Efficiency of Positive Event Dependence Models for Self-Controlled Case Series Designs

by

© **Gwangseop Shin**

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

Department of Mathematics and Statistics

Memorial University

June 2022

St. John's, Newfoundland and Labrador, Canada

Abstract

The self-controlled case series (SCCS) design is an outcome-dependent sampling design developed to investigate potential associations between time-varying exposures and adverse recurrent events in clinical settings. It is a case-only design, in which only cases, who are individuals experienced the event-of-interest at least one time, are included. The cases serve as their own controls. As a result, the SCCS design implicitly controls for all time-fixed confounders. The standard SCCS method is based on a conditional Poisson process model, which cannot be used to model the effects of past event occurrences. A new method, called positive event dependence self-controlled case series (PD-SCCS), has been proposed to deal with this issue. This method adjusts the baseline intensity function of an event process with the number of previous events, and maintains all features of the SCCS method. In this study, we consider rare recurrent events settings, where the events are generated from mixed nonhomogeneous pure birth models with immigration. We investigate the relative efficiency of the PD-SCCS design compared to other SCCS and cohort designs in the estimation of the relative incidence parameter, as well as impacts of some model misspecifications and violations of model assumptions required for the PD-SCCS design through extensive Monte Carlo simulation studies. We illustrate the methods by analyzing a dataset from a clinical vaccine study, as well as two synthetic datasets based on a postmarketing drug safety surveillance study.

This thesis is dedicated to my parents.

Lay summary

Study designs include methods and procedures used to collect and analyze data. They are widely used in medicine and epidemiology. In many such studies, individuals are potentially subject to experience a well-defined event-of-interest more than one time during their follow-ups. Such type of data is called recurrent event data. An individual experiencing the recurrent event at least one time is called a case. Otherwise, they are called a control. The self-controlled case series (SCCS) design is an outcome-dependent observational study design, in which only the cases are included. It has important advantages over classical study designs especially when the event of interest is rare. The main goal of a SCCS study is to investigate the effects of time-varying exposures or interventions on the intensity of adverse recurrent health outcome events. The SCCS design is self-controlled, which means that cases serve as their own controls. The positive event dependence SCCS (PD-SCCS) model has been recently proposed as an extension to the standard SCCS model. It satisfies all features of the SCCS design and allows dependence on the past event occurrences experienced by cases, which is not possible with the standard SCCS model. In this thesis, our main goal is to explore the relative efficiency of the PD-SCCS model by comparing it with the SCCS models and more established cohort models through extensive simulation studies. Furthermore, we investigate the impacts of model misspecification and violations of assumptions required for the PD-SCCS model.

Acknowledgements

First of all, I would like to thank my supervisor Dr. Candemir Cigsar for his continuous support and feedback on my research. I sincerely appreciate his kindness, patience and guidance. This thesis would not have been possible without his guidance and advice.

Besides my supervisor, I would like to thank my thesis committee members Dr. Zhaozhi Fan and Dr. Yanqing Yi for their insightful comments and suggestions.

I am also grateful to all the professors in our department who taught me subjects during my program.

Lastly, it is my great pleasure to thank my father Gilwan Shin and my mother Kiock Seo for their endless support and love.

Statement of contribution

Dr. Candemir Cigsar proposed the research question that was investigated throughout this thesis. Dr. Candemir Cigsar and Gwangseop Shin jointly designed the overall study. Gwangseop Shin implemented the algorithms, conducted the simulation study, and drafted the manuscript. Dr. Candemir Cigsar supervised the study and contributed to the final manuscript.

Table of contents

List of tables

List of figures

Chapter 1

Introduction

1.1 The Self Controlled Case Series Design

In epidemiology, observational study designs such as cohort or case-control design are used in order to find the effect of treatment or exposures. A cohort is a group of individuals sharing common characteristics or conditions. A nonexperimental cohort study includes two or more groups of individuals defined by their distinct level of exposure to a potential cause of disease (Lash et al., 2021). These groups are called the study cohorts. If two study cohorts are investigated, one is considered as the exposed or index cohort and the other one is referred to as the reference cohort. For example, in an epidemiology study to measure the effect of a specific vaccination, the cohort including individuals exposed to the vaccine is called the exposed. Individuals who have not been exposed to the vaccine constitute the unexposed study cohort. This design is a prospective study, which means that the individuals in the study cohorts are followed over time to observe the outcome of interest, called the "event", experienced by the members of the cohort. An important issue with the classical cohort designs is that some prominent statistical methods for their analysis require information on every member of the cohort. As a result, a study may become cost-inefficient especially when expensive covariates are of interest. It should also be noted that cohort designs may suffer confounders in determining the association between risk factors (i.e. exposures) and outcome of interest. A confounder is a variable considered as a risk factor for the outcome of interest in a study, which is not of direct interest to the researchers but related to the main risk factors (Lash et al., 2021). As discussed

by Keogh and Cox (2014), the success of any study design to assess relations between exposures and outcomes is highly based on creating groups of individuals so that all confounders equally operate on them. Then, the effect of exposures on the outcomes can be measured on each group and compared. In experimental studies, this is usually done with randomization to obtain cause-and-effect type of relations. Observational studies usually deal with confounders by conditioning on the values of explanatory variables for both exposures and outcome of interest. This approach is not always feasible especially when there is too much heterogeneity across the cohort members and limited sources to collect data about explanatory variables. As we discussed later in this section, the self-controlled case series design is an alternative to the classical cohort design to deal with these issues in some settings.

The case-control designs constitute another important class of study designs. They are retrospective observational studies and usually used to investigate the associations between rare events and explanatory variables (Keogh and Cox, 2014). In the case-control design, individuals who experience an event at least once are called a "case," and those at risk of having an event are called a "control." Once cases are observed, their exposures and other related available information are retrospectively collected.

The SCCS design is a relatively new study design, which was proposed by Farrington (1995). It is a specific case-control design. The main goal is to investigate the effect of time varying exposures or treatments on a rare outcome of interest. As the name implies, the SCCS design is a case-only design. That is, controls, in other words individuals who have not experienced an event, are not included in the study. The estimation of the relative incidence rate is solely based on the information obtained from the cases. Since the SCCS design only uses cases, it is cost-efficient. It also provides computational efficiency. An important characteristic of the SCCS design is that the individuals are included in the study if they have experienced the event of interest at least one time in their past. Once they are included, their exposure information is retrospectively obtained. Another important feature of the SCCS design is that the cases are used as their own controls. Therefore, the design is named as self-controlled. This is an important characteristic of this design because the ratio of cases and controls in the SCCS design is one to one. Such a ratio in a classical cohort design is usually very low especially when the outcome is a rare event. The estimation of the exposure-effect is carried out as follows. First, the observation periods of each case is divided into exposed and unexposed periods. Then, the rate of event occurrences

over the exposed periods are compared with that of the unexposed periods obtained from the same case. Since the comparison is made within-person, all time-invariant confounders such as gender, socio-economic status, genetic attributes or individual frailty are automatically adjusted in the SCCS design. This characteristic also helps to protect the data privacy as the SCCS design does not require information on time-invariant covariates. Because of these characteristics, there has been growing interest in the use of the SCCS design, especially in medicine and epidemiology (Hubbard et al., 2003). As a result, there have been many recent studies to generalize the SCCS model (Ghebremichael-Weldeselassie et al., 2017, 2016; Simpson, 2013)

In this thesis, we explore an important extension of the SCCS design. We next introduce this design.

1.1.1 A Positive Event Dependence Model for Self-Controlled Case Series

The standard SCCS model is based on a conditional Poisson process. Because of this reason, the SCCS model requires independence of the event occurrences from the history of the process. This property limits the use of the SCCS model when the probabilistic characteristics of a process are affected by the previous event occurrences. The positive event dependence model for the self-controlled case series design, shortly the PD-SCCS model, is developed as an extension to the standard SCCS model by Simpson (2013). The PD-SCCS model is based on a conditional nonhomogeneous pure birth process with immigration (Simpson, 2013). The PD-SCCS model positively adjusts the baseline intensity function of a process after each event occurrence so that it is named as positive event dependence. It satisfies all requirements of a SCCS design. Thus, it possesses all aforementioned advantages of the SCCS model.

The PD-SCCS design is the key model in this thesis. We discuss the details of the PD-SCCS model in Chapter 2, and study its relative efficiency in finite sample settings through an extensive Monte Carlo simulation study in Chapter 3. Furthermore, we present the results of other Monte Carlo simulation studies conducted to investigate the effects of the model misspecification and the violation of model assumptions in Chapter 4. Since the main purpose of the SCCS design is to estimate the effects of exposure on a recurrent event, we next introduce the recurrent event data types.

1.2 Data Types for Recurrent Event Processes

The SCCS design was first proposed by Farrington (1995) to investigate the associations between vaccinations and adverse health outcomes in vaccine safety studies. The outcome event is rare and of recurrent type, which means that individuals in a study may experience the event of interest more than one time during their followup. Recurrent event processes provide canonical models for the analysis of recurrent events. We therefore discuss the recurrent event processes in this thesis. In various areas such as science, medicine and technology, it is quite common to study the processes of recurrent events. In these studies, the events should be well-defined. For example, an event can be any condition that causes a stoppage of operations of a machine in a study from industrial engineering or a specific type of cancer recurrences experienced by individuals in a medical study. We consider such recurrent event data arise from stochastic processes, called recurrent event processes.

In this thesis, we consider recurrent events occurring in continuous time scale. As Amorim and Cai (2015) mentioned a recurrent event process in the continuous time scale is orderly, which means that at most one event can occur at any given time. Also, for any recurrent event process, the event occurrence times are necessarily ordered. In a recurrent event study, individuals in a cohort are longitudinally observed over a given time interval and occurrence times of a well-defined event and times of exposures along with their durations and the values of other explanatory variables are recorded. Event occurrence times are usually presented either in calendar times or as the elapsed times between event occurrences, which is called the gap times. In the prospective studies, the event data are collected over fixed observation periods for the members of the cohort. These time intervals are called the observation window. The observation window can be fixed for all individual or may vary for the individuals in a study. The individuals are either continuously or intermittently followedup over their observation windows. The start and end-of-followup times of individuals are also recorded in a recurrent event study.

As mentioned above, the values of some explanatory variables, in other words covariates, are usually included in the dataset. Covariates can be classified in a few different ways. A covariate, for example, is called time-fixed or time-invariant if its value does not change over the observation window. Otherwise, it is called a time-varying covariate. These are also external and internal covariates. As defined by

Kalbfleisch and Prentice (2002), an external covariate is a covariate whose value does not change by recurrent event processes but they may have some effects on recurrent event processes. All time-fixed covariates are external. For example, air pollution levels is a time-varying external covariate in a study involving hospital visits. Internal covariates are functions of the history of recurrent event processes. For example, number of previous events at any given time and elapsed time since the last event in a process are internal covariates.

Methods for the analysis of recurrent event data are usually considered in two broad approaches. The first one is based on the event counts. Poisson processes are considered as canonical models for the event counts in recurrent event processes. The gap time models are useful if there is a specific interest in the analysis of the specific gap times. Renewal processes and their extensions are usually applied to deal with gap time analysis in recurrent event settings. Both cases can be extended to deal with regression modeling. We discuss these issues in the next chapter, where we introduce fundamental models and their regression extensions. It should be noted that, since the SCCS method is based on conditional Poisson processes, our focus in this thesis is on the event count models for the Poisson processes perspective. However, some methods based on gap times can be useful in the analysis of event counts as well.

1.3 Illustrative Examples

In this section, we briefly introduce the datasets used in Chapter 5 to apply methods discussed in this thesis. The first dataset is given by Miller et al. (2001). These data were collected to study the association between measles, mumps and rubella (MMR) vaccine and idiopathic thrombocytopenic purpura (ITP). The MMR vaccine is implemented against the MMR diseases in children in two doses. The first dose is implemented at the age of one and the second dose is implemented at the age between 4 and 6. ITP is a blood disorder caused by a decreased number of platelets in the blood. This decrease causes internal bleeding, easy bruising and bleeding gums. It has been discussed that the administration of MMR vaccination in children may increase the relative incidence of ITP. That is, during the limited time after administration of the MMR vaccine, the risk of having ITP disorder increases. The dataset includes the times of follwoup of 35 children whose aged between 366 and 730 days as well as

the times of children's hospital admissions for ITP as events and the administration of MMR vaccination records.

The second and third datasets are synthetic datasets generated by considering a real life dataset analyzed by Simpson (2013). The original dataset was used to investigate the association between Vioxx and myocardial infarction (MI). Vioxx is a COX-2 selective nonsteroidal anti-inflammatory drug (NSAID). The primary purpose of this durg is to relieve signs and symptoms of arthritis, acute pain and painful menstrual cycles in adults. Vioxx was removed due to evidence of an elevated risk of cardiovascular events, including MI, in 2004 (Bresalier et al., 2005). Because the occurrence of MI is plausible that there exists positive event dependence between event occurrences, PD-SCCS model is a suitable model to analyse the association between Vioxx drug and occurrence of MI (Simpson, 2013). That is, the administration of Vioxx drug may increase the event occurrence of MI during certain risk periods after the implementation, and the feature of positive event dependence of MI occurrence may increase the future risk incidence of MI. An event is defined as the diagnosis records of MI for each individual. Although the analysis was based on the real longitudinal health insurance dataset, which includes the information of individuals such as diagnosis records and drug prescription records, Simpson (2013) did not provide the raw dataset. Therefore, we generated synthetic datasets using the summary of information given in Simpson (2013) in this thesis.

1.4 Literature Review

The self-controlled case series (SCCS) method was initially proposed by Farrington (1995) to estimate the relative occurrence of events between the fixed risk period after vaccination and the control period. Compared to the classical cohort design, the SCCS design only requires the data from cases. Therefore, it reduces a considerable amount of effort to collect data (Farrington, 1995). Whitaker et al. (2006) provided a tutorial to explain the basic properties of the SCCS method, such as origins, assumptions and limitation of its usages.

Although the SCCS method was used for finding the connection between vaccines adverse events at the beginning (Farrington, 1995; Farrington et al., 1995), it has been used in other studies as well. For example, Hubbard et al. (2003) with using a SCCS

design studied the association between exposure to tricyclic antidepressants and risk of hip fracture. Douglas and Smeeth (2008) used the case series method to investigate the relationship between atypical antipsychotic drugs and the occurrence of stroke in a patient. Minassian et al. (2010) used the SCCS mdethod to find the association between invasive dental treatment and transient risk for vascular events. The SCCS method was also used by Zenner et al. (2012) to find the association between the risk of tuberculosis diagnosis and after pregnancy. By using the SCCS design, Langan et al. (2014) showed that herpes zoster increases a stroke rate in the first 6 months.

The use and technical aspects of the SCCS model have been discussed in the literature. Weldeselassie et al. (2011) reviewed forty papers published in medical journals regarding vaccine safety studies based on the SCCS method. They discussed the issues related the use of the SCCS in those papers. Petersen et al. (2016) demonstrated the overview of the SCCS method with some applications in medical research. Suchard et al. (2013) studied the empirical performance of the SCCS method. There have been also studies discussing the conditions of the standard SCCS model and its extensions. For example, Musonda et al. (2008) discussed the performance of the SCCS model in small sample size datasets. Farrington et al. (2011) developed methods to deal with the situation in which event-dependent censoring is present. Whitaker et al. (2018) investigated the validity of the main assumptions required to apply the SCCS model using some designed tests and limited simulation studies. These assumptions include the independence between events and exposure, as well as the independence between event occurrences and observation periods in the SCCS method. They also discussed the robustness of the model when these assumptions are violated. Musonda et al. (2006) developed formulas for calculating the sample sizes for the case series model for obtaining a desired power. Farrington et al. (2009) applied the SCCS to the event occurrence that affects, curtails or censors post-event exposures. Although the SCCS method is based on a nonhomogeneous Poisson process, this model is applicable to a non-recurrent event in cases where these are rare (Farrington and Whitaker, 2006; Farrington, 1995). Whitaker et al. (2018) investigated this rare non-recurrent outcome assumption in the SCCS model to quantify the meaning of rare events analytically and by simulation studies. Farrington and Hocine (2010) discussed the method to test the assumption of independence of event occurrence within-individual required by the SCCS model and alternative approaches when this assumption is violated.

In addition to those studies, there have been recent studies about extensions of

the parametric SCCS model. Farrington and Whitaker (2006) developed a semi-parametric model for the SCCS model. This method is important because the standard SCCS model may be affected by the misspecification of the age groups.
Ghebremichael-Weldeselassie et al. (2014) replaced the piecewise constant age effects used in the standard SCCS method with a smooth spline functions. Ghebremichael-Weldeselassie et al. (2016) added flexibility to the exposure effects especially when the exposure effect last long using the regression spline model to overcome the limitation of modeling exposure effect with step functions. Ghebremichael-Weldeselassie et al. (2017) provided the non-parametric SCCS method by using spline functions to model both age and exposure effects. Simpson et al. (2013) proposed a way to apply the SCCS model to the longitudinal observational databases. Simpson (2013) developed a new SCCS model that allows for the positive event dependence from previous event occurrences to the risk of a future event occurrence, which is called the positive event dependence SCCS (PD-SCCS) model.

1.5 Goal and Outline of the Thesis

Discovering the effects of treatments or exposures on the occurrence of adverse events with a suitable modeling is important in medince and epidemiology. The SCCS design is an important alternative to the more established cohort designs to investigate associations between time varying exposures and adverse health outcomes in rare event settings. It simplifies data collection, increases computational efficiency and helps protect data privacy. The SCCS model also produces efficient estimates of the effect of exposures. However, it has its limitations, and it is important to generalize it to apply to various situations without losing its advantages. Positive event dependence models for the self-controlled case series design (PD-SCCS) provides an extension to the SCCS model in such a way that the occurrence of an event may potentially increase the risk of future event occurrences, while maintaining the advantages of the SCCS design. The PD-SCCS model is a relatively new model. To use the PD-SCCS model properly, it is important to test the performance of this model in various finite sample settings. Therefore, our main goal in this thesis is to explore the properties of some finite sample settings of the PD-SCCS model. We, therefore, first investigate the bias and relative efficiency of the parameter estimates in the PD-SCCS model by comparing it with other promising models under extensive finite sample scenarios through

Monte Carlo simulations. Then, we examine the robustness of the PD-SCCS model with respect to certain model misspecifications and violations of assumptions required for the PD-SCCS model on the estimation of relative incidence in finite sample settings with simulations. The outcome of this thesis provides a better understanding of the use of the PD-SCCS model in real life settings, and will be instrumental to understand the settings in which the PD-SCCS model performs better than other promising recurrent event models in order to estimate the effects of time-varying exposures on the adverse event occurrences.

The remainder of this thesis is given as follows. In Chapter 2, we present details of the notation used frequently in the rest of this thesis. We also introduce the canonical models for the analysis of recurrent events and development of the SCCS design. The framework for likelihood procedures is given as well. The last section of Chapter 2 includes the algorithms used for data generation in our simulation studies. In Chapter 3, we present the results of our first simulation study conducted to investigate the relative efficiency of the PD-SCCS model compared with other promising recurrent event models under various settings. In Chapter 4, we investigate the impact of certain model misspecification and violation of assumptions of the PD-SCCS model. We consider two assumptions of event-independent exposure times and event-independent observation periods. The results of two extensive Monte Carlo simulation studies are presented. Chapter 5 includes illustrative examples of the methods discussed in the previous chapters. We apply the SCCS methods explained to a dataset from medicine and to two synthetic datasets generated based on a postmarketing drug safety surveillance study. Finally, we present the summary of the previous chapters and conclusion of the thesis in Chapter 6, along with some possible future research topics.

Chapter 2

Technical Background and Notation

2.1 Introduction

The goal of this section is to introduce the common notation and technical background used in this thesis. Since counting processes provide models and methods suitable for the analysis of recurrent events, we discuss the notation of counting processes and their intensity functions in this chapter. Our goal is not to give a comprehensive treatment of counting processes but to introduce concepts that are frequently used in the remaining parts of this thesis. For simplicity, we first start with the single process case and extend our notation to the multiple processes cases whenever it is needed.

Suppose that there is an interest in the analysis of random occurrences of a well-defined recurrent event. Let $0 < T_1 < T_2 < \cdots < T_j < \cdots$ denote the event occurrence times, where T_j is the occurrence time of the jth event for $j = 1, 2, \ldots$, and by convention $T_0 = 0$. We let W_j, $j = 1, 2, \ldots$, denote the jth gap time or waiting time between the jth and $(j-1)$st events; that is, $W_j = T_j - T_{j-1}$. For any $t \geq 0$, the random variable $N(t)$ denotes the number of events occurred during the time interval $(0, t]$ with the convention $N(0) = 0$. For any $0 \leq s < t$, we also use the notation $N(s, t)$ to define the number of event occurrences over the interval $(s, t]$; that is, $N(s, t) = N(t) - N(s)$. Note that $N(t) = \sum_{j=1}^{\infty} I(T_j \leq t)$, where the indicator function $I(A)$ equals to 1 if the event A is true and 0, otherwise.

We use the notation $\{N(t);\ t \geq 0\}$ to denote a counting process specified for a well-defined recurrent event. The cumulative mean function or, shortly, the mean function of the process $\{N(t);\ t \geq 0\}$ is denoted by $\mu(t),\ t \geq 0$, which gives the expected number of events over the interval $(0, t]$; that is, $\mu(t) = E\{N(t)\}$. We also use the notation $\mu(s, t)$ to represent the expected number of event occurrences over the interval $(s, t]$, where $0 \leq s < t$. Another useful function of a counting process is its rate function $\rho(t)$, which is defined as

$$\rho(t) = \lim_{\Delta t \to 0} \frac{\Pr\{\Delta N(t) \geq 1\}}{\Delta t}, \qquad t \geq 0, \tag{2.1}$$

where Δt is a small positive quantity and $\Delta N(t)$ denotes the number of event occurrences in the interval $[t, t + \Delta t)$. For a small Δt, $\rho(t)\Delta t$ approximates the probability of one or more event occurrences in the interval $[t, t + \Delta t)$. If the derivative exists, it can be shown that $\rho(t) = d\mu(t)/dt$ for any $t > 0$. By using this notation, we denote the number of events in an infinitesimal time interval $[t, t + dt)$ by $dN(t)$. Following our previous notation, we also define $\rho(s, t) = \rho(t) - \rho(s)$ for any $0 \leq s < t$.

We next define the intensity function of a counting process, which is a crucial concept in modeling recurrent events. To do this, we first define the history of the counting process $\{N(t);\ t \geq 0\}$ at time t as

$$H(t) = \{N(s) : 0 \leq s < t\}, \qquad t \geq 0. \tag{2.2}$$

The history $H(t)$ includes all information about the counting process $N(t)$ in the interval $[0, t)$. We assume that two or more events cannot simultaneously occur at any given time t. The event intensity function is then defined as

$$\lambda(t|H(t)) = \lim_{\Delta t \to 0} \frac{\Pr\{\Delta N(t) = 1|H(t)\}}{\Delta t}, \qquad t \geq 0. \tag{2.3}$$

Given the history at time t, the intensity function (2.3) gives the instantaneous conditional probability of an event occurrence in the interval $[t, t + \Delta t)$ as Δt approaching 0. The intensity function (2.3) completely specifies a recurrent event process in the continuous time settings (Cook and Lawless, 2007, p.28).

The hazard function is useful especially to model the gap times of a recurrent event process. Let W denotes a continuous and non-negative random variable. The

hazard function of the random variable W is then defined by

$$h(w) = \lim_{\Delta w \to 0} \frac{\Pr(W < w + \Delta w | W \geq w)}{\Delta w} = \frac{f(w)}{S(w)}, \qquad w \geq 0, \tag{2.4}$$

where $f(w)$ is the probability density function (p.d.f), $F(w) = \int_0^w f(u)du$ is the cumulative density function (c.d.f), and $S(w) = \Pr(W > w) = 1 - \Pr(W \leq w) = 1 - F(w)$ is the survival function of W. More detailes about hazard functions can be found in Cox and Oakes (1984) and Lawless (2003).

Models and methods in the context of recurrent event processes are usually based on the intensity functions of the counting processes or the hazard functions for the gap times W_j. An important extension to them is the inclusion of covariates. The regression models for counting processes and gap times are extensively discussed by Cook and Lawless (2007, Chapters 3 and 4). In the next section, we introduce the fundamental models for both cases and briefly discuss how to extend them to include covariates.

2.2 Fundamental Models for Recurrent Event Process

Poisson and renewal processes provide flexible models to analyze recurrent event data. Poisson processes are canonical for modeling event counts in a recurrent event process. In a Poisson process, events in non-overlapping time intervals occur independently. Renewal processes, on the other hand, focus on the gap times between events. Renewal processes require the strong assumption of independent gap times. More care is needed if this assumption does not hold in applications. Both models can be extended to include covariates as well as random effects.

In this chapter, we briefly explain these two canonical models. However, since our focus in this thesis is on the self-controlled case series (SCCS) model, which is related to Poisson processes as described in Section 2.3, our main concern is Poisson processes. Many properties of the Poisson and renewal processes can be found in point process or stochastic processes books. Good examples include Cox and Isham (1980), Parzen (1962), Ross (1996) and Daley and Vere-Jones (2003).

2.2.1 Poisson Processes

We start with the definition of a Poisson process as follows. A counting process $\{N(t);\ t \geq 0\}$ is called a Poisson process with the rate function $\rho(t)$, $t \geq 0$, if it satisfies the following conditions.

1. $N(0) = 0$.
2. The numbers of event occurrences in disjoint time intervals are independent.
3. For any $0 \leq s < t$, and $n = 0, 1, 2 \ldots$, $\Pr\{N(s,t) = n\} = \frac{[\mu(s,t)]^n}{n!} e^{-\mu(s,t)}$, where

$$\mu(s,t) = E\{N(s,t)\} = \int_s^t \rho(u)du. \tag{2.5}$$

A Poisson process can be equivalently characterized by its intensity function as follows. A counting process is called a Poisson process if and only if its intensity function is equal to its rate function. That is, a counting process $\{N(t);\ t \geq 0\}$ with the intensity function $\lambda(t|H(t))$ and the rate function $\rho(t)$ is a Poisson process if and only if

$$\lambda(t|H(t)) = \lim_{\Delta t \to 0} \frac{\Pr\{\Delta N(t) = 1\}}{\Delta t} = \rho(t), \qquad t \geq 0. \tag{2.6}$$

A proof of the above assertion can be found in Daley and Vere-Jones (2003, Chapter 2). If the rate function $\rho(t)$ in (2.6) is a constant, say ρ, it is called a homogeneous Poisson process. Otherwise, it is called a nonhomogeneous Poisson process. Therefore, the rate function of nonhomogeneous Poisson processes can be used to model some systematic changes in the rate of occurrence of events over time. For example, time trends can be included in nonhomogeneous Poisson processes (Lawless et al., 2012).

Many general properties of Poisson processes can be found in stochastic process books. For example, in a homogeneous Poisson process $\{N(t);\ t \geq 0\}$ with the associated rate function ρ, the gap times W_j follow an exponential distribution with the p.d.f, $f(w;\rho) = \rho e^{-\rho w}, w > 0$, and in this case, the gap times W_j $(j = 1, 2, \ldots)$ are independent and identically distributed (i.i.d.). However, the gap times are not independent in nonhomogeneous Poisson processes (Thompson, 1988, Section 6.1). Another useful property of nonhomogeneous Poisson process is given as follows. Let $\{N(t);\ t \geq 0\}$ be a nonhomogeneous Poisson process with the associated rate function $\rho(t)$ and the mean function $\mu(t)$. By setting $s = \mu(t)$, we define a new time scale s. Let $\{N^*(s);\ s \geq 0\}$ denote a counting process on the time scale s. Then, it can be

shown that $N^*(s) = N(\mu^{-1}(s))$, $s > 0$, is a homogeneous Poisson process with the rate function $\rho^*(s) = 1$ (Daley and Vere-Jones, 2003, p. 258). We use this feature to generate realizations of nonhomogeneous Poisson processes for a given rate function in our simulation studies.

Poisson process models can be parametric or non-parametric. For the parametric case, models of the exponential or power-law forms are very useful. For example for any $t \geq 0$, the exponential model can be written as $\rho(t; \alpha, \beta) = \exp(\alpha + \beta t)$, $\alpha, \beta \in \mathbb{R}$ and the power-law model can be written as $\rho(t; \alpha, \beta) = \alpha\beta t^{\beta-1}$, $\alpha, \beta > 0$. Our focus in this thesis is on the exponential model because the derivation of the SCCS model is related to this form. It should be noted that both models can be used to model the time trends in the rate functions. Also, a test of $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ in the exponential form and $H_0 : \beta = 1$ against $H_1 : \beta \neq 1$ in the power-law form can be used for testing the absence of monotonic time trends in Poisson processes.

External covariates in a Poisson process can be easily included through the rate functions. For any time t, we let $x(t)$ be a $p \times 1$ vector of time-varying and/or time-fixed external covariates in the sense discussed by Kalbfleisch and Prentice (2002, Section 6.3.1). We next define the history of the covariates at time t as $x^{(t)} = \{x(u) : \ 0 \leq u \leq t\}$, and extend the history of the counting process $\{N(t);\ t \geq 0\}$ by including $x^{(t)}$ in the history; that is, $H(t) = \{N(s),\ x(u);\ 0 \leq s < t,\ 0 \leq u \leq t\}$. Then, all probabilities and related functions including the intensity function (2.3) are conditional on the extended $H(t)$. Note that, since the history $H(t)$ includes the values of $x(t)$ in the closed interval $[0, t]$, it is assumed that the values of $x(t)$ are known at time t. However, $N(t)$ is still a random variable at time t, since the history does not include its value at time t. There are also internal covariates as defined by Kalbfleisch and Prentice (2002, Section 6.3.2). We discuss their inclusion to general intensity functions in Section 2.2.3.

Once the history is defined to include external covariates, the intensity function of the Poisson process $\{N(t);\ t \geq 0\}$ is given by

$$\lambda(t|H(t)) = \rho(t|x^{(t)}) = \rho_0(t) \exp\{x'(t)\beta\}, \qquad t \geq 0, \tag{2.7}$$

where $\rho_0(t)$ is a positive-valued baseline rate function and β is a $p \times 1$ vector of regression parameters. The model (2.7) is usually called the multiplicative intensity model, where the exponential functional form ensures that $\rho(t|x^{(t)}) \geq 0$ for all $t \geq 0$.

There are additive models as well, but we do not discuss them here. The mean function of the Poisson process with the rate function (2.7) is then defined by

$$\mu(t|x^{(t)}) = E\{N(t)|x^{(t)}\} = \int_0^t \rho_0(u)\exp\{x'(u)\beta\}du, \qquad t \geq 0. \tag{2.8}$$

If all covariates are time-fixed such that $x(t) = x$ is a vector of fixed covariates, then the mean function in (2.8) is $\mu(t|x^{(t)}) = \mu_0(t)\exp\{x'(t)\beta\}$, where $\mu_0(t) = \int_0^t \rho_0(u)du$. When the covariate vector $x(t)$ includes external time-varying covariates, the process is still a nonhomogeneous Poisson process for the given values of $x(t)$ with the mean function given in (2.8). It should be noted that if the vector $x(t)$ includes internal covariates, such as backward recurrence time or the number of previous events, the process is not a Poisson process anymore, but it is sometimes called the modulated Poisson process. We discuss this situation in Section 2.2.3.

Both parametric and semi-parametric regression models are proposed for the analyses of recurrent event processes with Poisson processes. In parametric models, both the baseline rate function $\rho_0(t)$ and the multiplicative term $\exp\{x'(t)\beta\}$ are parametrically specified. Semi-parametric models impose a parameter specification only for the multiplicative form $\exp\{x'(t)\beta\}$ and leave the baseline rate function $\rho_0(t)$ without any parameter specification. Many parametric and semi-parametric models and methods for the analysis of recurrent events through Poisson processes have been discussed by Cook and Lawless (2007, Chapter 3).

An important limitation of the Poisson processes, which is frequently seen in applications with multiple processes, arises when there is extra variation in the rate functions of individuals. Such an excess heterogeneity cannot be handled by a Poisson process even after conditioning on the values of covariates. To explain this situation, suppose that there are N independent processes included in a study. We let $\{N_i(t);\ t \geq 0\}$, $i = 1, 2, \ldots, N$, denote the Poisson process for the ith individual with the corresponding rate function $\rho_i(t|x^{(t)})$. From the properties of the Poisson processes, the mean and variance functions should be the same at any given time. If this property is not satisfied, the random effects term u_i for individuals $i = 1, 2, \ldots, n$ can be included in the model to address this issue. Given the value of the random effect u_i and external covariates $x_i(t)$, the process $\{N_i(t);\ t \geq 0\}$ is assumed to be a

Poisson process with the rate function

$$\rho_i(t|x_i^{(t)}, u_i) = u_i \rho_0(t) \exp\{x_i'(t)\beta\}, \qquad t \geq 0, \tag{2.9}$$

where the u_i are i.i.d. random variables following a distribution with the c.d.f. $G(u)$. Since it is mathematically tractable, the random variables u_i are usually modeled with a gamma distribution with mean 1 and variance ϕ, where $\phi > 0$. The variance parameter ϕ is sometimes called the heterogeneity parameter because it represents the unexplained variability across the rate functions of the individuals. The random effects models are discussed by Cook and Lawless (2007, Section 3.5). We revisit the random effects models in Section 2.3, and consider them from the perspective of the SCCS design.

2.2.2 Renewal Processes

Another important class of models for recurrent events is based on renewal processes. A renewal process is characterized by its gap times. In particular, a renewal process is a point process, in which the gap times W_j are i.i.d. random variables. Renewal processes can be equivalently defined by their intensity functions as well. The process $\{N(t);\ t \geq 0\}$ is a renewal process if and only if its intensify function is given by

$$\lambda(t|H(t)) = h(B(t)), \qquad t \geq 0, \tag{2.10}$$

where $h(w)$ is the hazard function of the gap time W_j and $B(t)$, called the backward recurrence time, denotes the elapsed time between the latest event time and t; that is, $B(t) = t - T_{N(t^-)}$. These characterizations and many other properties of the renewal processes are discussed by Cook and Lawless (2007, Chapter 4).

