Estimating the Effects of Air Pollutants on Recurrent Hospital Admission for Respiratory Diseases

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon

By
Shan Qiao

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics
Room 142 McLean Hall
106 Wiggins Road
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5E6

Abstract

Recurrent data are widely encountered in many applications. This thesis work focuses on how the recurrent hospital admissions relate to the air pollutants. In particular, we consider the data for two major cities in Saskatchewan. The study period ranges from January 1, 2005 to December 30, 2011 and involves 20,284 patients aged 40 years and older. The hospital admission data is from the Canadian Institute for Health Information (CIHI). The air pollutants data is from the National Air Pollution Surveillance Program (NAPS) from Environment Canada. The data set has been approved by the Biomedical Research Ethics Board, University of Saskatchewan. The gaseous pollutants included in this study are carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), as well as particulate matter $PM_{2.5}$ (tiny particles in the air that are ≤ 2.5 microns in width).

In the data analysis, we applied three different existing models to all respiratory diseases and asthma, respectively. The three models are the Poisson process model (also called Andersen-Gill model), the Poisson process model with the number of previous events as a covariate and the Poisson process model with shared gamma distributed frailties (random effects). For all respiratory diseases, the Poisson process model with random effects provides the best fit in comparison to the other two models. The model output suggests that the increased risk of hospital readmission is significantly associated with increased CO and O_3. For asthma, the Poisson process model provides the best fit in comparison to the other two models. We found that only CO and O_3 have significant effects on recurrent hospital admissions due to asthma. We concluded this thesis with the discussion on the current and potential future work.

Acknowledgements

I wish to express my deepest appreciation to my M.Sc supervisors Dr. Shahedul A. Khan and Dr. Juxin Liu for their judicious guidance, unwavering support and caring patience all through the year of working on this M.Sc thesis. In spite of having a lot of work, they spend a lot of time listening to my presentation, discussing problems together and answering my questions at any time. They always provided me kind comments about the work I did, which encouraged me to keep working hard on my research. It was my honor to have Dr. Shahedul A. Khan and Dr. Juxin Liu as my supervisors.

I also would like to thank my committee members: Dr. Chris Soteros, Dr. Mik Bickis, Dr. Shahedul A. Khan and Dr. Juxin Liu for providing me their insightful comments.

I would like to thank Prof. Richard Cook for his valuable suggestions.

I am grateful to the University of Saskatchewan for providing me the Graduate Student Scholarship and Education Equity Scholarship to support my study in Canada.

I would like to thank the Canadian Institute for Health Information (CIHI) for providing me the data for this research through the CIHI Graduate Student Data Access Program. Without their help, it is impossible to complete this thesis.

I also wish to thank Environment Canada National Air Pollution Surveillance Program (NAPS) and Saskatchewan Ministry of Environment for allowing me to download the air pollutants and weather data from their websites.

Last but not least, I wish to thank my parents, Wenrui Qiao and Huili Sun, for their unconditionally support and love throughout the years.

Contents

List of Tables

Chapter 1

Introduction

The primary objective of my study is to investigate the effects of air pollutants on recurrent hospital admissions due to respiratory diseases in two major cities (Saskatoon and Regina) of Saskatchewan.

This chapter presents the motivation of this work, specific research objective, a review of related works, and a brief introduction to the methodology used in this thesis. A general discussion of the existing research, presented in Section 1.1, demonstrates that air pollutants can pose significant adverse effects on population respiratory system. An introduction to the concepts of recurrent event data and a review of the statistical methods for modeling recurrent event data are presented in Section 1.2. An outline of this thesis is presented in Section 1.3.

1.1 Background and Objective

Air pollutants are well-established risk factors that can cause severe health-related problems among humans. In particular, adverse effects in lung and respiratory systems are commonly observed across the world (Firket, 1936; Burnett et al., 1997; Cho et al., 2000; Xu et al., 1994; Wong et al., 1999). In the mid-twentieth century, a series of air pollution disasters in Meuse Valley (Firket, 1936), London (Logan, 1953) and Donora (Ciocco et al., 1961) resulted in high concentration of pollutants into the air and caused many human deaths. Firket (1936) mentioned that fog along Meuse Valley not only injured people, but also caused severe respiratory problems among many people, which resulted in a large number of deaths. The Great Smog (1952) in London is known to be the worst air pollution event for the history of the United Kingdom; Logan (1953) reported that during the following week of the fog, a

large number of people of all ages died from problems caused by the difficulty of breathing. The Donora smog in 1948 is considered to be one of the worst air pollution events in the United States; Ciocco and Thompson (1961) reported that individuals with acute illness during the smog period had subsequently higher rates of mortality and morbidity than the individuals living in the same community without such an illness at that time.

In Canada, numerous reports and articles have demonstrated a fairly consistent association between air pollution and respiratory diseases. Burnett et al. (1997) reported a significant association between ozone and respiratory hospitalization in 16 cities across Canada. Fung (2006) observed significant effects of air pollution on respiratory diseases hospital admissions among the elderly in Vancouver. Villeneuve (2007) showed that exposure to ambient levels of air pollution is one of the main causes of asthma hospitalization, particularly among young children and the elderly in northern Alberta. A large number of similar investigations in Korea (Cho et al., 2000), China (Xu et al., 1994; Wong et al., 1999) and Europe (Derriennic et al., 1989) also reported the adverse effects of air pollution on the human health systems.

The above discussion demonstrates that air pollution can trigger serious health-related problems to human beings due to elevated concentrations of toxic pollutants. These may lead to subtle biochemical changes in the human body system, difficulty in breathing or even death (Health Canada, 2006). The adverse health effects in turn can lead to an increase in health care costs, which comprise of an increase in medication use, doctor and emergency-room visits, hospital admissions, and so on. Consequently, it is of paramount importance to investigate the relationship between ambient air pollution concentrations and the rate of hospitalization due to respiratory diseases. Such findings can be extremely useful to us. For example, the atmospheric scientists and policy makers can work together to take appropriate measures to control the ambient air pollution levels by identifying the sources of pollution in the air; the health workers can suggest preventive actions against poor air quality, which may lead to a better living environment. Therefore, many researchers have paid attention to the relationship between ambient air pollution and respiratory diseases (e.g., Atkinson et al., 1999; Burnett et al., 1997, 2000; Gouveia et al., 2000; Fung, 2006; Villeneuve, 2007). The findings from this study can be very useful to assist the policy makers to adjust the regulations

which include changing and/or modifying the criteria for the emissions from factories. From our own perspectives, information on the effects of air pollutants on our respiratory system will definitely increase the awareness of protecting the environment as well as protecting ourselves from the adverse effects of air pollutants.

So far, no such studies focus on the situation in Saskatchewan. The purpose of my thesis is to fill in this blank. The data come from different resources in the real world. Available acute respiratory disease hospitalization data are from the Canadian Institute for Health Information (CIHI), Discharge Abstract Database (DAD). The air pollution data are from the Environment Canada National Air Pollution Surveillance Program (NAPS), which allowed us to carry out statistical analyses to explore the potential effects of air pollutants on repeated hospital admissions due to respiratory diseases. The study period ranges from January 1, 2005 to December 30, 2011. The hospital admission data include acute inpatient (i.e., a patient has a short period of time hospitalization) or day surgery records of the patients aged 40 years and above who were admitted to hospitals governed by the Regina Qu'Appelle Regional Health Authority and Saskatoon Regional Health Authority in Saskatchewan (see Section 1.2.1 and Chapter 3 for detail). The air pollution and weather data include the daily average concentrations of gaseous pollutants, particulate matters, temperature and relative humidity.

1.2 Recurrent Data

In the following, we introduce the concept of recurrent events in Section 1.2.1. Then, we present a review of some statistical methods to analyze recurrent events in Section 1.2.2.

1.2.1 Recurrent Event Data

In a large number of health and medical studies, interest lies in non-fatal events. Thus, it is possible to observe the interested events repeatedly over the study period for everyone (see Kalbfleisch and Prentice, 2002). Examples of such events include infections, repeated heart attacks of coronary patients during a heart disease treatment, recurrent hospital admissions due to respiratory diseases, and successive occurrences of tumours in cancer research-such

repeated events are often referred to as recurrent events.

For each individual, the number of hospital admissions due to respiratory diseases over time plays an important role in modeling the data; see Chapter 2. In the hospital admission data, some people experienced only one admission while other people had been hospitalized multiple times. Note that all hospitalizations due to respiratory diseases can be considered as events of the same type. Furthermore, note that the hospital admission data can be regarded as recurrent event data, as there are individuals with repeated hospitalizations over time. Therefore, the statistical problem of analyzing repeated hospital admission data can be considered as a recurrent event problem.

Cook and Lawless (2007) pointed out that processes that can generate events repeatedly over time (e.g., repeated hospital admissions due to respiratory diseases) are referred to as *recurrent event processes*. Recurrent event processes can naturally arise in many areas, including health sciences, biomedicine, equipment reliability, engineering and social sciences (Gail et al., 1980; Prentice et al., 1981; Allison, 1984; Andersen et al., 1993).

1.2.2 Modeling Recurrent Event Data: A Review

In classical survival analysis, the event of interest is usually fatal. In such case, the focus is time to the occurrence of at most one event for each individual. Example of such events includes death. The events are treated independent as they come from different individuals. Cox proportional hazards model (1972) is the most widely used regression model to analyze the survival data. The model is a semi-parametric regression model, and based on the specification of a hazard function. A brief description of the basic methods of survival analysis (including the definition of the hazard function and the Cox model) is presented in Section 2.1.

In the studies of recurrent events, the event of interest may occur repeatedly over time to each individual. In such cases, the successive occurrences of the event can induce autocorrelation into the time series of event counts within each individual/patient. The respiratory diseases may lead to impair the function of lung and respiratory system. Even though the patient becomes healthy and stable after the treatment, some risk factors (e.g., air pollution) may cause the disease again. After the treatment of disease, it is possible that the

patient experiences the disease again due to air pollution or other risk factors. Thus, once a patient is admitted into a hospital for a respiratory condition, there may be an increased probability of a subsequent readmission for the same condition. The statistical methods to analyze recurrent event data should take into account such dependency. For this reason, the Cox proportional hazards model for independent event occurrences cannot properly handle modeling the recurrent event data.

The statistical theory of the Cox proportional hazards model can be formulated using the counting process technique (see Andersen and Gill, 1982; Fleming and Harrington, 1991 and Andersen et al., 1993). The counting process is a stochastic process $\{N(t), t \geq 0\}$ which records the cumulative number of events experienced by an individual over the time interval $[0, t]$. This development makes it possible to extend the Cox proportional hazards model to take into account multiple observations (e.g., repeated hospital admissions due to respiratory diseases) (Therneau and Grambsch, 2000). By using the counting process theory, the recurrent event data can be modeled based on the intensity function, which is defined as the instantaneous probability of the occurrence of a new event in a short time interval (see Section 2.2.1 for the mathematical definition of intensity function). Aalen (1975) first studied the nonparametric statistical methods for survival data using the *counting process* technique. Later, Aalen (1978b) described the mathematical details of the general framework of the counting process technique.

Based on reformulating the Cox proportional hazards model into a counting process, several statistical techniques have been developed to analyze recurrent event data. One of the most widely used approaches is the Andersen-Gill (AG) model (Andersen and Gill, 1982). The AG model can be considered as an extension to the Cox proportional hazards model, which can be used to analyze the recurrent events. Details of this model can be found in Kelly et al. (2000), Therneau and Grambsch (2000) and Lim et al. (2007). Description of other widely used approaches can be found in Prentice et al. (1981) and Wei et al. (1989).

Compared with other common approaches for modeling recurrent event data, the AG model is the simplest to set up and is based on the Poisson process. The AG model is also appealing due to its greatly interpretable coefficients; the effects of the covariates can ne described using the relative risk parameters defined in terms of the coefficients. In addition,

the software for the Cox proportional hazards model has been adapted to deal with the AG model for recurrent event data. We use the AG model to analyze the recurrent hospital admission data in this study.

In the AG model, it is assumed that the recurrent events are not affected by previous events that happened to the same individual. In addition, the baseline intensities for all events are the same. The intensity function for the ith individual is

$$Y_i(t)\lambda_0(t)\exp(\mathbf{x}(t)'\boldsymbol{\beta})$$

where the at-risk process $Y_i(t)$ equals one when the individual is under observation and zero otherwise. Thus, the risk set at time t includes all the individuals under observation regardless of the number of recurrences experienced by each individual. The $\lambda_0(t)$ is called *baseline intensity* and $\boldsymbol{\beta}$ is the vector of regression parameters. Details of the notations used in the above equation can be found in Section 2.3.

When modeling recurrent event data, it is important to take into account the within-individual correlation due to repeated event occurrences for each individual and the heterogeneity across individuals. The heterogeneity across individuals arises because the level of association between air pollutants and respiratory hospital admission can differ from one individual to another due to some unobservable and/or unmeasured covariates (e.g., genetic and environmental factors). For instance, some patients may experience the relapse more quickly than others, because they might have a genetic disposition to develop the disease. Ignoring the heterogeneity may lead to some biased estimation. Firstly, the covariate effects may be underestimated. Individuals who have higher risks will have interested event occurrences earlier than others. As time goes on, those remaining in the risk set will have a lower average risk. This will "drag down" the risk rate and result in underestimating the covariate effects (Aalen, 2008). Secondly, the estimates of standard errors can be wrong (Box-Steffensmeier and Jones, 2004). Kelly and Lim (2000) reported that when incorporating heterogeneity, the standard errors were small. For this reason, ignoring the heterogeneity may result in incorrect conclusion in terms of statistical significance in the covariate effects.

