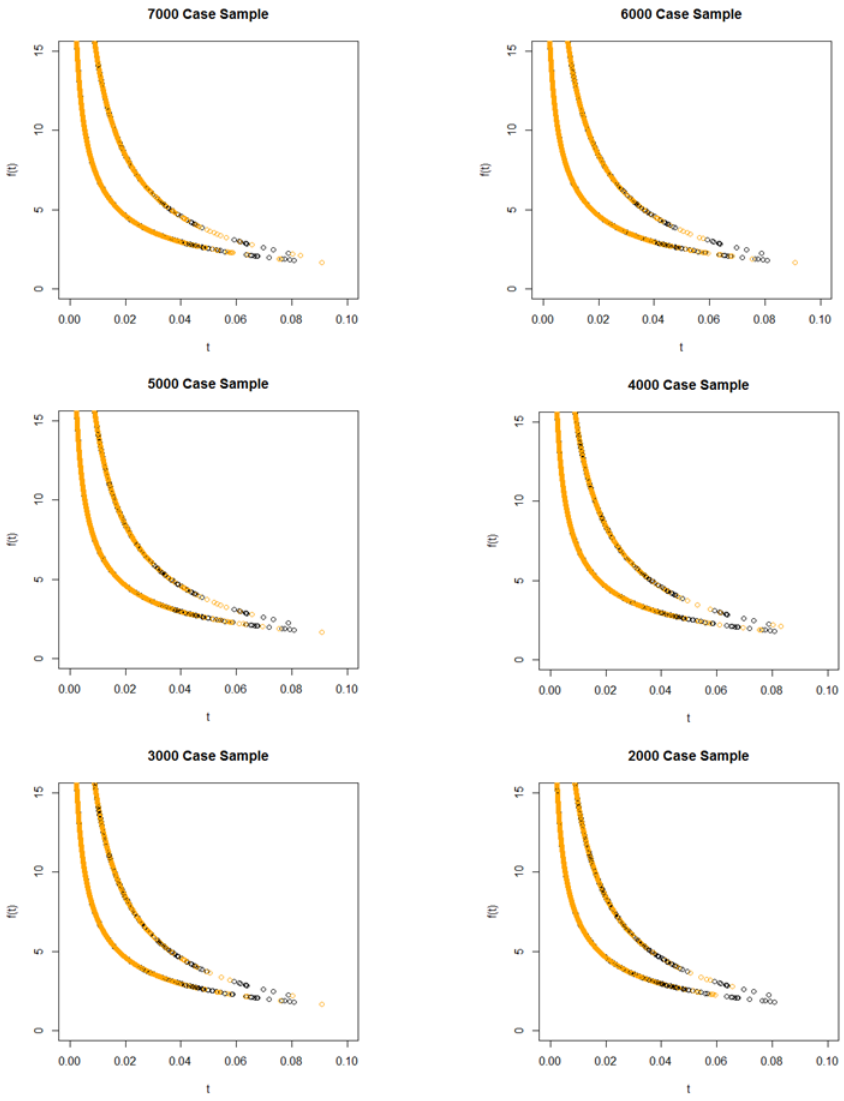

Figure C.1: The figures depict the empirical conditional pdf when $\gamma_1 = 1$ and $\alpha = 0.5$ in Eq. 1.7 for the case set. The black points represent the cases within the entire cohort. The orange points indicate sampled cases under each sampling scenario.