Many of the elementary properties of renewal processes can be found in good point process books. Here we discuss a few important ones, and refer to Daley and Vere-Jones (2003) for their proofs. In general, renewal processes are not equal to Poisson processes. An exception occurs when the gap times W_j are i.i.d exponential random variables with mean ρ^{-1}, where $\rho > 0$. In this case, the process $\{N(t);\ t \geq 0\}$ is a homogeneous Poisson process with the rate function ρ. It should be noted that the events "$N(t) \geq n$" and "$T_n \leq t$", where $T_n = W_1 + W_2 + \ldots + W_n$, are equivalent. In some cases, this relation between the counts and event times is useful to find the

distribution of $N(t)$ in renewal processes. Furthermore, the event "$N(t) = n$" is equivalent to "$N(t) = \max\{n : T_n \leq t\}$" and "$N(t) = \min\{n : T_n > t\}$". With the help of these relations, some of the properties of $N(t)$ in a renewal process can be stated in terms of T_n. For example, it can be shown that in a renewal process $\{N(t);\ t \geq 0\}$, the mean function $\mu(t) = E\{N(t)\} = \sum_{n=1}^{\infty} \Pr(T_n \leq t)$, where $\Pr(T_n \leq t) = \Pr(T_{n+1} \leq t) + \Pr(N(t) = n)$.

In some studies there is an interest in adjusting the renewal processes models with external covariates. In such situations, the intensity function in (2.10) is extended to include the values of covariates. Following our notation stated in the previous section, we let $x(t)$ denote a $p \times 1$ dimensional vector of covariates. Then, the intensity function of the renewal process $\{N(t);\ t \geq 0\}$ is given by

$$\lambda(t|H(t)) = h(B(t)|x^{(t)}), \qquad t \geq 0. \tag{2.11}$$

The multiplicative hazard function sometimes, referred to as the proportional hazards regression model, for the gap times W_j with time dependent covariate $x(t)$ is given by

$$h(w|x^{(t)}) = h_0(w) \exp\{x'(t)\beta\}, \qquad t \geq 0, \tag{2.12}$$

where $h_0(w)$ is a positive-valued baseline hazard function and $w = t - t_{N(t^-)}$. When the covariates are time fixed, that is, $x(t) = x$, the hazard function is given by $h(w|x) = h_0(w) \exp\{x'\beta\}$. If the covariate $x(t)$ includes functions of the history such as the elapsed time since the last event occurrence or the number of prior events at time t, the process is not a renewal process, but it is called a modulated renewal process. In this thesis, we do not discuss the renewal processes in detail. However, we use a specific renewal process to generate data in our simulations as explained in Section 2.6. Many of the statistical models and methods based on the renewal processes and their extension are given by Cook and Lawless (2007, Chapter 4).

2.2.3 General Intensity-Based Recurrent Event Models

In some studies, models based on Poisson and renewal processes are inadequate. For example, there might be a need for modeling the dependence of the model on the previous events or gap times. In such cases, models can be based on general intensity functions. For example, Lawless and Thiagarajah (1996) proposed a general

intensity-based model that can incorporate both calendar and local times together, and discussed hypothesis testing, interval estimation and model checking with applications to the reliability of repairable systems. Poisson and renewal processes are special cases of their models.

An important use of the general intensity-based models is the inclusion of covariates as a function of the history of a process. For example, Cigsar and Lawless (2012) discussed the use of the model

$$\lambda(t|H(t)) = \rho_0(t)\exp\{z(t)\beta\}, \qquad t \geq 0, \tag{2.13}$$

where $z(t) = I(N(t^-) > 0)I(B(t) \leq \Delta)$ and Δ is a positive pre-specified value. Note that the covariate $z(t)$ in the model (2.13) takes the value 1 for Δ time period after each event occurrence; otherwise, its value is 0. Thus, the intensity function (2.13) becomes $\rho_0(t)\exp\{\beta\}$ after each event occurrences for Δ time period. Otherwise, it is $\rho_0(t)$. A test for $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ gives whether there is any persistent effects of event occurrences over Δ time period. Since it is a function of the history, the covariate $z(t)$ is an internal covariate. Internal covariates may include vital information about a counting process. Miloslavsky et al. (2004) underline the importance of internal covariates in modeling of recurrent event processes.

As briefly aforementioned, modulated Poisson and renewal processes are also considered as general intensity-based models. These modulated processes are applied when there exist more complicated relationships between event occurrences and prior event histories. The intensity function of a modulated Poisson process under the multiplicative form can be given by

$$\lambda(t|H(t)) = \rho_0(t)\exp\{z'(t)\beta\}, \qquad t \geq 0, \tag{2.14}$$

where $z(t)$ is a vector of time-varying covariates allowed to include functions of the event history as well as external covariates $x(t)$. In the model (2.14), the history $H(t)$ is extended to include the covariate process $z^{(t)} = \{z(t),\ t \geq 0\}$ in $[0, t]$. With this analogy, the multiplicative models of the intensity function for the modulated renewal processes can be given as

$$\lambda(t|H(t)) = h_0(B(t))\exp\{z'(t)\beta\}, \qquad t \geq 0, \tag{2.15}$$

where t denotes chronological (global) time, $z(t)$ is a vector of time-varying covariates and $B(t)$ is the backward recurrence time.

General intensity-based models are discussed to some extent by Cook and Lawless (2007, Chapter 5). Another interesting application of the general intensity-based models for recurrent events is given by Simpson (2013). Since this model is the focal point of this thesis, we discuss it in Section 2.3.1 in more detail.

2.3 The Self-Controlled Case Series Model

We introduced the self-controlled case series (SCCS) model in Section 1.1. The SCCS model is based on a conditional Poisson process. This can be explained as follows. Let $\{N_i(t);\ t \geq 0\}$, $i = 1, 2, \ldots, N$, be a Poisson processes with the associated intensity function $\lambda_i(t|H_i(t))$ observed over the observation window $[a_i, b_i]$ for the ith individual, where $0 \leq a_i < b_i$. The number of events experienced by the ith individual over $[a_i, b_i]$ is denoted by n_i; that is, $N_i(a_i, b_i) = n_i$, $i = 1, 2, \ldots, N$. The intensity function of the ith individual in the SCCS model is given by

$$\lambda_i(t|x_i^{(t)}) = \alpha\psi(t)\exp\{\gamma_i + x_i'(t)\beta\}, \qquad t \geq 0, \tag{2.16}$$

where the parameter α represents the underlying age effect at time a_i, which is sometimes called the baseline age effect, $\psi(t)$ is age specific relative effect, γ_i represents the summation of all fixed covariates and random effects for the ith individual, where $i = 1, 2, \ldots N$. In the model (2.16), we assume that there are p exposures, so corresponding p covariates are included in the age-dependent external covariate vector at age t $x_i(t) = (x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))'$ with the corresponding $p \times 1$ vector of parameters $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$. In the SCCS model, covariates $x_i(t)$ are indicator functions for the exposed risk periods. For example, let e_{ij} denote the time of the start of the jth exposure for the ith individual, $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, p$. If the ith individual experiences the jth exposure at time t, then the covariate $x_{ij}(t)$ takes the value of 1 over the time interval $(e_{ij}, e_{ij} + \Delta_j]$; and $x_{ij}(t) = 0$ otherwise. This time interval is called the risk window, and the positive quantity Δ_j is often referred to as the length of the exposed risk window. In this thesis, we consider a single exposure, but the model can be extended to the multiple exposure cases in a straightforward way. The main goal in the SCCS design is to make inference about the parameter

β in (2.16), which represents the relative exposure effect. Thus, the SCCS model can be used to make inferences about the effect of an external exposure on the occurrences of an adverse recurrent event. Inference procedures for the parameter β in the SCCS model (2.16) is obtained by using a conditional probability of the observed event pattern given that the total number of events. We discuss the derivation of the conditional likelihood approach in Section 2.5.

Although the SCCS design has some attractive features, there exists some limitations. Firstly, the SCCS design can only provides estimates of the relative incidence. That is, it cannot estimate the absolute incidence. Another issue is that the occurrence of an event, as well as subsequent exposures, experienced by an individual in the SCCS design should be independent of the followup of this individual. That is, the process $\{N_i(t);\ t \geq 0\}$ should not affect the probabilistic characteristics of the followup periods and the covariate history $x_i^{(t)}$. Lastly, the events should occur independently. In other words, an earlier event should not influence the risk of occurrence of the subsequent events. For a disease such that the first event increases the risk of the subsequent events, such as myocardial infarction (MI), the SCCS model (2.16) may not be adequate. A positive event dependence model as the extension of the SCCS proposed by Simpson (2013) can be applied in some applications to relax this limitation. We next introduce this model.

2.3.1 The Positive Event Dependence Model for the SCCS Data

The positive event dependence self-controlled case series (PD-SCCS) model is an extension of the standard SCCS model. It allows the dependence on the past of a recurrent event process in a sense that the intensity function is adjusted with event occurrences. More formally, the PD-SCCS model is based on the intensity function

$$\lambda_i(t|H_i(t)) = \{\alpha_i + \delta N_i(t^-)\} \exp\{x_i'(t)\beta\}, \qquad t > 0, \tag{2.17}$$

where α_i is an individual specific baseline rate function at the age a_i, e^β is the relative risk, and $N_i(t^-)$ is the number of events just before time t for $i = 1, 2, \ldots, N$. As the name of the model implies, the positive-valued parameter δ represents the level of dependency of the model on the previous event occurrences. Therefore, the baseline

intensity function $\alpha_i + \delta N_i(t^-)$ in the model (2.17) increases at every event occurrences because of the increment in the term $N(t^-)$. A conditional likelihood function approach can be developed for making inferences and this approach retains all characteristic of a SCCS design as discussed in the previous section. This means that the estimation procedure requires only the inclusion of cases and the effects are estimated on a self-controlled manner. We discuss the details of the conditional likelihood approach in Section 2.5.3.

It should be noted that the model (2.17) is not a Poisson process because of the dependence on the history $H_i(t)$. As denoted by Simpson (2013), the model (2.17) is called nonhomogeneous pure birth process with immigration. As a special case, the model (2.17) becomes a nonhomogeneous Poisson process when $\delta = 0$. Therefore, a test of the null hypothesis $H_0 : \delta = 0$ against the alternative hypothesis $H_1 : \delta > 0$ can be developed to check whether there is such a dependency on the past event occurrences. It is worth noting that the baseline intensify function $\alpha_i + \delta N_i(t^-)$ is of additive form but the effects of the external covariates are modelled with a multiplicative term.

In the next section, we introduce the conditional likelihood approach to develop inference procedures in SCCS model and the PD-SCCS model.

2.4 Likelihood Based Inferences for the Cohort Model

Inference procedures for recurrent event processes have been extensively discussed by Cook and Lawless (2007). In this section, our goal is to introduce the likelihood function for the cohort model and briefly discuss some parametric and semi-parametric methods that will be used in the remaining part of this thesis. With the term cohort model, we mean a model for a large group of individuals followed-up over a time interval. An individual included in the cohort may be a case or control and may or may not be exposed to a condition.

Now, suppose that N independent individuals are included in the cohort. Let $\{N_i(t);\ t \geq 0\}$, $i = 1, 2, \ldots, N$, be a counting process with the associated intensify

function

$$\lambda_i(t|H_i(t)) = \rho_0(t)\exp\{x_i'(t)\beta\}, \qquad t \geq 0, \tag{2.18}$$

where $\rho_0(t)$ is the baseline rate function, β is a $p \times 1$ vector of unknown parameters and $x_i(t)$ is a subject specific $p \times 1$ vector of time varying covariates. Let the process $\{N_i(t);\ t \geq 0\}$, $i = 1, 2, \ldots, N$, be followed over the observation window $[a_i, b_i]$. The likelihood function of the outcome that "n_i events observed at times $t_{i1} < t_{i2} < \cdots < t_{in_i}$ over the interval $[a_i, b_i]$, $0 \leq a_i < b_i$" is given by

$$L_i = \left[\prod_{j=1}^{n_i} \rho_0(t_{ij})\exp(x_i'(t_{ij})\beta)\right] \exp\left\{-\int_{a_i}^{b_i} \rho_0(s)\exp(x_i'(s)\beta)ds\right\}. \tag{2.19}$$

For N independent process, the likelihood function is then

$$L = \prod_{i=1}^{N} L_i. \tag{2.20}$$

The derivation of the likelihood function (2.19) can be found in Cook and Lawless (2007, Section 2.1). Depending on the specification of the baseline rate function $\rho(t)$, the model (2.18) can be parametric, semi-parametric or weakly parametric with piecewise constant modeling. In the remaining parts of this section, we will introduce estimation methods based on these specifications of $\rho_0(t)$. We next start with the parametric cohort model.

2.4.1 Parametric Cohort Model

In the parametric cohort model, we specify the baseline rate function in the model (2.18) with a $k \times 1$ vector of parameters α. With this specification in the model (2.18), the baseline rate function is denoted by $\rho_0(t;\alpha)$ and the likelihood function is given by

$$L(\theta) = \prod_{i=1}^{N} L_i(\theta), \tag{2.21}$$

where

$$L_i(\theta) = \left[\prod_{j=1}^{n_i} \rho_0(t_{ij};\alpha)\exp(x_i'(t_{ij})\beta)\right] \exp\left\{-\int_{a_i}^{b_i} \rho_0(s)\exp(x_i'(s)\beta)ds\right\}, \tag{2.22}$$

and $\theta = (\alpha', \beta')'$. The log likelihood function is then

$$\begin{aligned} \ell(\theta) &= \log L(\theta), \\ &= \sum_{i=1}^{N} \ell_i(\theta), \end{aligned} \tag{2.23}$$

where

$$\ell_i(\theta) = \sum_{j=1}^{n_i} \left[\log \rho_0(t_{ij}; \alpha) + x_i'(t_{ij})\beta\right] - \int_{a_i}^{b_i} \rho_0(s; \alpha) \exp(x_i'(s)\beta) ds. \tag{2.24}$$

Let $U(\theta)$ denote the $(k+p)\times 1$ score vector with the components $U_\alpha(\theta)$ and $U_\beta(\theta)$, where $U_\alpha(\theta)$ is a $k \times 1$ vector with elements $\frac{\partial \ell(\theta)}{\partial \alpha_l}$, $l = 1, 2, \ldots, k$, and $U_\beta(\theta)$ is a $p \times 1$ vector with elements $\frac{\partial \ell(\theta)}{\partial \beta_q}$, $q = 1, 2, \ldots, p$. Then, for $l = 1, 2, \ldots, k$, we obtain

$$\frac{\partial \ell(\theta)}{\partial \alpha_l} = \sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} \frac{\partial \log \rho_0(t_{ij}; \alpha)}{\partial \alpha_l} - \int_{a_i}^{b_i} \frac{\partial \rho_0(s; \alpha)}{\partial \alpha_l} \exp(x_i'(s)\beta) ds \right], \tag{2.25}$$

and, for $q = 1, 2, \ldots, p$, we obtain

$$\frac{\partial \ell(\theta)}{\partial \beta_q} = \sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} x_{iq}(t_{ij}) - \int_{a_i}^{b_i} \rho_0(s; \alpha) x_{iq}(s) \exp(x_i'(s)\beta) ds \right]. \tag{2.26}$$

Let $I(\theta)$ denote the $(k+p)\times(k+p)$ information matrix. Then the information matrix is partitioned as follows.

$$I(\theta) = \begin{bmatrix} -\frac{\partial U_\alpha(\theta)}{\partial \alpha} & -\frac{\partial U_\alpha(\theta)}{\partial \beta} \\ -\frac{\partial U_\beta(\theta)}{\partial \alpha} & -\frac{\partial U_\beta(\theta)}{\partial \beta} \end{bmatrix} = \begin{bmatrix} I_{\alpha\alpha}(\theta) & I_{\alpha\beta}(\theta) \\ I_{\beta\alpha}(\theta) & I_{\beta\beta}(\theta) \end{bmatrix}, \tag{2.27}$$

where $I_{\alpha\alpha}(\theta)$ is a $k \times k$ matrix with elements $I_{\alpha_u\alpha_v}(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \alpha_u \alpha_v}$ for $u, v = 1, 2, \ldots, k$. Thus,

$$I_{\alpha_u\alpha_v}(\theta) = -\sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} \left(\frac{\partial^2 \log \rho(t_{ij}; \alpha)}{\partial \alpha_u \partial \alpha_v} \right) - \int_{a_i}^{b_i} \left(\frac{\partial^2 \rho(s; \alpha)}{\partial \alpha_u \partial \alpha_v} \right) \exp(x_i'(s)\beta) ds \right], \tag{2.28}$$

In (2.27), $I_{\alpha\beta}(\theta) = I_{\beta\alpha}(\theta)'$ is a $k \times p$ matrix with elements $I_{\alpha_u\beta_v}(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \alpha_u \beta_v}$, $u =$

$1, 2, \ldots, k$ and $v = 1, 2, \ldots, p$. Thus,

$$I_{\alpha_u\beta_v}(\theta) = \sum_{i=1}^{N} \left[\int_{a_i}^{b_i} x_{iv}(s) \left(\frac{\partial \rho(s;\alpha)}{\partial \alpha_u} \right) \exp(x_i'(s)\beta) ds \right]. \tag{2.29}$$

The remaining component, $I_{\beta\beta}(\theta)$ in (2.27) is a $p \times p$ matrix with elements $I_{\beta_u\beta_v}(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \beta_u \beta_v}$, where u and $v = 1, 2, \ldots, p$. Thus,

$$I_{\beta_u\beta_v}(\theta) = \sum_{i=1}^{N} \left[\int_{a_i}^{b_i} \rho(s;\alpha) x_{iu}(s)\, x_{iv}(s) \exp(x_i'(s)\beta) ds \right]. \tag{2.30}$$

The MLE, denoted by $\hat{\theta}$, of θ is the value of θ that maximizes $L(\theta)$, or equally $\ell(\theta)$. $\hat{\theta}$ can be obtained by solving the system of score equations $U(\theta) = 0$, where 0 is a $(k+p) \times 1$ vector of zeros. Under the usual regularity conditions (Andersen et al., 1993, pp. 420–421), $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{D} MVN(0, I^{-1}(\theta_0))$, where the notation $\xrightarrow{D}$ denotes the convergence in distribution. Furthermore, $\hat{\theta}$ is a consistent estimator of θ_0 so that asymptotically, $I^{-1}(\theta_0)$ can be replaced by $I^{-1}(\hat{\theta})$. The proofs of these results can be found in Andersen et al. (1993, Section VI.1.2). Our main interest of parameter is the relative exposure effect denoted by β in this thesis.

The maximization of $\ell(\theta)$ is usually done by an optimization software, such as `nlm` or `optim` function in R. They also provide asymptotic covariance matrix using the Hessian matrix $-I(\hat{\theta})$ so that we can obtain $I^{-1}(\hat{\theta})$. From this result, confidence intervals and hypothesis tests for θ can be developed. In the remaining chapters, we use the `optim` function in R to obtain estimates of parameters in the parametric cohort model.

2.4.2 The Semi-parametric Cohort Model

A semi-parametric regression model is defined by leaving the baseline rate function ρ_0 in the model (2.18) without any parametric specification. In the recurrent event framework, this model is usually called the Andersen-Gill model, which is an extended version of the Cox proportional hazards regression model for survival data. There are two methods to estimate the parameters in this case. These methods are based on either the profile likelihood function or the partial likelihood function (Cook and

Lawless, 2007, Section 3.42; Andersen et al., 1993, Section VII.2). Either methods can be used to develop inferences about β. Here, we consider the profile likelihood based approach. In addition to the notation stated in the previous section, we denote that $dN(t) = N(t) - N(t^-)$ for all $t > 0$. Because $N(t)$ is a step function, the first term of $\ell_i(\theta)$ given in the equation (2.24) can be written as

$$\sum_{j=1}^{n_i} [\log \rho_0(t_{ij}; \theta) + x_i'(t_{ij})\beta] = \int_{a_i}^{b_i} [\log \rho_0(t; \theta) + x_i'(t)\beta] \, dN_i(t). \tag{2.31}$$

Thus, the log likelihood function for individual i in (2.24) is re-written as

$$\ell_i(\theta) = \int_{a_i}^{b_i} \{[\log \rho_0(t; \theta) + x_i'(t)\beta] \, dN_i(t) - \exp\{x_i'(t)\beta\} d\mu_0(t)\}, \tag{2.32}$$

where $d\mu_0(t) = \rho_0(t)dt$. The log likelihood function is then given by

$$\ell(\theta) = \sum_{i=1}^{N} \ell_i(\theta). \tag{2.33}$$

In order to find $\hat{\beta}$ that maximizes (2.33), we solve the following substituted score equation (Cook and Lawless, 2007, Section 3.4.2) by treating $d\mu_0(t)$ as a parameter and considering β as fixed,

$$\sum_{i=1}^{N} [dN_i(s) - \exp\{x_i'(s)\beta\} d\mu_0(s)] = 0, \qquad s \geq 0. \tag{2.34}$$

The profile likelihood estimate of $d\mu_0(s)$ is then given by

$$d\tilde{\mu_0}(s; \beta) = \frac{\sum_{i=1}^{N} dN_i(s)}{\sum_{i=1}^{N} \exp\{x_i'(s)\beta\}}. \tag{2.35}$$

The $p \times 1$ vector of score functions $U_\beta(\theta) = \frac{\partial \ell(\theta)}{\partial \beta}$ becomes

$$U_\beta(\beta) = \sum_{i=1}^{N} \int_{a_i}^{b_i} x_i(s) \, [dN_i(s) - \exp\{x_i'(s)\beta\} d\mu_0(s)]. \tag{2.36}$$

By plugging (2.35) in (2.36) and solving the $p \times 1$ system of equations $U_\beta(\beta) = 0$, we can obtain the estimator $\hat{\beta}$ of β. As the number of individuals increases, $\sqrt{N}(\hat{\beta} - \beta)$

follows approximately a p-variate multivariate normal distribution with the $p \times 1$ mean vector 0 and a $p \times p$ efficient covariance matrix under some regularity conditions (Cook and Lawless, 2007).

The Andersen-Gill model can be fitted with the `coxph` function in `R`. The partial likelihood function obtained by maximizing the Cox partial likelihood function for the survival model regarding β is the same as the components of the score vector in (2.36) (Andersen et al., 1993, Section VII.2.1). Thus, we can obtain $\hat{\beta}$ and its covariance matrix using the `coxph` function in `R`. Therefore, in this thesis, we use the function called `coxph` in `survival` package to obtain the estimate of β and its estimated covariance matrix.

2.4.3 The Piecewise Constant Rates Models

Models with piecewise constant rates provide more flexibility in modeling recurrent events comparing with fully parametric cohort models discussed in Section 2.4.1. With the piecewise constant rates models, the baseline rate function $\rho_0(t)$ in (2.18) is weakly parameterized as explained below.

Suppose that the recurrent event process $\{N(t);\ t \geq 0\}$ is observed over a fixed observation window $(a, b]$. With the piecewise constant rates models, we consider a partition of the observation window into a prespecified number of pieces as follows. Let $(a, b]$ be partitioned into K non-overlapping intervals denoted by $(c_{k-1}, c_k]$ for $k = 1, 2, \ldots, K$, where $c_0 = a$ and $c_k = b$. Now, suppose that the intensify function of the recurrent event process $\{N(t);\ t \geq 0\}$ is defined by (2.18), where the baseline rate function is specified for $k = 1, 2, \ldots, K$ as

$$\rho_0(t; \alpha) = \alpha_k, \text{ if } t \in (c_{k-1}, c_k], \tag{2.37}$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)'$ is a $K \times 1$ vector of unknown parameters. With the specification (2.37), the baseline rate function is constant with the rate α_k over the interval $(c_{k-1}, c_k]$ for $k = 1, 2, \ldots K$. Such a specification may provide close approximations for various shapes of baseline rate functions, and results in flexibility in model fitting under reasonable simplicity because of constant rate specifications over pieces. It should be noted that the prespecification of the number of pieces K is an important issue because as K becomes large, the model will have more flexibility, although more

parameter estimations are needed. As recommended by Cook and Lawless (2007, Section 3.3), choosing the number of pieces K any number between 3 to 10 is appropriate for many situations in practice.

Now, suppose that N independent processes are included in a study with the intensity function of the ith process, $i = 1, 2, \ldots, N$, given as

$$\rho_0(t;\alpha)\exp\{x_i'(t)\beta\}, \qquad t \geq 0, \tag{2.38}$$

where $\rho_0(t;\alpha)$ is defined in (2.37), β is a $p\times 1$ vector of unknown parameters and $x_i(t)$ is a subject specific $p\times 1$ vector of time varying covariates. Let $(a_i, b_i]$ be the observation window of the ith process, $i = 1, 2, \ldots, N$. In the following development, we let $a = \min\{a_1, a_2, \ldots, a_N\}$, $b = \max\{b_1, b_2, \ldots, b_N\}$ and define $w_k(t)$ as the indicator function taking value of 1 when t is in $(c_{k-1}, c_k]$, $k = 1, 2, \ldots, K$. That is, $w_k(t) = I(c_{k-1} < t \leq c_k)$. Suppose that data set includes the event times and the values of covariates $\{t_{ij}, x_{iq}(t_{ij});\ i = 1, 2, \ldots, N,\ j = 1, 2, \ldots, n_i,\ q = 1, 2, \ldots p\}$. We also let n_{ik} denote the total number of events occurred in $(c_{k-1}, c_k]$, $k = 1, 2, \ldots, K$; that is, $n_{ik} = \sum_{j=1}^{n_i} w_k(t_{ij})$. Following (2.21), the likelihood function with the intensity function (2.38) can be written as

$$L(\theta) = \prod_{k=1}^{K}\left[\left\{\prod_{i=1}^{N}\alpha_k^{n_{ik}}\exp\left(\sum_{j=1}^{n_i}x_i'(t_{ij})\beta\right)\right\}\cdot\exp\left(-\alpha_k\sum_{i=1}^{N}\int_{c_{k-1}}^{c_k}\exp\{x_i'(s)\beta\}ds\right)\right], \tag{2.39}$$

where $\theta = (\alpha', \beta')'$ is a $(K + p) \times 1$ vector of unknown parameters. The log likelihood function is then given by

$$\ell(\theta) = \sum_{k=1}^{K} n_{\cdot k}\log(\alpha_k) + \sum_{i=1}^{N}\left\{\sum_{j=1}^{n_i}x_i'(t_{ij})\beta - \sum_{k=1}^{K}\alpha_k\int_{c_{k-1}}^{c_k}\exp\{x_i'(s)\beta\}ds\right\}, \tag{2.40}$$

where $n_{\cdot k} = \sum_{i=1}^{N} n_{ik}$ denotes the total number of events occurrence for all individuals over $(c_{k-1}, c_k]$ for $k = 1, 2, \ldots, K$.

Let $U(\theta)$ be a $(K + p) \times 1$ score vector with the components $U_\alpha(\theta)$ and $U_\beta(\theta)$, where $U_\alpha(\theta)$ is a $K \times 1$ vector with elements $\frac{\partial\ell(\theta)}{\partial\alpha_k}$, $k = 1, \ldots, K$, and $U_\beta(\theta)$ is a $p \times 1$

vector with elements $\frac{\partial \ell(\theta)}{\partial \beta_q}$, $q = 1, \ldots, p$. Then, for $k = 1, \ldots, K$, we obtain

$$\frac{\partial \ell(\theta)}{\partial \alpha_k} = \frac{n_{.k}}{\alpha_k} - \sum_{i=1}^{N} S_{ik}(\beta), \tag{2.41}$$

where $S_{ik}(\beta) = \int_{c_{k-1}}^{c_k} \exp\{x_i'(s)\beta\} ds$. Let $\tilde{\alpha}(\beta) = \{\tilde{\alpha}_1(\beta), \tilde{\alpha}_2(\beta), \ldots, \tilde{\alpha}_K(\beta)\}$ be a $K \times 1$ vector of estimators that maximizes $\ell(\theta)$ when β is considered as a $p \times 1$ vector of fixed constants. Solving $\frac{\partial \ell(\theta)}{\partial \alpha_k} = 0$ in (2.41) for the estimator $\tilde{\alpha}_k(\beta)$ of α_k gives

$$\tilde{\alpha_k}(\beta) = \frac{n_{.k}}{\sum_{i=1}^{N} S_{ik}(\beta)}, \qquad k = 1, \ldots, K. \tag{2.42}$$

For $q = 1, \ldots, p$, we obtain

$$\frac{\partial \ell(\theta)}{\partial \beta_q} = \sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} x_{iq}(t_{ij}) - \sum_{k=1}^{K} \alpha_k \int_{c_{k-1}}^{c_k} \exp\{x_i'(s)\beta\} x_{iq}(s) ds \right]. \tag{2.43}$$

By replacing α_k in (2.40) with $\tilde{\alpha}_k(\beta)$, we obtain the profile likelihood function $\ell(\tilde{\alpha}(\beta), \beta)$ for β, which is given by

$$\ell(\tilde{\alpha}(\beta), \beta) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left\{ x_i'(t_{ij})\beta - \sum_{k=1}^{K} w_k(t_{ij}) \sum_{l=1}^{N} \int_{c_{k-1}}^{c_k} \exp\{x_l'(s)\beta\} ds \right\}. \tag{2.44}$$

From the function (2.44), we can obtain the $p \times 1$ vector of partial score functions for β by taking the derivative $\frac{\partial \ell(\tilde{\alpha}(\beta), \beta)}{\partial \beta}$, which gives

$$\sum_{i=1}^{N} \sum_{j=1}^{n_i} \left[x_i(t_{ij}) - \frac{\sum_{k=1}^{K} w_k(t_{ij}) \int_{c_{k-1}}^{c_k} \sum_{l=1}^{N} \exp\{x_l'(s)\beta\} x_l(s) ds}{\sum_{k=1}^{K} w_k(t_{ij}) \int_{c_{k-1}}^{c_k} \sum_{l=1}^{N} \exp\{x_l'(s)\beta\} ds} \right]. \tag{2.45}$$

The value of β that maximizes $\ell(\tilde{\alpha}(\beta), \beta)$, denoted by $\hat{\beta}$, can be then obtained by solving the system of equations defined by $\frac{\partial \ell(\hat{\alpha}(\beta), \beta)}{\partial \beta} = 0$, where 0 is a $p \times 1$ vector of zeros. By replacing β in (2.42) with $\hat{\beta}$ gives the vector maximum likelihood estimates $\hat{\alpha}$ of the vector α. Inverse of the observed information matrix, $I^{-1}(\hat{\theta})$ is used to obtain the estimates of covariance matrix for $\hat{\theta} = (\hat{\alpha}', \hat{\beta}')$. We can use several optimization functions built in `R` to obtain the estimate of θ and $I^{-1}(\hat{\theta})$. In this thesis, we use `optim` function to estimate the parameter β in model (2.38) and its variance estimate.

2.5 Likelihood Based Inference for the SCCS Models

In this section, we introduce the likelihood function and the related inference procedures for the SCCS and PD-SCCS models. We discuss parametric and semi-parametric methods to estimate model parameters, as well as constructing confidence intervals and hypothesis testing.

2.5.1 The Parametric SCCS Model

The standard SCCS likelihood function is derived from a Poisson process based conditional likelihood of the cohort model. Suppose that the process $\{N_i(t);\ t \geq 0\}$, $i = 1, 2, \ldots, N$, with the intensity function

$$\lambda_i(t|H_i(t)) = \rho_i(t|x_i^{(t)}) = \alpha\psi(t)\exp\{\gamma_i + x_i'(t)\beta\}, \qquad t \geq 0, \tag{2.46}$$

is under observation over a fixed time interval $[a_i, b_i]$. As discussed in Section 2.2.1, the model (2.46) is a nonhomogeneous Poisson process. Thus, for $n_i = 0, 1, \ldots,$

$$\Pr\{N_i(a_i, b_i) = n_i\} = \frac{1}{n_i!}[\mu_i(a_i, b_i)]^{n_i}\ e^{-\mu_i(a_i,b_i)}, \tag{2.47}$$

where $\mu_i(a_i, b_i) = \int_{a_i}^{b_i} \rho_i(u)du$ is the mean function of the process. Note that all probabilities including (2.47) and the mean and the rate functions are conditional on the values of external covariates $x_i(t)$, but for simplicity we drop the conditional notation. Now, we let A_i denote the outcome "$N_i(a_i, b_i) = n_i$ events occurred at times $t_{i1} < t_{i2} < \cdots < t_{in_i}$ over the interval $[a_i, b_i]$, where $0 \leq a_i < b_i$" and B_i denote the outcome "$N_i(a_i, b_i) = n_i$". Then, for $i = 1, 2, \ldots, N$, we have

$$\Pr(A_i|B_i) = \frac{\Pr(A_i, B_i)}{\Pr(B_i)} = \frac{\Pr(A_i)}{\Pr(B_i)}, \tag{2.48}$$

where, from (2.19), the probability in the numerator of (2.48) is

$$\left[\prod_{j=1}^{n_i} \lambda_i(t|x_i^{(t)})\right]\ \exp\left\{-\int_{a_i}^{b_i} \lambda_i(u|x_i^{(u)})du\right\}, \tag{2.49}$$

and the probability in the denominator of (2.48) is given in (2.47). As a result, we obtain the conditional likelihood contribution of the ith individual process for the SCCS model as

$$n_i! \frac{\prod_{j=1}^{n_i} \lambda_i(t_{ij}|x_i(t_{ij}))}{\left(\int_{a_i}^{b_i} \lambda_i(u|x_i^{(u)})du\right)^{n_i}}. \tag{2.50}$$

The function $\psi(t)$ in (2.46) can be parametrically specified or left without any parametric specification, which leads to the parametric SCCS model or the semi-parametric SCCS, respectively. We first consider the parametric specification. The semi-parametric SCCS is discussed in the next section.