Incorporating individual-specific random effects (commonly called frailty) into the model is considered to be a useful way to take into account the heterogeneity across individuals.

In addition, these random effects are useful to incorporate correlation between the repeated events within each individual. Aalen (1988) discussed the impact of heterogeneity in statistical analyses and how the random effects take into account such a heterogeneity. Vaupel et al. (1979) can be regarded as an early contribution in the literature about using random effects models for survival data.

The simplest frailty model is a parametric model with shared gamma distributed frailty term (Duchateau and Janssen, 2008). In this model, all event occurrences within an individual share the same individual-specific frailty. The simplicity is due to two main reasons. Firstly, the gamma distributed frailty term makes it possible to obtain a simple expression for the marginal likelihood when integrating out the frailties from the conditional likelihood function; assuming other distributions for the frailty term can lead to an intractable integration with no closed-form expression for the marginal likelihood (Duchateau and Janssen, 2008). Secondly, the marginal likelihood is fully parametric with a parametric baseline hazard and hence the parameter estimation can be based on classical maximum likelihood method. Models without parametric assumption of the baseline intensity functions are appealing in some settings, such as time-to-event analysis (Cook and Lowless, 2007 and Aalen, 2008), Nielsen et al. (1992), Klein (1992) and Andersen et al. (1993) studied the semiparametric proportional hazards model with gamma frailty, in which the baseline intensity is completely unspecified. Gill (1985) pointed out that the Expectation-Maximisation (EM) algorithm (Dempster et al., 1977) can be used for maximum likelihood estimation for the semiparametric gamma frailty model. Klein (1992), Nielsen st al. (1992), Moeschberger et al. (2003) and Duchateau and Janssen (2008) further described the EM algorithm technique for the semiparametric frailty models. Therneau and Grambsch (2000), Therneau at al. (2003) and Duchateau and Janssen (2008) discussed an alternative approach called penalized partial likelihood method for the semiparametric frailty model; see Section 2.4.2 for detail.

In this study, we use the Anderson-Gill model to fit the recurrent hospital admission data. In order to take into account the heterogeneity across individuals, we also fit Anderson-Gill model with shared gamma frailties. Details about fitting different models can be found in Chapter 3. The method we use to estimate the regression parameters and variance of frailties is the penalized partial likelihood; see Section 2.4.2 for detail.

1.3 Thesis Overview

In Chapter 2, we introduce the notations in detail, concepts and statistical methodologies used in this thesis to analyze our recurrent hospital admission data. We proceed in Chapter 3 with results for the hospital admission data, and summary of our findings regarding the association between air pollution and respiratory hospital admission. We conclude the thesis in Chapter 4 together with some additional considerations relevant to this work.

Chapter 2

Statistical Methodology

In this chapter, we introduce the concepts, notations and the statistical methods used in this study. The basic concepts and ideas are presented in Sections 2.1 and 2.2. The statistical methods used to analyze the recurrent hospital admission data are described in Sections 2.3 and 2.4.

First, a brief review of the basic concept of classical survival analysis including the Cox proportional hazards model is presented in Section 2.1. Then, we introduce the mathematical details of the *counting process*, *intensity function* and *at-risk process* in Section 2.2. In Section 2.2, we also describe the Poisson process, which is one of the popular approaches of analyzing recurrent event data; the process is developed based on the counting process technique and forms the foundation of many other statistical models. The formulation of the multiplicative models based on the Poisson process is presented in Section 2.3. There, we also describe the likelihood based approach of estimation. In Section 2.4, we describe an extension of the multiplicative model by taking into account between-individual variation.

2.1 Survival Analysis

Survival analysis typically focuses on time to event data. The time variable is commonly referred to as *survival time* or *failure time*, which stands for the time to the occurrence of an event of interest. The survival time is often expressed in terms of years, months, days or age at which the event occurs to an individual. The term event is commonly known as the *survival event* or *failure*, and can be the death of an individual, cancer diagnosis, divorce or birth of a child. In survival analysis, the event of interest is typically a negative individual experience such as death, though the event can also be a positive experience in some situations (e.g., the

recovery from a particular disease after treatment). The classical survival analysis focuses on the time to the occurrence of a single event for each individual (e.g., death) and forms the foundation of many statistical models to analyze recurrent event data.

2.1.1 Survival Function and Hazard Function

The *survival function* and the *hazard function* are the two basic quantitative measures considered in any survival analysis. These two measures provide crucial summary information from survival data.

Suppose that the random variable T denotes the survival time and t denotes any specific value of interest for the random variable T. The survivor function, denoted by $S(t)$, is defined by

$$S(t) = \mathrm{P}(T \geq t) = 1 - F(t), \quad t \geq 0$$

where $F(t)$ is the cumulative distribution function of T. The survival function gives the probability of the survival of an individual longer than time t, and can be expressed in terms of the probability density function (pdf) $f(t)$ as follows:

$$S(t) = 1 - \int_0^t f(s)ds.$$

In contrast to the survival function, the hazard function, denoted by $h(t)$, gives the instantaneous rate per unit time for the event to occur (i.e., the instantaneous failure rate), given that the individual has survived up to time t. Mathematically, the hazard function can be expressed as

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{2.1}$$

where Δt denotes a small interval of time. Note that $h(t)\Delta t$ is the approximate probability of the event occurrence in the short time interval $[t, t + \Delta t)$ given that the individual has survived to time t. The mathematical formulation of the hazard function is equivalent to the formulation of the intensity function 2.7. A detailed description of the intensity function is presented in Section 2.2.1.

It can be shown from (2.1) that the occurrence of an event and survival are related to

each other as follows (see Collett, 2003):

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d\log[S(t)]}{dt} \tag{2.2}$$

from which the survival function can also be expressed as

$$S(t) = \exp\left[-\int_0^t h(s)ds\right] = \exp[-\Lambda(t)] \tag{2.3}$$

where $\Lambda(t)$ is called the *cumulative hazard function*. The cumulative hazard function measures the total amount of risk that has been accumulated up to time t. We use this frequently in plots to determine whether our methods satisfy their underly assumptions. Note that given one of the four functions, the other three are completely determined. For example, given a survival function, one may derive the hazard function and the probability density function using (2.2) and the cumulative hazard function from (2.3).

2.1.2 Censoring

In survival analysis, when one runs a study for a pre-specified length of time, it may happen that the event has not yet occurred for some subjects by the end of that time period (e.g., 5 months or 10 years). This is one of the distinguishing features of survival data: for some individuals the event of interest has occurred and therefore we know the exact survival time, whereas for others it has not occurred, and all we know is that the survival time exceeds the observation time. This phenomenon is called *censoring*. Survival analysis encompasses a wide variety of statistical methods to deal with time-to-event data in the presence of censoring.

In many applications, there exists various causes of censoring and the cause is typically known. Three common causes of censoring (Collett, 2003; Kleinbaum et al., 2012) are described below. (1) In studies with limited resources and/or with time constraints, it is impossible or impractical to wait for the event to occur for all individuals. Thus, for some individuals it is possible that the event has not yet occurred by the end of the study. As an example, in a study for patients with respiratory diseases, a patient may develop asthma after the study period and therefore we are not able to observe the exact survival time; only partial information is available that the patient has not developed asthma by the end of the study. (2) Censoring can also occur when an individual has been *lost to follow-up* during the

study period. For example, in a clinical trial, a patient moves to a different country after being recruited to the study and can not be traced since then. The only available information is up to his/her last clinical visit. (3) When a patient's failure is due to a cause that is known to be unrelated to the event of interest, the survival time is considered as censored. For instance, in a study of investigating the effects of air pollutants on recurrent hospitalization due to respiratory diseases, a patient may die due to a car accident.

Note that the type of censoring described above occurs after the individual has been entered into the study. This type of censoring occurs to the right of the last known survival time, and is therefore referred to as *right censoring* (Andersen et al., 1993; Collett, 2003; Aalen et al., 2008). Note that the right-censored survival time is less than the actual unknown survival time. In our study, the hospital admission time for each patient during January 1, 2005 to December 30, 2011 are of interest. Since we are interested in repeated hospital admissions and there can be more hospital admissions after the end of the study, the last observation of each individual can be regarded as right censored survival time. There are various other types of censoring can occur, with the most common type being right-censoring; readers may refer to Andersen et al. (1993) and Collett (2003) for a comprehensive description of various types censoring.

2.1.3 Cox Proportional Hazards Regression Model

One of the basic goals of survival analysis is to assess the relationship of explanatory variables (commonly called covariates) to survival time (Kleinbaum and Klein, 2012). For example, in a clinic trial of investigating the effectiveness of a treatment for lung cancer, a patient's survival time may not only depend on whether or not the treatment under study is administered to the patient, but also on other variables such as age, sex and smoking history of the patient; all these variables including the treatment can be considered as covariates in this study. Thus, when modeling survival time, it is pivotal to take into account the covariates to investigate their effects on survival time.

The Cox proportional hazards model is a popular regression model to analyze the survival data (Therneau and Grambsch, 2000; Kleinbaum and Klein, 2012). Below, we first describe the mathematical formulation of the Cox model. Then, we elaborate how to interpret the

regression coefficients using the hazard function and describe the meaning of the proportional hazards assumption. Later in this section, we introduce the maximum likelihood estimation of the Cox model.

2.1.4 Mathematical Formulation of the Cox Model

Let x_j be the jth covariate under study, $j = 1, 2, \ldots, p$. Also, let $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ be the $p \times 1$ vector of covariates. Then, the Cox model proportional hazards is defined using the hazard function as follows:

$$h(t) = h_0(t) \exp(\mathbf{x}'\boldsymbol{\beta}) \tag{2.4}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ is the vector of regression coefficients and $h_0(t)$ is an unspecified nonnegative function of time, called the *baseline hazard* function. Note that $h_0(t)$ is considered as the baseline hazard function as the Cox model reduces to $h_0(t)$ when no x's are in the model (i.e., all the x's are equal to zero). One important property of the Cox hazards model is that the baseline hazard function is not specified. Therefore, this model is considered as a *semi-parametric model*.

2.1.5 The Hazard Ratio

The hazard ratio provides an estimate for the effects of each variable adjusted for the other variables in the model (Kleinbaum and Klein, 2012). Generally, the hazard ratio (HR) can be defined as the hazard for one individual divided by the hazard for a different individual. Let $\mathbf{x}$ and $\mathbf{x}^*$ be two fixed covariate vectors for two different individuals. Then the hazard ratio is

$$\begin{aligned} HR &= \frac{h_0(t) \exp(\mathbf{x}'\boldsymbol{\beta})}{h_0(t) \exp(\mathbf{x}^{*\prime}\boldsymbol{\beta})} \\ &= \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{\exp(\mathbf{x}^{*\prime}\boldsymbol{\beta})} \\ &= \exp[(\mathbf{x} - \mathbf{x}^*)'\boldsymbol{\beta}]. \end{aligned} \tag{2.5}$$

The hazard ratio can be used to describe the relative risk of one covariate level to another. For instance, in study of investigating the effectiveness of a treatment for lung cancer, let $x_1 = 1$ stand for a new treatment and $x_1 = 0$ stand for the standard treatment. Also, let

$x_2 = age$. If the estimated regression coefficient for the treatment is $\hat{\beta}_1 = -0.2$, then the hazard ratio for the effects of the treatment adjusted for age is given by

$$\begin{aligned}\widehat{HR} &= \frac{h(t|x_1 = 1, x_2)}{h(t|x_1 = 0, x_2)} \\ &= \frac{\exp[(\hat{\beta}_1 \times 1 + \hat{\beta}_2 \times x_2]}{\exp[(\hat{\beta}_1 \times 0 + \hat{\beta}_2 \times x_2]} \\ &= e^{\hat{\beta}_1} \approx 0.82.\end{aligned}$$

This means the relative risk of death due to lung cancer for patients in the new treatment group is lower than patients in the standard treatment group controlling for age.

A hazard ratio equals to 1 indicates no difference in terms of risks between the two groups. A hazard ratio < 1 implies that the event is less likely to occur in the new treatment group compared to the standard treatment group. A hazard ratio > 1 indicates the event is more likely to occur in the new treatment group than in the standard treatment group. For a continuous variable such as age, the above interpretation applies to a unit difference in age.