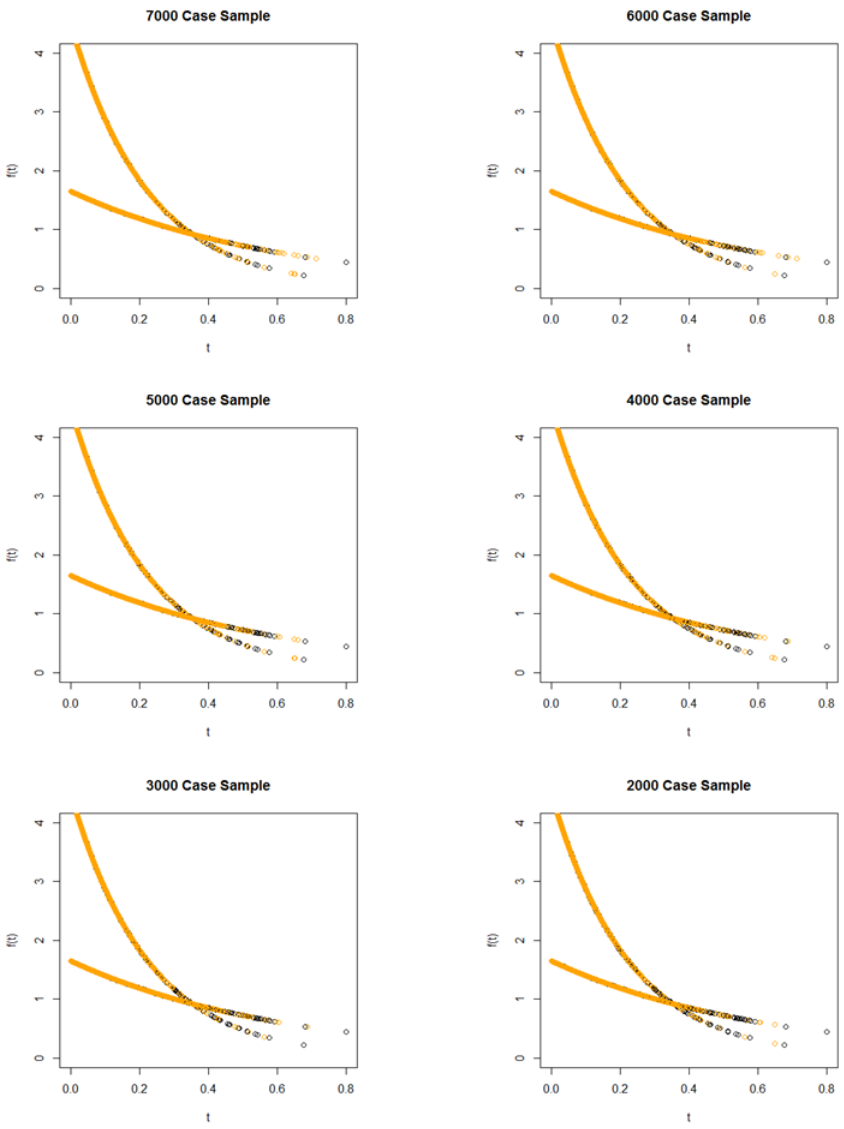

Figure C.2: The figures depict the empirical conditional pdf when $\gamma_1 = 1$ and $\alpha = 1.0$ in Eq. 1.7 for the case set. The black points represent the cases within the entire cohort. The orange points indicate sampled cases under each sampling scenario.

indicate sampled cases under each sampling scenario. From the plots we can conclude that sampling scenario 4 does not randomly select enough cases from the extremes of the distribution. Further, we see the case sample must be large enough that cases with short and long survival times are adequately represented within the mentioned sample.

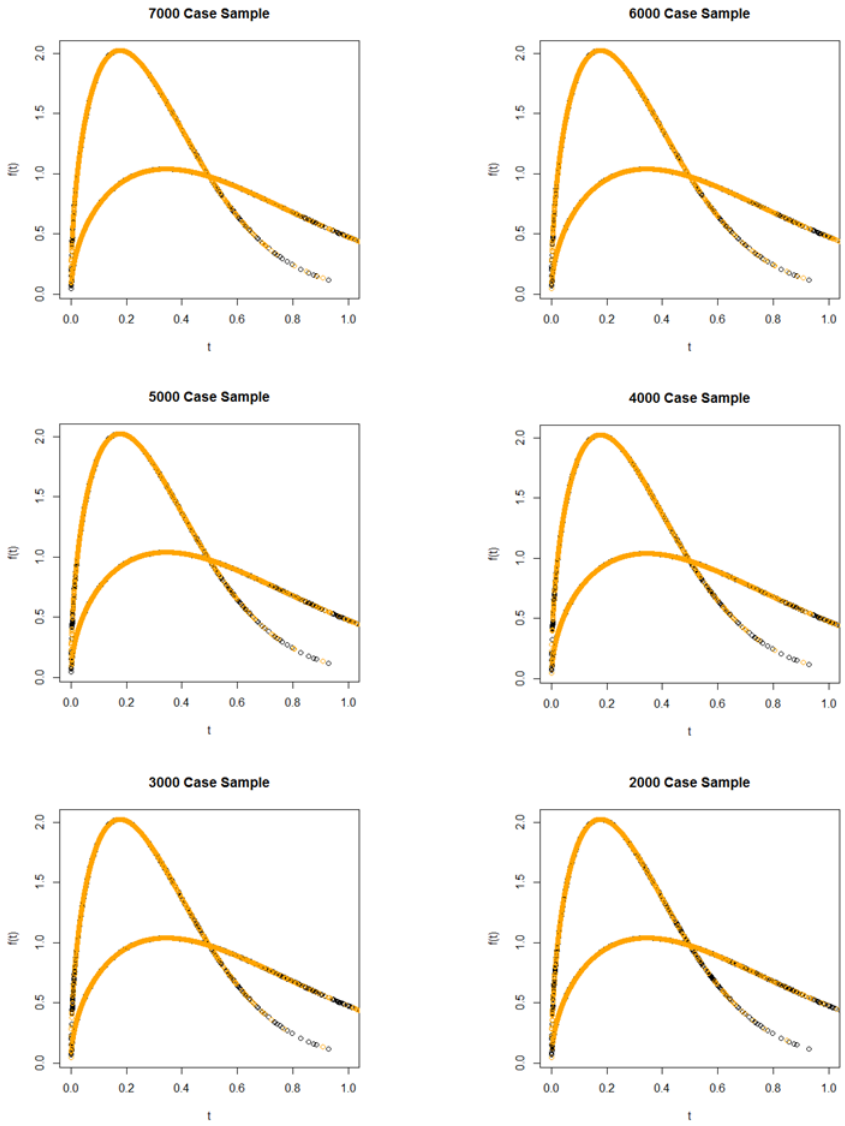

Figure C.3: The figures depict the empirical conditional pdf when $\gamma_1 = 1$ and $\alpha = 1.5$ in Eq. 1.7 for the case set. The black points represent the cases within the entire cohort. The orange points indicate sampled cases under each sampling scenario.