As discussed by Farrington (1995), the age-specific relative incidence function $\psi(t)$ can be defined as constant rate functions over a finite number of age intervals. It is assumed then that the age effect is the same for every individual over a given age group. For simplicity, we take $\psi(t) = \psi$. Then, from (2.46) and (2.50), the likelihood function of the SCCS model of N independent processes can be given as

$$\prod_{i=1}^{N} \frac{n_i! \prod_{j=1}^{n_i} \exp\{x_i'(t_{ij})\beta\}}{\left(\int_{a_i}^{b_i} \exp\{x_i'(s)\beta\}ds\right)^{n_i}}. \tag{2.51}$$

Note that the likelihood function (2.51) does not include the parameter γ_i, which means that all time fixed covariates over $[a_i, b_i]$ are automatically adjusted and thus the method is self-controlled. Also, if $N_i(a_i, b_i) = 0$, then the contribution of the ith individual to the product term over N individuals in (2.51) is one, which means that the controls can be ignored to make inference on the target parameter vector β. Let m denote the number of individuals in N with at least one event during their follow up; that is, the number of cases in N. Then, from (2.51), the likelihood function for the SCCS is given by

$$L_{SCCS}(\beta) = \prod_{i=1}^{m} \frac{n_i! \prod_{j=1}^{n_i} \exp\{x_i'(t_{ij})\beta\}}{\left(\int_{a_i}^{b_i} \exp\{x_i'(s)\beta\}ds\right)^{n_i}}. \tag{2.52}$$

The log likelihood function is then given by

$$\ell_{SCCS}(\beta) = \sum_{i=1}^{m} \left\{ \log n_i! + \sum_{j=1}^{n_i} x_i'(t_{ij})\beta - n_i \log\left(\int_{a_i}^{b_i} \exp\{x_i'(s)\beta\}ds\right)\right\}. \tag{2.53}$$

Let $U(\beta) = \frac{\partial}{\partial\beta}\ell_{SCCS}(\beta)$ denote the $p \times 1$ score vector, whose elements are given by $U_q(\beta) = \frac{\partial}{\partial\beta_q}\ell_{SCCS}(\beta)$, $q = 1, 2, \ldots, p$. Then, for $q = 1, 2, \ldots, p$, we obtain

$$\frac{\partial\ell_{SCCS}(\beta)}{\partial\beta_q} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{iq}(t_{ij}) - \sum_{i=1}^{m} \frac{n_i \int_{a_i}^{b_i} x_{iq}(s)\exp\{x_i'(s)\beta\}ds}{\int_{a_i}^{b_i}\exp\{x_i'(s)\beta\}ds}. \tag{2.54}$$

Let $I(\beta)$ denote the $p \times p$ information matrix with elements $I_{uv}(\beta) = \frac{-\partial^2}{\partial\beta_u\beta_v}\ell_{SCCS}(\beta)$ for $u, v = 1, 2, \ldots, p$. Thus, $I_{uv}(\beta)$ is given by

$$\sum_{i=1}^{m} n_i \left\{ \frac{\int_{a_i}^{b_i} x_{iu}(s)x_{iv}(s)e^{x_i'(s)\beta}ds \int_{a_i}^{b_i} e^{x_i'(s)\beta}ds - \int_{a_i}^{b_i} x_{iu}(s)e^{x_i'(s)\beta}ds \int_{a_i}^{b_i} x_{iv}(s)e^{x_i'(s)\beta}ds}{\left[\int_{a_i}^{b_i} e^{x_i'(s)\beta}ds\right]^2} \right\}. \tag{2.55}$$

The maximum likelihood estimator $\hat{\beta}$ of β can be obtained by solving the system of equations defined by $U(\beta) = 0$ where 0 is a $p \times 1$ vector of zeros . Under some regularity conditions $\sqrt{m}(\hat{\beta} - \beta) \xrightarrow{D} MVN(0, I_\beta^{-1}(\beta))$, where $I_\beta(\beta) = \lim_{m\to\infty} \frac{1}{m} I(\beta)$. (Farrington, 1995). As $m \to \infty$, $I_\beta^{-1}(\beta)$ can be replaced by $I_\beta^{-1}(\hat{\beta})$. In this thesis, we use the `standardsccs` function of the `SCCS` package in `R` to obtain $\hat{\beta}$ and $I_\beta^{-1}(\hat{\beta})$. From this result, we can develop confidence intervals and hypothesis tests for β such that $H_0 : \beta = 0$ and $H_1 : \beta \neq 0$.

2.5.2 The Semi-Parametric SCCS Model

In this section, we introduce the semi-parametric SCCS model proposed by Farrington and Whitaker (2006) to avoid bias in the estimation of the exposure effects caused by the misspecification of age groups in the parametric SCCS model. This semi-parametric SCCS model leaves the age effect $\psi(t)$ in (2.46) unspecified. Only requirement for $\psi(t)$ is that it is a non-negative and bounded function (Farrington and Whitaker, 2006). Since the underlying age effect α and summation of all covariates and random effect γ_i in (2.46) are cancelled out in the conditional likelihood function approach, we denote the intensity function as follows.

$$\lambda_i(t|x_i) = \psi(t)\exp\{x_i(t)\beta\}, \qquad t \geq 0. \tag{2.56}$$

Note that we consider a single exposure variable $x_i(t)$ for simplicity, but the procedure can be extended to the case for multiple exposures as well. Corresponding SCCS likelihood function is given by

$$L = \prod_{i=1}^{m} L_i(\psi, \beta) \propto \prod_{i=1}^{m} \frac{\prod_{j=1}^{n_i} \psi(t) \exp\{x_i(t_{ij})\beta\}}{\left[\int_{a_i}^{b_i} \exp\{x_i(s)\beta\} d\Psi(s)\right]^{n_i}}, \tag{2.57}$$

where $\Psi(t) = \int_a^t \psi(s)ds$ is the cumulative relative age effect to age t and $a = \min\{a_1, a_2, \ldots, a_m\}$. In the semi-parametric procedure, the function $\psi(t)$ in (2.57) is left unspecified. Let ζ denote a set of all distinct event times of t_{ij} for all m cases. Suppose the number of total distinct events is M such that $\zeta = \{s_1, s_2, \ldots, s_M\}$. The non-parametric maximum likelihood estimator of Ψ is a non-decreasing step function with jumping heights $\Delta\Psi(t)$ for $t \in \zeta$. Note that with the step function specification, the numerator of (2.57) is given by $\prod_{j=1}^{n_i} \Delta\Psi(t_{ij}) \exp\{x_i(t_{ij})\beta\}$. Thus, semi-parametric likelihood function from (2.57) for m cases can be written as

$$L(\Psi(t), \beta) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{\Delta\Psi(t_{ij}) \exp\{x_i(t_{ij})\beta\}}{\int_{a_i}^{b_i} \exp\{x(t)\beta\} d\Psi(t)}. \tag{2.58}$$

As explained by Farrington and Whitaker (2006), the maximum likelihood estimator $\hat{\beta}$ and $\hat{\Psi}(t)$ can be obtained as follows. First, for $r = 2, \ldots, M$, let $\Delta\Psi(s_r) = \exp(\alpha_r)$, and let $\alpha_1 = 0$ without loss of generality. The justification of this specification of $\Delta\Psi$ is given by Farrington and Whitaker (2006). Then, we define $w_{ir} = I_{(a_i, b_i]}(s_r)$. Note that, if s_r is in $(a_i, b_i]$, then $w_{ir} = 1$, and $w_{ir} = 0$, otherwise. Thus, the semi-parametric likelihood function (2.58) can be re-written as

$$L(\Psi(t), \beta) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{\exp\{\alpha_{ij} + x_i(t_{ij})\beta\}}{\sum_{r=1}^{M} w_{ir} \exp\{\alpha_r + x_i(s_r)\beta\}}, \tag{2.59}$$

where $\alpha_{ij} = \sum_{r=1}^{M} I_{\{t_{ij}\}}(s_r)\alpha_r$. Corresponding log likelihood function $\ell(\Psi(t), \beta)) = \log L(\Psi(t), \beta))$ is given by

$$\ell(\Psi(t), \beta) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[\alpha_{ij} + x_i(t_{ij})\beta - \log\left(\sum_{r=1}^{M} w_{ir} \exp\{\alpha_r + x_i(s_r)\beta\}\right)\right] \tag{2.60}$$

The components $U_{\alpha_r} = \frac{\partial \ell(\Psi(t),\beta)}{\partial \alpha_r}$ in the score vector $U_\alpha = (U_{\alpha_2}, U_{\alpha_3}, \ldots, U_{\alpha_M})'$ is given by

$$U_{\alpha_r} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} I_{\{t_{ij}\}}(s_r) - \sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{w_{ir}\exp\{\alpha_r + x_i(s_r)\beta\}}{\sum_{r=1}^{M} w_{ir}\exp\{\alpha_r + x_i(s_r)\beta\}}. \tag{2.61}$$

The score function $U_\beta = \frac{\partial \ell(\Psi(t),\beta)}{\partial \beta}$ is given by

$$U_\beta = \sum_{i=1}^{m}\sum_{j=1}^{n_i} x_i(t_{ij}) - \frac{\sum_{r=1}^{M} w_{ir}x_i(s_r)\exp\{\alpha_r + x_i(s_r)\beta\}}{\sum_{r=1}^{M} w_{ir}\exp\{\alpha_r + x_i(s_r)\beta\}}. \tag{2.62}$$

The maximum likelihood estimators $\hat{\Psi}$ and $\hat{\beta}$ can be calculated by solving both $U_\alpha = 0$, where 0 is an $(M-1)$ vector of zeroes, and $U_\beta = 0$ for α_r, $r = 2, 3, \ldots, M$, and β, simultaneously. Inverse of the information matrix $I^{-1}(\Psi, \beta)$ gives the asymptotic $M \times M$ covariance matrix of $\hat{\Psi}$ and $\hat{\beta}$. The components of $I^{-1}(\Psi, \beta)$ are the negative of the derivatives of U_α and U_β with respect to α_r, $r = 2, 3, \ldots, M$, and β. In this thesis, we use the `semisccs` function of the `SCCS` package in `R` to calculate the values of $\hat{\beta}$ and $\hat{\Psi}$, as well as their corresponding standard errors.

Farrington and Whitaker (2006) showed that the maximum likelihood estimators of β and Ψ are consistent. Also, as the number of cases m increases and thus corresponding number of events n increases, $\sqrt{n}(\hat{\beta} - \beta)$ converges to a normal distribution with mean 0 and an efficient asymptotic variance. Because we estimate M parameters in the semi-parametric SCCS model, fitting a semi-parametric SCCS model can be computationally demanding. Thus, this issue may restrict the usage of the semi-parametric SCCS model in some settings. We discuss this issue in Chapter 3 of our simulation study.

2.5.3 The Positive Event Dependence SCCS Model

We introduced the positive event dependent self-controlled case series (PD-SCCS) model in Section 2.3.1. In this section, we develop the likelihood function for the PD-SCCS model and discuss inference procedures based on it.

Suppose that $\{N_i(t);\ t \geq 0\}$, $i = 1, 2, \ldots, N$, are independent counting processes

with the intensity function

$$\lambda_i(t|H_i(t)) = \{\alpha_i + \delta N_i(t^-)\} \exp\{x_i'(t)\beta\}, \qquad t > 0, \tag{2.63}$$

where α_i is an individual heterogeneity baseline effect, e^β is the relative risk regarding the exposure, $N_i(t^-)$ is the number of events just before time t, and the parameter δ (≥ 0) represents the level of dependency of the model on the previous event occurrences. With the PD-SCCS model given in (2.63), the likelihood of the outcome that "$N(a_i, b_i) = n_i$, $a_i \leq t_{i1} < \cdots < t_{in_i} \leq b_i$ in $[a_i, b_i]$ for $i = 1, 2, \ldots, N$" is given as

$$L(\theta) = \prod_{i=1}^{N} L_i(\theta), \tag{2.64}$$

where θ is the vector of parameters including $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)'$, δ, and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ and

$$\begin{aligned} L_i =& \frac{\Gamma(\frac{\alpha_i}{\delta} + n_i)}{\Gamma(\frac{\alpha_i}{\delta})} \exp\left\{-\alpha_i \int_{a_i}^{b_i} \exp(x_i'(s)\beta) ds\right\} \\ & \cdot \exp\left\{-\delta \int_{a_i}^{b_i} N_i(u^-) \exp(x_i'(u)\beta) du\right\} \prod_{j=1}^{n_i} \delta e^{x_i'(t_{ij})\beta}. \end{aligned} \tag{2.65}$$

As discussed by Simpson (2013), by the factorization criterion for sufficiency, the number of events $N_i(a_i, b_i)$ is a sufficient statistic for the subject specific parameter α_i in the model (2.63). Since the parameters of interest are δ and parameters included in the vector β and the α_i are nuisance parameters, a conditional likelihood function given the values of $N_i(a_i, b_i) = n_i$ can be applied (Cox and Hinkley, 1974). To do this, we need to find the distribution of $N_i(a_i, b_i)$. This can be found by integrating $L_i(\theta)$ over all possible permutations of the set $\{t_{i1}, t_{i2}, \ldots, t_{in_i}\}$ in the interval $(a_i, b_i]$. Let $\Pr\{N_i(a_i, b_i) = n_i\} = p(n_i)$. Thus, integrating (2.65) over all event times $t_i = \{t_{i1}, t_{i2}, \ldots, t_{in_i}\}$, the probability mass function $p(n_i)$ can be written as

$$\begin{aligned} p(n_i) &= \int \ldots \int L_i dt_{i1} dt_{i2} \ldots dt_{in_i}, \\ &\propto \int \ldots \int \exp\left(-\delta \int_{a_i}^{b_i} N_i(s^-) \exp\{x_i'(s)\beta\} ds\right) \prod_{j=1}^{n_i} e^{x_i'(t_{ij})\beta} dt_{i1} dt_{i2} \ldots dt_{in_i}, \end{aligned} \tag{2.66}$$

where the multiple integrals in (2.66) are taken over the region $a_i \leq t_{i1} < t_{i2} <$

$\cdots < t_{in_i} \leq b_i$. Because $N_i(t^-)$ is a step function jumping at events, the term of $\int_{a_i}^{b_i} N_i(s^-)\exp\{x_i'(s)\beta\}ds$ in (2.66) is a function of ordered event times $t_{i1} < t_{i2} < \cdots < t_{in}$. As given by Simpson (2013), this integral can be written as

$$\int_{a_i}^{b_i} N_i(s^-)\exp\{x_i'(s)\beta\}ds = \sum_{j=1}^{n_i}\int_{t_{ij}}^{b_i}\exp\{x_i'(s)\beta\}ds. \tag{2.67}$$

By plugging the right-hand side of (2.67) in (2.66), the $p(n_i)$ can be written as

$$\begin{aligned} p(n_i) &\propto \int\ldots\int \exp\left(-\delta\sum_{j=1}^{n_i}\int_{t_{ij}}^{b_i}\exp\{x_i'(s)\beta\}ds\right)\prod_{j=1}^{n_i} e^{x_i'(t_{ij})\beta}dt_{i1}dt_{i2}\ldots dt_{in_i}, \\ &\propto \int\ldots\int \prod_{j=1}^{n_i} e^{x_i'(t_{ij})\beta}\cdot\exp\left(-\delta\int_{t_{ij}}^{b_i}\exp\{x_i'(s)\beta\}ds\right)dt_{i1}dt_{i2}\ldots dt_{in_i}. \end{aligned} \tag{2.68}$$

It should be noted that the elements of event times $\{t_{i1},\ldots,t_{in_i}\}$ to obtain integrand are not necessarily in increasing order anymore, and the integrand of a product in j in (2.68) is invariant with respect to the order of the t_{ij}. That is, for any sets of order statistics $\{t_{i(1)},\cdots,t_{i(n_i)}\}$, there are $n_i!$ cases to build the same marginal distribution. Also note that integrand $\prod_{j=1}^{n_i} e^{x_i'(t_{ij})\beta}$ is symmetric in $t_{i1}, t_{i2},\ldots,t_{in_i}$. Considering those steps, (2.68) can be re-written as,

$$\begin{aligned} p(n_i) &\propto \frac{1}{n!}\prod_{j=1}^{n_i}\int_{a_i}^{b_i} e^{x_i'(t_{ij})\beta}\cdot\exp\left(-\delta\int_{t_{ij}}^{b_j}\exp(x_i'(v)\beta)dv\right)dt_{ij}, \\ &= \frac{1}{n!\delta^{n_i}}\left[1-\exp\left\{-\delta\int_{a_i}^{b_i}\exp(x_i(v)\beta)dv\right\}\right]^{n_i}. \end{aligned} \tag{2.69}$$

Then, from (2.48), where the probability $\Pr(A_i)$ in the numerator is replaced by (2.65) and the probability $\Pr(B_i)$ in the denominator is replaced by (2.69), the conditional likelihood contribution of the ith individual with the PD-SCCS model (2.63) is given by

$$L_{PD} = \prod_{i=1}^{N} n_i!\exp\left(-\delta\int_{a_i}^{b_i} N_i(s^-)\exp\{x_i'(s)\beta\}ds\right)\cdot\prod_{j=1}^{n_i}\left(\frac{\delta\exp\{x_i'(t_{ij})\beta\}}{1-\exp\left\{-\delta\int_{a_i}^{b_i} e^{x_i'(v)\beta}dv\right\}}\right). \tag{2.70}$$

Similar to the SCCS design, there exists no contribution to the likelihood from controls in the study. Thus, only cases need to be sampled. Let m denote the number of cases,

then the likelihood function for the PD-SCCS model can be written as

$$L_{PD} = \prod_{i=1}^{m} n_i! \exp\left(-\delta \int_{a_i}^{b_i} N_i(s^-) \exp\{x_i'(s)\beta\} ds\right) \cdot \prod_{j=1}^{n_i} \left(\frac{\delta \exp\{x_i'(t_{ij})\beta\}}{1 - \exp\left\{-\delta \int_{a_i}^{b_i} e^{x_i'(v)\beta} dv\right\}}\right). \tag{2.71}$$

As discussed by Simpson (2013), in the limit as δ approaches 0, the likelihood function (2.71) converges to the SCCS likelihood function given in (2.52). From (2.71), we can drive the ML estimators of the parameters in the model (2.63). To do this, we write the log likelihood as follows.

$$\begin{aligned} \ell_{PD} = &\sum_{i=1}^{m} \Big[\log n! - \delta \int_{a_i}^{b_i} N_i(s^-) \exp\{x_i'(s)\beta\} ds \\ &+ \sum_{j=1}^{n_i} \log \delta + x_{ij}'(s)\beta - \log\left(1 - \exp\left\{-\delta \int_{a_i}^{b_i} e^{x_i'(v)\beta} dv\right\}\right) \Big]. \end{aligned} \tag{2.72}$$

For simplicity, we assume that there is a single exposure, and let $\theta = (\beta, \delta)'$. The 2 $\times$ 1 score vector $U(\theta) = (U_\beta(\theta), U_\delta(\theta))' = (\frac{\partial l_{PD}}{\partial \beta}, \frac{\partial l_{PD}}{\partial \delta})'$ is given by the components

$$\begin{aligned} U_\beta(\theta) &= \sum_{i=1}^{m} \Big[\sum_{j=1}^{n_i} x_i(t_{ij}) - \delta \int_{a_i}^{b_i} x_i(s)\{N_i(s^-) + n_i g_i(\theta)\} e^{x_i(s)\beta} ds \Big], \\ U_\delta(\theta) &= \sum_{i=1}^{m} \Big[\frac{n!}{\delta} - \int_{a_i}^{b_i} \{N_i(s^-) + n_i g_i(\theta)\} e^{x_i(s)\beta} ds \Big], \end{aligned} \tag{2.73}$$

where $g_i(\theta) = \exp\{-\delta \int_{a_i}^{b_i} e^{x_i(s)\beta} ds\} / (1 - \exp\{-\delta \int_{a_i}^{b_i} e^{x_i(v)\beta} dv\}$. Let $I(\theta)$ denote the 2 $\times$ 2 information matrix. Then, the information matrix is partitioned as follows.

$$I(\theta) = \begin{bmatrix} -\frac{\partial U_\beta(\theta)}{\partial \beta} & -\frac{\partial U_\beta(\theta)}{\partial \delta} \\ -\frac{\partial U_\delta(\theta)}{\partial \beta} & -\frac{\partial U_\delta(\theta)}{\partial \delta} \end{bmatrix} = \begin{bmatrix} I_{\beta\beta}(\theta) & I_{\beta\delta}(\theta) \\ I_{\delta\beta}(\theta) & I_{\delta\delta}(\theta) \end{bmatrix}, \tag{2.74}$$

where $I_{\beta\delta}(\theta) = I_{\delta\beta}(\theta)$. The components of the information function are given by

$$
\begin{aligned}
I_{\beta\beta}(\theta) &= \delta \sum_{i=1}^{m} \int_{a_i}^{b_i} x_i^2(v)\{N_i(v^-) + n_i g_i(\theta)\} e^{x_i(v)\beta} dv \\
&\quad - \delta^2 \sum_{i=1}^{N} n_i g_i(\theta)\{1 + g_i(\theta)\} \left(\int_{a_i}^{b_i} x_i(s) e^{x_i(s)\beta} ds \right)^2, \\
I_{\delta\beta}(\theta) &= \sum_{i=1}^{m} \int_{a_i}^{b_i} x_i(s) \Big(N_i(s^-) + n_i g_i(\theta) \\
&\quad \cdot \left[1 - \delta\{1 + g_i(\theta)\} \int_{a_i}^{b_i} e^{x_i(v)\beta} dv \right] \Big) e^{x_i(s)\beta} ds, \\
I_{\delta\delta}(\theta) &= \sum_{i=1}^{m} n! \left[\frac{1}{\delta^2} - g_i(\theta)\{1 + g_i(\theta)\} \left(\int_{a_i}^{b_i} e^{x_i(s)\beta} ds \right)^2 \right].
\end{aligned} \tag{2.75}
$$

The maximum likelihood estimators $\hat{\beta}$ and $\hat{\delta}$ can be obtained by solving the score equations $U_\beta(\theta) = 0$ and $U_\delta(\theta) = 0$, simultaneously. Under the regularity conditions, it can be shown that, as $m \to \infty$, $I_{\beta\beta}^{-\frac{1}{2}}(\theta)(\hat{\beta} - \beta) \xrightarrow{D} N(0, 1)$ (Simpson, 2013). Since $\hat{\theta}$ is a consistent estimator of θ, as the number of cases increases, $I_{\beta\beta}^{-1}(\theta)$ can be replaced by $I_{\beta\beta}^{-1}(\hat{\theta})$.

When the true value of a parameter is on the boundary point, the maximum likelihood estimator behaves differently (Moran 1971). Note that, since the parameter δ is non-negative, when testing the absence of dependence to the number of previous events $N(t^-)$ in the model (2.63) (i.e. testing $H_0 : \delta = 0$ vs. $H_1 : \delta > 0$), an adjustment is needed to find the asymptotic distribution of $(\hat{\delta} - \delta)$ as m approaches infinity. This issue is discussed by Moran (1971).

In our thesis, the `optim` function in `R` is used to obtain $\hat{\theta} = (\hat{\beta}, \hat{\delta})$ and $I^{-1}(\hat{\theta})$. For the adjustment of the parameter δ, we re-parametrize δ to e^{δ_1} so that δ_1 maps into the real line $\mathbb{R}$. The invariance property of maximum likelihood estimators is used to estimate δ from δ_1 and the Delta method is used to estimate $I_{\delta\delta}^{-1}(\theta)$. From this result, we can develop the hypothesis tests and confidence intervals for δ.

2.6 Simulation Procedures

In this section, we introduce simulation procedures used in the subsequent chapters. The following result is useful to generate realizations of a recurrent event process for a given intensity function.

Let $\{N(t);\ t \geq 0\}$ be a recurrent event process with the intensity function $\lambda(t|H(t))$. Then, for any $0 \leq s \leq t$,

$$\Pr\{N(s,t) = 0|H(s^+)\} = \exp\left\{-\int_s^t \lambda(u|H(u))du\right\}, \tag{2.76}$$

where $H(u)$ in the right hand side of (2.76) is given by $\{H(s^+);\ N(s,u) = 0\}$ and s^+ denotes the interval $(0, s]$. A proof of the result (2.76) can be found in Cook and Lawless (2007, Section 2.1). From the equivalence of the events "$N(t) \geq n$" and "$T_n \leq t$", we have the $\Pr\{N(t_{j-1}, t_{j-1} + w) = 0|H(t_{j-1}^+)\} = \Pr\{W_j > w|T_{j-1} = t_{j-1}, H(t_{j-1})\}$ for any $w > 0$. Thus, from the result (2.76), we can write

$$\Pr\{W_j > w|T_{j-1} = t_{j-1}, H(t_{j-1})\} = \exp\left\{-\int_{t_{j-1}}^{t_{j-1}+w} \lambda(u|H(u))du\right\}. \tag{2.77}$$

Now, if we define the random variable $E_j = \int_{t_{j-1}}^{t_{j-1}+W_j} \lambda(u|H(u))du$ for, $j = 1, 2\ldots$, where $t_0 = 0$, then the random variable E_j follows a standard exponential distribution (Cook and Lawless 2007, p. 44). This result can be used to generate the gap times W_j in a recurrent event process for any given intensity function in the continuous time settings.

First note that, for any $j = 1, 2, \ldots$, the left hand side of (2.77) is the conditional survival function of W_j given t_{j-1} and $H(t_{j-1})$. Therefore, we first generate a random variable U_j from a standard uniform distribution and let u_j denote its value. Then, from the relation $E_j = -\log(U_j)$, E_j follows a standard exponential distribution. Let e_j denote the value of E_j and w_j denote the value of W_j. Solving

$$e_j = \int_{t_{j-1}}^{t_{j-1}+w_j} \lambda(u|H(u))du, \tag{2.78}$$

for w_j, $j = 1, 2, \ldots$, we may obtain the value of the jth gap time of a recurrent event process $\{N(t);\ t \geq 0\}$ with the intensity function $\lambda(t|H(t))$. For $j = 1, 2, \ldots$, the

event times are then $t_j = \sum_{i=1}^{j} w_j$.

In order to generate event times with a general intensity function, $\lambda(t|H(t))$ over the observation window $[0, b]$, the following steps of a computer algorithm can be used.

1. Set $j = 1$, $t_0 = 0$.
2. Generate U_j from the standard uniform distribution, and let u_j be its value.
3. Let $e_j = -\log u_j$.
4. Let w_j be the value of W_j. Obtain w_j by solving $e_j = \int_{t_{j-1}}^{t_{j-1}+w_j} \lambda(u|\mathcal{H}(u))du$.
5. Set $t_j = t_{j-1} + w_j$. If $t_j \leq b$, then increase j by 1 and go to the step 2. Otherwise stop the loop, and $N(b) = j - 1$ and the event occurrence times are given by $\{t_1, \ldots, t_{j-1}\}$.

It should be noted that the integral in Step 4 of the above algorithm may not have a closed form. In such cases, numeric integration methods can be used.

The above algorithm is not restricted to generate only cases. In other words, the realization of the counting process $\{N(t);\ t \geq 0\}$ may have zero events over $[0, b]$. Therefore, we next consider a conditional approach to generate realizations, in which at least one event is observed in the interval $[0, b]$. Since the SCCS method is based on sampling of cases only, this algorithm is useful to generate data sets suitable for the SCCS design. The correct method to generate event times $\{t_1, \ldots, t_n\}$ in $[0, b]$ is then based on conditioning on the event "$N(b) = n$". In particular, when $\{N(t);\ t \geq 0\}$ is a Poisson process with the rate function $\rho(t)$, the conditional distribution of the event times $T_1, \ldots, T_n$, given $N(b) = n$, is equal to the distribution of n order statistics from a distribution with the c.d.f,

$$F(t) = \begin{cases} 0, & \text{if } t \leq 0, \\ \frac{\int_0^t \rho(u)du}{\int_0^b \rho(u)du}, & \text{if } 0 < t \leq b, \\ 1, & \text{if } t > b. \end{cases} \tag{2.79}$$

A proof of this statement is given by Rigdon and Basu (2000, pp. 59–60). We next provide a computer algorithm to generate events times in the SCCS design with a nonhomogeneous Poisson process with the rate function $\rho(t)$.

1. For any prespecified $n = 1, 2, \ldots$, generate the values of $U_1, \ldots, U_n$ from a standard uniform distribution, and let $u_1, \ldots, u_n$ be their values, respectively.

2. Obtain $k_1, \ldots, k_n$ by solving $u_j = F(k_j)$, $j = 1, \ldots, n$, where F is given in (2.79).

3. Arrange $\{k_1, \ldots, k_n\}$ in ascending order of magnitude and let $k_{(1)} \leq k_{(2)} \leq \cdots \leq k_{(n)}$ be the ordered values.

4. Let $t_j = k_{(j)}$, $j = 1, \ldots n$. Then, the event times are $t_1, \ldots, t_n$ given that $N(b) = n$.

The above procedure is recommended by Farrington and Whitaker (2006) to generate case series data for the SCCS model.

It should be noted that both event generation procedures given in this section should be extended to the case where multiple independent processes are included in a study. In such cases, we use the double subindices i and j notation as introduced in Section 2.4. The above data generation procedures can be adopted to this case in a straightforward manner. When the process is a homogeneous Poisson process with the rate function $\rho(t) = \rho$, the distribution in (2.79) is the uniform distribution over $[0, b]$. In this specific case, the event times $t_1, \ldots, t_n$, for a given $N(b) = n$ can be simply obtained by generalizing n realizations of the uniform distribution on $[0, b]$, and then by arranging them in ascending order of magnitude in order to find $t_1, \ldots, t_n$.

Other than algorithms mentioned above, we can also generate data using a function called `simulatesccsdata` built in the `SCCS` package in `R`. This function creates a dataset for a given set of parameter values in the SCCS design. Various specified inputs are available to create more sophisticated dataset such as the case where there exists multiple exposures, multiple risk periods of a single exposure, and age effects. However, unlike the two aforementioned data generation methods, this function can only generate a single event occurrence $N(0, b) = 1$ for all individuals.

Chapter 3

A Simulation Study for the Relative Efficiency of the PD-SCCS Model

In this chapter, we present the results of a simulation study conducted to investigate the estimation of the exposure effects with SCCS and cohort models under various scenarios and the relative efficiency of the PD-SCCS model compared to other cohort and SCCS models in finite sample settings. Our main objective is to study the "efficiency" of the PD-SCCS model. Various statistics such as the estimated mean square error and variance of estimates under various conditions are investigated.

We would like to note that the `SCCS` package in `R` software is relatively a new package. Since we utilized some functions in this package to estimate the effect of the exposures in this thesis, we first discuss its performance through a simulation study.

3.1 Validation of the Estimation Procedures

In this section, our goal is to validate the estimation procedures related to the relative effect of exposure in the models under various scenarios. As introduced in Chapter 2, the general model of interest is given as follows.

$$\lambda(t|H(t);\alpha,\beta) = \rho(t|x^{(t)};\alpha,\beta) = \alpha\psi(t)\exp\{\gamma + x'(t)\beta\}, \qquad t \geq 0, \tag{3.1}$$

where the parameter α represents the underlying age effect at time at a called the baseline age effect, $\psi(t)$ represents the age specific relative effect, γ is the summation of all time-fixed covariates and random effects. The $p \times 1$ vector $x(t) = (x_1(t), x_2(t), \ldots, x_p(t))'$ includes the values of the age-dependent external covariates $x_i(t)$ at age t and the $p \times 1$ vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ includes regression parameters. In this model the covariates $x_i(t)$ are indicator functions for the exposed risk periods. For example, let e_j denote the start time of the jth exposure for $j = 1, 2, \ldots, p$. If the jth exposure occurs at time t, then the covariate $x_j(t)$ takes the value of 1 when $t \in [e_j, e_j + \Delta_j]$; and $x_j(t) = 0$ otherwise. The quantity Δ_j is referred to as the length of the risk window for the jth exposure.

We generated $N = 1,000$ realizations of the recurrent event processes. To simplify the interpretation of the results, we only considered a single exposure without any age specific relative effect, as well as no fixed covariates and random effects were included. Thus, we considered the following intensity function of the nonhomogeneous Poisson process, $\{N_i(t);\ t \geq 0\}$, $i = 1, 2, \ldots, 1000$,

$$\lambda_i(t|H_i(t)) = \rho_i(t|x_i^{(t)}; \alpha, \beta) = \alpha \exp\{x_i(t)\beta\}, \qquad t \geq 0. \tag{3.2}$$

We generated the datasets in three different ways. We used (i) `simulatesccsdata` function in the `SCCS` package, (ii) the conditional distribution method, and (iii) the cohort model method. These methods are explained in Section 2.6. In each way, the data were generated over the fixed observation window $[a_i = 0, b_i = 500]$ for all individuals $i = 1, \ldots, N$. The length of risk period for ith individual was the same for everyone in the cohort and denoted by Δ. So everyone (N=1,000) had a single exposure time e_i, $i = 1, \ldots, N$ during their observation window, which were generated from a uniform distribution over [0, 500] so that $x_i(t) = 1$ if $t \in [e_i, e_i + \Delta]$ and $x_i(t) = 0$ otherwise. We conducted R=1,000 Monte Carlo simulation runs to estimate the effect of the log relative incidence rate represented by β in the model (3.2).

It should be noted that there exist several differences between the three aforementioned data generation methods. The function `simulatesccsdata` in the `SCCS` package of `R` generates only cases with a single event. That is, $N_i(0, b_i) = n_i = 1$. With this method, the choice of a high value of the relative exposure effect e^{β}, does not affect the total number of events, but it affects the ratio of events occurrence

during the exposed risk period Δ. Note that, in this method, there is no option to assign a value for the baseline rate function α in model (3.2). The second method is based on the use of the conditional distribution approach explained in Section 2.6. So we used the c.d.f in (2.79) to generate realizations. Comparing to the previous data generation method, the major difference is that we can set the average number of events for each individual process. Thus, we can generate case-only data in which the observed number of events over the followup can be larger than one; that is, $N_i(0, b_i) = n_i \geq 1$. Like the previous method, there is no option to assign a value for the baseline rate function α. We used a zero-truncated Poisson distribution to generate the number of events n_i in $[0, b_i]$ for the ith individual, $i = 1, \ldots, N$. In the third data generation method, we used the cohort model (3.2), which allowed us to generate both cases and controls in the cohort. The algorithm used in this method is given in Section 2.6. Unlike the two previous methods, both α and β were effective in the total number of observed events for individuals at the end of their observations. This data generation method was also used in Section 3.2 and Chapter 4.