2.1.6 The Meaning of the Proportionality Assumption

As shown in equation (2.5), the final expression for the hazard ratio is independent of time t. Thus, when the estimates of $\boldsymbol{\beta}$ are obtained and the values of $\mathbf{x}$ and $\mathbf{x}^*$ are specified, the exponential expression for the estimated hazard ratio is a constant, which is independent of t. If we use $\hat{\alpha}$ to denote this constant, then we can rewrite the hazard ratio as follows:

$$\hat{h}(t, \mathbf{x}) = \hat{\alpha}\hat{h}(t, \mathbf{x}^*).$$

This expression shows that the hazard functions between two individuals are proportional. In fact, the proportional hazards is a key assumption when using the Cox hazards regression model. For this reason, the Cox hazards model is also known as the Cox proportional hazards model. To test the assumption of hazards proportionality, we can use a graphical procedure and a procedure including time-dependent variables in the Cox regression model. In terms of the graphical approach, the log-cumulative hazard plot is the most widely used plot to assess the proportional hazards assumption. Specifically, we can plot the $\ln[-\ln \hat{S}(t)]$ against t over different categorical variables being investigated. Equation (2.3) together with equation (2.4)

gives

$$\begin{aligned}\ln[-\ln \hat{S}(t)] &= \ln \hat{H}(t) \\ &= \ln\left[\int_0^t h_0(s)\exp(\mathbf{x}'\hat{\boldsymbol{\beta}})ds\right] \\ &= \ln[H_0(t)] + \mathbf{x}'\hat{\boldsymbol{\beta}}\end{aligned}$$

where $H_0(t)$ is fixed at time t and $\mathbf{x}'\hat{\boldsymbol{\beta}}$ is independent of time t. When we compare the hazards between two different groups, the parallel curves for two groups demonstrate that the proportional hazards assumption is satisfied; see Collett (2003), Kleinbaum and Klein (2012) for details.

Compared with the graphic approach, adding time-dependent variables into the model to check the proportional hazards assumption is more formal. We can extend the Cox hazards model by including a product (i.e., interaction) term. This term is a function of time t and a time-independent variable of interest. For instance, suppose the proportional hazards assumption is being checked by two treatment groups indicating by x_1. Let $x_1 = 1$ stand for new treatment and $x_1 = 0$ stand for standard treatment. Then the Cox proportional hazards model can be extended by including the product term $x_2 = x_1 t$ which equals to t in new treatment group while equals to zero in standard treatment group. Then, the hazard ratio becomes

$$HR = \frac{h(t|x_1 = 1, x_2 = x_1 t = t)}{h(t|x_1 = 0, x_2 = x_1 t = 0)} = \exp(\beta_1 + \beta_2 t).$$

As can be seen in this expression, if the parameter β_2 for the product term is zero, then the proportional hazard assumption is satisfied. Consequently, the assumption of proportional hazards can be tested by carrying out a statistical test under the null hypothesis $H_0 : \beta_2 = 0$. If one fails to reject the null hypothesis H_0, then the time-dependent variable has no statistical significant effect in the model.

2.1.7 Maximum Likelihood Estimation of the Cox Model

Cox (1972) developed the *partial* likelihood function to estimate the unknown regression coefficients. The likelihood is partial in the sense that it considers probabilities only for those individuals for which complete information about the time to the occurrence of the

event of interest is available, and does not explicitly consider probabilities of those individual who are censored. Below, we describe the maximum likelihood estimation based on the partial likelihood function.

Suppose there are n ordered observed event times $t_{(1)} < t_{(2)} < \ldots < t_{(n)}$, so that $t_{(j)}$ is the jth ordered event time. Suppose there are no ties in the data, which means only one event can occur at each event time. Further reading about handling ties can be found in Collett (2003), Kalbfleisch and Prentice (2002). Let $x_{(j)}$ be the vector of covariates for the individual who has an event at time $t_{(j)}$. Let δ_i be the event indicator which equals to zero if the ith survival time is right-censored, and equals to one otherwise.

The method in Cox (1972) for $\boldsymbol{\beta}$ estimation is to consider the product of conditional probabilities that the individual has an event at some time $t_{(j)}$, conditional on one event occurred at time $t_{(j)}$. By Collett (2003), the expression of this conditional probability is

$$\mathrm{P}(\text{individual with } x_{(j)} \text{ had an event at } t_{(j)} \mid \text{one event occurred at } t_{(j)}).$$

According to the conditional probability theory, the above expression can be rewritten as

$$\frac{\mathrm{P}(\text{individual with } x_{(j)} \text{ had an event at } t_{(j)})}{\mathrm{P}(\text{one event occurred at } t_{(j)})}.$$

Here, events are treated as independent as they come from independent individuals. Then, the denominator part is the sum of the probabilities of having an event at time $t_{(j)}$, over all individuals who are at risk at $t_{(j)}$. Let $R(t_j)$ be the risk set, which includes the individuals who are still alive and uncensored up to time $t_{(j)}$. Then we can rewrite the above expression as follows:

$$\frac{\mathrm{P}(\text{individual with } x_{(j)} \text{ had an event at } t_{(j)})}{\sum_{l \in R(t_j)} \mathrm{P}(\text{individual } l \text{ had an event at } t_{(l)})}. \tag{2.6}$$

As mentioned above, we only consider the situation that no more than two events can occur at each event time. For this reason, the probabilities of having an event at time $t_{(j)}$ can be replaced by the probabilities of having an event in short time intervals $[t, t + \Delta t)$. Dividing both the numerator and denominator by Δt and letting Δt goes to zero, we can obtain the ratio of the corresponding hazards of having an event at time $t_{(j)}$. So that, equation (2.6) can be rewritten as

$$\frac{h_i(t_{(j)})}{\sum_{l \in R(t_j)} h_l(t_{(l)})}.$$

Substituting the hazard function for the Cox model into this equation gives

$$\frac{h_0(t_j)\exp(\mathbf{x}'_{(j)}\boldsymbol{\beta})}{\sum_{l\in R(t_j)} h_0(t_{(l)})\exp(\mathbf{x}'_{(l)}\boldsymbol{\beta})} = \frac{\exp(\mathbf{x}'_{(j)}\boldsymbol{\beta})}{\sum_{l\in R(t_j)}\exp(\mathbf{x}'_{(l)}\boldsymbol{\beta})}.$$

The equality holds by canceling out the baseline hazard function.

For n observed event times, the Cox partial likelihood without ties takes the following form

$$\prod_{i=1}^{n}\left\{\frac{\exp(\mathbf{x}'_i\boldsymbol{\beta})}{\sum_{l\in R(t_i)}\exp(\mathbf{x}'_l\boldsymbol{\beta})}\right\}^{\delta_i}.$$

More detail about the derivation of the Cox partial likelihood can be found in Collett (2003) and Anderson et al. (1993). The maximum likelihood estimation of $\boldsymbol{\beta}$ in the Cox proportional hazard model can be obtained by maximising the log partial likelihood (Cox, 1972; Therneau and Grambsch, 2000)

$$\log L(\beta) = \sum_{i=1}^{n}\delta_i\left\{(\mathbf{x}'_i\boldsymbol{\beta} - \log\sum_{l\in R(t_i)}\exp(\mathbf{x}'_l\boldsymbol{\beta})\right\}.$$

2.2 Models for Recurrent Events

In survival analysis, the time from the beginning of the study to the event occurrence is observed over the study period. An alternative way to formulate a model is by counting the number of event occurrences. In this case, the observations become nonnegative integers. If each individual only has one event (e.g., death), then the number of event occurrence changes from zero to one. If each individual has more than one events (e.g., recurrent hospital admissions due to respiratory diseases), then the number of events is a step function with only one unit jump when an event occurs. To this end, the *counting process* that records the number of event generated by the process can be used in survival analysis. Note that, the partial likelihood method for the Cox proportional hazards model can also be formulated using the counting process technique.

In this thesis, the counting process technique is used to analyze the recurrent hospital admission data and it is assumed that the events occur in continuous time. The consideration of discrete time models can be found in Cook and Lawless (2007). An introduction to the

mathematical details of *counting process* and *intensity function* which are two fundamental concepts throughout this thesis is given in Section 2.2.1.

2.2.1 Notation

A *counting process* $\{N(t), t \geq 0\}$ is a stochastic process that represents the cumulative number of events experienced by an individual over the time interval $[0, t]$. Specifically, suppose there is a recurrent event process starts at $t = 0$, let $0 \leq T_1 < T_2 < \cdots$ be the event times, where T_k is the time when the kth event occurs. Then, the counting process can be written as $N(t) = \sum_{k=1}^{\infty} I(T_k \leq t)$, where $I(A)$ equals to 1 if event A occurs and 0 otherwise. The sample path of $N(t)$ is a nondecreasing and integer valued step function with one unit jump whenever an event occurs. To be more general, for a specific individual, $N(s, t) = N(t) - N(s)$ indicates the number of events occurred over $(s, t]$ where it is assumed that $N(0) = 0$ and $N(t) = N(0, t)$ for $t > 0$. In addition, $N(t)$ is right continuous with $N(t) = N(t^+)$ since its value is updated precisely at event time (Cook and Lawless, 2007). Here, t^- (t^+) represent times that are infinitesimally smaller (larger) than t.

Cook and Lawless (2007) pointed out that models for recurrent events can be generally specified by considering the probability distribution for the number of event occurrences over short intervals $[t, t + \Delta t)$, conditional on the process history before time t. Let $\Delta N(t) = N(t + \Delta t^-) - N(t^-)$ be the number of events occurring over the short time interval $[t, t + \Delta t)$. Let $H(t) = \{N(s) : 0 \leq s < t\}$ be the process *history*. Then, conditional on the process history $H(t)$, the *intensity function* is defined as

$$\lambda(t|H(t)) = \lim_{\Delta t \to 0} \frac{\Pr\{\Delta N(t) = 1|H(t)\}}{\Delta t}, \quad t > 0, \tag{2.7}$$

which gives the instantaneous rate per unit time for the event to occur, given the process history up to time t.

As mentioned in Section 2.1.2, censoring is one of the distinguishing features of survival data. The event indicator δ equals 1 if the event is observed and 0 if the observation is censored. Similarly, in counting process, the *observation* or *at-risk* process can be used to indicate whether the individual is under observation. Suppose that an individual's events are recorded during the period $[\tau_0, \tau]$. The time τ_0 and τ are referred to as a *starting time* and

a *right-censoring time*, respectively. The time τ can be fixed (e.g., an individual is observed for one year) and also can be random (e.g., the subject dropped off the study or moved to another place that one cannot be observed any more). In our study, the study period ranges from January 1, 2005 (i.e., τ_0=0) to December 30, 2011 (i.e., τ=2554 days). For some individuals, the events may have not yet occurred by December 30, 2011, and all we know is that the event times exceed the observation time. This phenomenon is regarded as censoring.

The observation or at-risk process is defined as $Y(t) = I(\tau_0 \leq t \leq \tau)$. To explain, $Y(t) = 1$ indicates an individual is under observation at time t and hence he/she is "at risk" of having observed event at time t, and $Y(t) = 0$ otherwise. It is assumed that $Y(t)$ is a left continuous process with $Y(t) = Y(t^-)$ whose value at time t is known infinitesimally before t. This is because whether the individual is under observation at time t must be known before an event occurs at time t. Further detail about the at-risk process can be found in Cook and Lawless (2007).

By using the at-risk process, we can write the observed part of the counting process as $\bar{N}(t) = \int_0^t Y(u)dN(u)$. The history of the observable process can be written as $\bar{H}(t) = \{\bar{N}(s), Y(s), 0 \leq s < t\}$. Consequently, the intensity function of the observable process takes the following form

$$\bar{\lambda}(t|\bar{H}(t)) = \lim_{\Delta t \to 0} \frac{\mathrm{P}\{\Delta\bar{N}(t) = 1|\bar{H}(t)\}}{\Delta t}.$$

In terms of further developments, it is assumed that the $Y(t)$ is conditionally independent of $\Delta N(t)$, given the process history before t. Then, $\bar{\lambda}(t|\bar{H}(t)) = Y(t)\lambda(t|H(t))$. This expression shows that when $Y(t) = 0$ the individual is not under observation at time t; therefore it is impossible to observe an event at time t. Thus, the intensity of the observable process $\bar{\lambda}(t|\bar{H}(t))$ equals to zero. The censoring mechanism under this assumption is referred to as *conditionally independent* (Cook and Lawless, 2007).

In survival analysis, the observed data consist of (T_i, δ_i), where T_i is the min {event time, censoring time} and δ_i is the event indicator. In the counting process formulation, the pair of variables (T_i, δ_i) is replaced by $(N_i(t), Y_i(t))$, where $N_i(t)$ represents the number of event occurrences in the time interval $[0, t]$ and $Y_i(t)$ indicates whether the individual is under observation at time t. As a special case, the univariate right-censored survival data can be

expressed as

$$N_i(t) = I(\{T_i(t) \leq t, \delta_i = 1\})$$

and

$$Y_i(t) = I(\{T_i(t) \leq t\}).$$

By using the counting process formulation, it is possible to generalize the survival analysis for single event to recurrent event analysis. Consequently, the emphasis changes from modeling the hazard of a survival function to modeling the intensity or rate of a point process (Therneau and Grambsch, 2000). In the following sections, the model and method used for analyzing recurrent event data in this study are introduced.