In each simulation run, we considered scenarios with various values of the length of exposed risk period Δ, the relative exposure effect e^{β}, the average number of events for each individuals $E[N_i(500)]$ and the baseline age effect α. We selected the following values for those factors: $\Delta = 20, 40, 60, 80, 100$; $e^{\beta} = 0.5$, 1, 2, 3, 4; $E[N_i(500)] = 1$, 1.5, 2, 2.5; and $\alpha = 1/2000$, $1/1000$, $1/500$, $1/250$. At the end of each simulation run, we calculated the estimates $\hat{\beta}$ of β under each combination of $(\Delta, e^{\beta}, E[N(500)], \alpha)$ from the same generated data using two different methods (Method 1 and Method 2). "Method 1" fits the generated data with the `standardsccs` function of the `SCCS` package in `R`, which uses the conditional logistic regression model to find the values of parameters that maximize the likelihood function. "Method 2" is using the `optimize` function in `R` to obtain the value of parameter that maximizes the likelihood function. Also, we obtained the number of events that occurred during the exposed risk periods ($x(t) = 1$) and nonexposed risk periods ($x(t) = 0$). We conducted R=1,000 Monte Carlo simulation runs. We computed the mean and empirical estimated variances of $\hat{\beta}$. In addition to that, the empirical mean squared error (MSE) and mean bias for $\hat{\beta}$, respectively defined by the following equations, were calculated.

$$\text{MSE}(\hat{\beta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_r - \beta)^2, \tag{3.3}$$

Table 3.1: Simulation results when the "SCCS" package was used to generate data.

					Mean($\hat{\beta}$)		B$\bar{\text{i}}$as($\hat{\beta}$)($\times 10^{-3}$)		$\widehat{\text{var}}(\hat{\beta})(\times 10^{-3})$		MSE($\hat{\beta}$)($\times 10^{-3}$)	
Δ	e^{β}	β	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
20	0.5	-0.693	979.9	20.1	-0.683	-0.712	10.120	-18.685	51.920	56.112	51.971	56.405
	1	0.000	960.9	39.1	-0.011	-0.014	-10.564	-14.148	25.862	27.232	25.948	27.405
	2	0.693	924.8	75.2	0.677	0.684	-16.483	-9.612	14.508	14.997	14.766	15.075
	3	1.099	891.8	108.2	1.078	1.087	-21.000	-11.185	9.488	9.676	9.920	9.791
	4	1.386	861.1	138.9	1.361	1.373	-24.888	-13.630	9.238	9.498	9.848	9.674
40	0.5	-0.693	960.3	39.7	-0.696	-0.713	-2.951	-19.465	27.101	28.040	27.083	28.391
	1	0.000	923.2	76.8	-0.002	-0.006	-1.792	-5.905	13.979	14.246	13.969	14.267
	2	0.693	857.9	142.1	0.687	0.688	-5.781	-5.398	8.171	8.287	8.197	8.308
	3	1.099	802.2	197.8	1.088	1.089	-10.650	-10.055	6.015	6.171	6.122	6.266
	4	1.386	753.2	246.8	1.376	1.376	-10.664	-10.241	5.486	5.433	5.594	5.533
60	0.5	-0.693	940.2	59.8	-0.692	-0.703	0.975	-10.300	16.802	16.981	16.786	17.070
	1	0.000	887.5	112.5	-0.002	-0.007	-2.398	-6.726	9.960	10.192	9.956	10.227
	2	0.693	798.4	201.6	0.691	0.689	-1.946	-3.675	6.542	6.617	6.540	6.624
	3	1.099	727.4	272.6	1.091	1.089	-7.368	-9.326	5.134	5.212	5.183	5.294
	4	1.386	668.9	331.1	1.376	1.373	-10.574	-13.019	4.413	4.368	4.520	4.533
80	0.5	-0.693	919.8	80.2	-0.683	-0.693	10.535	0.005	13.309	13.684	13.407	13.670
	1	0.000	853.1	146.9	0.000	-0.005	-0.487	-5.130	8.346	8.402	8.338	8.420
	2	0.693	746.7	253.3	0.687	0.684	-6.323	-9.387	5.131	5.132	5.166	5.215
	3	1.099	664.4	335.6	1.094	1.090	-4.915	-8.716	4.657	4.704	4.677	4.775
	4	1.386	600.7	399.3	1.377	1.373	-9.052	-13.706	4.133	4.164	4.211	4.348
100	0.5	-0.693	901.2	98.8	-0.697	-0.706	-3.908	-13.087	11.394	11.526	11.398	11.686
	1	0.000	820.3	179.7	0.001	-0.004	1.494	-4.344	6.863	6.961	6.858	6.973
	2	0.693	699.1	300.9	0.691	0.687	-1.734	-6.239	4.839	4.902	4.837	4.936
	3	1.099	611.8	388.2	1.093	1.088	-5.470	-10.803	4.430	4.450	4.456	4.562
	4	1.386	546.0	454.0	1.378	1.371	-7.945	-15.174	4.162	4.111	4.221	4.337

and

$$\bar{\text{Bias}}(\hat{\beta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_r - \beta). \tag{3.4}$$

We present the results in Tables 3.1 to 3.3 for the selected scenarios in this section, and the remaining scenarios in Appendix A. In the tables, $\overline{N(\Delta)}$ and $\overline{N(\neg\Delta)}$ denote the average number of events in the exposed risk periods ($x(t) = 1$) and nonexposed risk periods ($x(t) = 0$), respectively.

The result when data were generated from the `SCCS` package are presented in Table 3.1. Our conclusions about Mean($\hat{\beta}$) obtained from by fitting the generated data with both methods are similar. As $\overline{N(\Delta)}$ increases, the values of $\widehat{\text{var}}(\hat{\beta})$ and MSE($\hat{\beta}$) decreases. Given the same Δ, the absolute values of $\bar{\text{Bias}}(\hat{\beta})$ obtained from Method 1 is smallest when $e^{\beta} = 1$ in most of the scenarios in Table 3.1. The absolute $\bar{\text{Bias}}(\hat{\beta})$ increases as β increases. For instance, in Table 3.1, given $\Delta = 40$ in Method 1, when $e^{\beta} = 1$, $\bar{\text{Bias}}(\hat{\beta})$ is -0.0017; when $e^{\beta} = 4$ $\bar{\text{Bias}}(\hat{\beta})$ is -0.0106. However, such a pattern is not observed in Method 2 where we fitted the data with the `optimize` function in `R`. Comparing both methods, overall, $\widehat{\text{var}}(\hat{\beta})$ obtained from Method 1 is slightly smaller than that of obtained from Method 2. For instance, in the same aforementioned scenarios with (Δ, e^{β}) = (20, 0.5) above, $\widehat{\text{var}}(\hat{\beta})$ from both Method 1 and Method 2

Table 3.2: Simulation results when the conditional distribution approach was used to generate data with scenarios $E[N_i(500)] = 1$.

					Mean($\hat{\beta}$)		B$\bar{\text{i}}$as($\hat{\beta}$)($\times 10^{-3}$)		$\widehat{\text{var}}$($\hat{\beta}$)($\times 10^{-3}$)		MSE($\hat{\beta}$)($\times 10^{-3}$)	
Δ	e^{β}	β	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
20	0.5	-0.693	980.0	20.0	-0.666	-0.718	26.752	-24.399	51.697	56.394	52.361	56.933
	1	0.000	961.1	39.0	-0.014	-0.019	-14.010	-18.900	26.637	28.088	26.807	28.417
	2	0.693	925.0	75.0	0.659	0.681	-33.882	-12.161	14.640	15.001	15.774	15.134
	3	1.099	890.3	109.7	1.072	1.102	-26.847	3.544	10.526	10.661	11.236	10.663
	4	1.386	859.7	140.3	1.350	1.385	-36.005	-1.258	8.275	8.271	9.563	8.265
40	0.5	-0.693	959.9	40.2	-0.673	-0.699	20.198	-6.199	24.821	25.866	25.204	25.879
	1	0.000	923.6	76.5	-0.008	-0.011	-7.774	-10.648	14.301	14.484	14.347	14.583
	2	0.693	857.1	143.0	0.684	0.695	-8.815	1.926	7.804	7.974	7.874	7.970
	3	1.099	800.5	199.6	1.085	1.100	-13.673	1.141	6.467	6.514	6.647	6.509
	4	1.386	752.0	248.0	1.365	1.382	-21.467	-4.069	5.496	5.512	5.951	5.523
60	0.5	-0.693	940.5	59.6	-0.687	-0.707	6.040	-14.102	18.057	18.516	18.075	18.696
	1	0.000	887.0	113.1	0.001	-0.001	1.196	-1.039	9.526	9.877	9.518	9.868
	2	0.693	798.2	201.8	0.684	0.691	-9.394	-2.553	5.934	5.955	6.017	5.955
	3	1.099	726.2	273.9	1.086	1.096	-12.306	-2.585	5.484	5.532	5.630	5.533
	4	1.386	666.1	334.0	1.374	1.385	-12.383	-0.940	4.795	4.839	4.944	4.835
80	0.5	-0.693	920.2	79.9	-0.682	-0.699	11.057	-5.584	14.177	14.323	14.285	14.340
	1	0.000	853.6	146.4	-0.006	-0.009	-6.460	-8.777	7.477	7.567	7.511	7.637
	2	0.693	745.1	255.0	0.687	0.692	-6.166	-1.160	5.585	5.670	5.617	5.666
	3	1.099	662.7	337.4	1.091	1.098	-7.277	-0.514	4.960	4.983	5.008	4.978
	4	1.386	597.0	403.1	1.381	1.389	-5.785	2.878	4.149	4.139	4.178	4.143
100	0.5	-0.693	900.4	99.7	-0.685	-0.698	8.056	-4.839	11.554	11.944	11.607	11.956
	1	0.000	819.6	180.5	0.003	0.001	3.403	0.877	6.759	6.786	6.763	6.780
	2	0.693	698.4	301.6	0.688	0.691	-5.595	-2.272	4.901	4.905	4.927	4.905
	3	1.099	608.8	391.2	1.096	1.101	-2.442	2.869	4.195	4.214	4.197	4.218
	4	1.386	542.2	457.9	1.379	1.386	-7.291	-0.475	4.370	4.369	4.419	4.364

Table 3.3: Simulation results when the cohort model approach was used to generate data with $\alpha = \frac{1}{2000}$.

					Mean($\hat{\beta}$)		B$\bar{\text{i}}$as($\hat{\beta}$)($\times 10^{-2}$)		$\widehat{\text{var}}$($\hat{\beta}$)($\times 10^{-2}$)		MSE($\hat{\beta}$)($\times 10^{-2}$)	
Δ	e^{β}	β	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
20	0.5	-0.693	239.7	4.9	-0.828	-0.793	-13.440	-10.033	190.929	23.854	192.545	24.837
	1	0.000	241.3	9.8	-0.047	-0.054	-4.710	-5.375	11.533	12.422	11.743	12.698
	2	0.693	240.6	19.6	0.648	0.670	-4.550	-2.361	5.381	5.492	5.582	5.543
	3	1.099	240.3	29.5	1.058	1.087	-4.067	-1.145	3.623	3.728	3.785	3.737
	4	1.386	240.3	39.3	1.341	1.376	-4.501	-1.021	3.269	3.327	3.468	3.334
40	0.5	-0.693	231.4	9.7	-0.717	-0.744	-2.408	-5.118	13.401	13.029	13.446	13.278
	1	0.000	230.8	19.1	-0.030	-0.034	-2.970	-3.410	6.493	6.747	6.575	6.856
	2	0.693	230.7	38.5	0.673	0.683	-2.027	-0.975	3.267	3.271	3.305	3.277
	3	1.099	231.5	57.9	1.080	1.095	-1.874	-0.348	2.198	2.220	2.231	2.219
	4	1.386	231.1	76.6	1.360	1.378	-2.622	-0.836	1.729	1.720	1.796	1.725
60	0.5	-0.693	221.9	14.1	-0.708	-0.728	-1.506	-3.453	8.076	8.451	8.091	8.562
	1	0.000	222.7	28.3	-0.014	-0.017	-1.444	-1.736	4.195	4.260	4.212	4.286
	2	0.693	222.3	56.2	0.673	0.680	-1.966	-1.270	2.207	2.206	2.243	2.220
	3	1.099	222.3	84.6	1.084	1.093	-1.427	-0.551	1.672	1.677	1.691	1.678
	4	1.386	221.6	113.6	1.381	1.392	-0.493	0.532	1.463	1.464	1.464	1.465
80	0.5	-0.693	213.3	18.6	-0.693	-0.709	0.014	-1.630	5.922	6.082	5.916	6.103
	1	0.000	213.1	36.8	-0.010	-0.011	-0.962	-1.082	3.307	3.364	3.313	3.373
	2	0.693	213.5	72.9	0.672	0.677	-2.108	-1.627	2.025	2.042	2.067	2.066
	3	1.099	212.8	110.6	1.094	1.100	-0.511	0.138	1.381	1.381	1.382	1.380
	4	1.386	213.1	146.4	1.373	1.381	-1.343	-0.571	1.197	1.200	1.214	1.202
100	0.5	-0.693	204.5	22.6	-0.694	-0.708	-0.051	-1.521	5.430	5.522	5.425	5.540
	1	0.000	204.8	44.9	-0.007	-0.009	-0.657	-0.943	2.614	2.637	2.616	2.644
	2	0.693	204.9	90.3	0.690	0.693	-0.309	0.034	1.649	1.652	1.648	1.650
	3	1.099	204.4	134.5	1.093	1.098	-0.598	-0.084	1.331	1.325	1.334	1.324
	4	1.386	205.8	179.8	1.376	1.382	-1.055	-0.430	1.061	1.068	1.072	1.069

are 0.0519 and 0.0561, respectively. Since the values of $\bar{\text{Bias}}(\hat{\beta})$ obtained by fitting both methods are close to 0, $\text{MSE}(\hat{\beta})$ from both methods showed a similar pattern with the pattern observed for $\widehat{\text{var}}(\hat{\beta})$ in Table 3.1.

Table 3.2 shows the results when the data were generated using the conditional distribution approach. As $\overline{N(\Delta)}$ increases, the values of $\widehat{\text{var}}(\hat{\beta})$ and $\text{MSE}(\hat{\beta})$ decreased in both Methods 1 and 2. As for the $\bar{\text{Bias}}(\hat{\beta})$, in most of the scenarios in Table 3.2, given the same Δ, $\bar{\text{Bias}}(\hat{\beta})$ is smallest when $e^{\beta} = 1$ in Method 1. For instance, in Table 3.2, given $\Delta = 60$ and $e^{\beta} = 1$, $\bar{\text{Bias}}(\hat{\beta})$ is 0.0012. However, as e^{β} increases to 2, 3, 4, the values of $\bar{\text{Bias}}(\hat{\beta})$ are -0.0094, -0.0123, -0.0124, respectively. However, such a pattern was not observed in Method 2. Comparing Method 1 and Method 2, overall, the values of $\widehat{\text{var}}(\hat{\beta})$ obtained under Method 1 is slightly lower than those obtained under Method 2. As for $\bar{\text{Bias}}(\hat{\beta})$, even though the values of $\bar{\text{Bias}}(\hat{\beta})$ obtained in both methods are close to 0, there are some systematic differences. When $e^{\beta} = 1$, for most of the scenarios, the values of $\bar{\text{Bias}}(\hat{\beta})$ obtained from Method 1 are smaller than that of obtained from Method 2. However, when $\beta \neq 1$, the values of $\bar{\text{Bias}}(\hat{\beta})$ obtained from Method 2 are smaller than that of obtained from Method 1. For instance, when $\Delta = 40$ and $e^{\beta} = 1$, the values of $\bar{\text{Bias}}(\hat{\beta})$ from Method 1 and Method 2 in Table 3.2 are -0.0078 and -0.0106, respectively. However, when $\beta = 0.5, 2, 3, 4$, those values are (0.0202, -0.0062),(-0.0088, 0.0019), (-0.0137, 0.0011), (-0.0215, -0.0041) from (Method 1, Method 2), respectively. For most of the scenarios with $e^{\beta} = 1$, the values of $\text{MSE}(\hat{\beta})$ obtained from Method 1 are smaller than those obtained from Method 2. For example, in Table 3.2, when $(\Delta, e^{\beta}) = (20, 1)$, the values of $\text{MSE}(\hat{\beta})$ from Method 1 and Method 2 are 0.0268 and 0.0284, respectively. We present the results for our simulation study in Tables 1, 2, and 3 in Appendix A when $E[N_i(500)] = 1.5$, 2 and 2.5, respectively. In these tables, we observe similar results with those obtained in Table 3.2

Table 3.3 gives the results when we used the cohort data generation approach. The results show some similarities to those obtained from two previous data generation methods, especially with the results obtained from the second data generation method. As $\overline{N(\Delta)}$ increases, the values of $\widehat{\text{var}}(\hat{\beta})$ and $\text{MSE}(\hat{\beta})$ decrease regardless of methods. In comparison of the results based on Method 1 and Method 2, overall, the values of $\widehat{\text{var}}(\hat{\beta})$ obtained from Method 1 are slightly lower than those obtained from Method 2. The values of $\bar{\text{Bias}}(\hat{\beta})$ obtained from both methods are close to 0. However, some patterns are observed in Table 3.3. Overall, Method 1 provides smaller bias

when $\beta = 1$. However, when $\beta > 1$, the values of $\bar{\text{Bias}}(\hat{\beta})$ obtained from Method 2 are smaller than those obtained from Method 1. For instance, in Table 3.3, when $\Delta = 100$ and $e^{\beta} = 1$, the values of $\bar{\text{Bias}}(\hat{\beta})$ from Method 1 and Method 2 are -0.0066 and -0.0094, respectively. On the other hand, when $\Delta = 100$ and $e^{\beta} = 3$, the values of $\bar{\text{Bias}}(\hat{\beta})$ from Method 1 and 2 are -0.0060 and -0.0008, respectively. As for MSE($\hat{\beta}$) comparisons, when $\beta \leq 1$, the values of MSE($\hat{\beta}$) obtained from Method 1 are smaller than those obtained from Method 2. When $\beta > 1$, the values of MSE($\hat{\beta}$) are larger in Method 1 comparing with those in Method 2. For example, when (Δ, e^{β}) = (100, 1), the values of MSE($\hat{\beta}$) obtained from Method 1 and Method 2 are 0.0262, 0.0264, respectively. However, when (Δ, e^{β}) = (100, 3), the values of MSE($\hat{\beta}$) obtained from Method 1 and Method 2 are 0.0133 and 0.0132, respectively. We obtained similar results for the other scenarios when $\alpha = 1/1000$, $1/500$ and $1/250$ and the results are respectively presented in Tables 4, 5 and 6 given in Appendix A.

We can sum up the results of this simulation study as follows. Firstly, the more number of events occurrences in the exposed risk period resulted in more precise and accurate results for estimating the relative exposure effects. With the scenarios considered in this study, the values of $\bar{\text{Bias}}(\hat{\beta})$ obtained from either method were close to 0, and negligible. However, given the same length of exposed risk periods, overall, Method 1 resulted in the smaller estimated bias in the estimation of β when $e^{\beta} = 1$, and this bias increased as β increases. Such a pattern was not observed in Method 2. Overall, the values of $\widehat{\text{var}}(\hat{\beta})$ obtained from Method 1 were slightly smaller than those obtained from Method 2 regardless of the data generation methods. The accuracy in the estimation of β with Method 1 was slightly higher than Method 2 for the small values of β, but it decreased as β increases when we used Method 1. However, the accuracy was improved by increasing the length of risk period Δ. Note that in Method 1, we fitted the case series data with the `SCCS` package. We used the `SCCS` package in the remaining parts of the thesis to fit the SCCS models. To obtain more accurate results, the values of β and Δ should be carefully selected if the `SCCS` package is used to fit the case series data.

3.2 Relative Efficiency of the PD-SCCS method

In this section, our goal is to investigate the relative efficiency in the estimation of the exposure effect in the PD-SCCS model. Therefore, we conducted a Monte Carlo simulation study with various settings. As mentioned in the previous chapters, the SCCS design is a good alternative to the classical cohort design especially when the event of interest is rare. Because of this reason, we considered settings in which the percentage of cases was small relative to the total number of individuals in the cohort. To do this, we used a binary random indicator function z_i, which follows the Bernoulli distribution with the success probability of 0.05.

We considered two settings (Setting A and Setting B) with $N = 1{,}000$ independent counting processes $\{N_i(t);\ t \geq 0\}, i = 1, 2, \ldots, N$. The model from which the data were generated defined the settings as given below.

Setting A: Data generated without the positive event dependence parameter δ from the following model.

$$(\alpha_0 + \gamma z_i) \exp\{x_i(t)\beta\}, \qquad t > 0, \tag{3.5}$$

Setting B: Data generated with the positive event dependence parameter δ from the following model.

$$(\alpha_0 + \gamma z_i + \delta N_i(t^-)) \exp\{x_i(t)\beta\}, \qquad t > 0. \tag{3.6}$$

In Models (3.5) and (3.6), if z_i equals 0, the baseline age effect of individuals remains α_0; when z_i equals 1, their baseline age effect becomes $\alpha_0 + \gamma$. In these settings, we fixed the parameter α_0 at a small value 0.000001 and $\alpha_0 + \gamma$ at 0.004. Therefore, if the ith individual has $z_i(t) = 0$, the probability that $N_i(500) > 0$ was very small comparing to that probability with $z_i(t) = 1$. Note that, since $z_i \sim$ Bernoulli (0.05), individuals with z_i constitutes 5% of the total populations. We considered three levels of proportion p of exposed individuals in the population N=1,000. We took a low level ($p = 0.3$), a mid level ($p = 0.6$), and a high level ($p = 0.9$) of proportion of exposed individual. A single exposure time e_i for those who were exposed was generated from a Uniform distribution over [0, 500]. Two different values of the positive event dependence parameter ($\delta = 0.001$ and 0.002) were used in Setting B. It should be noted that the parameter δ was 0 in Setting A. Lastly, we specified the value of the

relative exposure effect e^{β} at 1, 2, and 3.

The following issues affected the selection of the factors of the simulation and how to generate event times. Firstly, since cohorts models are included to measure the relative efficiency of PD-SCCS model, we used the cohort data generation method discussed in Section 2.6. Secondly, the performance of the SCCS package is affected by the choice of parameters to generate a data. As discussed in Section 3.1, when we fitted the model using the SCCS package, the larger $\bar{\text{Bi}}\text{as}(\hat{\beta})$ was obtained as higher value of the parameter β was used to generate the data. We, therefore, selected not too large values of β was selected to generate the data. Lastly, fitting semi-parametric SCCS model is computationally demanding. Thus, we could not consider the large size of cohort. The following models were used to fit the same generated data in each simulation run.

SCCS: The standard SCCS model with the intensity function

$$(\alpha_0 + \gamma z_i) \exp\{x_i(t)\beta\}, \qquad t > 0. \tag{3.7}$$

SP-SCCS: The semi-parametric SCCS model with the intensity function

$$\psi(t) \exp\{x_i(t)\beta\}, \qquad t > 0. \tag{3.8}$$

AG: Andersen-Gill model with intensity function

$$\rho_{0k}(t) \exp\{x_i(t)\beta\}, \qquad t > 0, \tag{3.9}$$

where $k = 1, 2$.

PD-SCCS: Positive event dependence SCCS with the intensity function

$$(\alpha_0 + \gamma z_i + \delta N_i(t^-)) \exp\{x_i(t)\beta\}, \qquad t > 0. \tag{3.10}$$

PD-Cohort: Positive event dependence cohort model with the intensity function

$$(\alpha_0 + \gamma z_i + \delta N_i(t^-)) \exp\{x_i(t)\beta\}, \qquad t > 0. \tag{3.11}$$

It should be noted that the baseline rate function $\rho_0(t)$, in the Andersen-Gill model given in Section 2.4.2 is assumed to be the same for all individuals. However, because of the disease indicator z_i, there will be two groups having different baseline incidences.

Thus, we used a stratified Andersen-Gill model with two strata (Cook and Lawless, 2007, Section 3.4.3). Thus, the baseline rate functions $\rho_{01}(t)$ and $\rho_{02}(t)$ in the model (3.9) are given for individuals with $z_i = 0$ and $z_i = 1$, respectively. We used the `coxph` function in `survival` package of `R` to fit the stratified Andersen-Gill model to obtain $\hat{\beta}$.

In Setting A, we considered the scenarios with various combinations of (p, Δ, e^{β}), where the values of these factors were $p = 0.3,\ 0.6,\ 0.9$, $\Delta = 40,\ 60,\ 80$ and $e^{\beta} = 1,\ 2,\ 3$. In Setting B, the scenarios included the combination of (δ, p, Δ, e^{β}). Here we took $\delta = 0.001$ and 0.002. We used the same values of p, Δ and e^{β} used in the scenarios of Setting A. For each scenario, we conducted $R = 1{,}000$ Monte Carlo simulation runs. After each simulation run, we fitted the models (3.7) to (3.11) using the same generated data, and obtained the maximum likelihood estimator $\hat{\beta}_r$ of β for $r = 1,\ 2,\ \ldots,\ R$ simulation runs. We also obtained the mean and empirical estimated variance of $\hat{\beta}$ based on $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_R$. In addition to that, the mean squared error (MSE) given in (3.3) and the mean bias for $\hat{\beta}$ defined in (3.4) were calculated. The values of $\overline{N(\Delta)}$ and $\overline{N(\neg\Delta)}$ denote the average number of events in the exposed risk periods ($x(t) = 1$) and nonexposed risk periods ($x(t) = 0$), respectively. Finally, the average number of cases $\bar{m}$ out of $N = 1{,}000$ individuals from $R = 1{,}000$ Monte Carlo simulation runs was obtained.

The results of Setting A are presented in Tables 3.4, 3.5 and 3.6. Overall, the two cohort models AG in (3.9) and PD-Cohort in (3.11) performed better compared to the SCCS models (SCCS in (3.7), SP-SCCS in (3.8) and PD-SCCS in (3.10)) in terms of MSE($\hat{\beta}$) and $\widehat{\text{var}}(\hat{\beta})$. For instance, in Setting A under the scenario $p = 0.3$, $\Delta = 60$ and $e^{\beta} = 2$, the smallest value of $\widehat{\text{var}}(\hat{\beta})$ among the results obtained by fitting SCCS models ((3.7), (3.8) and (3.10)) is 0.249. However, $\widehat{\text{var}}(\hat{\beta})$ from the Andersen-Gill model and PD-cohort model are 0.218 and 0.210, respectively. As expected in many scenarios, we obtained similar results because the cohort models (3.9) and (3.11) benefitted the information from the controls and cases, whereas the other SCCS models used the information only from cases. However, the values of $\bar{m}$ for most of the scenarios in Setting A are around 50 out of 1000 individuals. In other words, the SCCS models ((3.7), (3.8) and (3.10)) only used information from around 5 percent of individuals. Thus, the differences between the values of MSE($\hat{\beta}$) and $\widehat{\text{var}}(\hat{\beta})$ from SCCS models and cohort models are not too big given the fact that only around 5 percent of individuals are used in SCCS models. The average bias $\bar{\text{Bi}}\text{as}(\hat{\beta})$ in all scenarios are close to 0. In

Table 3.4: The values of Mean($\hat{\beta}$) and $\widehat{\text{var}}(\hat{\beta})$ calculated by fitting the SCCS, SP-SCCS, AG, PD-SCCS and PD-Cohort models in Setting A. The data were generated from the model (3.5), where $\delta = 0$

$p = 0.3$						Mean($\hat{\beta}$)					$\widehat{\text{var}}(\hat{\beta})$				
Δ	e^{β}	β	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	$\bar{m}$	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort
40	1	0.0000	102.7	2.7	47.9	-0.001	0.004	-0.017	-0.011	-0.021	0.326	0.355	0.323	0.333	0.309
	2	0.6931	102.1	5.0	48.4	0.602	0.612	0.608	0.610	0.604	0.284	0.302	0.265	0.287	0.259
	3	1.0986	101.1	7.2	48.4	1.011	1.034	1.021	1.027	1.008	0.230	0.252	0.202	0.230	0.193
60	1	0.0000	101.3	3.7	48.1	-0.050	-0.049	-0.062	-0.061	-0.069	0.344	0.354	0.309	0.348	0.307
	2	0.6931	101.6	7.1	48.6	0.610	0.617	0.605	0.610	0.599	0.249	0.267	0.218	0.251	0.210
	3	1.0986	100.6	10.7	48.7	1.052	1.066	1.052	1.059	1.049	0.167	0.178	0.141	0.168	0.135
80	1	0.0000	100.0	4.6	48.1	-0.084	-0.083	-0.097	-0.089	-0.103	0.331	0.344	0.283	0.335	0.284
	2	0.6931	100.2	9.4	48.8	0.659	0.669	0.649	0.660	0.643	0.188	0.203	0.158	0.194	0.152
	3	1.0986	99.9	13.7	49.1	1.072	1.090	1.070	1.075	1.059	0.141	0.157	0.107	0.143	0.101
$p = 0.6$															
40	1	0.0000	100.9	4.8	48.5	-0.122	-0.124	-0.134	-0.134	-0.134	0.299	0.315	0.296	0.305	0.289
	2	0.6931	99.2	9.5	48.6	0.623	0.631	0.631	0.634	0.626	0.151	0.159	0.144	0.150	0.141
	3	1.0986	101.1	14.4	49.8	1.028	1.042	1.042	1.042	1.037	0.106	0.118	0.101	0.108	0.096
60	1	0.0000	97.4	7.1	48.0	-0.084	-0.083	-0.087	-0.091	-0.091	0.217	0.228	0.221	0.221	0.214
	2	0.6931	96.7	14.3	48.8	0.672	0.683	0.681	0.673	0.674	0.107	0.116	0.105	0.108	0.101
	3	1.0986	97.2	20.9	49.9	1.068	1.080	1.079	1.074	1.074	0.071	0.078	0.066	0.071	0.060
80	1	0.0000	95.7	9.3	48.0	-0.056	-0.058	-0.062	-0.065	-0.066	0.150	0.157	0.148	0.153	0.144
	2	0.6931	96.2	18.6	49.8	0.667	0.671	0.668	0.665	0.667	0.078	0.083	0.070	0.077	0.067
	3	1.0986	95.4	27.8	50.5	1.081	1.088	1.086	1.083	1.084	0.062	0.070	0.054	0.063	0.051
$p = 0.9$															
40	1	0.0000	97.2	7.3	47.8	-0.066	-0.065	-0.067	-0.071	-0.070	0.200	0.207	0.202	0.202	0.199
	2	0.6931	97.5	14.5	49.3	0.652	0.665	0.667	0.661	0.660	0.088	0.095	0.090	0.088	0.087
	3	1.0986	98.7	21.8	50.7	1.056	1.067	1.068	1.068	1.069	0.054	0.061	0.056	0.054	0.053
60	1	0.0000	94.5	10.7	48.1	-0.034	-0.035	-0.038	-0.043	-0.041	0.117	0.122	0.118	0.119	0.116
	2	0.6931	94.3	21.2	49.6	0.662	0.668	0.670	0.664	0.666	0.062	0.069	0.063	0.063	0.061
	3	1.0986	94.9	32.6	51.5	1.097	1.108	1.102	1.102	1.101	0.043	0.049	0.046	0.044	0.043
80	1	0.0000	91.2	14.1	48.2	-0.026	-0.027	-0.028	-0.032	-0.030	0.103	0.109	0.103	0.104	0.100
	2	0.6931	89.6	27.5	49.7	0.678	0.687	0.683	0.678	0.679	0.051	0.056	0.052	0.051	0.049
	3	1.0986	90.6	41.6	52.0	1.088	1.099	1.095	1.089	1.092	0.038	0.043	0.038	0.038	0.035

Table 3.5: The values of MSE($\hat{\beta}$) and $\bar{\text{Bias}}(\hat{\beta})$ calculated by fitting the SCCS, SP-SCCS, AG, PD-SCCS and PD-Cohort models in Setting A. The data were generated from the model (3.5), where $\delta = 0$

$p = 0.3$						MSE($\hat{\beta}$)					$\bar{\text{Bias}}(\hat{\beta})$				
Δ	e^{β}	β	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	$\bar{m}$	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort
40	1	0.0000	102.7	2.7	47.9	0.326	0.355	0.323	0.332	0.309	-0.001	0.004	-0.017	-0.011	-0.021
	2	0.6931	102.1	5.0	48.4	0.292	0.308	0.272	0.294	0.267	-0.091	-0.081	-0.086	-0.083	-0.090
	3	1.0986	101.1	7.2	48.4	0.237	0.256	0.208	0.235	0.201	-0.088	-0.064	-0.078	-0.071	-0.091
60	1	0.0000	101.3	3.7	48.1	0.346	0.357	0.313	0.352	0.311	-0.050	-0.049	-0.062	-0.061	-0.069
	2	0.6931	101.6	7.1	48.6	0.256	0.272	0.226	0.257	0.218	-0.083	-0.077	-0.089	-0.083	-0.094
	3	1.0986	100.6	10.7	48.7	0.169	0.179	0.143	0.170	0.137	-0.046	-0.033	-0.047	-0.039	-0.050
80	1	0.0000	100.0	4.6	48.1	0.338	0.351	0.292	0.342	0.295	-0.084	-0.083	-0.097	-0.089	-0.103
	2	0.6931	100.2	9.4	48.8	0.189	0.204	0.160	0.195	0.154	-0.034	-0.024	-0.044	-0.034	-0.050
	3	1.0986	99.9	13.7	49.1	0.142	0.156	0.107	0.143	0.102	-0.027	-0.009	-0.028	-0.023	-0.040
$p = 0.6$															
40	1	0.0000	100.9	4.8	48.5	0.314	0.330	0.313	0.323	0.307	-0.122	-0.124	-0.134	-0.134	-0.134
	2	0.6931	99.2	9.5	48.6	0.156	0.163	0.148	0.154	0.146	-0.070	-0.062	-0.062	-0.059	-0.067
	3	1.0986	101.1	14.4	49.8	0.111	0.121	0.104	0.112	0.100	-0.070	-0.057	-0.057	-0.057	-0.061
60	1	0.0000	97.4	7.1	48.0	0.224	0.235	0.228	0.229	0.223	-0.084	-0.083	-0.087	-0.091	-0.091
	2	0.6931	96.7	14.3	48.8	0.107	0.116	0.105	0.108	0.101	-0.021	-0.011	-0.012	-0.020	-0.019
	3	1.0986	97.2	20.9	49.9	0.071	0.079	0.066	0.072	0.061	-0.030	-0.019	-0.020	-0.025	-0.024
80	1	0.0000	95.7	9.3	48.0	0.153	0.160	0.152	0.157	0.149	-0.056	-0.058	-0.062	-0.065	-0.066
	2	0.6931	96.2	18.6	49.8	0.078	0.083	0.070	0.078	0.067	-0.026	-0.022	-0.025	-0.028	-0.026
	3	1.0986	95.4	27.8	50.5	0.063	0.070	0.054	0.063	0.051	-0.018	-0.010	-0.012	-0.016	-0.015
$p = 0.9$															
40	1	0.0000	97.2	7.3	47.8	0.204	0.211	0.206	0.207	0.204	-0.066	-0.065	-0.067	-0.071	-0.070
	2	0.6931	97.5	14.5	49.3	0.089	0.095	0.090	0.089	0.088	-0.041	-0.028	-0.027	-0.032	-0.033
	3	1.0986	98.7	21.8	50.7	0.055	0.062	0.057	0.055	0.054	-0.043	-0.032	-0.031	-0.030	-0.030
60	1	0.0000	94.5	10.7	48.1	0.118	0.123	0.119	0.121	0.118	-0.034	-0.035	-0.038	-0.043	-0.041
	2	0.6931	94.3	21.2	49.6	0.063	0.069	0.064	0.064	0.061	-0.031	-0.026	-0.023	-0.030	-0.027
	3	1.0986	94.9	32.6	51.5	0.043	0.049	0.046	0.044	0.043	-0.001	0.010	0.004	0.004	0.003
80	1	0.0000	91.2	14.1	48.2	0.104	0.109	0.104	0.105	0.101	-0.026	-0.027	-0.028	-0.032	-0.030
	2	0.6931	89.6	27.5	49.7	0.051	0.056	0.052	0.051	0.049	-0.015	-0.006	-0.010	-0.015	-0.014
	3	1.0986	90.6	41.6	52.0	0.038	0.043	0.038	0.038	0.035	-0.011	0.001	-0.003	-0.010	-0.006

Table 3.5, the largest value of $\bar{\text{Bias}}(\hat{\beta})$ is found when p = 0.6, $\Delta = 40$ and $e^{\beta} = 1$. We also observe that the estimated bias reduces as the number of events increases. Overall, in Setting A, regardless of which model was fitted, if the length of the risk period Δ and/or the relative exposure effect e^{β} increases, we observe smaller values of $\text{MSE}(\hat{\beta})$, $\widehat{\text{var}}(\hat{\beta})$ and $\bar{\text{Bias}}(\hat{\beta})$. It should be noted that in these cases the number of events over the exposed risk periods increases.

The results of the relative efficiency of PD-SCCS model in Setting A are presented in Table 3.7. In particular, the relative efficiency of the estimation of β in the PD-SCCS model with respect to the estimation of that in other models is given by

$$\widehat{RE} = \frac{\frac{1}{R}\sum_{r=1}^{R} \widehat{\text{var}}_r(\hat{\beta}_{PD-SCCS})}{\frac{1}{R}\sum_{r=1}^{R} \widehat{\text{var}}_r(\hat{\beta}_{other})}. \tag{3.12}$$

In the process of finding the average estimated variance of β from Monte Carlo simulation runs by using the optimization method, we experienced issues with computational efficiency because of finding the inverse of the Hessian matrix. Therefore, we used the empirical variance of $\hat{\beta}$ obtained from 1,000 Monte Carlo simulation runs to calculate the relative efficiency (3.12). Although it is not presented here, our analysis based on a small simulation study showed that the estimated model-based variance and values of empirical variance of $\hat{\beta}$ are quite close. The results in Table 3.7 are based on $R =$ 1,000 Monte Carlo simulation runs. $\widehat{RE}$ is greater than 1 means that the comparison model performed better than the PD-SCCS model in the estimation of the variance of $\hat{\beta}$. Otherwise, the PD-SCCS model performed better than the comparison model. Note that, the SCCS model (3.7) is the model from which we generated data in Setting A. Therefore, as expected, $\widehat{RE}$ for the SCCS model is greater than 1 in most of the scenarios in Table 3.7. Regarding the comparison with cohort models (AG and PD-Cohort), when the proportion of exposed individuals p is small, $\widehat{RE}$ is greater than 1. However, as p becomes bigger, $\widehat{RE}$ tends to be closer to 1. For example, in Table 3.7, $\widehat{RE}$ compared to PD-Cohort is 1.195 in the scenario with (p, Δ, e^{β}) = (0.3, 60, 2). When $p = 0.6$ and 0.9 under the same values of Δ and e^{β}, $\widehat{RE}$ compared to PD-Cohort are 1.067 and 1.038, respectively. Also, the results in the same table shows that, as the length of risk period Δ and β increase, $\widehat{RE}$ increases. For example in Table 3.7, under the fixed values of (p, e^{β}) = (0.3, 2) when $\Delta = 40$, 60, 80, $\widehat{RE}$ compared to AG are 1.082, 1.149 and 1.224 , respectively. Another example is given in the same table. Under the fixed values of (p, Δ) = (0.6, 80), when $e^{\beta} = 1$, 2, 3, $\widehat{RE}$

compared to PD-Cohort are 1.060, 1.161 and 1.229, respectively. Note that Table 3.6 shows the average of estimated dependence parameter δ and its estimated empirical variance. In this table, Mean($\hat{\delta}_1$) and $\widehat{\text{var}}(\hat{\delta}_1)$ were calculated when we fitted the PD-SCCS model, and Mean($\hat{\delta}_2$) and $\widehat{\text{var}}(\hat{\delta}_2)$ are obtained when we fitted the PD-Cohort model. In either case, the estimated value of δ is close to 0. Because of this reason, the relative efficiency of the PD-SCCS model comparing with the SCCS model is close to 1 in all scenarios considered in Table 3.7.

The results of Setting B are displayed in Tables 3.8 to 3.11. Tables 3.8 and 3.9 show the results when $\delta = 0.001$ in the data generation with the model (3.6). Tables 3.10 and 3.11 show the results when $\delta = 0.002$. Similarly, we present the $\widehat{RE}$ results in Table 3.12 when $\delta = 0.001$ and in Table 3.13 when $\delta = 0.002$. Since the intensity function to generate the data in Setting B is the same as the one in the PD-SCCS model (3.10), overall MSE($\hat{\beta}$) and $\widehat{\text{var}}(\hat{\beta})$ values obtained from PD-SCCS are lower compared to those obtained from the SCCS and SP-SCCS models. We also observe that, overall, the values of $\bar{\text{Bi}}\text{as}(\hat{\beta})$ become smaller in all scenarios for all models considered as the average number of events in the exposed risk period $\overline{N(\Delta)}$ increases. For example, in Tables 3.8 and 3.9, MSE($\hat{\beta}$) and $\widehat{\text{var}}(\hat{\beta})$ from fitting PD-SCCS model are, respectively, 0.188 and 0.188 in the scenario with (δ, p, Δ, e^{β}) = (0.001, 0.3, 60, 2). Under the same scenario, the values of MSE($\hat{\beta}$) and $\widehat{\text{var}}(\hat{\beta})$ from fitting the SCCS model are 0.190 and 0.190, respectively. In other scenarios, similar results are observed. In Table 3.11, when (δ, p, Δ, e^{β}) = (0.002, 0.3, 60, 2), the value of $\bar{\text{Bi}}\text{as}(\hat{\beta})$ obtained by fitting the PD-SCCS model is -0.038, and the value of $\overline{N(\Delta)}$ is 13.2. When we increase p from 0.3 to 0.6 while keeping other factors the same in the previous scenario, (that is, (δ, p, Δ, e^{β}) = (0.002, 0.6, 60, 2)), the value of $\bar{\text{Bi}}\text{as}(\hat{\beta})$ is -0.009 and $\overline{N(\Delta)} = 26.8$. Similarly, in scenario, (δ, p, Δ, e^{β}) = (0.002, 0.3, 80, 2), the values of $\bar{\text{Bi}}\text{as}(\hat{\beta})$ and $\overline{N(\Delta)}$ are is -0.022 and 17.7, respectively, and in scenario with (δ, p, Δ, e^{β}) = (0.002, 0.3, 60, 3), the values of $\bar{\text{Bi}}\text{as}(\hat{\beta})$ and $\overline{N(\Delta)}$ is 0.003 and 21.5, respectively. Therefore, an increase in any of the factors p, Δ and e^{β} while keeping the others fixed results in a lower absolute value of $\bar{\text{Bi}}\text{as}(\hat{\beta})$ and a larger value of $\overline{N(\Delta)}$. In Setting B, the SCCS model tends to have a larger bias than the PD-SCCS model for large values of δ and/or β. For example, in Table 3.11 in the scenario with (Δ, e^{β}) = (80, 3) under p = 0.3, 0.6, 0.9, the values of $\bar{\text{Bi}}\text{as}(\hat{\beta})$ from fitting (SCCS and PD-SCCS) are (0.003, 0.002), (0.008, 0.003), (0.006, -0.002), respectively. Another general pattern observed in the tables is that the higher exposure effects β

and the higher length of risk period Δ result in smaller values of MSE($\hat{\beta}$) and $\widehat{\text{var}}(\hat{\beta})$ regardless of the fitted model. This result is caused by the fact that the increase in β or Δ leads to an increases in the number of events in the exposed risk periods. Also, as expected, the PD-Cohort model (3.11) performed better than the PD-SCCS method (3.9) in terms of MSE($\hat{\beta}$) and $\widehat{\text{var}}(\hat{\beta})$ values because the former model uses information from the both controls and cases. For instance, in Table 3.9, the values of MSE($\hat{\beta}$) obtained from fitting (PD-SCCS, PD-Cohort) are (0.059, 0.052) in the scenario (p, Δ, e^{β}) = (0.6, 80, 2). In all other scenarios, similar results are obtained. It should be noted that the average number of cases was around 50 out of 1,000 individuals in all scenarios in Setting B, and the values of estimated MSE($\hat{\beta}$) and $\widehat{\text{var}}(\hat{\beta})$ obtained from the PD-SCCS and PD-Cohort models are close even for such a small number of cases.

Tables 3.12 and 3.13 show the results of the relative efficiency of the PD-SCCS model for the estimation of β in Setting B for the values of $\delta = 0.001$ and 0.002, respectively. The relative efficiency of the PD-SCCS model compared to the PD-Cohort model gets closer to 1 as p increases and the combination of β and Δ decreases. Compared to the SCCS and SP-SCCS models given in (3.7) and (3.8), respectively, the relative efficiency $\widehat{RE}$ is less than 1 in most of the scenarios in Table 3.12 and 3.13. When $\delta = 0.002$, the efficiency of the PD-SCCS model increases comparing with the same scenarios with $\delta = 0.001$. For instance, when (δ, p, Δ, e^{β}) = (0.001, 0.6, 60, 3), in Table 3.12, $\widehat{RE}$ compared to SCCS is 0.982. Under the same scenario with $\delta =$ 0.002 in Table 3.13, the value of $\widehat{RE}$ is 0.851.

It should be noted that the relative efficiency of the AG model decreases as the dependency parameter δ increases from 0.001 to 0.002. All values of under the AG columns in Table 3.13 are less that those given in Table 3.12. This shows that for the scenarios considered in this study, the AG model, even though it is stratified, loose efficiency in the estimation of the relative exposure effect as the dependency on the previous number of events increases. However, values of the bias in the estimation of β with the stratified AG model as given in Tables 3.9 and 3.11 are still small and does not significantly change with respect to the change in δ.

An interesting comparison between the SCCS and SP-SCCS models can be based on the values given in Tables 3.12 and 3.13. Since the data were generated from the

PD-SCCS model, the SCCS model is misspecified. However, the baseline of the SP-SCCS model takes its shape from the data as it is a semi-parametric model. We note that the values under SCCS in Table 3.12, where $\delta = 0.001$, are higher than those under SP-SCCS for all scenarios. However, in Table 3.13, where $\delta = 0.002$, the same comparison reveals that the SCCS model is loosing more efficiency comparing with the SP-SCCS model. As explained, this result shows that as the dependence parameter δ increases, the misspecification of the baseline rate function is more pronounced in the parametric SCCS model. In such cases, the SP-SCCS model performs better than the SCCS model.

Finally, we would like to remark about the computational burden caused by fitting the SP-SCCS model, which limited us in the choices of the values of parameters in our simulation studies. For example, the `pro.time()` function in `R` used to measure the consuming time showed that under Setting B, when $e^{\beta} = 3$, $p = 0.9$, $\delta = 0.002$, and $\Delta = 80$, the average time spent to obtain results with the SP-SCCS model for a single Monte Carlo run was 466 seconds, whereas the average consuming time of the PD-SCCS model was 2 seconds for a single Monte Carlo run with the same scenario. It took approximately around 129 hours to fully complete a simulation (with $R = 1{,}000$ runs) of fitting SP-SCCS to the data under this scenario. In such cases, we recommend to use weakly parametric SCCS models explained in Section 2.4.3, instead of using the SP-SCCS model.

Table 3.6: The mean of the estimated values of the dependence parameter δ and their estimated variances. The values of Mean($\hat{\delta_1}$) and $\widehat{\text{var}}(\hat{\delta_1})$ are obtained by fitting the PD-SCCS model. The values of Mean($\hat{\delta_2}$) and $\widehat{\text{var}}(\hat{\delta_2})$ are obtained by fitting the PD-Cohort model.

Δ	e^β	β	Mean($\hat{\delta_1}$)($\times 10^{-4}$)	Mean($\hat{\delta_2}$)($\times 10^{-4}$)	$\widehat{\text{var}}(\hat{\delta_1})(\times 10^{-7})$	$\widehat{\text{var}}(\hat{\delta_2})(\times 10^{-7})$
$p = 0.3$						
40	1	0.0000	2.71	0.80	1.61	0.16
	2	0.6931	2.69	0.80	1.47	0.16
	3	1.0986	2.74	0.69	1.54	0.12
60	1	0.0000	2.53	0.84	1.48	0.18
	2	0.6931	2.72	0.80	1.52	0.16
	3	1.0986	2.33	0.76	1.23	0.14
80	1	0.0000	2.92	0.84	1.90	0.19
	2	0.6931	2.46	0.75	1.45	0.14
	3	1.0986	2.31	0.77	1.19	0.15
$p = 0.6$						
40	1	0.0000	2.56	0.73	1.63	0.14
	2	0.6931	2.74	0.84	1.45	0.19
	3	1.0986	2.49	0.74	1.27	0.13
60	1	0.0000	2.58	0.79	1.53	0.16
	2	0.6931	2.44	0.79	1.30	0.15
	3	1.0986	2.41	0.69	1.17	0.10
80	1	0.0000	2.63	0.90	1.66	0.19
	2	0.6931	2.58	0.71	1.33	0.13
	3	1.0986	2.29	0.73	1.05	0.12
$p = 0.9$						
40	1	0.0000	2.55	0.84	1.52	0.17
	2	0.6931	2.41	0.74	1.36	0.14
	3	1.0986	2.47	0.74	1.25	0.13
60	1	0.0000	2.59	0.81	1.54	0.17
	2	0.6931	2.38	0.74	1.32	0.13
	3	1.0986	2.22	0.69	1.01	0.11
80	1	0.0000	2.79	0.90	1.55	0.20
	2	0.6931	2.21	0.72	1.16	0.13
	3	1.0986	1.97	0.63	0.81	0.09

Table 3.7: The relative efficiency $\widehat{RE}$ of the PD-SCCS model compared to the SCCS, SP-SCCS, AG and PD-Cohort models in Setting A, where the data were generated from the model (3.5)

$p = 0.3$						
Δ	$\exp(\beta)$	β	SCCS	SP-SCCS	AG	PD-Cohort
40	1	0.0000	1.019	0.937	1.031	1.076
	2	0.6931	1.011	0.951	1.082	1.109
	3	1.0986	1.003	0.914	1.138	1.194
60	1	0.0000	1.012	0.982	1.126	1.134
	2	0.6931	1.006	0.940	1.149	1.195
	3	1.0986	1.009	0.947	1.196	1.247
80	1	0.0000	1.010	0.972	1.184	1.178
	2	0.6931	1.028	0.953	1.224	1.276
	3	1.0986	1.011	0.914	1.342	1.421
$p = 0.6$						
40	1	0.0000	1.020	0.970	1.033	1.057
	2	0.6931	0.996	0.944	1.044	1.064
	3	1.0986	1.019	0.916	1.076	1.131
60	1	0.0000	1.018	0.969	1.001	1.031
	2	0.6931	1.004	0.925	1.028	1.067
	3	1.0986	1.012	0.911	1.089	1.181
80	1	0.0000	1.019	0.976	1.033	1.060
	2	0.6931	0.998	0.933	1.106	1.161
	3	1.0986	1.004	0.899	1.163	1.229
$p = 0.9$						
40	1	0.0000	1.010	0.976	1.002	1.017
	2	0.6931	1.001	0.928	0.977	1.012
	3	1.0986	1.011	0.884	0.974	1.025
60	1	0.0000	1.020	0.977	1.014	1.027
	2	0.6931	1.017	0.916	0.993	1.038
	3	1.0986	1.015	0.895	0.957	1.032
80	1	0.0000	1.007	0.958	1.012	1.038
	2	0.6931	0.988	0.906	0.968	1.042
	3	1.0986	0.993	0.893	0.999	1.076

Table 3.8: The values of Mean($\hat{\beta}$) and $\widehat{\text{var}}(\hat{\beta})$ calculated by fitting the SCCS, SP-SCCS, AG, PD-SCCS and PD-Cohort models in Setting B. The data were generated from the model (3.6), where $\delta = 0.001$

$p = 0.3$						Mean($\hat{\beta}$)					$\widehat{\text{var}}(\hat{\beta})$				
Δ	e^{β}	β	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	$\bar{m}$	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort
40	1	0.0000	132.5	3.2	47.8	-0.062	-0.072	-0.089	-0.073	-0.087	0.305	0.314	0.303	0.311	0.289
	2	0.6931	134.2	6.6	48.6	0.627	0.626	0.638	0.632	0.623	0.234	0.245	0.235	0.230	0.214
	3	1.0986	133.9	9.9	48.7	1.054	1.058	1.075	1.062	1.049	0.171	0.186	0.167	0.168	0.144
60	1	0.0000	132.1	4.7	48.2	-0.099	-0.112	-0.124	-0.113	-0.125	0.291	0.303	0.293	0.293	0.275
	2	0.6931	132.0	9.7	48.5	0.666	0.658	0.667	0.666	0.648	0.190	0.200	0.184	0.188	0.163
	3	1.0986	134.5	14.9	49.2	1.087	1.086	1.094	1.091	1.059	0.119	0.122	0.110	0.116	0.093
80	1	0.0000	129.8	6.2	48.2	-0.034	-0.057	-0.086	-0.059	-0.084	0.248	0.253	0.241	0.253	0.227
	2	0.6931	131.8	12.8	48.9	0.679	0.666	0.678	0.665	0.649	0.144	0.152	0.133	0.142	0.115
	3	1.0986	131.7	19.9	49.3	1.108	1.102	1.133	1.101	1.080	0.093	0.097	0.082	0.091	0.065
$p = 0.6$															
40	1	0.0000	130.2	6.4	48.0	-0.059	-0.068	-0.079	-0.076	-0.081	0.223	0.229	0.230	0.229	0.224
	2	0.6931	130.3	12.9	48.5	0.667	0.661	0.669	0.671	0.665	0.096	0.104	0.100	0.096	0.091
	3	1.0986	132.8	20.1	49.6	1.087	1.091	1.103	1.097	1.089	0.074	0.080	0.072	0.073	0.067
60	1	0.0000	126.5	9.3	47.9	-0.036	-0.052	-0.055	-0.053	-0.054	0.137	0.141	0.144	0.141	0.137
	2	0.6931	130.1	19.2	49.4	0.662	0.651	0.669	0.655	0.656	0.071	0.073	0.070	0.071	0.064
	3	1.0986	130.8	29.5	49.9	1.093	1.091	1.106	1.093	1.083	0.054	0.058	0.053	0.053	0.046
80	1	0.0000	123.6	12.6	48.0	0.012	-0.006	-0.014	-0.012	-0.017	0.112	0.115	0.111	0.113	0.106
	2	0.6931	126.1	25.6	49.4	0.700	0.684	0.699	0.685	0.682	0.058	0.062	0.058	0.059	0.052
	3	1.0986	128.3	39.7	50.3	1.108	1.100	1.125	1.102	1.094	0.043	0.047	0.042	0.043	0.036
$p = 0.9$															
40	1	0.0000	126.8	9.5	48.1	-0.043	-0.051	-0.052	-0.055	-0.055	0.128	0.133	0.135	0.133	0.132
	2	0.6931	127.5	19.0	48.7	0.662	0.659	0.667	0.665	0.664	0.070	0.074	0.071	0.070	0.069
	3	1.0986	131.0	29.7	50.4	1.082	1.082	1.090	1.095	1.093	0.050	0.055	0.051	0.050	0.048
60	1	0.0000	122.2	14.0	48.2	-0.013	-0.029	-0.031	-0.031	-0.032	0.091	0.095	0.095	0.094	0.092
	2	0.6931	125.2	28.8	49.7	0.687	0.678	0.682	0.681	0.680	0.046	0.051	0.048	0.046	0.045
	3	1.0986	128.8	44.9	51.3	1.102	1.097	1.101	1.100	1.098	0.033	0.034	0.033	0.032	0.030
80	1	0.0000	118.2	18.4	48.3	-0.006	-0.026	-0.025	-0.028	-0.026	0.070	0.073	0.072	0.072	0.070
	2	0.6931	122.2	38.5	50.2	0.702	0.685	0.689	0.686	0.685	0.040	0.042	0.040	0.040	0.038
	3	1.0986	125.2	59.4	51.8	1.111	1.098	1.103	1.102	1.099	0.027	0.028	0.028	0.026	0.025

Table 3.9: The values of MSE($\hat{\beta}$) and $\bar{\text{Bias}}(\hat{\beta})$ calculated by fitting the SCCS, SP-SCCS, AG, PD-SCCS and PD-Cohort models in Setting B. The data were generated from the model (3.6), where $\delta = 0.001$

$p = 0.3$						MSE($\hat{\beta}$)					$\bar{\text{Bias}}(\hat{\beta})$				
Δ	e^{β}	β	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	$\bar{m}$	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort
40	1	0.0000	132.5	3.2	47.8	0.309	0.319	0.310	0.317	0.297	-0.062	-0.072	-0.089	-0.073	-0.087
	2	0.6931	134.2	6.6	48.6	0.238	0.250	0.238	0.233	0.219	-0.066	-0.067	-0.055	-0.061	-0.070
	3	1.0986	133.9	9.9	48.7	0.172	0.188	0.167	0.169	0.146	-0.045	-0.041	-0.024	-0.036	-0.049
60	1	0.0000	132.1	4.7	48.2	0.300	0.315	0.308	0.306	0.290	-0.099	-0.112	-0.124	-0.113	-0.125
	2	0.6931	132.0	9.7	48.5	0.190	0.201	0.185	0.188	0.165	-0.028	-0.035	-0.026	-0.027	-0.045
	3	1.0986	134.5	14.9	49.2	0.119	0.122	0.110	0.116	0.095	-0.011	-0.013	-0.005	-0.008	-0.039
80	1	0.0000	129.8	6.2	48.2	0.249	0.256	0.249	0.256	0.234	-0.034	-0.057	-0.086	-0.059	-0.084
	2	0.6931	131.8	12.8	48.9	0.145	0.152	0.133	0.142	0.117	-0.014	-0.027	-0.016	-0.028	-0.044
	3	1.0986	131.7	19.9	49.3	0.093	0.097	0.083	0.091	0.065	0.010	0.004	0.034	0.003	-0.018
$p = 0.6$															
40	1	0.0000	130.2	6.4	48.0	0.227	0.234	0.236	0.235	0.230	-0.059	-0.068	-0.079	-0.076	-0.081
	2	0.6931	130.3	12.9	48.5	0.096	0.105	0.101	0.097	0.092	-0.026	-0.032	-0.024	-0.022	-0.028
	3	1.0986	132.8	20.1	49.6	0.074	0.080	0.072	0.073	0.067	-0.011	-0.008	0.004	-0.002	-0.009
60	1	0.0000	126.5	9.3	47.9	0.138	0.144	0.147	0.144	0.140	-0.036	-0.052	-0.055	-0.053	-0.054
	2	0.6931	130.1	19.2	49.4	0.072	0.075	0.071	0.072	0.066	-0.031	-0.042	-0.024	-0.038	-0.037
	3	1.0986	130.8	29.5	49.9	0.054	0.058	0.053	0.053	0.047	-0.005	-0.007	0.007	-0.006	-0.015
80	1	0.0000	123.6	12.6	48.0	0.112	0.114	0.112	0.113	0.106	0.012	-0.006	-0.014	-0.012	-0.017
	2	0.6931	126.1	25.6	49.4	0.058	0.062	0.058	0.059	0.052	0.007	-0.009	0.005	-0.008	-0.011
	3	1.0986	128.3	39.7	50.3	0.043	0.047	0.043	0.043	0.036	0.009	0.001	0.026	0.003	-0.004
$p = 0.9$															
40	1	0.0000	126.8	9.5	48.1	0.130	0.135	0.138	0.136	0.135	-0.043	-0.051	-0.052	-0.055	-0.055
	2	0.6931	127.5	19.0	48.7	0.071	0.075	0.072	0.071	0.070	-0.031	-0.035	-0.027	-0.028	-0.029
	3	1.0986	131.0	29.7	50.4	0.050	0.055	0.051	0.050	0.047	-0.016	-0.017	-0.008	-0.004	-0.006
60	1	0.0000	122.2	14.0	48.2	0.091	0.096	0.096	0.095	0.093	-0.013	-0.029	-0.031	-0.031	-0.032
	2	0.6931	125.2	28.8	49.7	0.046	0.051	0.048	0.046	0.045	-0.006	-0.015	-0.012	-0.012	-0.013
	3	1.0986	128.8	44.9	51.3	0.033	0.034	0.033	0.032	0.030	0.004	-0.002	0.003	0.002	0.000
80	1	0.0000	118.2	18.4	48.3	0.070	0.074	0.073	0.073	0.071	-0.006	-0.026	-0.025	-0.028	-0.026
	2	0.6931	122.2	38.5	50.2	0.040	0.042	0.040	0.040	0.038	0.009	-0.008	-0.005	-0.007	-0.008
	3	1.0986	125.2	59.4	51.8	0.027	0.028	0.028	0.026	0.025	0.013	-0.001	0.005	0.004	0.000

Table 3.10: The values of Mean($\hat{\beta}$) and $\widehat{\text{var}}(\hat{\beta})$ calculated by fitting the SCCS, SP-SCCS, AG, PD-SCCS and PD-Cohort models in Setting B. The data were generated from the model (3.6), where $\delta = 0.002$

$p = 0.3$						Mean($\hat{\beta}$)					$\widehat{\text{var}}(\hat{\beta})$				
Δ	e^{β}	β	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	$\bar{m}$	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort
40	1	0.0000	177.5	4.3	48.3	-0.095	-0.106	-0.124	-0.108	-0.123	0.303	0.311	0.323	0.296	0.282
	2	0.6931	176.1	8.8	48.2	0.640	0.632	0.651	0.647	0.629	0.206	0.204	0.225	0.198	0.186
	3	1.0986	180.3	13.4	48.6	1.034	1.045	1.093	1.058	1.045	0.131	0.129	0.142	0.117	0.105
60	1	0.0000	172.7	6.2	47.9	-0.057	-0.087	-0.099	-0.081	-0.097	0.236	0.235	0.241	0.225	0.208
	2	0.6931	177.1	13.2	48.5	0.669	0.648	0.674	0.656	0.637	0.137	0.134	0.149	0.126	0.113
	3	1.0986	180.8	21.5	49.0	1.094	1.093	1.160	1.102	1.082	0.092	0.085	0.094	0.078	0.062
80	1	0.0000	171.3	8.2	47.9	-0.024	-0.062	-0.087	-0.068	-0.086	0.205	0.200	0.209	0.192	0.177
	2	0.6931	175.5	17.7	48.6	0.695	0.671	0.703	0.671	0.655	0.105	0.097	0.105	0.092	0.075
	3	1.0986	180.3	28.4	49.3	1.101	1.090	1.174	1.100	1.077	0.082	0.069	0.080	0.065	0.049
$p = 0.6$															
40	1	0.0000	171.5	8.5	48.0	-0.040	-0.059	-0.068	-0.063	-0.068	0.159	0.162	0.166	0.164	0.157
	2	0.6931	175.0	17.5	48.6	0.663	0.657	0.677	0.666	0.662	0.086	0.089	0.090	0.081	0.076
	3	1.0986	181.1	27.6	49.8	1.081	1.082	1.111	1.096	1.089	0.053	0.053	0.055	0.048	0.045
60	1	0.0000	167.6	12.5	47.8	-0.004	-0.034	-0.040	-0.036	-0.039	0.117	0.119	0.125	0.116	0.111
	2	0.6931	175.0	26.8	49.5	0.690	0.673	0.692	0.684	0.679	0.054	0.054	0.055	0.051	0.045
	3	1.0986	181.7	42.5	50.3	1.085	1.081	1.130	1.093	1.092	0.038	0.036	0.039	0.033	0.030
80	1	0.0000	162.9	16.4	47.8	0.025	-0.016	-0.033	-0.015	-0.026	0.089	0.086	0.092	0.083	0.079
	2	0.6931	171.2	35.6	49.3	0.712	0.686	0.715	0.694	0.689	0.048	0.045	0.047	0.040	0.035
	3	1.0986	181.5	58.2	50.8	1.106	1.093	1.146	1.102	1.097	0.039	0.032	0.037	0.029	0.024
$p = 0.9$															
40	1	0.0000	166.1	12.4	47.7	-0.033	-0.050	-0.051	-0.050	-0.050	0.105	0.108	0.109	0.105	0.104
	2	0.6931	175.1	26.6	49.4	0.680	0.674	0.681	0.682	0.680	0.047	0.047	0.049	0.044	0.044
	3	1.0986	182.3	41.4	50.4	1.081	1.084	1.090	1.097	1.096	0.036	0.037	0.036	0.033	0.033
60	1	0.0000	162.2	18.8	48.2	-0.003	-0.029	-0.033	-0.029	-0.030	0.071	0.071	0.073	0.069	0.068
	2	0.6931	170.0	39.6	49.8	0.694	0.676	0.680	0.685	0.684	0.040	0.038	0.039	0.035	0.034
	3	1.0986	181.5	63.5	51.2	1.092	1.087	1.090	1.097	1.097	0.029	0.027	0.026	0.025	0.024
80	1	0.0000	154.2	24.3	47.7	0.012	-0.028	-0.027	-0.024	-0.023	0.059	0.059	0.058	0.055	0.053
	2	0.6931	168.9	53.4	50.5	0.703	0.680	0.683	0.685	0.685	0.030	0.028	0.029	0.025	0.024
	3	1.0986	181.5	86.9	52.0	1.105	1.091	1.088	1.097	1.095	0.025	0.022	0.022	0.020	0.019

Table 3.11: The values of MSE($\hat{\beta}$) and $\bar{\text{Bias}}(\hat{\beta})$ calculated by fitting the SCCS, SP-SCCS, AG, PD-SCCS and PD-Cohort models in Setting B. The data were generated from the model (3.6), where $\delta = 0.002$

$p = 0.3$						MSE($\hat{\beta}$)					$\bar{\text{Bias}}(\hat{\beta})$				
Δ	e^{β}	β	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	$\bar{m}$	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort
40	1	0.0000	177.5	4.3	48.3	0.312	0.322	0.338	0.308	0.297	-0.095	-0.106	-0.124	-0.108	-0.123
	2	0.6931	176.1	8.8	48.2	0.209	0.207	0.227	0.200	0.190	-0.053	-0.061	-0.042	-0.047	-0.064
	3	1.0986	180.3	13.4	48.6	0.135	0.132	0.142	0.118	0.108	-0.065	-0.053	-0.005	-0.041	-0.054
60	1	0.0000	172.7	6.2	47.9	0.239	0.242	0.251	0.231	0.217	-0.057	-0.087	-0.099	-0.081	-0.097
	2	0.6931	177.1	13.2	48.5	0.137	0.136	0.149	0.127	0.116	-0.025	-0.045	-0.019	-0.038	-0.056
	3	1.0986	180.8	21.5	49.0	0.092	0.085	0.097	0.078	0.062	-0.004	-0.006	0.061	0.003	-0.016
80	1	0.0000	171.3	8.2	47.9	0.205	0.203	0.216	0.196	0.185	-0.024	-0.062	-0.087	-0.068	-0.086
	2	0.6931	175.5	17.7	48.6	0.105	0.098	0.105	0.092	0.077	0.002	-0.023	0.010	-0.022	-0.038
	3	1.0986	180.3	28.4	49.3	0.082	0.069	0.086	0.065	0.050	0.003	-0.009	0.075	0.002	-0.022
$p = 0.6$															
40	1	0.0000	171.5	8.5	48.0	0.160	0.166	0.171	0.167	0.161	-0.040	-0.059	-0.068	-0.063	-0.068
	2	0.6931	175.0	17.5	48.6	0.087	0.090	0.090	0.081	0.077	-0.031	-0.036	-0.017	-0.027	-0.032
	3	1.0986	181.1	27.6	49.8	0.053	0.053	0.055	0.048	0.045	-0.018	-0.016	0.013	-0.003	-0.010
60	1	0.0000	167.6	12.5	47.8	0.117	0.120	0.126	0.117	0.112	-0.004	-0.034	-0.040	-0.036	-0.039
	2	0.6931	175.0	26.8	49.5	0.054	0.055	0.055	0.051	0.046	-0.003	-0.021	-0.001	-0.009	-0.014
	3	1.0986	181.7	42.5	50.3	0.038	0.036	0.040	0.033	0.030	-0.013	-0.018	0.032	-0.006	-0.006
80	1	0.0000	162.9	16.4	47.8	0.090	0.086	0.093	0.084	0.079	0.025	-0.016	-0.033	-0.015	-0.026
	2	0.6931	171.2	35.6	49.3	0.049	0.045	0.048	0.040	0.035	0.018	-0.007	0.022	0.000	-0.004
	3	1.0986	181.5	58.2	50.8	0.039	0.032	0.039	0.029	0.024	0.008	-0.005	0.048	0.003	-0.002
$p = 0.9$															
40	1	0.0000	166.1	12.4	47.7	0.106	0.110	0.111	0.107	0.106	-0.033	-0.050	-0.051	-0.050	-0.050
	2	0.6931	175.1	26.6	49.4	0.047	0.047	0.049	0.044	0.044	-0.013	-0.019	-0.012	-0.011	-0.013
	3	1.0986	182.3	41.4	50.4	0.036	0.037	0.036	0.033	0.033	-0.018	-0.015	-0.009	-0.002	-0.002
60	1	0.0000	162.2	18.8	48.2	0.071	0.072	0.074	0.070	0.069	-0.003	-0.029	-0.033	-0.029	-0.030
	2	0.6931	170.0	39.6	49.8	0.040	0.038	0.039	0.035	0.034	0.001	-0.018	-0.014	-0.008	-0.009
	3	1.0986	181.5	63.5	51.2	0.029	0.027	0.026	0.025	0.024	-0.007	-0.012	-0.008	-0.001	-0.001
80	1	0.0000	154.2	24.3	47.7	0.059	0.060	0.059	0.056	0.053	0.012	-0.028	-0.027	-0.024	-0.023
	2	0.6931	168.9	53.4	50.5	0.030	0.028	0.029	0.025	0.024	0.010	-0.013	-0.010	-0.008	-0.009
	3	1.0986	181.5	86.9	52.0	0.025	0.022	0.022	0.020	0.019	0.006	-0.008	-0.011	-0.002	-0.004