2.2.2 Poisson Process

The counting process notation introduced in Section 2.2.1 provides a convenient framework in terms of modeling recurrent event data. From a practical perspective, it is convenient to interpret the regression coefficients as relative risks between two levels of covariates. Two commonly used ways in terms of describing and modeling event occurrences are *event counts* and *gaps* or *waiting times* between two successive events. The Poisson process and renewal process are the two canonical frameworks to model the recurrent event data based on event counts and gap times, respectively. Models based on event counts are commonly used when event occurrence rates (e.g., the rate of recurrent hospital admissions due to respiratory diseases) in populations or groups of individuals are of interest (Cook and Lawless, 2007). In this study, the Poisson process is used since it is the most widely used approach to model the recurrent event data based on event counts. An introduction to the mathematical details of the Poisson process is given in this section. Details about the renewal process can be found in Cook and Lawless (2007).

One way to characterize the Poisson process is through the following three postulates:

(i) $N(0) = 0$;

(ii) The process $\{N(t); t \geq 0\}$ has the independent increment property. Specifically, suppose $(a, b]$ and $(c, d]$ are any two non-overlapping time intervals with $0 \leq a < b \leq c < d$, then the random variable $N(a, b)$ and $N(c, d)$ are independent random variables;

(iii) No more than one event can occur in any short time interval.

Following the definition of intensity function in equation (2.7) and the assumption that no more than one event can occur in $[t, t+\Delta t)$ gives

$$\mathrm{P}\{N(t+\Delta t^-) - N(t^-) = 1|H(t)\} = \lambda\{t|H(t)\}\Delta t + o(\Delta t),$$

$$\mathrm{P}\{N(t+\Delta t^-) - N(t^-) = 0|H(t)\} = 1 - \lambda\{t|H(t)\}\Delta t + o(\Delta t),$$

and

$$\mathrm{P}\{N(t) \geq 2|H(t)\} = o(\Delta t),$$

where $\lim_{\Delta t\to 0} \frac{o(\Delta t)}{\Delta t} = 0$. This indicates when Δt approaches to zero, the probability of more than one event occurrence over a short time interval becomes negligible.

From what has been mentioned above, one can see that the Poisson process can be used to describe situations where events occur randomly in such a way that the number of event occurrences in nonoverlapping time intervals are independent. The probability of an event occurrence in the short time interval $[t, t+\Delta t)$ may depend on time t but is independent of the process history $H(t)$. Then the intensity function of a Poisson process without any covariates takes the following form

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

where $\rho(t)$ is the *rate function* which is a nonnegative integrable function and presents the unconditional instantaneous probability of an event occurrence at time t. If the rate function $\rho(t) = \rho$ is a constant, then the process is referred to as *homogeneous Poisson process*; otherwise it is called *nonhomogeneous Poisson process*.

Dewanji and Moolgavkar (2000) proposed a non-homogeneous Poisson process model to analyze the recurrent event data for a given individual during the study period. They illustrated the proposed model by investigating the associations between the recurrent hospitalization due to chronic respiratory diseases in King country and air pollution indices. In our study, we use the non-homogeneous Poisson process model to carry out the analysis of our recurrent hospital admission data. We describe such models in the following section based on the Poisson process technique.

In situations where independent variables or covariates are involved, one needs to consider regression models to account for the effects of the covariates.

2.3 Multiplicative Model for Recurrent Events

As mentioned in Chapter 1, the goal of this study is to investigate the impact of air pollutants on recurrent hospital admissions due to respiratory diseases. To evaluate the extent to which the air pollutants are associated with hospital admissions, we also consider age and sex of the patients, which are not of primary interest. Cook and Lawless (2007) pointed out that one can incorporate covariates by broadening the event history $H(t)$ to include information of fixed or time-dependent covariates, and then let the event intensity function depend on such covariates. Typically, let $\mathbf{z}$ be the fixed covariates, and $\mathbf{z}(t)$ be the time-dependent covariates.

By far, the most commonly used framework to specify covariate effects is through the multiplicative model. Suppose we have a $p \times 1$ covariates vector $\{\mathbf{z}_i(t), t \geq 0\}$ with $\mathbf{z}_i(t) = (z_{i1}(t), z_{i2}(t), \ldots, z_{ip}(t))'$. The *history of covariate* over interval $[0, t]$ can be denoted as $\mathbf{z}^{(t)} = \{z(s) : 0 \leq s \leq t\}$ and the *complete covariate path* is $z^{(\infty)} = \{z(s) : 0 \leq s\}$. The extension of the process history is given by $H(t) = \{N(s) : 0 \leq s < t; z^{(\infty)}\}$, which includes the history of the response process and the covariate process.

Suppose there are m individuals under study and consider a specified process time scale t with a well-defined origin. Let t_{ij} be the jth event time for the ith individual. Also, let the study end at time τ_i for the ith individual, so that we can observe exact event time only up to τ_i. In this respect, τ_i can be considered as the censoring time for the ith individual. For our data, we have the dates of respiratory hospital admissions for each patient, and the study period ranges from January 1, 2005 to December 30, 2011. So, the follow-up for each patient begins at January 1, 2005 (i.e., $\tau_0 = 0$) and ends at December 30, 2011 (i.e., $\tau_i = 2,554$ days for all individuals). The intensity function for individual i can be written as

$$\lambda_i(t|H(t)) = \lambda_0(t) g(\mathbf{z}_i(t)'; \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ denotes the vector of unknown regression parameters and $g(\mathbf{z}_i(t)'; \boldsymbol{\theta})$ is a nonnegative function which specifies the relationship between the covariates and the intensity function. This multiplicative model ensures positive-valued multiplicative effects of $\mathbf{z}_i(t)$ for $\boldsymbol{\theta}$. The positive-valued function $\lambda_0(\cdot)$ denotes the non-parametric *baseline rate* or baseline intensity function.

The model is referred to as fully parametric when $\lambda_0(\cdot)$ is specified parametrically, while it is referred to as semiparametric when $\lambda_0(\cdot)$ is nonparametric. In this thesis, it is assumed that $\lambda_0(\cdot)$ is nonparametric. In addition, we consider an exponential form for $g(\mathbf{z}_i(t)'; \boldsymbol{\theta})$ for which the regression coefficients are easily interpretable using the hazard (or risk) ratio as described in Section 2.1.5. This yields the conditional intensity function as follows:

$$\lambda_i(t|H(t)) = \lambda_0(t) \exp(\mathbf{z}_i(t)'\boldsymbol{\theta}). \tag{2.8}$$

The cumulative intensity function for the ith individual is given by

$$\Lambda_i(t) = \int_0^t \lambda_0(s) \exp(\mathbf{z}_i(s)'\boldsymbol{\theta}) ds,$$

and the baseline cumulative intensity function is given by

$$\Lambda_0(t) = \int_0^t \lambda_0(s) ds.$$

For the Poisson process, the probability of a new event occurrence at time t dose not depend on the process history $H(t)$ up to time t. Thus, the Poisson multiplicative model can be expressed as

$$\lambda_i(t|H(t)) = \rho_i(t) = \rho_0(t) \exp(\mathbf{z}_i'(t)\boldsymbol{\theta}) \tag{2.9}$$

where ρ_0 is called the *baseline rate* or *intensity* when all $z_i(t)$'s are equal to zero.

Model (2.9) can be extended by allowing $\mathbf{z}(t)$ to include components based on arbitrary features of previous event history, such as the time since the most recent event or the number of previous events. In this situation, $\mathbf{z}(t)$ includes components that based on event history and the process is no longer Poisson. To explain, for the Poisson process, the independent increments property implies that the process history at time t has no influence on the intensity function at time t. However, this property will be invalid when incorporating the information of previous history in the intensity function. This process is usually referred to as *modulated Poisson process* and such processes may incorporate dependence on prior event history (Cook and Lawless, 2007).

Note that the expression of (2.8) is identical to the Cox proportional hazards model for univariate survival data. The difference lies in the definition of the event indicator. For recurrent event analysis, an individual with more than one event occurrences may remain

in the risk set with $Y(t) = 1$ until the censoring occurs (i.e., Y(t)=0), whereas for classical survival analysis, an individual is removed from the risk set as soon as the event or censoring occurs.

Given the functional form of the regression model, one needs to estimate the regression parameters. The derivation of the likelihood function by using the product-integration is presented in the following section.

2.3.1 Likelihood Function

Instead of considering a transition from one event time to another event time during the study period, one could consider a transition from one infinitesimal time interval to another infinitesimal time interval with the probability of an event occurrence in the infinitesimal time interval, given the history. Then, the product integration can be used to represent the likelihood function of a process observed over a time interval $[\tau_0, \tau]$ as an infinite product of conditional likelihoods of the process in each infinitesimal time interval, conditional on the previous history. This provides an alternative way of writing down the probability densities which may be easier to interpret (Anderson er al., 1993).

Aalen and Johansen (1978) introduced the product-integral as the canonical transformation from hazard function or intensity function to distribution function. The *product integral* is a generalization of ordinary products. One simple and intuitive way to think the product integration is to conceptualize it as a product of many terms that all or most of those terms are very close to one. Consider the partition of the interval $[a, b]$ as $a = u_0 < u_1 < \cdots < u_R = b$ with $\Delta u_r = u_{r+1} - u_r$, $r = 0, 1, \ldots, R$ and $u_{R+1} = u_R^+$. The product integral of a continuous integrable function $g(u)$ over $[a, b]$ can be defined as

$$\prod_{[a,b]} \{1 + g(u)du\} = \lim_{R\to\infty} \prod_{r=0}^{R} \{1 + g(u_r)\Delta u_r\}. \tag{2.10}$$

The left-hand side of equation (2.10) shows the formula of product integral, while the right-hand side shows how actually this product integral is calculated. When R approaches to infinity, the size of Δu_r terms approaches to zero. According to the Taylor expansion we have $\log\{1 + g(u)\Delta u_r\} = g(u)\Delta u_r + o(\Delta u_r)$. Then, the log of (2.10) approaches to the

Riemann integral $\int_a^b g(u)du$ in the limit, and we have

$$\prod_{[a,b]}\{1+g(u)du\} = \exp\left\{\int_a^b g(u)du\right\}.$$

In order to obtain the full likelihood function based on the probability distribution

$$\mathrm{P}\{N(t+\Delta t^-) - N(t^-) = 0|H(t)\} = 1 - \lambda\{t|H(t)\}\Delta t + o(\Delta t),$$

we also have the following equation

$$\prod_{[a,b]}\{1+g(u)du+o(du)\} = \exp\left\{\int_a^b g(u)du\right\}. \tag{2.11}$$

Further details about the product integration can be found in Cook and Lawless (2007) and Andersen et al. (1993).

Below, we show the derivation of the likelihood function based on the derivation in Cook and Lawless (2007). Let first only consider one event process which is observed over the time interval $[\tau_0, \tau]$, given its history $H(\tau_0)$. Suppose there are n observable event occurrences at times $t_1, \ldots, t_n$. The time interval $[\tau_0, \tau]$ can be equally divided into a number of small time intervals $\tau_0 = u_0 < u_1 < \cdots < u_R = \tau$, each of length $\Delta u_r = u_{r+1} - u_r$. Let $H(t)$ be the history at time t, which provides all the available information to the researcher just before t. In addition, let $\Delta N(u_r)$ be the number of event occurrence in the short interval $[u_r, u_{r+1})$. We can derive the joint probability distribution for all data in $[u_0, u_R)$ by decomposing it into the product of the conditional probability distribution for $\Delta N(u_r)$ over Δu_r, given the previous history. Then we have the following expression

$$\begin{aligned}
\mathrm{P}\{N(u_1), N(u_2), \ldots, N(u_R)|H(u_0)\} &= \Pr\{\Delta N(u_R)|\Delta N(u_1), \Delta N(u_2), \\
&\quad \ldots, \Delta N(u_{R-1}), H(u_0)\} \\
&\quad \times \mathrm{P}\{\Delta N(u_{R-1})|\Delta N(u_1), \Delta N(u_2), \\
&\quad \ldots, \Delta N(u_{R-2}), H(u_0)\} \\
&\quad \vdots \\
&\quad \times \mathrm{P}\{\Delta N(u_0)|H(u_0)\} \\
&= \prod_{r=0}^{R} \mathrm{P}\{\Delta N(u_r)|H(u_r)\}.
\end{aligned} \tag{2.12}$$

According to the definition of intensity function, the probability of a new event occurs within the time interval Δu_r and no event occurs in this interval can be written as follows:

$$\mathrm{P}\{\text{a new event in } [u_r, u_{r+1})|H(u_r)\} = \lambda(u_r|H(u_r))\Delta u_r + o(\Delta u_r),$$

$$\mathrm{P}\{\text{no events in } [u_r, u_{r+1})|H(u_r)\} = 1 - \lambda(u_r|H(u_r))\Delta u_r + o(\Delta u_r).$$

As mentioned in Section 2.2.2, it is assumed that no more than one event can occur over the time interval Δu_r. This gives

$$\Pr\{\Delta N(u_r) \geq 2|H(u_r)\} = o(\Delta u_r).$$

Inserting the above three equations into (2.12), the joint probability distribution can be rewritten as

$$\begin{aligned}\prod_{r=0}^{R} \mathrm{P}\{\Delta N(u_r)|H(u_r)\} &= \prod_{r=0}^{R}\{\lambda(u_r|H(u_r))\Delta u_r + o(\Delta u_r)\}^{\Delta N(u_r)} \\ &\quad \times\{1 - \lambda(u_r|H(u_r))\Delta u_r + o(\Delta u_r)\}^{1-\Delta N(u_r)}.\end{aligned}$$

When R increases, the length of the short interval approaches to zero, and the finite product will approach a product-integral.