Table 3.12: The relative efficiency $\widehat{RE}$ of the PD-SCCS model compared to the SCCS, SP-SCCS, AG and PD-Cohort models in Setting B, where the data were generated from the model (3.6) with $\delta = 0.001$

$p = 0.3$						
Δ	$\exp(\beta)$	β	SCCS	SP-SCCS	AG	PD-Cohort
40	1	0.0000	1.021	0.992	1.029	1.077
	2	0.6931	0.982	0.937	0.979	1.073
	3	1.0986	0.987	0.903	1.009	1.173
60	1	0.0000	1.007	0.966	0.998	1.067
	2	0.6931	0.991	0.941	1.018	1.153
	3	1.0986	0.980	0.953	1.061	1.247
80	1	0.0000	1.018	0.999	1.046	1.111
	2	0.6931	0.980	0.935	1.067	1.230
	3	1.0986	0.982	0.941	1.105	1.397
$p = 0.6$						
40	1	0.0000	1.026	1.001	0.997	1.025
	2	0.6931	1.009	0.930	0.963	1.055
	3	1.0986	0.998	0.919	1.022	1.100
60	1	0.0000	1.029	0.998	0.976	1.027
	2	0.6931	0.994	0.968	1.012	1.102
	3	1.0986	0.982	0.911	1.005	1.144
80	1	0.0000	1.008	0.984	1.011	1.066
	2	0.6931	1.006	0.951	1.010	1.125
	3	1.0986	0.998	0.930	1.035	1.210
$p = 0.9$						
40	1	0.0000	1.041	1.003	0.985	1.011
	2	0.6931	1.005	0.950	0.988	1.019
	3	1.0986	1.000	0.918	0.989	1.054
60	1	0.0000	1.033	0.987	0.986	1.020
	2	0.6931	0.992	0.909	0.957	1.031
	3	1.0986	0.974	0.923	0.963	1.043
80	1	0.0000	1.028	0.989	0.999	1.033
	2	0.6931	0.998	0.958	1.009	1.059
	3	1.0986	0.978	0.932	0.946	1.041

Table 3.13: The relative efficiency $\widehat{RE}$ of the PD-SCCS model compared to the SCCS, SP-SCCS, AG and PD-Cohort models in Setting B, where the data were generated from the model (3.6) with $\delta = 0.002$

$p = 0.3$						
Δ	$\exp(\beta)$	β	SCCS	SP-SCCS	AG	PD-Cohort
40	1	0.0000	0.978	0.952	0.916	1.051
	2	0.6931	0.957	0.969	0.878	1.063
	3	1.0986	0.893	0.905	0.822	1.113
60	1	0.0000	0.955	0.959	0.932	1.082
	2	0.6931	0.921	0.941	0.845	1.117
	3	1.0986	0.849	0.924	0.832	1.259
80	1	0.0000	0.937	0.960	0.917	1.081
	2	0.6931	0.871	0.943	0.872	1.214
	3	1.0986	0.788	0.939	0.805	1.315
$p = 0.6$						
40	1	0.0000	1.029	1.008	0.983	1.044
	2	0.6931	0.932	0.908	0.900	1.056
	3	1.0986	0.919	0.910	0.878	1.076
60	1	0.0000	0.986	0.974	0.929	1.042
	2	0.6931	0.937	0.936	0.921	1.121
	3	1.0986	0.851	0.904	0.826	1.095
80	1	0.0000	0.936	0.976	0.903	1.059
	2	0.6931	0.822	0.878	0.841	1.124
	3	1.0986	0.748	0.914	0.790	1.181
$p = 0.9$						
40	1	0.0000	1.002	0.974	0.966	1.011
	2	0.6931	0.937	0.933	0.900	0.997
	3	1.0986	0.933	0.910	0.919	1.028
60	1	0.0000	0.977	0.975	0.956	1.017
	2	0.6931	0.875	0.925	0.912	1.033
	3	1.0986	0.858	0.936	0.949	1.058
80	1	0.0000	0.936	0.927	0.950	1.038
	2	0.6931	0.831	0.918	0.881	1.040
	3	1.0986	0.788	0.891	0.923	1.069

Chapter 4

Model Misspecification and Violation of Assumptions

In this chapter, our goal is to investigate the effect of certain model misspecifications and assumption violations on the estimation of the relative exposure effect in the PD-SCCS model through simulations. As denoted by Farrington and Whitaker (2006), estimates of exposure effects in the SCCS design can be biased by the misspecification of age effects. This motivation leads to the development of other methods to explain the age effects such as piecewise constant baseline rate and semi-parametric SCCS models. Therefore, we conducted a Monte Carlo simulation study to estimate the magnitude of bias in the estimation of relative exposure effect resulted from the misspecification of age effects in the PD-SCCS model.

There are two important assumptions required by the SCCS design. These assumptions are event-independent exposure times and event-independent observation periods. We investigated the effects of violation of these assumptions in this chapter with Monte Carlo simulations.

The remainder of this chapter is organized as follows. In the next section, we discuss the effect of age misspecification on the estimation of the relative exposure with prominent cohort and SCCS models under various scenarios. In Section 4.2, we investigate the impact of the violation of the aforementioned assumptions about the SCCS design under various scenarios.

4.1 Effect of Age Misspecification on the Estimation of the Exposure Effect

In this section, our goal is to investigate the effect of the misspecification of age groups on the estimation of the relative exposure in the SCCS designs. Therefore, we conducted a Monte Carlo simulation study with various scenarios.

The general model of interest is the PD-SCCS model with the given intensity function.

$$\lambda(t|x^{(t)};\alpha,\beta,\delta) = (\alpha + \delta N(t^-))\exp\{x'(t)\beta\}, \qquad t > 0, \tag{4.1}$$

where α is the baseline age effect at age a, $N(t^-)$ gives the number of events in $[0,\ t)$, and the positive-valued δ represents the level of dependency on $N(t^-)$. The $p\times 1$ vector $x'(t) = (x_1(t), x_2(t), \ldots, x_p(t))'$ includes the age-dependent external covariates $x_i(t)$ at age t and the $p \times 1$ vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ includes the regression coefficients. As defined in the previous chapters, the covariates $x(t)$ in the model (4.1) are 0-1 valued indicator functions for the risk period. We let e_j denote the time of the start of the jth exposure for $j = 1, 2, \ldots, p$. If the jth exposure occurs at time t, then the covariate $x_j(t)$ takes the value of 1 over the time interval $(e_j,\ e_j + \Delta_j]$ and $x_j(t) = 0$ otherwise. The time interval $(e_j,\ e_j + \Delta_j]$ is called the (exposure) risk window, and Δ_j is referred to as the length of the risk window for the jth exposure, $j = 1,\ 2,\ \ldots,$ p.

We conducted $R = 1{,}000$ Monte Carlo simulation runs to estimate the effect of the relative exposure effect in model (4.1). For simplicity, we only included a single exposure, but considered various age effects. We used the cohort data generation method given in Section 2.6. We generated $N = 1{,}000$ realizations of the two models given below. These models defined the settings.

Setting A: Data were generated without the dependence parameter δ (i.e. $\delta = 0$) from the model

$$\lambda_i(t|x_i^{(t)};\alpha,\beta) = \alpha\rho(t)\exp\{x_i(t)\beta\}, \qquad t > 0, \tag{4.2}$$

Setting B: Data were generated with the dependence parameter δ (i.e. $\delta > 0$) from

the model

$$\lambda_i(t|x_i^{(t)};\alpha,\beta,\delta) = (\alpha\rho(t) + \delta N_i(t^-))\exp\{x_i(t)\beta\}, \qquad t > 0, \tag{4.3}$$

where $i = 1, 2, \ldots, N = 1{,}000$. Since we are interested in the misspecification of age effects, we used various form of the function $\rho(t)$ in models (4.2) and (4.3). This function defined the age effect after $t > a$. Therefore, we considered four scenarios with the specification of $\rho(t)$ as follows.

1. Scenario 1 (S1) - Constant age effect:

$$\rho(t) = 1, \qquad \text{if } 0 \le t < 500. \tag{4.4}$$

2. Scenario 2 (S2) - Monotonically increasing age effect:

$$\begin{aligned} \rho(t) &= 1, && \text{if } 0 \le t < 100, \\ &= 2, && \text{if } 100 \le t < 200, \\ &= 3, && \text{if } 200 \le t < 300, \\ &= 4, && \text{if } 300 \le t < 400, \\ &= 5, && \text{if } 400 \le t < 500. \end{aligned} \tag{4.5}$$

3. Scenario 3 (S3) - Monotonically decreasing age effect:

$$\begin{aligned} \rho(t) &= 5, && \text{if } 0 \le t < 100, \\ &= 4, && \text{if } 100 \le t < 200, \\ &= 3, && \text{if } 200 \le t < 300, \\ &= 2, && \text{if } 300 \le t < 400, \\ &= 1, && \text{if } 400 \le t < 500. \end{aligned} \tag{4.6}$$

4. Scenario 4 (S4) - Bell-shaped age effect:

$$\begin{aligned} \rho(t) &= 1, && \text{if } 0 \leq t < 100, \\ &= 3, && \text{if } 100 \leq t < 200, \\ &= 5, && \text{if } 200 \leq t < 300, \\ &= 3, && \text{if } 300 \leq t < 400, \\ &= 1, && \text{if } 400 \leq t < 500. \end{aligned} \tag{4.7}$$

We generated the data over the fixed observation window $[a_i = 0; b_i = 500]$ for all individuals $i = 1, 2, \ldots, N$. Every individual had a single exposure time, denoted by e_i over their observation window, which was generated from a uniform distribution over [0, 500]. The length of the exposed risk period Δ was the same for everyone in the cohort. The exposed risk indicator $x_i(t)$ took the value of 1 if $t \in [e_i, e_i + \Delta]$, otherwise, $x_i(t) = 0$. The following models were used to fit the same generated data in each simulation run.

SCCS: The standard SCCS model based on the intensity function

$$\alpha_0 \exp\{x_i(t)\beta\}, \qquad t > 0. \tag{4.8}$$

SP-SCCS: The semi-parametric SCCS model based on the intensity function

$$\psi_1(t) \exp\{x_i(t)\beta\}, \qquad t > 0. \tag{4.9}$$

AG: The Andersen-Gill model with the intensity function

$$\psi_2(t) \exp\{x_i(t)\beta\}, \qquad t > 0. \tag{4.10}$$

PD-SCCS: The positive event dependence SCCS model based on the intensity function

$$(\alpha_0 + \delta N_i(t^-)) \exp\{x_i(t)\beta\}, \qquad t > 0. \tag{4.11}$$

PD-Cohort: The positive event dependence cohort model with the intensity function

$$(\alpha_0 + \delta N_i(t^-)) \exp\{x_i(t)\beta\}, \qquad t > 0. \tag{4.12}$$

In the data generation process, we selected values of the baseline rate function α,

the relative age effect β, the positive dependency δ, and the length of risk period Δ as follows. When data were generated from the SCCS model (4.2) in Setting A, we chose $\delta = 0$ and $\alpha = 1/10000$. In Setting B, when the intensity function of the PD-SCCS model (4.3) was used to generate the data, we specified $\delta = 0.003$ and $\alpha = 1/15000$. In both settings, the values of β and Δ were the same in data generation ($\beta = \log(2) = 0.693$ and $\Delta = 60$) and the exposed proportion of individuals was $p = 0.6$. For each scenario, we conducted $R = 1{,}000$ Monte Carlo simulation runs. After each simulation run, we fitted the models (4.8) to (4.12) using the same generated data, and obtained the maximum likelihood estimator $\hat{\beta}_r$ of β for $r = 1, 2, \ldots, R$ simulation runs. We also obtained the mean and empirical estimated variance of $\hat{\beta}$ based on $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_R$. In addition to that, we obtained the mean square error $\text{MSE}(\hat{\beta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_r - \beta)^2$, and the average of the bias $\bar{\text{Bias}}(\hat{\beta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_r - \beta)$. We let $\overline{N(\Delta)}$ and $\overline{N(\neg\Delta)}$ denote the average number of events in the exposed risk periods ($x(t) = 1$) and nonexposed risk periods ($x(t) = 0$), respectively. Finally, we used the notation $\bar{m}$ to represent the average number of cases among $N = 1{,}000$ individuals. Since fitting the semi-parametric SCCS model (4.9) becomes computationally demanding as the number of events increases, we selected those values in the simulations to have a suitable number of events in the context of the previous chapter so that we still maintain to obtain useful results without losing computational efficiency.

We present the results of our simulation study in Table 4.1 for Setting A, where we generated the dataset from the model (4.2) with the above scenarios defined by the age effects. In all scenarios, the values of $\widehat{\text{var}}(\hat{\beta})$ and hence $\text{MSE}(\hat{\beta})$ in Table 4.1 are affected by the values of $\overline{N(\Delta)}$ and $\bar{m}$. Because of this reason, the values of $\widehat{\text{var}}(\hat{\beta})$ and $\text{MSE}(\hat{\beta})$ under Scenario 1 in Table 4.1 are significantly bigger than those given under other scenarios. Considering this, the magnitude of the bias is smallest with the SP-SCCS model. Other models under Scenario 1 provide similar bias in terms of its magnitude. When we generated the data with the monotonically increasing age effect (4.5) in Scenario 2, the average number of cases and events in total over the exposed risk periods are about 148 and 21, respectively. The SP-SCCS model provides the smallest magnitude of the bias in this scenario (0.015). However, the results obtained from the SCCS, AG and PD-Cohort models are close. The magnitude of the bias in the PD-SCCS is the largest (0.041). The values of $\widehat{\text{var}}(\hat{\beta})$ are similar for all models. It should be noted that the standard normal distribution based on 95% confidence interval for β include $\beta = \log(2) = 0.693$ in all models. However, the

Table 4.1: The values of Mean($\hat{\beta}$), $\bar{\text{Bi}}\text{as}(\hat{\beta})$, $\widehat{\text{var}}(\hat{\beta})$ and MSE($\hat{\beta}$) obtained by fitting the data with SCCS, SP-SCCS, AG, PD-SCCS and PD-Cohort models given in (4.8) to (4.12), respectively. The data were generated from the model (4.2) ($\delta = 0$) under scenarios S1, S2, S3, S4. The values of $\overline{N(\neg\Delta)}$, $\overline{N(\Delta)}$ and $\bar{m}$ are given.

	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	$\bar{m}$
Scenario 1	46.314	6.750	51.607
Scenario 2	139.577	21.273	148.669
Scenario 3	140.136	19.344	147.596
Scenario 4	120.577	18.175	129.651

Mean($\hat{\beta}$)

	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort
Scenario 1	0.633	0.652	0.629	0.637	0.628
Scenario 2	0.720	0.678	0.674	0.652	0.716
Scenario 3	0.613	0.663	0.667	0.619	0.621
Scenario 4	0.700	0.661	0.664	0.705	0.705

$\bar{\text{Bi}}\text{as}(\hat{\beta})$

	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort
Scenario 1	-0.060	-0.041	-0.064	-0.056	-0.065
Scenario 2	0.027	-0.015	-0.020	-0.041	0.022
Scenario 3	-0.081	-0.030	-0.026	-0.074	-0.072
Scenario 4	0.007	-0.032	-0.029	0.012	0.012

$\widehat{\text{var}}(\hat{\beta})$

	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort
Scenario 1	0.203	0.220	0.198	0.205	0.196
Scenario 2	0.068	0.070	0.062	0.065	0.061
Scenario 3	0.065	0.072	0.061	0.065	0.060
Scenario 4	0.075	0.083	0.068	0.076	0.068

MSE($\hat{\beta}$)

	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort
Scenario 1	0.206	0.222	0.202	0.208	0.200
Scenario 2	0.069	0.071	0.062	0.066	0.062
Scenario 3	0.072	0.073	0.062	0.071	0.065
Scenario 4	0.075	0.084	0.069	0.076	0.068

Table 4.2: The values of Mean($\hat{\beta}$), $\bar{\text{Bias}}(\hat{\beta})$, $\widehat{\text{var}}(\hat{\beta})$ and MSE($\hat{\beta}$) obtained by fitting the data with SCCS, SP-SCCS, AG, PD-SCCS and PD-Cohort models given in (4.8) to (4.12), respectively. The data were generated from the model (4.3) ($\delta = 0.003$) under scenarios S1, S2, S3, S4. The values of $\overline{N(\neg\Delta)}$, $\overline{N(\Delta)}$ and $\bar{m}$ are given.

	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	$\bar{m}$
Scenario 1	76.336	11.729	34.768
Scenario 2	181.407	29.099	100.967
Scenario 3	276.739	42.663	100.964
Scenario 4	189.740	30.324	88.351

	Mean($\hat{\beta}$)				
	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort
Scenario 1	0.656	0.650	0.646	0.654	0.637
Scenario 2	0.735	0.691	0.715	0.675	0.699
Scenario 3	0.675	0.663	0.708	0.683	0.676
Scenario 4	0.722	0.672	0.711	0.705	0.702

	$\bar{\text{Bias}}(\hat{\beta})$				
	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort
Scenario 1	-0.037	-0.043	-0.048	-0.039	-0.056
Scenario 2	0.042	-0.002	0.022	-0.018	0.006
Scenario 3	-0.018	-0.030	0.015	-0.010	-0.017
Scenario 4	0.029	-0.022	0.017	0.012	0.009

	$\widehat{\text{var}}(\hat{\beta})$				
	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort
Scenario 1	0.173	0.173	0.183	0.152	0.137
Scenario 2	0.060	0.055	0.061	0.046	0.043
Scenario 3	0.036	0.033	0.042	0.031	0.029
Scenario 4	0.053	0.049	0.059	0.045	0.041

	MSE($\hat{\beta}$)				
	SCCS	SP-SCCS	AG	PD-SCCS	PD-Cohort
Scenario 1	0.174	0.175	0.185	0.153	0.140
Scenario 2	0.061	0.055	0.061	0.046	0.043
Scenario 3	0.036	0.034	0.043	0.032	0.029
Scenario 4	0.054	0.050	0.060	0.045	0.041

semi-parametric models SP-SCCS and AG provide better overall results. In Scenario 3, we generated the data from a monotonically decreasing baseline rate function given in (4.6). Note that in this case, more events are generated in the early periods of observation. Comparing with the results in Scenario 2, the values of $\bar{\text{Bi}}\text{as}(\hat{\beta})$ are larger than in this scenario. The semi-parametric models SP-SCCS and AG provide similar bias in terms of its magnitude, whereas the magnitudes of bias in the parametric models SCCS, PD-SCCS and PD-Cohort are larger. In scenario 4, the magnitude of the bias is the smallest in the SCCS model and the largest in the SP-SCCS model. In terms of MSE($\hat{\beta}$), we obtained smaller results in all scenarios, except the first scenario, where the values of $\overline{N(\Delta)}$ and $\bar{m}$ were significantly smaller than others. We also obtained the mean of maximum likelihood estimators of δ (Mean($\hat{\delta}$)) in the PD-SCCS model given in (4.11). The values of Mean($\hat{\delta}$) are 0.000385, 0.003080, 0.00000 and 0.000140 under Scenarios 1, 2, 3, and 4, respectively. The values of Mean($\hat{\delta}$) is overestimated in Scenario 2.

In Setting B, we generated the data from the intensity function (4.3) with $\rho(t)$ defined under scenarios given above. The results are presented in Table 4.2. In Scenario 1, since the values of $\overline{N(\Delta)}$ and $\bar{m}$ were small, the values of $\widehat{\text{var}}(\hat{\beta})$ and MSE($\hat{\beta}$) were large. In this scenario, the magnitude of bias was close under each fitted model. The PD-Cohort model performed the best in terms of MSE($\hat{\beta}$) and the AG model was the worst. In the second scenario, overall, the values of $\bar{\text{Bi}}\text{as}(\hat{\beta})$, $\widehat{\text{var}}(\hat{\beta})$ and MSE($\hat{\beta}$) were smaller than those values in the previous scenario. The SP-SCCS and PD-Cohort models provided the smallest values of the magnitude of the bias, respectively. The PD-Cohort and PD-SCCS models performed the best in terms of their MSE($\hat{\beta}$) values (MSE($\hat{\beta}$) = 0.043 and 0.046, respectively). In Scenario 3, the data generating model included a monotonically decreasing trend. In this case, the magnitude of bias arose from the PD-SCCS model was the smallest. The PD-Cohort model and the PD-SCCS model performed the best in terms of their MSE($\hat{\beta}$) values. However, the SP-SCCS and SCCS models provided close MSE($\hat{\beta}$) values. In the last scenario, the baseline rate function of the data generating process included a bell-shaped specification. Once again, the PD-Cohort and PD-SCCS models provided the smallest magnitude of bias in terms of the estimation of β. The values of MSE($\hat{\beta}$) obtained from the PD-Cohort and PD-SCCS were smallest, and very close to each other. The values of $\bar{\text{Bi}}\text{as}(\hat{\beta})$ and MSE($\hat{\beta}$) in Table 4.2 indicate that, overall, the models including the term $N(t^-)$ (i.e. the PD-SCCS and PD-Cohort models) perform

Table 4.3: The results of relative efficiency ($\widehat{RE}$) of the PD-SCCS model compared with the SCCS, SP-SCCS, AG and PD-Cohort models under the scenarios S1, S2, S3 and S4 when $\delta = 0$ and $\delta = 0.003$

	$\delta = 0$				$\delta = 0.003$			
	SCCS	SP-SCCS	AG	PD-Cohort	SCCS	SP-SCCS	AG	PD-Cohort
S1	1.010	0.930	1.037	1.046	0.878	0.878	0.830	1.108
S2	0.952	0.921	1.047	1.061	0.775	0.844	0.759	1.073
S3	1.002	0.906	1.074	1.087	0.869	0.955	0.740	1.088
S4	1.022	0.914	1.114	1.116	0.837	0.904	0.750	1.087

better if the age groups are misspecified comparing with models not including $N(t^-)$ term (i.e the SCCS, SP-SCCS and AG models). Since the values of $\overline{N(\Delta)}$ and $\bar{m}$ were effective in the values of $\bar{\text{Bi}}\text{as}(\hat{\beta})$ and $\text{MSE}(\hat{\beta})$, it is not easy to make comparison how the shape of $\rho(t)$ affects this conclusion. We will investigate this in a more extensive simulation study as a future work. We also obtained the mean of maximum likelihood estimators of δ ($\text{Mean}(\hat{\delta})$) in the PD-SCCS model given in (4.11). The values of $\text{Mean}(\hat{\delta})$ are 0.00303, 0.00480, 0.00196 and 0.00327 under Scenarios 1, 2, 3, and 4, respectively. In this setting, the true value of δ is 0.003 in the data generation. The values of $\text{Mean}(\hat{\delta})$ showed that the parameter δ is overestimated when there is a monotonically increasing trend and underestimated when there is a monotonically decreasing trend.

Table 4.3 shows the results of the relative efficiency of the PD-SCCS model for both settings. When $\delta = 0$, as expected from the results in the previous chapter, except for the SP-SCCS model (4.9), the values of $\widehat{RE}$ were greater than 1 in most of the scenarios. In Scenario 2 with the monotonically increasing age effect, it was less than 1 in comparison to the SCCS model. This result supports that the PD-SCCS model was efficient even when the true data generating model was the cohort model (4.2) in which $\delta = 0$. When $\delta = 0.003$, the values of $\widehat{RE}$ were less than 1 except for the PD-Cohort model (4.12). It should be noted that, comparing to the data generation when $\delta = 0$, the total number of observed events increased more than three times when $\delta = 0.003$. We note that the $\widehat{RE}$ of the PD-SCCS model was higher compared to the SCCS, SP-SCCS and AG models in all scenarios. For example, the largest value of $\widehat{RE}$ compared to the AG model is 0.83 in Scenario 1. The relative efficiency of the PD-SCCS model compared to the standard SCCS is as low as 0.775 in Scenario

2 when $\delta = 0.003$. The $\widehat{RE}$ of the PD-SCCS model comparing with the PD-Cohort model was lower but the difference was small when $\delta = 0.003$. In addition to that, we also observe that, overall, the SP-SCCS model performed better than the SCCS model when $\delta = 0.003$. For instance, in Setting B, except the result in Scenario 1, all $\widehat{RE}$ obtained by comparing to SP-SCCS model was greater than the one obtained by comparing to SCCS model.

To sum up, the results presented in Tables 4.1, 4.2 and 4.3 show that the observed number of events ($\overline{N(\Delta)}$ and $\bar{m}$) was an important factor in the relative efficiency of the PD-SCCS model for the estimation of β. Because of the reason, it was difficult to assess the performance of the PD-SCCS model across scenarios considered in our simulation. Regarding the results of Setting A given in Table 4.1, two cohort models (AG and PD-Cohort models) performed better than the SCCS models in terms of MSE($\hat{\beta}$) in all scenarios. This result is not surprising because the cohort models use information from the entire population to obtain the estimates as well as no variation of baseline rate functions exists in Setting A. However, as seen from the results in Table 4.3, the AG model does not perform well in terms of estimating the variance of $\hat{\beta}$ when there exists significant variation of the baseline rate functions between individuals. Finally, under the scenarios considered in our simulation study, we observed that the PD-SCCS model is more robust to the misspecification of an age effect when there is positive dependency in the data, as compared with other models.

4.2 Effects of the Violation of the Assumptions

Two essential assumptions required for the SCCS models are the assumptions of the event-independent observation periods and event-independent exposures. Whitaker et al. (2018) denote that if these two assumptions are violated the estimates of relative exposure β might be seriously biased. In this section, we discuss the effects of violation of these assumptions on the estimation of β in the PD-SCCS model through Monte Carlo simulation studies.

4.2.1 Event-Dependent Observation Periods

We first consider the effect of the violation of the "event-independent observation periods" assumption. As discussed by Whitaker et al. (2018), it is not easy to predict the effects of the violation of this assumption on the estimation of the relative incidence in the SCCS models. To investigate this issue, we generated event dependent observation periods by following the method given in Whitaker et al. (2018). In the data generation, we first fixed an administrative censoring time b at 500 days for everyone in the cohort of size N = 1,000. Then, we generated an exposure time e_i, $i = 1, 2, \ldots, N$, from the uniform distribution over (0,500) for everyone in the cohort. After that, we generated an early termination date b_i from a uniform distribution between the first event time t_{i1} and $b = 500$ for every case in the cohort. That is, if the ith individual was a case with the first event occurrence at time t_{i1}, then the followup of that individual was over [0, b_i], where $b_i \sim$ Uniform (t_{i1}, 500). Note that this procedure creates a situation in which the observation period of a case depends on the occurrence time of the first event, which is a violation of the event-independent observation periods. If the termination time b_i was earlier than the generated exposure times e_i, then the individual was considered not to be exposed. The followup of controls were censored at $b = 500$ days.

Based on the general model of interest given in (4.1), we considered a single exposure model with the intensity function for $i = 1, 2, \ldots, N$=1,000,

$$(\alpha_0 + \delta N_i(t^-)) \exp\{x_i(t)\beta\}, \qquad t > 0, \tag{4.13}$$

where the parameter α_0 denotes the underlying age effect at the start of the followup, δ represents the level of dependency of the model on the number of previous events, the covariate $x_i(t)$ is a binary valued time varying exposure indicator, and β is the regression parameter. If the exposure occurs at time t, then the covariate $x_i(t)$ takes the value of 1 over the time interval (e_i, $e_j + \Delta_i$]; and $x_i(t) = 0$ otherwise. The length of risk period for ith individual is denoted by Δ_i.

To investigate the effect of violation of assumption on estimating the relative exposure effect β in the PD-SCCS model (4.13), we fitted the PD-SCCS model (4.13) with the generated dataset in each Monte Carlo simulation run. We let p denote the proportion of cases with terminated observation periods. We chose four different

Table 4.4: The values of Mean($\hat{\beta}$), $\bar{\text{Bias}}(\hat{\beta})$, $\widehat{\text{var}}(\hat{\beta})$ and MSE($\hat{\beta}$) when some cases have event-dependent observations. For different proportion of cases with event-dependent observations (p = 0, 0.3, 0.6, 0.9) are considered.

p	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	$\bar{m}$	Mean($\hat{\beta}$)	$\bar{\text{Bias}}(\hat{\beta})$	$\widehat{\text{var}}(\hat{\beta})$	MSE($\hat{\beta}$)
0	808.693	210.431	425.797	0.693	0.000	0.006	0.006
0.3	729.659	191.717	426.904	0.617	-0.076	0.008	0.013
0.6	649.680	173.602	426.743	0.539	-0.154	0.009	0.033
0.9	570.055	154.265	427.563	0.453	-0.241	0.012	0.070

Table 4.5: The values of Mean($\hat{\delta}$), $\bar{\text{Bias}}(\hat{\delta})(\times 10^{-3})$, $\widehat{\text{var}}(\hat{\delta})(\times 10^{-7})$ and MSE($\hat{\delta}$)$(\times 10^{-7})$ when some cases have event-dependent observations. For different proportion of cases with event-dependent observations (p = 0, 0.3, 0.6, 0.9) are considered.

p	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	$\bar{m}$	Mean($\hat{\delta}$)	$\bar{\text{Bias}}(\hat{\delta})$	$\widehat{\text{var}}(\hat{\delta})$	MSE($\hat{\delta}$)
0	808.693	210.431	425.797	0.0020	0.000	0.402	0.402
0.3	729.659	191.717	426.904	0.0024	0.434	0.576	2.460
0.6	649.680	173.602	426.743	0.0031	1.100	0.831	12.90
0.9	570.055	154.265	427.563	0.0043	2.280	1.200	53.10

values of p (p = 0, 0.3, 0.6, 0.9). Note that in the first scenario with $p = 0$, the assumption of event-independent observation periods in the PD-SCCS model holds. To generate data from the model (4.13), we selected the following values. The relative exposure effects $\beta = \log(2) = 0.693$, the length of risk period $\Delta = 60$ and the positive event dependence parameter $\delta = 0.002$. We generated the dataset using the cohort data generation method as explained in Section 2.6. We conducted a Monte Carlo simulation with R = 1,000 runs. After each run, we used `optim` function in `R` to find the estimates $\hat{\beta}$ of β and $\hat{\delta}$ of δ that maximize the log likelihood function, and calculated the mean and estimated variance of $\hat{\beta}$. In addition to that, mean square error MSE($\hat{\beta}$), and mean bias $\bar{\text{Bias}}(\hat{\beta})$, defined in (3.3) were obtained. $\overline{N(\Delta)}$ and $\overline{N(\neg\Delta)}$ denote the average number of events in the exposed risk periods ($x(t) = 1$) and nonexposed risk periods ($x(t) = 0$), respectively, and the average number of cases in each scenario is denoted by $\bar{m}$.

The results regarding β and δ are shown in Tables 4.4 and 4.5, respectively. Table 4.4 shows that the estimate of the relative incidence β is not robust with respect to the violation of event-independent observation periods assumption. First, we note that the values of $\widehat{\text{var}}(\hat{\beta})$ and MSE($\hat{\beta}$) in Table 4.4 decrease as the average of the

total number of events in the risk period, $\overline{N(\Delta)}$ increases. For example, the values of $\overline{N(\Delta)}$ are 210.4 and 191.7 when $p = 0$ and $p = 0.3$, respectively, and the values of $\widehat{\text{var}}(\hat{\beta})$ are 0.006 and 0.008. The values of $\text{MSE}(\hat{\beta})$ in those scenarios are 0.006 and 0.013, respectively. The values of $\bar{\text{Bi}}\text{as}(\hat{\beta})$ decrease as p increases. That is, the higher proportion of individuals violates the assumption, the parameter β is underestimated. Although it was not our main interest, we also investigated the bias in the estimation of the dependence parameter δ. We present the results in Table 4.5. Note that the data were generated from the model (4.13) with $\delta = 0.002$. The results in Table 4.5 show that the estimate of the positive event dependence parameter δ has an increasing bias, as p increases. Based on the results given in Tables 4.4 and 4.5, we conclude that if the assumption of event independent observation periods is violated, the relative exposure may be underestimated and the positive dependence parameter may be overestimated with the PD-SCCS model given in (4.13).