The likelihood function is regarded as a function of $\boldsymbol{\theta}$, given the realization of $\Delta N(t)$. Then the likelihood function is based on the probability of observed data which is proportional to the above joint probability distribution function

$$\begin{aligned}L^*(\theta) &\propto \lim_{R\to\infty}\prod_{r=0}^{R}\{\lambda(u_r|H(u_r))\Delta u_r + o(\Delta u_r)\}^{\Delta N(u_r)} \\ &\quad \times\{1 - \lambda(u_r|H(u_r))\Delta u_r + o(\Delta u_r)\}^{1-\Delta N(u_r)}. \qquad (2.13)\end{aligned}$$

When one event happens in the short interval $[u_r, u_{r+1})$, $\Delta N(u_r) = 1$ and $\sum_{r=0}^{R} \Delta N(u_r) = n$. Then, the first part in (2.13) turns on and becomes a contribution component in the likelihood function. In fact, the first part in (2.13) is the joint probability distribution of one event occurrence in $[u_r, u_{r+1})$. Since there are n event occurrences in the study period $[\tau_0, \tau]$, there will only be n intervals that include the event times $t_1, \ldots, t_n$; for all other time intervals we have $\Delta N(u_r) = 0$. When $\Delta N(u_r) = 0$, the second part in (2.13) turns on and also makes contribution to the likelihood function. Since the counting process will have a finite number

of event occurrences, the product integral of the first part in the above expression is just an ordinary finite product. The values of the exponent part (i.e., $1 - \Delta N(u_r)$) in the seconde part are 1 when there are no event occurrences. Thus, the second part can be considered as a product of many terms that are close to one. In addition, the exponent can be omitted without altering the value of the product integral (Aalen, 2008). Dividing the first part in (2.13) by $\prod_{r=0}^{R}(\Delta u_r)^{\Delta N(u_r)}$, one can obtain the following intensity function

$$\begin{aligned}\lim_{R\to\infty}\left[\frac{\lambda(u_r|H(u_r))\Delta u_r + o(\Delta u_r)}{\Delta u_r}\right] &= \lim_{\Delta u_r\to 0}\left[\frac{\Pr\{\Delta N(u_r)=1|H(u_r)\}}{\Delta u_r}\right] \\ &= \prod_{j=1}^{n}\lambda(t_j|H(t_j)).\end{aligned}$$

Dividing (2.13) by $\prod_{r=0}^{R}(\Delta u_r)^{\Delta N(u_r)}$, inserting $g(u) = -\lambda(u|H(u))$ into equation (2.11) and letting R approach to infinity we can obtain the following likelihood function for n observed events at times $t_j (j = 1, 2, \ldots, n)$

$$L^*(\theta) = \prod_{j=1}^{n}\lambda(t_j|H(t_j)) \cdot \exp\left\{-\int_{\tau_0}^{\tau}\lambda(s|H(s))ds\right\}.$$

When it comes to a group of m independent individual processes and each individual has $n_i \geq 0$ observed events at time $t_{ij} (j = 1, 2, \ldots n_i)$, the likelihood becomes

$$L(\theta) = \prod_{i=1}^{m} L_i^* = \prod_{i=1}^{m}\left[\prod_{j=1}^{n_i}\lambda_i(t_{ij}|H(t_{ij})) \cdot \exp\left\{-\int_{\tau_0}^{\tau_i}\lambda_i(s|H(s))ds\right\}\right]. \tag{2.14}$$

More detail of the above derivation can be found in Andersen et al. (1993), Fleming & Harrington (1991), Cook and Lawless (2007) and Aalen et al. (2008).

An alternative way to write down the likelihood function is based on the observation or at-risk process $\{Y(t), t \geq 0\}$ (Cook and Lawless, 2007). The definition of the observation or at-risk process $\{Y(t), t \geq 0\}$ has been introduced in Section 2.2.1. Under the conditionally independent censoring mechanism, the intensity of the observable process is $\bar{\lambda}(t|\bar{H}(t)) = Y(t)\lambda(t|H(t))$. Then the likelihood for the observable data can be written as

$$L^* = \prod_{j=1}^{n}\lambda(t_j|H(t_j)) \cdot \exp\left\{-\int_{0}^{\infty}Y(s)\lambda(s|H(s))ds\right\}.$$

Accordingly, equation (2.14) can be rewritten as

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{m} L_i^* \\ &= \prod_{i=1}^{m} \left[\prod_{j=1}^{n_i} \lambda_i(t_{ij}|H(t_{ij})) \cdot \exp\left\{ -\int_0^{\infty} Y_i(s)\lambda(s|H(s))ds \right\} \right] \\ &= \left\{ \prod_{i=1}^{m} \prod_{j=1}^{n_i} \lambda_i(t_{ij}|H(t_{ij})) \right\} \cdot \exp\left\{ -\int_0^{\infty} \sum_{\ell=1}^{m} Y_\ell(s)\lambda_\ell(s|H(s))ds \right\} \end{aligned} \tag{2.15}$$

When the baseline intensity function $\lambda_0(t)$ is assumed to be nonparametric, the standard likelihood method cannot be used to estimate $\boldsymbol{\theta}$. Cox (1975) proposed partial likelihood based inference to estimate models with high-dimensional nuisance parameters, after he used the same idea to deal with proportional hazards regression model. The likelihood is partial since it only considers probabilities of those individuals whose complete information about the time to the event occurrence is available, and does not consider probabilities of those individuals who are censored.

2.3.2 Estimation for the Semiparametric Regression Model

In this section, we first introduce how to estimate the cumulative baseline intensity (i.e., $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$) from the standpoint of the martingale theory. The estimation of cumulative baseline intensity is useful to estimate the regression parameters $\boldsymbol{\theta}$. In addition, it is important in terms of obtaining the estimation of the variance for random effects; see Section 2.4.2 for detail. Then, we describe the estimation procedure for the semiparametric regression model (2.9) in which the baseline intensity is assumed to have no particular parametric form.

By Therneau et al. (2000), the *counting process martingale* is the differences between the observed counting processes and expected number of events for the ith individual by time t, which takes the following expression

$$M_i(t) = \bar{N}_i(t) - \int_0^t Y_i(s)\lambda_0(s)\exp\{\mathbf{z}_i'(t)\boldsymbol{\theta}\}ds. \tag{2.16}$$

According to the *Doob-Meyer decomposition* theorem, we can decompose the counting process as the sum of a martingale and a *compensator*. The *compensator* is a right-continuous

process with value zero at time zero. In terms of fitting a model to data, the previous decomposition is analogous to the decomposition: observed counting process = estimated compensator + martingale residual process. Then we have

$$\hat{M}_i(t) = \bar{N}_i(t) - \int_0^t Y_i(s) e^{\mathbf{z}_i'(t)\hat{\boldsymbol{\theta}}} d\hat{\Lambda}_0(s) \tag{2.17}$$

where the expectation value of $M_i(t)$ is zero. Then the estimate of cumulative baseline intensity is

$$\hat{\Lambda}_0(t) = \int_0^t \frac{d\bar{N}.(s)}{\sum_{i=1}^m Y_i(s) \exp(\mathbf{z}_i'(s)\hat{\boldsymbol{\theta}})} \tag{2.18}$$

where $d\bar{N}.(s) = \sum_{i=1}^m Y_i(s) dN_i(s)$ is the total number of observed events over short interval $[s, s+ds)$. Further details can be found in Anderson et al. (1993) and Therneau et al. (2000). The above baseline intensity estimator is useful for estimating the regression parameters later.

The logarithm of (2.15) gives the following log likelihood function

$$\ell(\theta) = \sum_{i=1}^m \left[\sum_{j=1}^{n_i} \log \lambda_i(t_{ij}|H(t_{ij})) - \int_0^\infty Y_i(s) \lambda(s|H(s)) ds \right] \tag{2.19}$$

where $dN(t) = N(t) - N(t^-)$ denotes the number of event occurrence over $[t, t+\Delta t)$. Note that $N(t)$ is a step function with one unit jump at event times. $Y(t) = I(\tau_0 \leq t < \tau)$ equals to 1 when the individual is under observation and 0 otherwise. Then, the sum for the ith individual at observed event times $t_{i1}, \ldots, t_{in_i}$ can be expressed as (Cook and Lawless, 2007):

$$\sum_{j=1}^{n_i} \log \lambda_i(t_{ij}|H(t_{ij})) = \int_0^\infty Y_i(s) \log \lambda_i(s|H(s)) dN_i(s).$$

Thus, we can rewrite the equation (2.19) as

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^m \left[\int_0^\infty Y_i(s) \log \lambda_i(s|H(s)) dN_i(s) - \int_0^\infty Y_i(s) \lambda(s|H(s)) ds \right] \\ &= \sum_{i=1}^m \left[\int_0^\infty Y_i(s) \log \left\{ \lambda_0(s) \exp(z_i'(s)\beta) \right\} dN_i(s) \right. \\ &\quad \left. - \int_0^\infty Y_i(s) \lambda_0(s) \exp(z_i'(s)\beta) ds \right]. \end{aligned} \tag{2.20}$$

Differentiating $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ gives the score function for $\boldsymbol{\theta}$

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \sum_{i=1}^{m}\left[\int_0^{\infty} Y_i(s)\mathbf{z}_i(s)dN_i(s) - \int_0^{\infty} Y_i(s)\lambda_0(s)\exp(\mathbf{z}_i'(s)\boldsymbol{\theta})\mathbf{z}_i(s)ds\right] \\
&= \sum_{i=1}^{m}\left[\int_0^{\infty} Y_i(s)\mathbf{z}_i(s)dN_i(s) - \int_0^{\infty} Y_i(s)\exp(\mathbf{z}_i'(s)\boldsymbol{\theta})\mathbf{z}_i(s)d\Lambda_0(s)\right] \\
&= \sum_{i=1}^{m}\int_0^{\infty} Y_i(s)\mathbf{z}_i(s)\left[dN_i(s) - \frac{d\bar{N}.(s)}{\sum_{l=1}^{m} Y_l(s)\exp(\mathbf{z}_l(s)\theta)}\exp(\mathbf{z}_i'(s)\boldsymbol{\theta})\right]. \quad (2.21)
\end{aligned}$$

The third line is obtained by inserting equation (2.18) into the second line.

By replacing the $d\bar{N}.(s)$ by $\sum_{i=1}^{m} Y_i(s)dN_i(s)$ we can rewrite the score function (2.21) as

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{m}\int_0^{\infty} Y_i(s)W_i(s;\boldsymbol{\theta})dN_i(s),$$

where

$$W_i(s;\boldsymbol{\theta}) = \mathbf{z}_i(t) - \frac{\sum_{l=1}^{m} Y_l(s)\exp(\mathbf{z}_l(s)\boldsymbol{\theta})\mathbf{z}_l'(t)}{\sum_{l=1}^{m} Y_l(s)\exp(\mathbf{z}_l(t)\boldsymbol{\theta})}. \quad (2.22)$$

Anderson et al. (1993) and Cook and Lawless (2007) showed that the covariance matrix estimates can base on $I_{\theta\theta}(\boldsymbol{\theta}) = E\Big[\{\partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}\{\partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}'\Big]$. Then,

$$\begin{aligned}
I_{\theta\theta}(\boldsymbol{\theta}) &= E\left\{\sum_{i=1}^{m}\int_0^{\infty} Y_i(s)W_i(s;\boldsymbol{\theta})dN_i(s) \times \sum_{i=1}^{m}\int_0^{\infty} Y_i(s)W_i'(s;\boldsymbol{\theta})dN_i(s)\right\} \\
&= \sum_{i=1}^{m} cov\left\{\int_0^{\infty} Y_i(s)W_i(s;\boldsymbol{\theta})dN_i(s),\ \int_0^{\infty} Y_i(s)W_i'(s;\boldsymbol{\theta})dN_i(s)\right\} \\
&= \sum_{i=1}^{m}\int_0^{\infty}\int_0^{\infty} Y_i(s)Y_i(t)W_i(s;\boldsymbol{\theta})W_i'(t;\boldsymbol{\theta})cov\{dN_i(s), dN_i(t)\}.
\end{aligned}$$

The variance estimates for $\hat{\boldsymbol{\theta}}$ can be taken from $I_{\theta\theta}^{-1}(\boldsymbol{\theta})$. In the following sections, a more recent development that use random effect or frailty within the general framework of the statistical models based on counting process is introduced.

2.4 Frailty Model

In studies of recurrent events, it is common to encounter the diversity in observed data. As shown in the multiplicative model (2.9), the regression parameters $\boldsymbol{\theta}$ are assumed to be common for all individuals. This means all individuals in the study have the same relationship

between observable covariates and recurrent rate. However, in survival analysis some patients may experience their relapse more quickly than other patients because they have a genetic disposition to develop a disease. Such unobservable covariate effects can lead to heterogeneity across individuals which means the relationship between observable covariates and recurrent rate is different from one individual to another. Thus, an individual-specific random intercept u_i is included in the model to show an individual's deviation from the population average, after the effects of the observable covariates have been accounted for.