4.2.2 Event-Dependent Exposures

The impact of the violation of the event-independent exposures on the estimation of the relative incidence in the standard SCCS model has been discussed by Whitaker et al. (2018). In this section, we investigate the issue in the PD-SCCS model through a Monte Carlo simulation study. We used a similar cohort data generation method explained in the previous section. We generated event times from the model $(\alpha_0 + \delta N_i(t^-)) \exp\{x_i(t)\beta\}$, over the observation window $[a_i = 0, b_i = 500]$, $i = 1, 2, \ldots, N = 1{,}000$. We then fitted the PD-SCCS model with the intensity function.

$$(\alpha_0 + \delta N_i(t^-)) \exp\{x_i(t)\beta\}, \qquad t > 0. \tag{4.14}$$

We generated a single exposure time e_i, $i = 1, 2, \ldots, N$, from a uniform distribution over $[0, 500]$, so that $x_i(t) = 1$ if $t \in [e_i, e_i + \Delta]$, and $x_i(t) = 0$ otherwise. To create event-dependent exposures, we followed a similar method applied by Whitaker et al. (2018). We first fixed two values for η ($\eta = 25$ and 50). If the generated exposure time e_i was within η days after an event occurrence, the exposure time was delayed by η days so that the exposure time became $e_i^* = e_i + \eta$. Otherwise, the generated exposure time e_i remained the same. If the delayed exposure time was after the end of the observation period $b = 500$, that individual was considered not to be exposed.

Table 4.6: The values of Mean($\hat{\beta}$), $\bar{\text{Bias}}(\hat{\beta})$, $\widehat{\text{var}}(\hat{\beta})$ and MSE($\hat{\beta}$) when some individuals have event-dependent exposures for η ($\eta = 25$ or 50) time units of delay. The value p ($p = 0$, 0.3, 0.6, 0.9) denotes the proportion of individuals with event-dependent exposures

η	p	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	$\bar{m}$	Mean($\hat{\beta}$)	$\bar{\text{Bias}}(\hat{\beta})$	$\widehat{\text{var}}(\hat{\beta})$	MSE($\hat{\beta}$)
25	0	809.611	210.975	427.398	0.694	0.0007	0.006	0.006
	0.3	809.245	209.720	425.937	0.694	0.0009	0.006	0.006
	0.6	810.240	209.861	426.687	0.698	0.0046	0.006	0.006
	0.9	808.124	208.992	426.181	0.703	0.0096	0.006	0.006
50	0	809.611	210.975	427.398	0.694	0.0007	0.006	0.006
	0.3	811.718	208.586	427.729	0.703	0.0096	0.006	0.006
	0.6	812.521	204.753	426.242	0.707	0.0139	0.006	0.007
	0.9	813.022	202.826	426.935	0.720	0.0273	0.006	0.007

Other than $\eta = 25$ and 50, we selected the following values in our scenarios. We chose four different values of the proportions of individuals who violate the assumption of event-independent exposure ($p = 0$, 0.3, 0.6 and 0.9). We fixed the relative exposure effects β at log(2) (= 0.693), the length of risk period Δ at 60, and the positive event dependence parameter δ at 0.002. In each Monte Carlo simulation run, we used the `optim` function in `R` to obtain the values of $\hat{\beta}$ and $\hat{\delta}$, which were the maximizers of the likelihood function. Based on $R = 1{,}000$ Monte Carlo simulation runs, we obtained the results of $\widehat{\text{var}}(\hat{\beta})$, MSE($\hat{\beta}$) and $\bar{\text{Bias}}(\hat{\beta})$ as well as the values of the average number of event in nonexposed risk periods $\overline{N(\neg\Delta)}$, in exposed risk periods $\overline{N(\Delta)}$, and the average number of cases $\bar{m}$. The results are presented in Table 4.6.

Note that when $p = 0$ in Table 4.6, the event-independent exposure times assumption holds. Since the total number of events is quite large for all scenarios, the values of $\widehat{\text{var}}(\hat{\beta})$ are small in all scenarios considered in Table 4.6. The values of $\bar{\text{Bias}}(\hat{\beta})$ indicates an increasing bias as the value of p increases. That is, the higher proportion of cases with event-dependent exposure, the higher the bias in the estimated relative exposure effects. For instance when $\eta = 25$ days and $p = 0$, 0.3, 0.6 and 0.9, the values of $\bar{\text{Bias}}(\hat{\beta})$ are 0.007, 0.009, 0.0046 and 0.0096, respectively. We also observe that $\bar{\text{Bias}}(\hat{\beta})$ increases as the delayed days η increases. For instance, when $p = 0.6$, $\bar{\text{Bias}}(\hat{\beta})$ is 0.0046 and 0.0139 when $\eta = 25$ and $\eta = 50$, respectively. Based on our simulation study, we observe that both η and p affect the bias in the estimation of β with the PD-SCCS model (4.14).

Our overall conclusion is that the violation of the event-independent exposures may induce significant bias in the estimation of the relative exposure effect when the PD-SCCS model is used. We would like to note that the calculated average bias $\bar{\text{Bias}}(\hat{\beta})$ in the estimation of δ with the scenarios considered in this section was very close to 0 regardless of the values of η and p. Therefore, we did not present those results here.

Chapter 5

Applications

The purpose of this chapter is to illustrate PD-SCCS method through datasets. We applied the SCCS methods to three datasets. The first dataset was collected to analyze the association between measles, mumps and rubella (MMR) vaccine and hospital admissions for idiopathic thrombocytopenic purpura (ITP). This dataset is referenced in Miller et al. (2001). Our analysis showed that there is no significant positive event dependence in this dataset. The second dataset is a synthetic dataset created to include positive event dependence. To generate this dataset, we used some information on a real life dataset discussed by Simpson (2013). We analyzed this synthetic dataset to fit the PD-SCCS model in Section 5.2. The last dataset is another synthetic dataset based on the information provided by Simpson (2013). However, we generated the data so that the assumptions of event-independent exposures and event-independent observation periods were violated. We analyzed this dataset in Section 5.3.

5.1 Data Analysis 1: MMR Dataset

MMR vaccine is implemented against the measles, mumps and rubella diseases. ITP is a blood disorder caused by a decreased number of platelets in the blood, which results in internal bleeding, and bleeding gums. It has been debated that the implementation of the MMR vaccine in children is associated with an increasing risk of the ITP disorder for a limited time after administering MMR vaccine. An event is defined

as the hospital admissions for the ITP disorder. The MMR dataset (presented in Appendix B) includes recurrent event data on 35 children whose ages are between 366 and 730 days. Those children have experienced at least one event (i.e. the number of cases $m = 35$). There are 44 ITP admissions among those children during the observation periods. That is, $\sum_{i=1}^{35} n_i = 44$, where n_i denotes the number of events for the ith case. The dataset includes the ages at the vaccination and gender information (sex = 1 for male, and 2 for female). Since the SCCS model adjusts time-invariant confounders such as gender and other time-fixed effects, those time-invariant confounders are automatically controlled.

Figure 5.1 shows the plot of the cumulative sample mean function $\hat{\mu}(t)$ vs. the time t, where

$$\hat{\mu}(t) = \frac{1}{m} \sum_{i=1}^{m} N_i(t). \tag{5.1}$$

The concave down shape of the plot indicates that there is a monotonically decreasing trend in the rate of event occurrences as time increases. In order to support this conclusion, we next applied a trend test given by Cook and Lawless (2007, Section 3.7.1). Before we conducted the trend test, we tested whether there is heterogeneity in the baseline rate functions across individuals to choose a suitable form of a trend test. A mixed-Poisson regression model can be utilized for testing the heterogeneity. Such a test was proposed by Dean and Lawless (1989). To apply their method, we used the following intensity function.

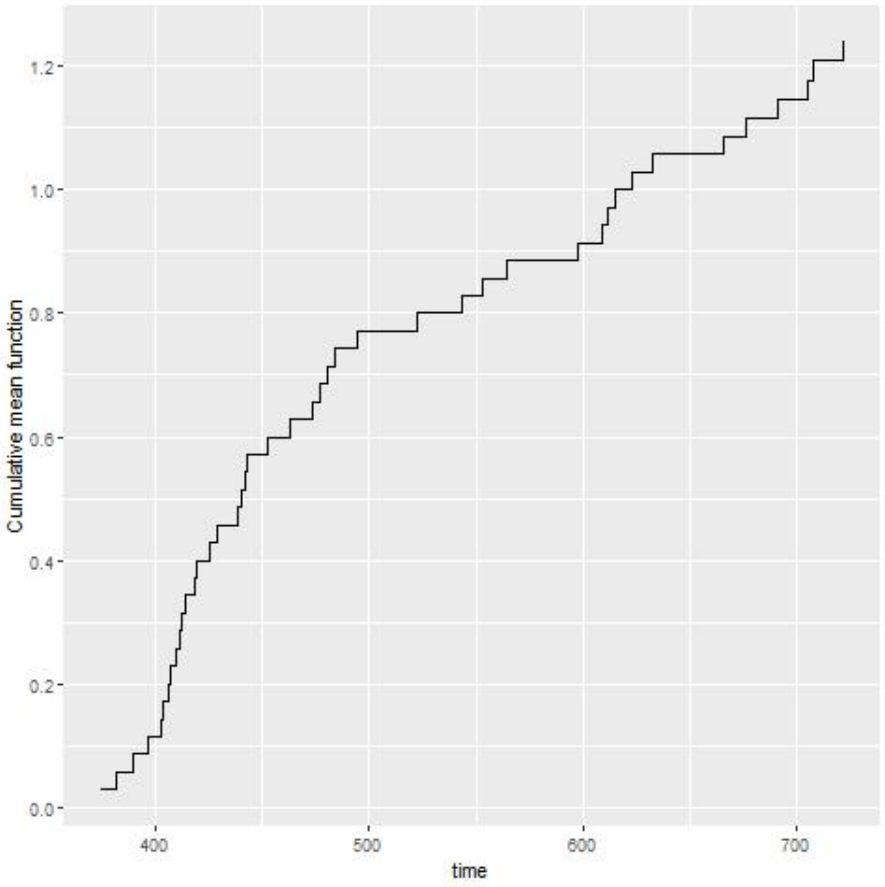
$$\lambda_i(t|H_i(t), u_i) = u_i \rho_0(t) \exp(x_i(t)\beta), \qquad t > 0, \tag{5.2}$$

where the u_i are unbiased random effects following a gamma distribution with mean 1 and variance ψ. A test of extra-Poisson variation can be developed by considering the null hypothesis $H_0 : \psi = 0$ against the alternative hypothesis $H_a : \psi > 0$. Dean and Lawless (1989) proposed the following test statistic for this purpose.

$$Z_1 = \frac{\sum_{i=1}^{m} [(n_i - \hat{\mu}_i(b_i - a_i))^2 - \hat{\mu}_i(b_i - a_i)]}{[2 \sum_{i=1}^{m} \hat{\mu}_i{}^2 (b_i - a_i)]^{1/2}}, \tag{5.3}$$

where $\hat{\mu}_i(b_i - a_i) = \int_{a_i}^{b_i} [\sum_{i=1}^{m} n_i / \sum_{i=1}^{m} (b_i - a_i)] dt$ and a_i and b_i denote the start and end of followup times of the ith individuals ($i = 1, \ldots, 35$). The test statistic Z_1 follows asymptotically a standard normal distribution under the null hypothesis. The

Figure 5.1: The plot of the cumulative mean function $\hat{\mu}(t)$ versus time t for the MMR dataset.



value of test statistic Z_1 is 1.726. Corresponding p-value obtained from the standard normal distribution is 0.0421. Thus, we reject the null hypothesis at 0.05 level of significance and conclude that there is excess heterogeneity present in the baseline rate functions. We next applied a trend test provided by Cook and Lawless (2007, Section 3.7.1), which assumes that there exists heterogeneity baseline rate functions across individuals. The robust variance estimation method for the trend test was used because it allows to deal with over-dispersion of the Poisson processes so that even when the total number of events are not following the Poisson process, the results provide correct information about the presence of trend (Cook and Lawless, 2007, Section 3.7.1). We consider the following intensity function. For $i = 1, 2, \ldots, m$,

$$\lambda_i(t|H(t); \alpha_i, \alpha^*) = \exp(\alpha_i + \alpha^* t), \qquad t > 0, \tag{5.4}$$

where α_i is the baseline rate function of the ith individual. If $\alpha^* > 0$ ($\alpha^* < 0$), there is an increasing (decreasing) trend in the rate function of event occurrence as time t increases. Note that a test of $H_0 : \alpha^* = 0$ against $H_a : \alpha^* \neq 0$ can be used for testing the absence of monotonic trends in a Poisson process. A conditional likelihood function given the total number of events for each individual n_i leads to the following score test statistic for testing $H_0 : \alpha^* = 0$ (Cook and Lawless, 2007, Section 3.2.1).

$$Z = \frac{U_c(0)}{[\widehat{asvar}\{U_c(0)\}]^{1/2}}, \tag{5.5}$$

where

$$U_c(0) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} t_{ij} - \frac{1}{2}\sum_{i=1}^{m} n_i(b_i - a_i), \tag{5.6}$$

$$\widehat{asvar}(U_c(0)) = \sum_{i=1}^{m}\left[\int_{a_i}^{b_i}(t - G_./b_.)(dN_i(t) - \rho dt)\right]^2, \tag{5.7}$$

where $G_. = \sum_{i=1}^{m}\int_{a_i}^{b_i} udu$ and $b_. = \sum_{i=1}^{m}(b_i - a_i)$. The value of test statistic Z is -2.352. Thus, the robust trend test rejects the null hypothesis $H_0 : \alpha^* = 0$ at 0.05 level based on the standard normal distribution. Thus, we conclude that there is a mild monotonically decreasing trend. Based on our preliminary analyses, we conclude that there is a mild decreasing trend in the rate of event occurrences. Therefore, we consider the following three models to fit the MMR data.

SCCS: The standard SCCS model with the intensity function

$$\alpha_0 \exp\{x_i(t)\beta\}, \qquad t > 0. \tag{5.8}$$

SP-SCCS: The semi-parametric SCCS model with the intensity function

$$\psi(t) \exp\{x_i(t)\beta\}, \qquad t > 0 \tag{5.9}$$

PD-SCCS: The positive event dependence SCCS model with the intensity function

$$(\alpha_0 + \delta N_i(t^-)) \exp\{x_i(t)\beta\} \qquad t > 0. \tag{5.10}$$

In those three models, the parameter α_0 represents the baseline age effect and the covariate $x_i(t)$, $i = 1, 2, \ldots, m$, is a 0-1 valued indicator function for the exposed risk period. In this data, $x_i(t)$ becomes 1 after the MMR vaccination for the risk period

Table 5.1: The ML estimates $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ of β in the SCCS, SP-SCCS and PD-SCCS models, respectively, are presented under different length of exposed risk periods Δ. The standard errors $(s.e)$ of the ML estimates are given. The maximum values of the negative log-likelihood function for the SCCS and PD-SCCS models are given.

Δ	$\overline{N(\Delta)}$	$\hat{\beta}_1$	$s.e(\hat{\beta}_1)$	$\hat{\beta}_2$	$s.e(\hat{\beta}_2)$	$\hat{\beta}_3$	$s.e(\hat{\beta}_3)$	$-\ell_c^{max}(\hat{\beta}_1)$	$-\ell_c^{max}(\hat{\beta}_3)$
21	5	0.933	0.483	0.501	0.521	0.980	0.483	256.974	256.969
42	13	1.512	0.351	1.103	0.397	1.536	0.351	250.722	250.720
63	15	1.319	0.345	1.015	0.405	1.337	0.345	251.932	251.930
84	17	1.229	0.344	1.022	0.401	1.244	0.344	252.448	252.448

Δ days. The parameter β represents the effects of the MMR vaccination.

In order to find the maximum likelihood estimate $\hat{\beta}$ of β and their standard errors, we used the `standardsccs` and `semisccs` functions in the SCCS package of `R` to fit the SCCS model (5.8) and the SP-SCCS model (5.9). For the PD-SCCS model (5.10), we used the `optim` function in `R`.

The results are presented in Table 5.1. The estimates $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ of β are given for the SCCS, SP-SCCS and PD-SCCS models, respectively, and $s.e$ indicates their standard errors. The notation $\overline{N(\Delta)}$ denotes the number of events in the exposed risk period. We also present the values of the log likelihood function maximized at $\hat{\beta}_1$ and $\hat{\beta}_3$ for four different values of Δ. Based on that, the data supports the choice of Δ at 42 days. It should be noted that based on a Wald type test, the parameter δ in the PD-SCCS model (5.10) is not significant. Therefore, the results based on the models (5.8) and (5.10) are very similar in Table 5.1. Based on the likelihood and a Wald type of tests (p-values = 1.67 $\times 10^{-5}$, 0.0054 and 6.19 $\times 10^{-6}$, respectively) when Δ= 42 days, we conclude that the exposure significantly increases the risk of adverse ITP event occurrence over the exposed risk windows in all models considered.

It should be noted that 35 cases with 44 events might not be large enough to use the asymptotic normal approximation for the test statistics as explained before. To solve this concern, we conducted the parametric bootstrapping using the PD-SCCS model given in (5.10). The estimate of the positive event dependence $\hat{\delta}$ for this data is $e^{-12.870} \simeq 0$ so that we ignore it for simplicity and generate the data using the conditional distribution approach explained in Section 2.6 for this bootstrapping study. We generated the dataset under the null hypothesis $\beta = 0$ while keeping the information such as the number of events, the exposure time and the observation

windows the same as given in the dataset. Thus, the only randomness arises from the event occurrence times of cases $\{t_{ij};\ i = 1, 2, \ldots, 35,\ j = 1, 2, \ldots, n_i\}$. We next fitted the PD-SCCS model (5.10) and obtained estimates $\hat{\beta}$ of β. We calculate p-value as the number of cases that satisfy $|W_{bj}| > |W_{ori}|, i = j, \ldots, 1,000$ out of 1,000 bootstrap samples, where $|W_{bj}|$ is the value of the Wald test statistic using the jth bootstrap samples and $|W_{ori}|$ is the value of the Wald test statistic from the original dataset when $\Delta = 42$ days. The calculated bootstrap p-value is 0.006, which leads to the same conclusion based on the asymptotic standard normal distribution.

It should be noted that the estimate $\hat{\beta}_2$ obtained with the SP-SCCS model (5.9) is significantly different than those obtained from the SCCS and PD-SCCS models. Also, the presence of monotonic trend affected the estimate of β in the SCCS models. That is, in terms of the best choice of the estimate of β, $\hat{\beta}_2$ is the most appropriate one because the SP-SCCS model (5.9) is the only model reflecting the decreasing trend. Moreover, based on the simulation results in Section 4.1, the estimate using SP-SCCS model is likely to have the lowest bias when there is a decreasing age effect compared to the other two models.

5.2 Data Analysis 2: Vioxx and MI Dataset 1

Vioxx is a COX-2 selective nonsteroidal anti-inflammatory drug (NSAID). The primary purpose is to relieve signs and symptoms of arthritis, acute pain in adults, and painful menstrual cycles. Vioxx was removed due to evidence of an elevated risk of cardiovascular events, including myocardial infarction (MI), in 2004 (Bresalier et al., 2005). To investigate the risk of MI associated with the use of Vioxx, a case series study was conducted. A PD-SCCS model was applied to address the positive dependence of the risk on the cumulative number of MI events.

For the data generation, we considered $N = 10{,}000$ independent counting processes observed over the interval $[0,b_i]$, where b_i is generated from a uniform distribution over (162,182). For $i = 1, 2, \ldots, N$, the intensity functions of the individuals are given by

$$(\alpha_0 + \delta N_i(t^-)) \exp\{x_i(t)\beta\}, \qquad t > 0, \tag{5.11}$$

where α_0 is the baseline age effect, δ represents the level of dependency on the previous number of events, $x_i(t)$ denotes the age-dependent external covariate of the ith individual at time t with the corresponding parameter β. Every individual who is exposed has a single exposure time e_i, which is generated from a uniform distribution over their observation window $[0, b_i]$. Again, if the exposure occurs at time t, then the covariate $x_i(t)$ takes the value of 1 over the time interval $(e_j, e_j + \Delta]$, and $x_i(t) = 0$ otherwise. The time interval Δ is called the length of risk period. We generated a dataset using the following values to approximate the study given by Simpson (2013). We chose $\alpha_0 = 1/600$, $\beta = \log(2)$, $\Delta = 30$ and $\delta = 2.74 \times 10^{-3}$. Lastly, the exposure rate of subjects is denoted by p and we took $p = 0.0719$.

The generated synthetic dataset includes $m = 2{,}550$ cases; 199 of them (7.80 %) experienced a single exposure at some point during their observation period $[0, b_i]$. Out of 2,550 cases, 1,700 (66.7%) had a single event. 565 (22.1%) had two events, 180 (7.05%) cases had three events, and 105 (4.11 %) had four or more events. Figure 5.2 shows the plot of the cumulative sample mean against time t, which reveals a mild monotonically increasing trend. The value of the test statistic (5.3) for testing the excess heterogeneity is 17.72 (p-value $\simeq 0$), which indicates that there is a strong baseline heterogeneity across the rate functions of individuals. The value of the trend test statistic (5.5) is 8.216 (p-value $\simeq 0$), showing that there is a strong increasing trend. It should be noted that the roughly straight line of the plot given in Figure 5.2 suggests a constant rate function. However, the large sample size of the data results in a significant trend in the data.

Because of the computational demand, we did not consider the semi-parametric SCCS model for this dataset. We fitted the standard SCCS and PD-SCCS models using following intensity functions.

SCCS: The standard SCCS model with the intensity function

$$\alpha_0 \exp\{x_i(t)\beta\}, \qquad t > 0. \tag{5.12}$$

PD-SCCS: positive event dependence SCCS model with the intensity function

$$(\alpha_0 + \delta N_i(t^-)) \exp\{x_i(t)\beta\} \qquad t > 0. \tag{5.13}$$

Figure 5.2: The plot of the cumulative mean function $\hat{\mu}(t)$ against time t for the Vioxx and MI Dataset 1

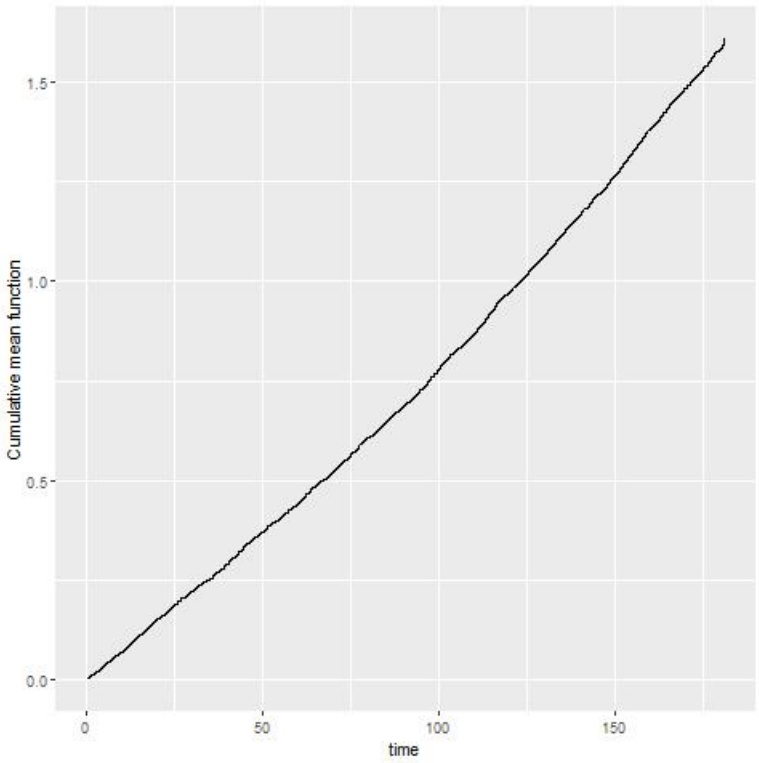


Table 5.2: The ML estimates $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\delta}$ of β and δ in the SCCS model (5.12) and PD-SCCS model (5.13) are presented under different length of exposed risk periods Δ. The standard errors ($s.e$) of the ML estimates are given. The maximum values of the negative log-likelihood function for the SCCS and PD-SCCS models are given.

Δ	$\overline{N(\Delta)}$	$\hat{\beta}_1$	$s.e(\hat{\beta}_1)$	$\hat{\beta}_2$	$s.e(\hat{\beta}_2)$	$\hat{\delta}(\times 10^{-3})$	$s.e(\hat{\delta})(\times 10^{-4})$	$-\ell_c^{max}(\hat{\beta}_1)$	$-\ell_c^{max}(\hat{\beta}_2)$
20	54	0.462	0.150	0.544	0.150	2.773	3.244	19721.670	19684.940
30	90	0.701	0.126	0.740	0.125	2.731	3.217	19711.530	19675.270
40	103	0.594	0.122	0.596	0.120	2.700	3.224	19714.780	19679.360
50	114	0.508	0.120	0.484	0.117	2.682	3.234	19717.490	19682.610

The estimates $\hat{\beta}_1$ of β in the SCCS model (5.12) was obtained by using the `standardsccs` function of `R`. We used the `optim` function in `R` to obtain the maximum likelihood estimates $\hat{\beta}_2$ of β and $\hat{\delta}$ of δ in the PD-SCCS model (5.13). The results are presented in Table 5.2. The estimates of the exposure effect parameters and their standard errors are given for various different length of exposed risk periods Δ. The corresponding numbers of events during the exposed risk periods are denoted by $\overline{N(\Delta)}$. We use a Wald type of test for the null hypothesis $H_0 : \beta = 0$ against $H_a : \beta \neq 0$. Regardless of the values of Δ, we conclude that all tests reject the null hypothesis at 0.05 level of significance. Thus, the exposure increases the risk of events for any given Δ values in Table 5.2. The data support the length of the risk period Δ at 30, which gives the largest value of log likelihood function for both models. We also conducted a Wald type test for the null hypothesis, $H_0 : \delta = 0$ against $H_a : \delta > 0$. Regardless of the values of Δ, we reject $H_0 : \delta = 0$ at 0.05 level of significance. It should be noted AIC supports the PD-SCCS model over the SCCS model for the dataset (when $\Delta = 30$, AIC of PD-SCCS is 19,678.27 and AIC of SCCS is 19,715.53). Overall, our conclusion is that Vioxx vaccination is associated with an increase of adverse MI events, and that there is evidence of positive dependence of the risk of the future MI events on the previous number of MI events.

5.3 Data Analysis 3: Vioxx and MI Dataset 2

In this section, we analyze a synthetic dataset to illustrate the impact of the violations of event-independent exposures and event-independent observation periods of the estimation of model parameters. For the data generation, we considered $N =$

10,000 independent counting processes with the intensity function given in (5.11). Once again, we chose $\alpha_0 = 1/600$, $\beta = \log(2) = 0.693$, $\Delta = 30$, and $\delta = 2.74 \times 10^{-3}$, and $p = 0.0719$. The data were generated over the observation window $[0,\ b_i]$, where b_i is generated from Uniform(162,182) distribution.

The generated dataset includes $m = 2{,}543$ individuals who have at least one events. Out of $m = 2{,}543$ cases, 1,876 (73.7%) have a single event, 456 (17.9%) have two events, 140 (5.5%) have three events, and 71 (2.8%) have four or more events. Out of $m = 2{,}543$ cases, 167 (6.6%) cases are exposed to a single exposure at some point during their observation period, $[0,\ b_i]$. We conducted a preliminary analysis for the excess heterogeneity and trend as explained in the previous section. The results are not shown here. We concluded that there is a strong heterogeneity and mild increasing trend in the rate function of processes. It should be noted that our data generation method created three groups of individuals with at least one event (cases). "Group 1" includes cases who violated event independent observation period. "Group 2" contains the cases violating the assumption of event-independent exposure. The last group "Group 3" includes cases who satisfied the assumptions of the SCCS model. We have the following breakdown of the number of cases in each group: Group 1 with 817 cases (32.1 %), Group 2 with 863 cases (33.9 %) and Group 3 with 863 cases (33.9 %).

We generated the data so that individuals in Group 1 were censored at time b_i^* (for $i = 1,\ 2,\ \ldots,\ 817$), where b_i^* was generated from a uniform distribution over $[t_{i1}, b_i]$. Thus, the censoring time b_i^* of the ith individual in Group 1 depends on his or her first event time t_{i1}. If an individual is in the second group, we generated event dependent exposure as follows. For the ith individual in Group 2, if the elapsed time between the event occurrence time t_{ij} $(j = 1,\ 2,\ \ldots,\ n_i)$ and the exposure time e_i is less than 30 days, we extended the exposure time by delaying it 30 days; that is, the new exposure time of the ith individual in Group 2 becomes $e_i^* = e_i + 30$.

For this data, we fit the PD-SCCS models using the following intensity function.

$$(\alpha_0 + \delta N_i(t^-)) \exp\{x_i(t)\beta\} \qquad t > 0. \tag{5.14}$$

We used `optim` function in `R` to find the ML estimates of β and δ in the model (5,14). The estimates of the model parameters and their standard errors under the PD-SCCS model, as well as the values of $\overline{N(\Delta)}$ and $\overline{N(\neg\Delta)}$, are given in Table 5.3.

Table 5.3: The ML estimates of $\hat{\beta}$ and $\hat{\delta}(\times 10^{-2})$ of β and δ and their standard errors ($s.e.(\hat{\beta})$ and $s.e.(\hat{\delta})(\times 10^{-4})$) are given for subgroups and full cohort. A 95% confidence interval for β is presented.

	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	m	$\hat{\beta}$	$s.e.(\hat{\beta})$	$\hat{\delta}$	$s.e.(\hat{\delta})$	95% CI of β
Group 1	1007	26	817	0.576	0.222	1.442	8.75	(0.140, 1.012)
Group 2	1193	29	863	0.910	0.224	0.211	5.678	(0.470, 1.350)
Group 3	1246	31	863	0.691	0.211	0.246	5.578	(0.278, 1.103)
Overall	3446	86	2543	0.725	0.128	0.462	3.595	(0.475, 0.975)

We conducted separate analyses on different groups of cases. Based on a Wald type of test for the null hypothesis H_0: $\beta = 0$, we conclude that $H_0 : \beta = 0$ is rejected for all groups and overall at 0.05 level of significance. Also, a Wald type of test rejects the null hypothesis $H_0 : \delta = 0$ vs. the alternative hypothesis $H_1 : \delta > 0$ in all groups and overall. It should be noted that comparing with the true values of parameters ($\beta = 0.693$ and $\delta = 2.74 \times 10^{-3}$), the estimates obtained in Group 1 and Group 2, as well as in Overall, are significantly biased. Group 3 satisfies the assumptions required for the SCCS method. Therefore, the estimates $\hat{\beta}$ and $\hat{\delta}$ in Group 3 are close to the true values of β and δ.

Chapter 6

Summary and Future Work

In this chapter, we present the summary and conclusion of the thesis. We also discuss some of the possible future works briefly.

6.1 Summary and Conclusion

The self-controlled case series (SCCS) design compared with the more established study designs such as cohort or case-control designs is a relatively new design that can investigate the association between time-varying exposures and the adverse events in recurrent events settings. The most important advantage of this method is that all time-invariant confounders that act multiplicatively on the event intensity function are inherently adjusted. The SCCS method is an outcome dependent design in a sense that only cases, who are individuals experienced the event-of-interest at least one time, are included. As a result, it simplifies the data collection, increases the computational efficiency and helps protect the data privacy. Because of these reasons, the SCCS design is considered as an important alternative to the more established classical cohort design especially when the event of interest is rare. Despite the inference is only based on the data from cases, the SCCS method provides a consistent estimate of the relative incidence regarding time-varying exposure effects. The SCCS design has been recently extended to deal with the positive event dependence (PD). The PD-SCCS model allows the dependence on the previous event occurrences in individuals. It maintains all advantages of the standard SCCS method (Simpson, 2013).

Since the PD-SCCS model has been recently proposed, there is a limited number of studies on the features of it. The main goal of this thesis is to study the relative efficiency of the PD-SCCS model in the estimation of the exposure effects through simulations. For this purpose, we conducted Monte Carlo simulation studies. We compared the results obtained from fitting the PD-SCCS model with those obtained from fitting the parametric and semi-parametric SCCS models, as well as the Andersen-Gill and cohort models. We also studied the effect of some model misspecifications and violation of model assumptions required for the SCCS design through Monte Carlo simulation studies. Lastly, we apply the PD-SCCS model to analyze the real and synthetic datasets.

In Chapter 1, we introduce the SCCS and PD-SCCS models and give the goal and outline of this thesis. Chapter 2 introduces the notation and other technical background required in this thesis. In Chapter 3, we first presented the results of a simulation study conducted to check the performance of the `SCCS` package in `R`. Since this package is a recent package and was used in the rest of the thesis, it was necessary to understand how this package performs. We next investigated the relative efficiency of the PD-SCCS model through a Monte Carlo simulation study. Since the SCCS design is mainly used to model rare event occurrences, we considered the scenarios with rare events. Our simulation study shows overall that as the number of events during the exposed risk periods increases, the precision of the estimation of the exposure effect increases regardless of fitted models. The two cohort models performed better than the three other SCCS models in terms of mean square error and variance estimation, especially when there is no excess variation in the baseline rate functions across individuals. However, when there is an excess variation in the baseline rate functions caused by the positive event dependence on the previous number of events, the semi-parametric Andersen-Gill model performed worse than the PD-SCCS model. In this simulation study, we kept the proportion of cases around five percent as it is in-line with the definition of a rare event (Lash et al., 2021). Even though the three SCCS models (SCCS, SP-SCCS, and PD-SCCS) only used the information from cases, they still provided reasonable estimates of the relative exposure effect in terms of the bias and mean square error. It should be noted that, even when there is no positive event dependence in the generated data, the PD-SCCS model performed as good as the standard SCCS model for the estimation of the relative incidence. Few factors affected the relative efficiency of the PD-SCCS model compared to the cohort models.

As the proportion of exposed individuals increased, the efficiency of the PD-SCCS model was closer to the efficiency of the cohort models. Also, as the relative exposure effect and/or the length of the risk period decreased, the relative efficiency of the PD-SCCS model increased when we compared it with the cohort models.