The mixed Poisson process in which the intensity function for the recurrent event partly depends on unobservable random effects is considered in this thesis. Nielsen et al. (1992) and Anderson et al. (1993) mentioned that if the outcome event of the counting process is failure or death, the underlying random variable can be considered as a *frailty*. The term frailty indicates that some individuals are more or less likely to experience the event of interest than others.

There are several other ways to incorporate the frailty in the multiplicative model (Therneau and Grambsch, 2000; Cook and Lawless, 2007; Duchateau and Janssen, 2008). The shared frailty model is one useful model in terms of reflecting the heterogeneity across individuals caused by some unobservable covariates which has multiplicative effects on the intensity function (Andersen et al., 1993). Repeated events within individual i share the same frailty u_i, which is why the model is called shared frailty model. Suppose for each individual there exists an unobservable positive-valued subject-specific frailty u_i. Given u_i, the intensity function becomes

$$\begin{aligned} \lambda_i(t|H_i(t);u_i) &= \lim_{\Delta t\to 0}\frac{\Pr\{\Delta N(t)=1|H(t),u_i\}}{\Delta t} \\ &= u_i\lambda_0(t)\exp(\mathbf{z}_i'(t)\boldsymbol{\theta}) \end{aligned} \tag{2.23}$$

where $u_1, u_2, \ldots, u_m$ are independent and identical distributed. In fact, the most widely assumed distribution for frailties is the gamma distribution (Klein, 1992; Andersen et al., 1993; Cook and Lawless, 2007) with mean 1 and variance ϕ. Then the density function for u_i is

$$f_U(u) = \frac{u^{1/\phi-1}\exp(-u/\phi)}{\phi^{1/\phi}\Gamma(1/\phi)}. \tag{2.24}$$

And the cumulative intensity function becomes

$$\Lambda_i(t) = \int_0^t u_i(s)\lambda_0(s)\exp(\mathbf{z}_i'(s)\boldsymbol{\theta})ds. \tag{2.25}$$

It can be seen from (2.23) that the process is still a Poisson process with rate $u_i\lambda_i(t)$, conditional on u_i . When integrating out the frailties u_i, the process is no longer a Poisson process; see the next section for detail.

2.4.1 Likelihood Function

Taking into account the individual-specific shared gamma frailty in the multiplicative model, the complete or full data likelihood function for the ith individual can be obtained from the joint density of $\mathbf{z}$ and $\mathbf{u}$ as follows:

$$L_c(\boldsymbol{\theta}, \phi) = \prod_{i=1}^{m} L_i(\lambda_0(\cdot)\,,\,\boldsymbol{\theta}\,,\,u_i) \tag{2.26}$$

where

$$\begin{aligned} L_i(\lambda_0(\cdot)\,,\,\boldsymbol{\theta}\,,\,u_i) &= \prod_{j=1}^{n_i}\{u_i\lambda_0(t_{ij})\exp(\mathbf{z}_i'(t_{ij})\boldsymbol{\theta})\} \\ &\quad \exp\left(-\int_0^{\tau_i} Y_i(s)u_i\lambda_0(s)\exp(\mathbf{z}_i'(s)\boldsymbol{\theta})ds\right) \\ &\quad \times\frac{u_i^{1/\phi-1}\exp(-u_i/\phi)}{\phi^{1/\phi}\Gamma(1/\phi)}. \end{aligned} \tag{2.27}$$

Since the random effects u_i are unobserved, integrating out u_i from (2.27) yields the following observed data likelihood

$$L_m(\boldsymbol{\theta}, \phi) = \prod_{i=1}^{m} L_{marg,i}\left(\lambda_0(\cdot)\,,\,\boldsymbol{\theta}\,,\,\phi\right)$$

where

$$\begin{aligned} L_{marg,i}\left(\lambda_0(\cdot)\,,\,\boldsymbol{\theta}\,,\,\phi\right) &= \int_0^{\infty}\left[\prod_{j=1}^{n_i}\{u_i\lambda_0(t_{ij})\exp(\mathbf{z}_i'(t_{ij})\boldsymbol{\theta})\}\right. \\ &\quad \exp\left\{-\int_0^{\tau_i} Y_i(s)u_i\lambda_0(s)\exp(\mathbf{z}_i'(s)\boldsymbol{\theta})ds\right\} \\ &\quad \left.\times\frac{u_i^{1/\phi-1}\exp(-u_i/\phi)}{\phi^{1/\phi}\Gamma(1/\phi)}\right] d(u_i). \end{aligned} \tag{2.28}$$

Below, we show the derivation of the intensity function after integrating out u_i and the derivation of the closed form of the above marginal likelihood. Suppose we only consider one individual, then the probability of one event occurrence that does not depend on u_i is

$$\begin{aligned}
\mathrm{P}\{\Delta N(t)=1|H(t)\} &= \int_0^\infty \mathrm{P}\{\Delta N(t)=1|H(t),u\}du \\
&= \frac{\int_0^\infty \mathrm{P}\{\Delta N(t)=1|H(t),u\}\mathrm{P}\{H(t)|u\}f(u)du}{\int_0^\infty \mathrm{P}\{H(t)|u\}f(u)du} \\
&= \{\lambda_0(t)\exp(\mathbf{z}_i'(t)\boldsymbol{\theta})\Delta t + o(\Delta t)\}\frac{\int_0^\infty u\mathrm{P}\{H(t),u\}du}{\int_0^\infty \mathrm{P}\{H(t)\}du} \\
&= \{\lambda_0(t)\exp(\mathbf{z}_i'(t)\boldsymbol{\theta})\Delta t + o(\Delta t)\}\int_0^\infty u\mathrm{P}\{u|H(t)\}du \\
&= \{\lambda_0(t)\exp(\mathbf{z}_i'(t)\boldsymbol{\theta})\Delta t + o(\Delta t)\}E\{u|H(t)\}.
\end{aligned}$$

From the second line to the third line we use the equation (2.23). Dividing the last line of the above equation by Δt and letting $\Delta t \to 0$, we obtain the following intensity function:

$$\lambda_i(t|H_i(t)) = \lambda_0(t)\exp(\mathbf{z}_i'(t)\boldsymbol{\theta})E\{u_i|H(t)\}. \tag{2.29}$$

Next, it is necessary to derive the conditional probability distribution of the random effects given history $H(t)$. By Duchateau and Janssen (2008), we can derive the marginal likelihood and the conditional distribution of u_i as follows. After some simplifications, the marginal likelihood (2.28) can be expressed as

$$\begin{aligned}
L_{marg,i}\left(\lambda_0(\cdot)\,,\,\theta\,,\,\phi\right) &= \frac{\prod_{j=1}^{n_i}\lambda_0(t_{ij})\exp(\mathbf{z}_i'(t_{ij})\boldsymbol{\theta})}{\phi^{1/\phi}\,\Gamma(1/\phi)} \\
&\quad \int_0^\infty \left[(u_i)^{n_i}\exp\left\{-u_i\sum_{j=1}^{n_i}\int_0^{\tau_i} Y_i(s)u_i\lambda_0(s)\exp(\mathbf{z}_i'(s)\boldsymbol{\theta})ds\right\}\right. \\
&\quad \left.(u_i)^{(1/\phi)-1}\exp(-u_i/\phi)\right]du_i.
\end{aligned} \tag{2.30}$$

To obtain a closed expression for this integral, we let

$$A_i(\tau_i) = \sum_{j=1}^{n_i}\int_0^{\tau_i} Y_i(s)\exp(\mathbf{z}_i'(s)\boldsymbol{\theta})d\Lambda_0(s)$$

and $r = 1/\phi + A_i$. Substituting A_i and r into equation (2.30), the marginal likelihood function

becomes

$$\begin{aligned} L_{marg,i} &= \frac{\prod_{j=1}^{n_i} \lambda_0(t_{ij}) \exp(\mathbf{z}_i'(t_{ij})\boldsymbol{\theta})}{\phi^{1/\phi}\,\Gamma(1/\phi)} \\ &\quad \int_0^\infty \left[(u_i)^{n_i} \exp\left\{-u_i(r-1/\phi)\right\} (u_i)^{(1/\phi)-1} \exp(-u_i/\phi) \right] du_i \\ &= \frac{\prod_{j=1}^{n_i} \lambda_0(t_{ij}) \exp(\mathbf{z}_i'(t_{ij})\boldsymbol{\theta})}{r^{(n_i+1/\phi)}\phi^{1/\phi}\,\Gamma(1/\phi)} \int_0^\infty (ru_i)^{(n_i+1/\phi)-1} \exp(-ru_i) d(ru_i). \end{aligned}$$

Using the fact that $\Gamma(\alpha) = \int_0^\infty x^{(\alpha-1)} e^{-x} dx$ with $\alpha > 0$ gives the following expression of the marginal likelihood function:

$$L_{marg,i}\left(\lambda_0(\cdot)\,,\,\theta\,,\,\phi\right) = \frac{\prod_{j=1}^{n_i} \lambda_0(t_{ij}) \exp(\mathbf{z}_i'(t_{ij})\boldsymbol{\theta})\,\Gamma(n_i+1/\phi)}{\Big(1/\phi + A_i(\tau_i)\Big)^{(n_i+1/\phi)} \phi^{1/\phi}\,\Gamma(1/\phi)} \tag{2.31}$$

where n_i is the total number of observed events for the ith individual during the study period $[0, \tau_i)$. According to the Bayes theorem, we can obtain the following conditional distribution for u_i given history

$$\begin{aligned} f(u_i \,|\, \mathbf{z}) &= \frac{L_i(\lambda_0(\cdot)\,,\,\phi \mid u_i)\, f_U(u_i)}{L_{marg,i}\left(\lambda_0(\cdot)\,,\,\theta\,,\,\phi\right)} \\ &= \frac{u_i^{n_i+1/\phi-1} \exp\{-u_i(1/\phi + A_i(\tau_i)\}(1/\phi + A_i(\tau_i))^{n_i+1/\phi}}{\Gamma(n_i+1/\phi)}. \end{aligned}$$

This expression corresponds to a gamma distribution with shape $(n_i + 1/\phi)$ and scale $(1/\phi + A_i)$. Then,

$$E(u_i | H_i(\tau_i); \theta) = \frac{1+\phi n_i}{1+\phi A_i(\tau_i)}.$$

Under independent censoring, the intensity function that does not depend on u_i takes the following form

$$\lambda_i(t|H_i(t)) = \left\{ \frac{1+\phi N_i(t-)}{1+\phi A_i(t-)} \right\} \lambda_0(t) \exp(\mathbf{z}_i'(t)\boldsymbol{\theta}). \tag{2.32}$$

where $A_i(t) = A_i(\lambda_0(\cdot)\,,\,\theta) = \sum_{j=1}^{n_i} \int_0^t Y_i(s) \exp(\mathbf{z}_i'(s)\boldsymbol{\theta}) d\Lambda_0(s)$.

As shown in equation (2.32), when $\phi = 0$, the process is a Poisson process and there is no unobservable heterogeneity across individuals. However, when $\phi > 0$ the intensity function at any time t depends both on ϕ and on the process history before t and therefore the process is no longer a Poisson process. The likelihood ratio test of $H_0 : \phi = 0$ and $H_1 : \phi > 0$ can be conducted to check whether there is heterogeneity across individuals or not. The likelihood

ratio test statistic is the twice the difference between the log-likelihood of the full model and the reduced model according to $H_0 : \phi = 0$. The parameter ϕ is nonnegative and on the boundary of the parameter space under the null hypothesis $\phi = 0$. Thus, the asymptotic null distribution of the likelihood ratio statistic is the mixture distribution of 50% point mass at zero and 50% χ^2_1 (the chi-squared distribution with 1 degree of freedom), that is, $0.5\chi^2_0 + 0.5\chi^2_1$ (Cook and Lawless, 2007). This nonstandard limiting distribution can be used when the parameter of interest is on the boundary of the parameter space under the null hypothesis (Self and Liang, 1987).

It can be seen from the equation (2.32) that the intensity function depends on ϕ and the number of previous events. When the number of previous events $N_i(t-)$ increases, the intensity at time t increases as well. This may be due to the reason that individuals who have many events before time t are more likely to have a new event than others in a process beyond time t, which means they may have a higher recurrent rate.

In the following section, penalized partial likelihood approach used for estimating regression parameters (i.e., $\boldsymbol{\theta}$) and the variance of random effects (i.e., ϕ) is described.

2.4.2 Penalized Partial Likelihood

When including the gamma shared frailties in the model, the ordinary maximum likelihood estimation method cannot be used due to the incomplete information about the unobservable frailties. The Expectation-Maximisation (EM) algorithm (Dempster et al., 1977) is an algorithm that typically used to solve problems with unobserved information. Gill (1985) suggested that the EM algorithm can be used to do the maximum likelihood estimation for the semiparametric gamma frailty model. Klein (1992) and Nielsen st al. (1992) further discussed this method in terms of fitting semiparametric frailty models for clustered survival data. The EM algorithm depends on the marginal likelihood after integrating out the gamma random effects from the complete likelihood. This method iterates between an expectation and maximization step. In the expectation step, the expected values of the unobserved frailties given the observed information and the current parameter estimates are obtained. In the maximisation step, new estimates of the parameters can be obtained by treating the expected values as fixed values or *offset* (Dempster et al., 1977; Klein, 1992; Nielsen et al., 1992). The

main drawback of the EM algorithm is time-consuming in some cases. Furthermore, the EM algorithm dose not automatically provide an estimate of the covariance matrix of the parameter estimates. This can be a drawback when these estimates are desired.