In Chapter 4, we investigated the effect of age misspecification and the effect of the violation of model assumptions on estimating parameters in the PD-SCCS model. To do this, we conducted Monte Carlo simulations. When we generated data using the intensity function with positive event dependence, the value of estimates of the positive event dependence parameter increases (decreases) when there is a monotonically increasing (decreasing) age effect. As a result, these changes of estimates of the positive event dependence parameter based on the age effect resulted in less bias in the estimation of the relative exposure effects. The results of our simulation studies show that the PD-SCCS model is mildly robust to the misspecification of age effects regarding estimating the relative exposure effect with the scenarios considered in our simulation study. When positive event dependence was not present in the data, the estimated variance of the relative exposure effect in the PD-SCCS model was smaller than that of in the standard SCCS model when an increasing age effect existed. The results of our simulation study in Chapter 4 revealed that the violation of event-independent observation period and event-independent exposure time assumptions affects the bias in the estimation of the relative exposure effects in the PD-SCCS model. Overall, the higher proportion of cases that violate the event-independent observation periods assumption in the data, the higher the upward bias in the estimation of the relative exposure effect. As for the event-independent exposure time assumption, the higher proportion of cases that violates the assumption in the data, the higher the downward bias in the estimation of the relative exposure effect.

In Chapter 5, we analyzed three datasets. Our goal was to illustrate the methods discussed in the previous chapters. The first dataset was from a vaccination study in children. The second and third datasets were generated by using information from a study in postmarketing surveillance. Through these datasets, we illustrated the analysis under the positive dependency on the previous number of events and assumption violations.

6.2 Future Work

The SCCS designs have important advantages. However, they also have their restrictive limitations. In this section, we discuss some future work topics about the SCCS designs.

6.2.1 Misspecification of the Length of the Exposed Risk Periods

In this thesis, we considered the performance and robustness of PD-SCCS model in the estimation of the relative incidence when the age effect was misspecified. Another important issue is that the misspecification of the exposed risk periods Δ. Based on the results of Chapter 5, the misspecification of the value of the length of risk window Δ may affect the performance of the parameter estimation based on the PD-SCCS model. We will investigate this issue as a future work.

6.2.2 More Complicated Models

In this thesis, we focused on the PD-SCCS model, which handles only the positive event dependence on the previous number of events in a process. More complex dependence on the previous number of events on counting processes may be needed in some studies. In the PD-SCCS model, the positive event dependence is of the additive form on the baseline rate function. Since the additive form of the baseline with the positive event dependence in the PD-SCCS model would be considered as an unspecified age effect in the semi-parametric SCCS model, the age effect and positive event dependence would be confounded in the semi-parametric method applied to the PD-SCCS model. The multiplicative intensity function of the form $\rho_0(t)\exp\{x'(t)\beta + \delta N(t^-)\}$ could be developed to include the effects of $N(t^-)$ in a study. This model can be investigated in the framework of weakly parametric piecewise constant rate models. This multiplicative form of the PD-SCCS model can include negative dependence on the previous number of events, and appropriately approximates the semi-parametric models for a suitable number of pieces. We will investigate this approach as a future work.

In this thesis, we assumed that all individuals have the same constant baseline age effects over the observation periods. Even though our study showed that the positive dependence parameter adjusts the variation of the age effect across individuals up to a certain level, in some studies, especially when the followup periods are significantly long, time-varying age effects could be needed. A possible extension to the PD-SCCS model is given with the intensity function $(\alpha_i \rho_0(t) + \delta N(t^-) \exp\{x(t)\beta)\}$, $t \geq 0$. That is, including the age effect in this PD-SCCS model would allow fitting to the data having a decreasing trend without causing bias in the estimation of the positive event dependence parameter. We will investigate such models as a future work.

Bibliography

[1] Amorim, L. D., and Cai, J. Modelling recurrent events: a tutorial for analysis in epidemiology. *International Journal of Epidemiology 44*, 1 (2015), 324–333.

[2] Andersen, P. K., Borgan, O., Gill, R., and Keiding, N. *Statistical models based on counting processess*. Springer-Verlag, New York, 1993.

[3] Bresalier, R. S., Sandler, R. S., Quan, H., Bolognese, J. A., Oxenius, B., Horgan, K., Lines, C., Riddell, R., Morton, D., Lanas, A., Konstam, M. A., and Baron, J. A. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. *The New England Journal of Medicine 352* (2005), 1092–1102.

[4] Cigsar, C., and Lawless, J. F. Assessing transient carryover effects in recurrent event process, with application to chronic health conditions. *The Annals of Applied Statistics 6*, 4 (2012), 1641–1663.

[5] Cook, R. J., and Lawless, J. F. *The Statistical Analysis of Recurrent Events*. Springer, 2007.

[6] Cox, D., and Hinkley, D. *Theoretical Statistics*. Chapman and Hall, London, 1974.

[7] Cox, D. R., and Isham, V. *Point Processes*. Chapman and Hall/CRC, 1980.

[8] Cox, D. R., and Oakes, D. *Analysis of Survival Data*. Chapman and Hall/CRC, 1984.

[9] Daley, D. J., and Vere-Jones, D. *An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods*. Springer, New York, 2003.

[10] Dean, C., and Lawless, J. Tests for detecting overdispersion in poisson regression models. *Journal of the American Statistical Association 84*, 406 (1989), 467–472.

[11] Douglas, I. J., and Smeeth, L. Exposure to antipsychotics and risk of stroke: self controlled case series study. *The BMJ 337* (2008).

[12] Farrington, C. P. Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics 51*, 1 (1995), 228–235.

[13] Farrington, C. P., Anaya-Izquierdo, K., Whitaker, H. J., Hocine, M. N., Douglas, I., and Smeeth, L. Self-controlled case series analysis with event-dependent observation periods. *Journal of the American Statistical Association 106*, 494 (2011), 417–426.

[14] Farrington, C. P., and Whitaker, H. J. Semiparametric analysis of case series data. *Royal Statistical Soceity.Series C : Applied Statistics 55*, 5 (2006), 553–594.

[15] Farrington, C. P., Whitaker, H, J., and Hocine, M. N. Case series analysis for censored, perturbed, or curtailed post-event exposures. *Biostatistics 10*, 1 (2009), 3–16.

[16] Farrington, C. P. F., and Hocine, M. N. Within-individual dependence in self-controlled case series models for recurrent events. *Royal Statistical Soceity.Series C : Applied Statistics 59*, 3 (2010), 457–475.

[17] Farrington, P., Pugh, S., Colville, A., Flower, A., Nash, J., Morgan-Capner, P., Rush, M., and Miller, E. A new method for active surveillance of adverse events from diphtheria/tetanus/pertussis and measles/mumps/rubella vaccines. *The Lancet 345*, 8949 (1995), 567–569.

[18] Ghebremichael-Weldeselassie, Y., Whitaker, H. J., and Farrington, C. P. Self-controlled case series method with smooth age effect. *Statistics in Medicine 33*, 4 (2014), 639–649.

[19] Ghebremichael-Weldeselassie, Y., Whitaker, H. J., and Farrington, C. P. Flexible modelling of vaccine effect in self-controlled case series models. *Biometrical Journal 58*, 3 (2016), 607–622.

[20] Ghebremichael-Weldeselassie, Y., Whitaker, H. J., and Farrington, C. P. Spline-based self-controlled case series method. *Statistics in Medicine 36*, 19 (2017), 3022–3038.

[21] Hubbard, R., Farrington, P., Smith, C., Smeeth, L., and Tattersfield, A. Exposure to tricyclic and selective serotonin reuptake inhibitor antidepressants and the risk of hip fracture. *American Journal of Epidemiology 158*, 1 (2003), 77–84.

[22] Kalbfleisch, J. D., and Prentice, R. L. *The Statistical Analysis of Failure Time Data, Second Edition*. John Wiley and Sons, Inc., 2002.

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

[24] Langan, S. M., Minassian, C., Smeeth, L., and Thomas, S. L. Risk of stroke following herpes zoster: A self-controlled case-series study. *Clinical Infectious Diseases 58*, 11 (2014), 1497–1503.

[25] Lash, T. L., VanderWeele, T. J., Haneause, S., and Rothman, K. *Modern Epidemiology*. Wolters Kluwer Health, 2021.

[26] Lawless, J. F. *Statistical Models and Methods for Lifetime Data, Second Edition*. Wiley, 2003.

[27] Lawless, J. F., Cigsar, C., and Cook, R. J. Testing for monotone trend in recurrent event processes. *Technometrics 54*, 2 (2012), 147–158.

[28] Lawless, J. F., and Thiagarajah, K. A point process model incorporating renewals and time trends. *Technometrics 38* (1996), 158–168.

[29] Miller, E., Waight, P., Farrington, P., Andrews, N., Stowe, J., and Taylor, B. Idiopathic thrombocytopenic purpura and MMR vaccine. *Archives of Disease in Childhood 84* (2001), 227–229.

[30] Miloslavsky, M., Keles, S., van Der Laan, M. J., and Butter, S. Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *Journal of the Royal Statistical Society. Series B 66* (2004), 239–257.

[31] Minassian, C., D'Aiuto, F., Hingorani, A. D., and Smeeth, L. Invasive dental treatment and risk for vascular events. *Annals of Internal Medicine 153*, 8 (2010), 499–506.

[32] Moran, P. Maximum-likelihood estimation in non-standard conditions. *Proceedings of the Cambridge Philosophical Society 70* (1971), 441–450.

[33] Musonda, P., Farrington, C. P., and Whitaker, H. J. Sample sizes for self-controlled case series studies. *Statistics in Medicine 25*, 15 (2006), 2618–2631.

[34] Musonda, P., Hocine, M. N., Whitaker, H., and Farrington, C. Self controlled case series analyses: small sample performance. *Compulational Statistics and Data Analysis 52*, 4 (2008b), 1942–1957.

[35] Parzen, E. *Stochastic Processes*. Holden-Day, San Francisco, 1962.

[36] Petersen, I., Douglas, I., and Whitaker, H. Self controlled case series methods: an alternative to standard epidemiological study designs. *The BMJ 354* (2016).

[37] Rigdon, S. E., and Basu, A. P. *Statistical methods for the reliability of repariable systems*. John Wiley and Sons, New York, 2000.

[38] Ross, S. M. *Stochastic Processes*. Wiley, 1996.

[39] Simpson, S. E. A positive event dependence model for self-controlled case series with applications in postmarketing surveillance. *Biometrics 69*, 1 (2013), 128–136.

[40] Simpson, S. E., Madigan, D., Zorych, I., Schuemie, M. J., Ryan, P. B., and Suchard, M. A. Multiple self-controlled case series for large-scale longitudinal observational databases. *Biometrics 69*, 4 (2013), 893–902.

[41] Suchard, M. A., Zorych, I., Simpson, S. E., Schuemie, M. J., Ryan, P. B., and Madigan, D. Empirical performance of the self-controlled case sereis design: lessons for developing a risk identification and analysis system. *Drug Safety 36* (2013), 83–93.

[42] Thompson, Jr., W. A. *Point Process Models with Applications to Safety and Reliability*. Chapman and Hall, 1988.

[43] Weldeselassie, Y. G., Whitaker, H. J., and Farrington, C. P. Use of the self-controlled case-series method in vaccine safety studies: review and recommendations for best practice. *Epidemiology and Infection 139*, 12 (2011), 1805–1817.

[44] Whitaker, H. J., Farrington, C. P., Spiessens, B., and Musonda, P. Tutorial in biostatistics: the self-controlled case series method. *Statistica in Medicine 25*, 10 (2006), 1768–1797.

[45] Whitaker, H. J., Steer, C. D., and Farrington, C. P. Self-controlled case series studies: Just how rare does a rare non-recurrent outcome need to be? *Biometrical Journal 60*, 6 (2018), 1110–1120.

[46] Whitaker, H. J., Weldeselassie, Y. G., Douglas, I. J., Liam, S., and Farrington, C. Investigating the assumptions of the self-controlled case series method. *Statistica in Medicine 37*, 4 (2018), 643–658.

[47] Zenner, D., Kruijshaar, M. E., Andrews, N., and Abubakar, I. Risk of tuberculosis in pregnancy a national, primary care–based cohort and self-controlled case series study. *American Journal of Respiratory and Critical Care Medicine 185*, 7 (2012), 779–784.

Appendix A

In this appendix, we present the remaining results of our simulation study explained in Section 3.1. More details can be found there. The rest of simulation results in Section 3.1.

Table 1: Simulation results when the conditional distribution approach was used to generate data with scenarios $E[N_i(500)] = 1.5$..

					Mean($\hat{\beta}$)		B$\bar{\text{i}}$as($\hat{\beta}$)($\times 10^{-3}$)		$\widehat{\text{var}}$($\hat{\beta}$)($\times 10^{-3}$)		MSE($\hat{\beta}$)($\times 10^{-3}$)	
Δ	e^{β}	β	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
20	0.5	-0.693	1681.0	34.3	-0.654	-0.707	38.840	-13.818	24.694	26.938	26.178	27.102
	1	0.000	1647.2	67.1	-0.008	-0.010	-7.735	-9.580	15.100	16.033	15.145	16.108
	2	0.693	1586.4	129.6	0.670	0.691	-23.470	-2.150	8.665	8.903	9.207	8.898
	3	1.099	1529.3	186.4	1.062	1.093	-36.899	-5.802	6.157	6.307	7.512	6.334
	4	1.386	1475.8	240.2	1.348	1.383	-37.843	-2.865	4.513	4.576	5.941	4.580
40	0.5	-0.693	1647.6	69.0	-0.668	-0.695	25.364	-1.675	15.009	15.461	15.638	15.448
	1	0.000	1584.0	131.4	-0.005	-0.007	-4.914	-7.086	9.273	9.463	9.288	9.504
	2	0.693	1472.9	243.9	0.678	0.689	-14.652	-4.530	4.602	4.659	4.812	4.675
	3	1.099	1374.1	342.2	1.085	1.100	-13.979	1.474	3.378	3.392	3.570	3.391
	4	1.386	1288.6	426.3	1.368	1.386	-18.202	-0.500	3.129	3.137	3.458	3.134
60	0.5	-0.693	1613.4	102.7	-0.678	-0.699	14.722	-6.048	10.762	11.244	10.968	11.269
	1	0.000	1523.5	193.4	-0.002	-0.004	-1.679	-3.592	5.749	5.919	5.747	5.926
	2	0.693	1370.6	345.9	0.683	0.690	-9.754	-3.163	3.760	3.803	3.852	3.809
	3	1.099	1245.7	470.1	1.087	1.097	-11.146	-1.589	2.938	2.919	3.059	2.918
	4	1.386	1143.3	572.8	1.374	1.385	-12.497	-1.311	2.738	2.745	2.891	2.744
80	0.5	-0.693	1580.0	136.4	-0.685	-0.701	8.415	-7.416	7.422	7.676	7.486	7.724
	1	0.000	1463.4	252.5	0.000	-0.002	0.211	-2.010	4.157	4.183	4.152	4.183
	2	0.693	1279.0	437.2	0.687	0.692	-6.008	-1.524	3.018	3.051	3.051	3.050
	3	1.099	1137.8	578.6	1.090	1.098	-8.140	-1.093	2.829	2.881	2.892	2.879
	4	1.386	1026.5	689.3	1.375	1.383	-11.598	-3.040	2.607	2.644	2.739	2.650
100	0.5	-0.693	1546.0	171.1	-0.682	-0.696	11.053	-2.353	6.907	7.083	7.022	7.081
	1	0.000	1407.8	307.9	-0.002	-0.005	-2.465	-5.025	4.148	4.174	4.150	4.195
	2	0.693	1198.3	517.3	0.687	0.690	-5.961	-2.676	2.823	2.836	2.856	2.840
	3	1.099	1045.2	670.1	1.093	1.099	-5.351	0.254	2.756	2.773	2.782	2.770
	4	1.386	930.1	785.6	1.379	1.386	-6.898	-0.171	2.404	2.423	2.450	2.420

Table 2: Simulation results when the conditional distribution approach was used to generate data with scenarios $E[N_i(500)] = 2$.

					Mean($\hat{\beta}$)		Bi$\bar{a}$s($\hat{\beta}$)($\times 10^{-3}$)		$\widehat{\text{var}}(\hat{\beta})(\times 10^{-3})$		MSE($\hat{\beta}$)($\times 10^{-3}$)	
Δ	e^{β}	β	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
20	0.5	-0.693	2144.8	43.6	-0.656	-0.707	37.485	-14.190	21.468	24.026	22.852	24.203
	1	0.000	2100.7	86.2	0.001	0.000	1.342	-0.408	12.273	12.892	12.262	12.879
	2	0.693	2023.7	164.8	0.667	0.689	-26.527	-4.205	6.612	6.833	7.309	6.844
	3	1.099	1950.2	237.8	1.063	1.094	-35.233	-4.397	4.734	4.827	5.971	4.842
	4	1.386	1881.3	307.3	1.353	1.388	-33.649	1.376	3.772	3.801	4.900	3.799
40	0.5	-0.693	2101.4	87.7	-0.670	-0.696	23.513	-3.101	11.255	11.790	11.797	11.788
	1	0.000	2021.1	169.2	0.005	0.003	4.707	3.247	6.562	6.760	6.578	6.764
	2	0.693	1878.1	311.5	0.680	0.691	-12.728	-2.426	3.692	3.704	3.850	3.706
	3	1.099	1753.9	436.0	1.083	1.098	-15.806	-0.474	2.802	2.839	3.049	2.836
	4	1.386	1645.8	543.9	1.368	1.386	-18.256	-0.775	2.426	2.446	2.757	2.444
60	0.5	-0.693	2056.7	130.3	-0.683	-0.703	10.095	-9.748	7.883	8.161	7.977	8.248
	1	0.000	1942.9	247.0	0.000	-0.002	0.201	-2.246	4.307	4.318	4.303	4.319
	2	0.693	1747.3	442.0	0.685	0.692	-7.696	-0.991	2.777	2.764	2.834	2.762
	3	1.099	1587.4	600.0	1.089	1.099	-9.608	-0.079	2.198	2.208	2.289	2.206
	4	1.386	1457.7	731.5	1.376	1.387	-10.576	0.733	2.056	2.070	2.166	2.068
80	0.5	-0.693	2015.3	175.3	-0.678	-0.693	15.254	0.353	6.809	7.038	7.035	7.031
	1	0.000	1866.8	322.3	0.001	-0.001	0.722	-1.110	3.546	3.560	3.543	3.558
	2	0.693	1630.0	558.7	0.690	0.694	-3.564	1.255	2.522	2.526	2.532	2.525
	3	1.099	1451.1	737.0	1.090	1.097	-8.833	-1.753	2.130	2.155	2.206	2.155
	4	1.386	1309.9	878.8	1.375	1.383	-11.529	-3.151	2.008	2.035	2.139	2.043
100	0.5	-0.693	1971.5	218.2	-0.682	-0.695	11.612	-1.397	4.852	5.007	4.982	5.004
	1	0.000	1793.5	394.9	0.004	0.001	3.558	1.073	3.212	3.196	3.222	3.194
	2	0.693	1527.2	659.7	0.688	0.691	-5.213	-2.144	2.020	2.026	2.045	2.029
	3	1.099	1335.7	854.1	1.090	1.095	-8.567	-3.155	1.995	1.984	2.067	1.992
	4	1.386	1185.2	1001.5	1.381	1.388	-5.143	1.567	1.881	1.889	1.906	1.889

Table 3: Simulation results when the conditional distribution approach was used to generate data with scenarios $E[N_i(500)] = 2.5$.

					Mean($\hat{\beta}$)		Bi$\bar{a}$s($\hat{\beta}$)($\times 10^{-3}$)		$\widehat{\text{var}}(\hat{\beta})(\times 10^{-3})$		MSE($\hat{\beta}$)($\times 10^{-3}$)	
Δ	e^{β}	β	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
20	0.5	-0.693	2595.8	52.7	-0.655	-0.708	38.228	-14.481	17.005	18.517	18.449	18.708
	1	0.000	2543.2	103.8	-0.001	-0.005	-1.302	-4.756	10.116	10.638	10.108	10.650
	2	0.693	2448.4	200.1	0.670	0.693	-23.070	-0.618	5.487	5.639	6.013	5.634
	3	1.099	2358.3	288.5	1.066	1.098	-32.174	-1.095	3.798	3.893	4.829	3.890
	4	1.386	2277.3	371.2	1.350	1.386	-35.891	-0.721	3.251	3.272	4.536	3.270
40	0.5	-0.693	2539.3	105.8	-0.670	-0.697	22.827	-3.889	9.824	10.259	10.335	10.264
	1	0.000	2446.5	203.8	0.001	-0.001	0.594	-0.894	5.358	5.451	5.353	5.446
	2	0.693	2272.0	376.8	0.681	0.691	-12.359	-1.939	3.254	3.241	3.403	3.241
	3	1.099	2122.9	526.5	1.081	1.096	-17.870	-2.747	2.334	2.336	2.651	2.341
	4	1.386	1987.7	658.5	1.370	1.387	-16.276	1.101	1.883	1.862	2.146	1.861
60	0.5	-0.693	2490.7	159.1	-0.675	-0.694	17.824	-0.766	6.511	6.767	6.823	6.761
	1	0.000	2350.4	298.2	-0.002	-0.004	-1.701	-3.711	4.016	4.076	4.015	4.086
	2	0.693	2115.4	533.1	0.682	0.688	-10.886	-4.650	2.311	2.296	2.428	2.315
	3	1.099	1920.3	725.8	1.089	1.099	-9.565	0.048	1.871	1.874	1.961	1.872
	4	1.386	1765.1	883.8	1.373	1.385	-12.841	-1.589	1.801	1.806	1.964	1.807
80	0.5	-0.693	2433.7	210.7	-0.681	-0.697	12.274	-3.605	5.367	5.593	5.512	5.600
	1	0.000	2258.7	389.2	0.000	-0.003	-0.101	-2.782	3.027	3.054	3.024	3.059
	2	0.693	1973.6	674.5	0.687	0.692	-6.004	-1.478	2.121	2.147	2.155	2.147
	3	1.099	1754.3	893.9	1.092	1.099	-6.600	0.421	1.694	1.699	1.736	1.698
	4	1.386	1583.5	1065.0	1.377	1.385	-9.682	-1.392	1.673	1.675	1.765	1.676
100	0.5	-0.693	2384.6	263.5	-0.682	-0.696	10.670	-2.635	3.930	3.977	4.040	3.980
	1	0.000	2169.9	477.4	0.004	0.001	3.940	1.203	2.550	2.566	2.563	2.565
	2	0.693	1849.7	800.8	0.690	0.693	-3.093	0.182	1.763	1.758	1.771	1.757
	3	1.099	1614.7	1033.7	1.092	1.098	-6.351	-0.897	1.714	1.709	1.753	1.708
	4	1.386	1437.6	1211.7	1.378	1.385	-8.205	-1.494	1.564	1.572	1.630	1.573

Table 4: Simulation results when the cohort model approach was used to generate data with $\alpha = \frac{1}{1000}$.

					Mean($\hat{\beta}$)		$\widetilde{\text{Bias}}(\hat{\beta})(\times 10^{-2})$		$\widehat{\text{var}}(\hat{\beta})(\times 10^{-2})$		MSE($\hat{\beta}$)($\times 10^{-2}$)	
Δ	e^{β}	β	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
20	0.5	-0.693	480.4	10.0	-0.698	-0.731	-0.470	-3.825	44.634	12.026	44.591	12.160
	1	0.000	480.3	19.5	-0.026	-0.030	-2.557	-3.039	5.337	5.604	5.397	5.690
	2	0.693	481.1	39.4	0.663	0.685	-3.011	-0.803	2.888	2.957	2.976	2.960
	3	1.099	481.7	59.1	1.062	1.093	-3.679	-0.559	1.920	1.930	2.053	1.932
	4	1.386	480.8	78.5	1.346	1.381	-4.020	-0.533	1.451	1.459	1.611	1.460
40	0.5	-0.693	462.0	19.4	-0.682	-0.708	1.154	-1.503	5.102	5.370	5.111	5.387
	1	0.000	461.1	38.2	-0.014	-0.016	-1.435	-1.645	2.866	2.954	2.884	2.978
	2	0.693	462.4	76.5	0.671	0.682	-2.231	-1.131	1.513	1.536	1.561	1.547
	3	1.099	462.7	115.8	1.083	1.098	-1.580	-0.071	1.103	1.119	1.127	1.118
	4	1.386	460.5	153.9	1.371	1.389	-1.516	0.232	0.821	0.820	0.843	0.820
60	0.5	-0.693	443.7	28.3	-0.686	-0.710	0.712	-1.695	4.173	4.319	4.174	4.344
	1	0.000	443.8	56.3	-0.009	-0.011	-0.912	-1.107	2.086	2.128	2.092	2.138
	2	0.693	443.4	113.0	0.685	0.692	-0.832	-0.145	1.121	1.113	1.127	1.113
	3	1.099	444.0	169.6	1.088	1.098	-1.014	-0.020	0.843	0.850	0.853	0.849
	4	1.386	443.5	225.9	1.376	1.387	-1.049	0.042	0.679	0.682	0.690	0.682
80	0.5	-0.693	426.8	37.0	-0.685	-0.702	0.811	-0.856	2.864	2.937	2.867	2.942
	1	0.000	425.1	73.5	-0.001	-0.004	-0.100	-0.398	1.603	1.654	1.602	1.654
	2	0.693	426.0	147.4	0.690	0.694	-0.300	0.067	0.909	0.913	0.909	0.912
	3	1.099	426.9	221.4	1.092	1.099	-0.639	0.058	0.701	0.703	0.704	0.703
	4	1.386	427.2	293.5	1.373	1.381	-1.324	-0.512	0.557	0.558	0.574	0.560
100	0.5	-0.693	409.7	45.2	-0.685	-0.700	0.771	-0.667	2.497	2.574	2.500	2.576
	1	0.000	410.3	90.2	-0.001	-0.004	-0.116	-0.358	1.448	1.440	1.446	1.440
	2	0.693	409.9	179.9	0.688	0.691	-0.518	-0.228	0.829	0.830	0.831	0.830
	3	1.099	409.6	270.3	1.097	1.102	-0.210	0.308	0.620	0.618	0.619	0.618
	4	1.386	410.9	359.3	1.376	1.382	-1.051	-0.428	0.539	0.539	0.550	0.540

Table 5: Simulation results when the cohort model approach was used to generate data with $\alpha = \frac{1}{500}$.

					Mean($\hat{\beta}$)		$\widetilde{\text{Bias}}(\hat{\beta})(\times 10^{-2})$		$\widehat{\text{var}}(\hat{\beta})(\times 10^{-2})$		MSE($\hat{\beta}$)($\times 10^{-2}$)	
Δ	e^{β}	β	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
20	0.5	-0.693	958.5	19.4	-0.677	-0.728	1.602	-3.446	5.331	5.648	5.351	5.761
	1	0.000	963.4	39.1	-0.014	-0.017	-1.393	-1.709	2.478	2.653	2.495	2.679
	2	0.693	960.4	78.3	0.665	0.687	-2.831	-0.647	1.374	1.431	1.452	1.434
	3	1.099	961.0	117.2	1.061	1.092	-3.788	-0.703	0.929	0.937	1.071	0.941
	4	1.386	960.4	156.5	1.347	1.382	-3.946	-0.432	0.724	0.727	0.879	0.728
40	0.5	-0.693	924.4	38.2	-0.685	-0.712	0.855	-1.890	2.757	2.965	2.762	2.998
	1	0.000	923.8	77.1	-0.001	-0.002	-0.063	-0.239	1.379	1.400	1.378	1.399
	2	0.693	924.1	153.7	0.681	0.691	-1.249	-0.259	0.760	0.774	0.775	0.774
	3	1.099	922.5	229.8	1.080	1.095	-1.901	-0.386	0.608	0.617	0.643	0.618
	4	1.386	924.0	307.3	1.367	1.385	-1.910	-0.177	0.421	0.421	0.457	0.421
60	0.5	-0.693	887.7	56.3	-0.682	-0.703	1.147	-0.943	1.761	1.833	1.773	1.840
	1	0.000	888.1	113.0	-0.001	-0.003	-0.100	-0.316	1.011	1.043	1.010	1.043
	2	0.693	884.6	226.2	0.691	0.697	-0.230	0.387	0.557	0.557	0.557	0.558
	3	1.099	887.6	337.1	1.084	1.094	-1.445	-0.503	0.391	0.392	0.411	0.394
	4	1.386	888.0	450.8	1.374	1.385	-1.276	-0.175	0.332	0.331	0.348	0.331
80	0.5	-0.693	852.1	73.8	-0.679	-0.695	1.403	-0.224	1.307	1.318	1.326	1.317
	1	0.000	852.1	147.5	0.002	0.000	0.195	-0.047	0.823	0.830	0.823	0.830
	2	0.693	851.8	294.8	0.690	0.695	-0.304	0.149	0.473	0.473	0.474	0.473
	3	1.099	852.7	441.5	1.091	1.098	-0.771	-0.097	0.347	0.350	0.352	0.350
	4	1.386	853.1	589.4	1.379	1.387	-0.770	0.040	0.301	0.298	0.307	0.298
100	0.5	-0.693	820.2	90.0	-0.686	-0.698	0.702	-0.505	1.172	1.195	1.176	1.197
	1	0.000	820.6	179.8	-0.001	-0.004	-0.131	-0.357	0.682	0.694	0.681	0.694
	2	0.693	819.9	360.0	0.689	0.693	-0.383	-0.059	0.399	0.399	0.400	0.398
	3	1.099	819.1	541.1	1.097	1.102	-0.202	0.327	0.347	0.347	0.347	0.348
	4	1.386	820.4	720.8	1.381	1.387	-0.514	0.091	0.282	0.281	0.284	0.281

Table 6: Simulation results when the cohort model approach was used to generate data with $\alpha = \frac{1}{250}$.

					Mean($\hat{\beta}$)		$\widehat{\text{Bias}}(\hat{\beta})(\times 10^{-2})$		$\widehat{\text{var}}(\hat{\beta})(\times 10^{-2})$		MSE($\hat{\beta}$)($\times 10^{-2}$)	
Δ	e^{β}	β	$\overline{N(\neg\Delta)}$	$\overline{N(\Delta)}$	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
20	0.5	-0.693	1921.6	39.2	-0.656	-0.705	3.740	-1.178	2.378	2.582	2.516	2.593
	1	0.000	1921.9	78.2	-0.007	-0.009	-0.680	-0.874	1.337	1.398	1.340	1.404
	2	0.693	1920.6	157.7	0.674	0.696	-1.940	0.319	0.685	0.702	0.722	0.702
	3	1.099	1921.3	235.3	1.067	1.097	-3.183	-0.142	0.485	0.495	0.585	0.494
	4	1.386	1922.5	313.9	1.350	1.385	-3.587	-0.081	0.359	0.362	0.487	0.362
40	0.5	-0.693	1846.6	76.7	-0.673	-0.701	1.972	-0.784	1.375	1.460	1.412	1.465
	1	0.000	1845.2	152.8	-0.006	-0.008	-0.574	-0.759	0.696	0.726	0.699	0.731
	2	0.693	1845.9	306.9	0.681	0.691	-1.214	-0.193	0.395	0.399	0.409	0.399
	3	1.099	1846.0	460.5	1.083	1.097	-1.608	-0.126	0.269	0.271	0.295	0.271
	4	1.386	1844.3	614.3	1.370	1.387	-1.652	0.066	0.209	0.209	0.236	0.209
60	0.5	-0.693	1773.6	112.3	-0.680	-0.701	1.299	-0.758	0.945	0.967	0.961	0.971
	1	0.000	1775.8	225.3	-0.001	-0.004	-0.105	-0.388	0.538	0.537	0.537	0.538
	2	0.693	1773.3	451.8	0.688	0.694	-0.558	0.093	0.294	0.295	0.297	0.295
	3	1.099	1773.0	676.9	1.089	1.099	-0.919	0.055	0.212	0.214	0.220	0.214
	4	1.386	1773.0	903.0	1.376	1.388	-0.987	0.143	0.154	0.156	0.164	0.156
80	0.5	-0.693	1704.6	147.2	-0.681	-0.695	1.226	-0.229	0.696	0.706	0.710	0.706
	1	0.000	1705.4	295.1	0.004	0.001	0.427	0.150	0.383	0.390	0.385	0.390
	2	0.693	1704.6	588.9	0.689	0.694	-0.403	0.042	0.232	0.232	0.233	0.231
	3	1.099	1704.5	882.9	1.092	1.099	-0.671	0.020	0.183	0.183	0.187	0.183
	4	1.386	1705.2	1179.1	1.379	1.387	-0.722	0.096	0.145	0.145	0.150	0.145
100	0.5	-0.693	1638.5	179.5	-0.683	-0.697	0.990	-0.407	0.637	0.660	0.646	0.661
	1	0.000	1639.0	360.1	0.002	0.000	0.235	-0.038	0.331	0.335	0.332	0.335
	2	0.693	1639.3	719.3	0.689	0.692	-0.443	-0.114	0.222	0.223	0.223	0.223
	3	1.099	1638.8	1080.2	1.095	1.100	-0.403	0.147	0.154	0.155	0.156	0.155
	4	1.386	1639.8	1437.8	1.378	1.385	-0.790	-0.158	0.152	0.152	0.159	0.153

Appendix B

MMR dataset: case is an individual identifier, itp is the age at admission for ITP, sta and end are the age on the first day and last day of observation, respectively. mmr is the age at MMR vaccination. sex is 1 for males and 2 for females.

case	itp	sta	end	mmr	sex
1	691	454	730	670	1
2	722	366	730	868	2
3	442	366	730	540	1
4	429	366	730	378	2
5	414	366	730	710	1
5	418	366	730	710	1
6	708	439	730	487	1
7	615	366	730	461	1
8	463	366	730	526	1
9	440	366	730	529	1
9	473	366	730	529	1
10	477	366	730	458	1
11	396	366	730	374	2
12	676	366	730	428	2
13	480	366	730	446	1
14	633	366	730	423	1
15	403	366	730	365	1
16	419	366	730	369	2
16	443	366	730	369	2
17	553	366	730	889	1
17	666	366	730	889	1
18	705	366	730	389	1
19	419	366	730	389	2
20	402	366	730	385	1
21	406	366	730	458	2
22	494	366	730	468	2
23	374	366	730	819	2
23	389	366	730	819	2
23	452	366	730	819	2
23	522	366	730	819	2
23	564	366	730	819	2
24	598	366	730	430	2
25	409	366	730	384	2
26	612	366	730	398	1
27	381	366	723	427	1
28	438	366	730	427	1
29	425	366	677	647	1
30	543	366	677	422	1
31	609	366	674	860	1
32	412	366	730	387	1
33	407	366	730	396	2
34	484	366	730	408	1
34	623	366	730	408	1
35	411	366	730	383	2