Consequently, an alternative method called penalized likelihood approach is used to estimate the unknown parameters and the variance of frailty. This approach is based on the partial likelihood and a penalty term. The partial likelihood can be used to estimate the regression parameters, while the penalty term is used to obtain the estimation of frailties and to avoid large differences between the frailties for different individuals. Both the regression parameters and the frailties are included in the penalized likelihood function. Then the maximum likelihood method can be used to do the estimation.

The penalized likelihood approach was introduced by Goodd and Gaskins (1971) in the context of nonparametric probability density estimation. McGilchrist and Aisbett (1991) and McGilchrist (1993) used this method in Cox regression model estimation by assuming that frailties follow a log-normal distribution. More detail about the penalized partial likelihood can be found in Therneau and Grambsch (2000), Therneau et al. (2003) and Duchateau and Janssen (2008). Therneau et al. (2003) developed the penalized likelihood approach in the context of the shared frailty model. The penalized partial likelihood formulation of the frailty model can be easily developed by reparameterizing $u_i = \exp(b_i)$. In order to distinguish between u_i and b_i, we will call u_i the frailty and b_i the random effect. Then, the alternative representation of the random effects model takes the following expression

$$\lambda_i(t|H(t)) = \lambda_0(t)\exp(\mathbf{z}_i'(t)\boldsymbol{\theta} + b_i) \tag{2.33}$$

which is equivalent to function (2.23). This model includes the shared gamma frailty model as a special case as described in Anderson et al. (1993). In addition, according to the density function for the u_i's in (2.24), the density function for the B_i's can be expressed as

$$f_B(b) = \frac{\exp(b)^{\phi^{-1}}\exp(-\exp(b)/\phi)}{\phi^{\phi^{-1}}\Gamma(\phi^{-1})},$$

and the expectation of b_i is zero. According to equation (2.27), the logarithm of the complete likelihood function takes the following form when $u_i = \exp(b_i)$:

$$\begin{aligned}\ell_{full}(\lambda_0(\cdot)\,,\,\boldsymbol{\theta}\,,\,\phi) &= \log f(\mathbf{z},\mathbf{b}\,|\,\lambda_0(\cdot)\,,\,\boldsymbol{\theta}\,,\,\phi)\\ &= \log f(\mathbf{z}\,|\,\lambda_0(\cdot)\,,\,\boldsymbol{\theta}\,,\,\mathbf{b}) + \log f(\mathbf{z}\,|\,\phi)\end{aligned}$$

The first part is the conditional log likelihood of the data given the random effects, whereas the second part is the logarithm of the distribution of the random effects. In this approach, the second term of the likelihood is referred to as a penalty term.

Following the method proposed by Therneau et al. (2003), the estimation of the involving parameters in model (2.33) can be conducted by maximizing the penalized partial log-likelihood (ppl)

$$\ell_{ppl}(\phi,\, \boldsymbol{\theta},\, \mathbf{b}) = \ell_{part}(\boldsymbol{\theta},\, \mathbf{b}) - \ell_{pen}(\phi,\, \mathbf{b}) \tag{2.34}$$

over both parameters $\boldsymbol{\theta}$ and random effects $\mathbf{b}$. In this equation, $\ell_{part}(\boldsymbol{\theta},\, \mathbf{u})$ is the log of usual Cox partial likelihood (Cox 1972) that expressed in counting process notation, conditional on $\mathbf{b}_i$ are fixed. The expression of the partial likelihood is

$$L_{part} = \prod_{i=1}^{m}\prod_{j=1}^{n_i}\left\{\frac{\exp(\mathbf{z}_i'(t_{ij})\boldsymbol{\theta} + b_i)}{\sum_{\ell=1}^{m} Y_\ell(t_{\ell j})\exp(\mathbf{z}_\ell'(t_{\ell j})\boldsymbol{\theta}) + b_\ell)}\right\}.$$

Then the log of the above equation is

$$\ell_{part}(\theta,\, \mathbf{b}) = \sum_{i=1}^{m}\int_0^\infty Y_i(s)\left[(\mathbf{z}_i'(s)\boldsymbol{\theta} + b_i) - \log\left\{\sum_{\ell=1}^{m} Y_\ell(s)\exp(\mathbf{z}_\ell'(s)\boldsymbol{\theta} + b_\ell)\right\}\right] dN_i(s). \tag{2.35}$$

We can cancel out the baseline intensity since it appears in both numerator and denominator in the partial likelihood. The second part in (2.34) is referred to as a penalty term which is used to avoid large differences between $\mathbf{b}_i$ for different individuals and takes the following expression

$$\ell_{pen}(\phi,\, \mathbf{b}) = -\frac{1}{\phi}\sum_{i=1}^{m}(b_i - \exp(b_i)).$$

Usually, choosing the penalty function to "shrink" b_i towards its mean value (i.e., zero) is of interest. Specifically, when using the maximum likelihood method, we need to maximize $\ell_{part}(\theta,\, \mathbf{b})$ and minimize $\ell_{pen}(\phi,\, \mathbf{b})$ to maximize the log penalized partial likelihood. When b_i approaches zero, the absolute value of $\frac{1}{\phi}\sum_{i=1}^{m}(b_i - \exp(b_i))$ at this point will have a small negative contribution to the penalized partial likelihood.

The variance of the random effects ϕ can be considered as a nuisance parameter which is used to control the amount of shrinkage. In the coxph function, the variance ϕ can be chosen based on the profile likelihood. The estimation procedure would be to follow McGilchrist and Aisbett (1991), who derived the penalized partial likelihood approach from logarithm

of complete data likelihood (2.26) for the frailty model with normally distributed random effects.

Suppose that ϕ is known, we can obtain the estimates for $\boldsymbol{\theta}$ and $\mathbf{b}$ via solving the following score functions based on the first partial derivatives of equation (2.34). Since $\boldsymbol{\theta}$ is not included in the penalty function, the estimating equations for $\boldsymbol{\theta}$ are identical to those for an usual Cox model with the $\log(b)$'s treated as fixed offset terms; see Therneau and Grambsch (2000) for further details. Thus, the score function deriving from function (2.35) takes the following form

$$\frac{\partial \ell_{part}}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{m} \int_0^{\infty} Y_i(s) \left[z_i(s) - \frac{\sum_{\ell=1}^{m} Y_\ell(s) \exp(b_\ell + \mathbf{z}'_\ell(s)\boldsymbol{\theta}) z_\ell(s)}{\sum_{\ell=1}^{m} Y_\ell(s) \exp(b_\ell + \mathbf{z}'_\ell(s)\boldsymbol{\theta})} \right] dN_i(s). \tag{2.36}$$

In terms of random effects $\mathbf{b}$, the score function is

$$\begin{aligned} \frac{\partial \ell_{ppl}}{\partial b_i} &= \frac{\partial \ell_{part}}{\partial b_i} - \frac{\partial \ell_{pen}}{\partial b_i} \\ &= \sum_{i=1}^{m} \int_0^{\infty} \left[Y_i(s) - \frac{\sum_{i=1}^{m} Y_i(s) \exp(b_i + \mathbf{z}'_i(s)\boldsymbol{\theta})}{\sum_{\ell=1}^{m} Y_\ell(s) \exp(b_\ell + \mathbf{z}'_\ell(s)\boldsymbol{\theta})} \right] dN_i(s) \\ &\quad -\frac{1}{\phi} \sum_{i=1}^{m} (1 - \exp(b_i)) \end{aligned} \tag{2.37}$$

Inserting equation (2.18) into (2.37), the simplified score function for b_i takes the following form

$$\begin{aligned} \frac{\partial \ell_{ppl}}{\partial b_i} &= \sum_{i=1}^{m} \left[\int_0^{\infty} Y_i(s) dN_i(s) - \int_0^{\infty} Y_i(s) \exp(\mathbf{z}'_i(s)\boldsymbol{\theta}) e^{b_i} d\Lambda_0(s) \right] - \frac{1}{\phi} \sum_{i=1}^{m} (1 - \exp(b_i)) \\ &= \sum_{i=1}^{m} \left[n_i - \widehat{A}_i e^{b_i} \right] - \frac{1}{\phi} \sum_{i=1}^{m} (1 - \exp(b_i)) \end{aligned} \tag{2.38}$$

where n_i is the total number of observed events for the ith individual.

By Therneau and Grambsch (2000), the estimation of the variance parameter ϕ for random effects can be carried out by maximizing a profile marginal likelihood function $\ell^{(l)}_{marg,i}$ in the (ℓ)th iteration. Taking the logarithm of the marginal likelihood (2.31) and summing

over m individuals gives the following marginal log-likelihood

$$\begin{aligned}\ell_{marg}\left(\lambda_0(\cdot)\,,\,\boldsymbol{\theta}\,,\,\phi\right) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\log\lambda_0(t_{ij})+\mathbf{z}_i'(t_{ij})\boldsymbol{\theta}) \\ &\quad+\sum_{i=1}^{m}\Bigg[\log\left(\frac{\Gamma(n_i+1/\phi)}{\Gamma(1/\phi)}\right)+n_i\log\phi \\ &\quad-(n_i+1/\phi)\log\left(1+\phi A_i\right)\Bigg]\end{aligned}$$

The profiled marginal likelihood is obtained as follows. For a particular $\phi^{(l)}$, the estimates $\widehat{\boldsymbol{\theta}}_{\phi^{(l)}}$ and $\widehat{\mathbf{u}}_{\phi^{(l)}}$ can be obtained by solving score functions (2.36) and (2.38). Equation (2.18) together with equation (2.25) gives the following estimates of cumulative baseline intensity

$$\hat{\Lambda}_0(t)=\int_0^t\frac{d\bar{N}_{\cdot}(s)}{\sum_{i=1}^{m}u_iY_i(s)\exp(\mathbf{z}_i'(t)\hat{\boldsymbol{\theta}})}=\int_0^t\hat{\lambda}_0(s)ds,$$

where $d\bar{N}_{\cdot}(s)=\sum_{i=1}^{m}Y_i(s)dN_i(s)$ is the total number of observed events over short interval $[s,s+ds)$. Then, by replacing $\mathbf{u}$ and $\widehat{\boldsymbol{\theta}}$ by $\widehat{\mathbf{u}}_{\phi^{(l)}}$ and $\widehat{\boldsymbol{\theta}}_{\phi^{(l)}}$ in the above equation, we can obtain the estimates for the baseline intensity function and cumulative baseline intensity function. This leads to the following expression

$$\begin{aligned}\ell_{marg}^{(l)} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\log\widehat{\lambda}_{0,\phi^{(l)}}(t_{ij})+\mathbf{z}_i'(t_{ij})\widehat{\boldsymbol{\theta}}_{\phi^{(l)}}) \\ &\quad+\sum_{i=1}^{m}\Bigg[\log\left(\frac{\Gamma(n_i+1/\phi^{(l)})}{\Gamma(1/\phi^{(l)})}\right)+n_i\log\phi^{(l)} \\ &\quad-(n_i+1/\phi^{(l)})\log\left(1+\phi^{(l)}\widehat{A}_{i,\phi^{(l)}}\right)\Bigg]. \qquad (2.39)\end{aligned}$$

In summary, the maximisation of penalized partial log-likelihood consists of a doubly iterative process which alternates between an inner loop and an outer loop. In step 1, an initial value of the variance for random effect ϕ is guessed. In step 2, we solve score functions (2.36) and (2.37) by using Newton-Raphson procedure for fixed ϕ. In step 3, we insert the current values of $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{u}}$ into (2.39) to find a new value for ϕ. Then step 2 and step 3 are iterated until convergence, which means the difference in estimates between two successive iterations falls below a desired tolerance (Therneau and Grambsch, 2000 and Duchateau and Janssen, 2008).

Chapter 3

Data Analysis

The aim of this chapter is to present the statistical analysis of the recurrent hospital admission data. In Section 3.1, we describe the respiratory hospitalization data and air pollutants data. In Section 3.2, we present the results for the hospital admission data, and summary of our findings regarding the association between air pollution and respiratory hospital admission.

3.1 Data

The data under consideration consist of two parts: (1) hospital admission due to respiratory diseases and (2) air pollutants. Each part will be described in the following two subsections.

3.1.1 Study population and Hospital Admission

The hospital admission data for patients were obtained from the Canadian Institute for Health information (CIHI), Discharge Abstract Database (DAD, CIHI 2011). The study period is from January 1, 2005 to December 30, 2011. The data set consists of all patients age 40 years and older with primary diagnoses of respiratory diseases (ICD-9 codes 460-519, ICD-10-CA codes J00-J99). The ICD stands for the International Classification of Diseases which is an international standard for reporting clinical diagnoses and health management developed by the World Health Organization (WHO, 2013). The ICD-9 is the 9th revision. The ICD-10-CA is an enhanced version of ICD-10 developed by CIHI for morbidity classification in Canada. Except the ICD codes, the data set also includes the date of admission, age, sex and admit category for each patient. In addition, each patient has a unique patient ID which is

used to identify readmission due to respiratory diseases for the same patient over the study period. In this study, the hospital admission time is recorded in calendar day.

As mentioned in Chapter 1, the emphasis throughout this study is to investigate the effects of air pollutants on recurrent hospital admissions due to respiratory diseases in two major cities (Saskatoon and Regina) of Saskatchewan. Thus, we only use the hospital admission data from the hospitals governed by Regina Qu'Appelle Regional Health Authority and Saskatoon Regional Health Authority , Saskatchewan, Canada.

3.1.2 Air Pollution and Weather data

The subset of air pollutants that can cause smog and acid rain are sometimes referred to as the Criteria Air Contaminants (Environment Canada, 2013). The Criteria Air Contaminants are commonly used in studies for investigating the effects of air pollutants on respiratory hospitalization (Braun et al., 1992; Roemer et al., 1993; Luginaah et al., 2005; Fung et al., 2006). Thus, the gaseous pollutants included within this study are carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2) and ozone (O_3). And the particulate matter is $PM_{2.5}$ (tiny particles in the air that are ≤ 2.5 microns in width). The data of PM_{10} (tiny particles in the air that are ≤ 10 microns in width) is not available in Saskatoon from 2006 to 2011. For this reason, PM_{10} is not included into the analysis. Temperature and relative humidity are two weather variables included into the analysis in order to take into account the seasonal effects. Daily average air pollutants and weather data are used for this study. Daily mean gaseous pollutants and particulate matters concentrations in Regina and Saskatoon monitoring stations were obtained from the National Air Pollution Surveillance Program (NAPS) from Environment Canada. Weather data include daily average temperature, and relative humidity were obtained from the Environment Canada (2013).

In the dataset, some information are missing due to power failure or other unavoidable reasons. Missing data can be replaced by data from other available data set. For instance, the hourly air pollutants' concentrations in Regina and Saskatoon can be obtained from the Saskatchewan Ministry of Environment. Taking the average of 24 hours recordings give the daily average values of air pollutants and weather variables. If there is no longer available data from other data set, missing data can be replaced by the mean of nearby 6 points: three

days earlier and three days after. In this study, the percentage of missing values for all air pollutants is small (1.54% for Regina and 1.46% for Saskatoon).

In order to carry out statistical analyses, the first thing is to link the respiratory hospital admission data to the air pollutants and weather data. Then, the regression analysis includes the following covariates:

- x_{i1}: Patient's age (≥ 40) in years ;
- x_{i2}: Patient's sex (0=Female,1=Male);
- x_{i3}, x_{i4}, x_{i5}, x_{i6} and x_{i7} : Average daily gaseous pollutants (CO, NO_2, SO_2, O_3) and particulate matters ($PM_{2.5}$) concentrations, respectively;
- x_{i8} and x_{i9}: Average daily temperature and relative humidity, respectively.

The patient ID is also included in the analysis in order to count the number of recurrent events for each patient; see Section 3.1.1 for detail.

3.2 Statistical Analysis

In this section, we first analyze the recurrent hospital admission data for all respiratory diseases. Secondly, we only consider asthma, which is one of the most common respiratory diseases in Canada.

For comparison, three different models are fitted. Model 1 is a Poisson process model (Anderson-Gill Model) with intensity

$$\lambda_i(t|H(t)) = \lambda_0(t)\exp(\mathbf{x}_i(t)'\boldsymbol{\beta}).$$

The vector of covariates is $\mathbf{x}_i(t) = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}, x_{i7}, x_{i8}, x_{i9})'$. Model 2 includes a time-dependent variable $N_i(t-)$ that indicates the number of previous events. Then the intensity takes the following expression

$$\begin{aligned}\lambda_i(t|H(t)) &= \lambda_0(t)\exp(\mathbf{x}_i(t)'\boldsymbol{\beta} + \gamma N_i(t-)) \\ &= \lambda_0(t)\exp(\mathbf{z}_i(t)'\boldsymbol{\theta}).\end{aligned}$$

The vector of covariates is $\mathbf{z}_i(t) = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}, x_{i7}, x_{i8}, x_{i9}, N_i(t-))'$ and the vector of regression coefficients is $\boldsymbol{\theta} = (\boldsymbol{\beta}', \gamma)'$. Model 3 is a Poisson model with shared gamma distributed random effects. This model is referred to as mixed Poisson model as it includes both random terms (e.g., u_i) and fixed parameters (e.g., β) (Cook and Lawless, 2007). The intensity function is

$$\lambda_i(t|H(t); u_i) = u_i \lambda_0(t) \exp(\mathbf{x}_i(t)'\boldsymbol{\beta}).$$

The vector of covariates is the same as that in Model 1.

For Model 1 and Model 2, the maximum likelihood estimation can be conducted to estimate the regression coefficients; see Section 2.3.2 for detail. For Model 3, the penalized partial likelihood method can be used to estimate the regression coefficients and the variance of random effects ϕ; see Section 2.4.2 for detail.

When we fit the statistical model to the available recurrent event data, it is critical to assess how well the model fits the data. The statistical tests based on the martingale residuals have been suggested for checking the adequacy of multiplicative models for recurrent event data (e.g., Cook and Lawless, 2007; Lin et al., 2000). The correlation between repeated events within the same individuals lead to additional technical challenges when checking the goodness of fit for recurrent event data. In this thesis, we did not pursue the formal test for the goodness of fit for recurrent event data.

3.2.1 Statistical Analysis for All Respiratory Diseases

Summary statistics of daily average concentrations of air pollutants and weather variables during the study period is provided in Table 3.1. Those values are the mean of the monitoring stations in Regina and Saskatoon.

During the study period from January 1, 2005 to December 30, 2011, a total of 20,284 patients (Male n=10,643, Female n=9,641) age 40 years and older admitted into the hospitals governed by Regina Qu'Appelle Reginal Health Authority and Saskatoon Reginal Health Authority with primary diagnoses of respiratory disease (ICD-9 codes 460-519, ICD-10-CA codes J00-J99). Table 3.2 shows the summary of hospital admission data. The total number of hospital admissions was 51,008, including 30,744 readmissions. Most patients (75.88%)

Table 3.1: Summary statistics of the daily mean concentrations of air pollutants and weather variables, Saskatoon and Regina, January 1, 2005 to December 30, 2011

Variables (units)	Minimum	Maximum	Mean	SD	SE
Air pollutants					
CO(ppm)	0.000	1.300	0.306	0.162	0.003
NO_2(ppb)	2.000	40.000	11.114	5.34	0.106
O_3(ppb)	2.000	52.000	21.301	7.712	0.153
SO_2 (ppb)	0.000	5.000	0.627	0.618	0.012
$PM_{2.5}(\mu g/m^3)$	0.000	73.000	5.206	3.381	0.067
Weather					
Temperature(°C)	-34.400	27.400	2.638	13.371	0.265
Relative humidity(%)	30.000	98.000	73.474	12.18	0.241

had no readmission, while only 0.46% patients had more than 10 admissions. According to Table 3.2, there are 13 patients have no hospital admission during the study period and their " at-risk" indicator equal to zero.

Table 3.3 displays the results of fitting three models. Note that Model 1 is nested within Model 2. Thus, a likelihood ratio test can be used to check which model is better. The test statistic is twice the difference between the log likelihood values of Model 1 and Model 2. The distribution of this test statistic under the null hypothesis is approximately a Chi-squared distribution. The degrees of freedom is the difference between the number of free parameters of two models. When comparing Model 1 and Model 2, the degree of freedom is 1.

The **coxph** function in R provides null log likelihood and fitted log likelihood values. The first one is the log likelihood under the null hypothesis that $H_0 : \beta = 0$, while the second one is the log likelihood after plugging in the estimated coefficients in the log likelihood function (`http://cran.r-project.org/web/packages/survival/survival.pdf`). Thus, the fitted log likelihood values can be used to carry out a likelihood ratio test. A likelihood ratio test of Model 1 versus Model 2 is -2(-297717.8+298550.4)=1665.2 on one degree of freedom and p-value < 0.0001. Nielsen et al., (1992) showed that the Chi-squared approximation with one degree of freedom is valid for models based on counting process technique. The result

Table 3.2: Number of admissions due to all respiratory diseases (ICD1-10-CA codes J00-J99) in Regina Qu'Appelle Reginal Health Authority and Saskatoon Reginal Health Authority, January 1, 2005 to December 30, 2011 (2554 days)

Number of admissions	Number of individuals	percent
0	13	0.06
1	15392	75.88
2	2824	13.92
3	918	4.53
4	474	2.34
5	238	1.17
6	120	0.59
7	80	0.39
8	63	0.31
9	37	0.18
10	23	0.11
> 10	102	0.46

suggested that the introduction to the number of previous events ($N_i(t-)$) can improve the fit significantly. In addition, the significant effect of ($N_i(t-)$) implies the dependence on the previous event occurrences. It suggests that the same baseline intensities for different recurrence times is not appropriate (Lim et al., 2007). For this reason, Model 1 which assumed common baseline intensities for different recurrent times might not be appropriate for the hospital admissions due to all respiratory diseases. In Model 2, for the ith individual, $N_i(t_{ij}-)$ denotes the number of events have occurred before the jth event. We can rewrite Model 2 as follows:

$$\begin{aligned} \lambda_i(t|H(t)) &= \lambda_0(t)\exp(\mathbf{x}_i(t)'\boldsymbol{\beta} + \gamma N_i(t-)) \\ &= \exp(\gamma N_i(t-))\lambda_0(t)\exp(\mathbf{x}_i(t)'\boldsymbol{\beta}). \end{aligned}$$

Then different recurrent times have different baseline intensity for the same individual. Thus, Model 2 is preferred over Model 1.

In Model 3, the point estimate for the variance of the random effects is $\hat{\phi} = 0.158$,

Table 3.3: Results for recurrent hospital admission due to all respiratory diseases

	Model 1		Model 2		Model 3	
Covariate	EST.	P	EST.	p	EST.	p
age	-0.0018	0.0000	-0.0031	0.0000	-0.0035	0.0000
sex	0.0331	0.0038	0.0275	0.0162	0.0360	0.0058
Event Counts						
$(N(t-))$	-	-	0.1424	0.0000	-	-
CO	0.0022	0.0000	0.0022	0.0000	0.0024	0.0000
NO_2	-0.0191	0.0000	-0.0186	0.0000	-0.0214	0.0000
O_3	0.0249	0.0000	0.0265	0.0000	0.0297	0.0000
SO_2	-0.0843	0.0000	-0.0754	0.0000	-0.0776	0.0000
$PM_{2.5}$	-0.0104	0.0000	-0.0084	0.0000	-0.0073	0.0007
Temp	0.0079	0.0000	0.0085	0.0000	0.0081	0.0000
RH	-2.645	0.0000	-2.719	0.0000	-3.0121	0.0000
Variance (ϕ)	-		-		0.158	
Fitted Loglik	-298550.4		-297717.8		-293989.5	
I-likelihood	-		-		-297995.7	

1. Abbreviations: Temp: Temperature; RH: Relative humidity; EST: Estimated parameters; P: P value.
2. Null log likelihood= -304896.3.
3. I-likelihood is the log partial-likelihood with the frailty terms integrated out.
4. Model 1: $\lambda_i(t|H(t)) = \lambda_0(t)\exp(\mathbf{x}_i(t)'\boldsymbol{\beta})$.
5. Model 2: $\lambda_i(t|H(t)) = \lambda_0(t)\exp(\mathbf{x}_i(t)'\boldsymbol{\beta} + \gamma N_i(t-))$.
6. Model 3: $\lambda_i(t|H(t);u_i) = u_i\lambda_0(t)\exp(\mathbf{x}_i(t)'\boldsymbol{\beta})$.
7. Round to 4 decimal places.

suggesting there is heterogeneity across individuals. Since model 1 is nested within Model 3, a likelihood ratio test can be used to test whether random effects are needed. The null hypothesis is $H_0 : \phi = 0$, while the alternative hypothesis is $H_1 : \phi \neq 0$. The test statistic is twice the difference between the I-likelihood of Model 3 and the fitted log likelihood of Model 1 (see Therneau and Grambsch, 2000). The I-likelihood is the log partial likelihood after integrating out the random effects. As shown in (2.31), ϕ is the only different parameter between log marginal likelihood from Model 3 and the fitted log likelihood from Model 1 after integrating out random effects. For this reason, the degree of freedom of the Chi-squared distribution is 1. Since $H_0 : \phi = 0$ is on the boundary to the parameter space, so that the distribution on the likelihood ratio test is a 50:50 mixture of a point mass at zero and a χ^2 distribution (Self and Liang, 1987); see Section 2.4 for detail. Then, the likelihood ratio test of Model 1 versus Model 3 is -2(-297995.7+293989.5)=8012.4 on one degree of freedom with p-values < 0.0001. This result also indicates the presence of heterogeneity across individuals. Section 1.2.2 and Section 2.4 introduced that incorporating the random effect in the model is useful to reflect the heterogeneity across individuals due to some unobservable fixed covariates. Thus, Model 3 is preferred over Model 1.

Now let's compare Model 2 and Model 3. In Model 2, the time-dependent variable $N_i(t-)$ is also referred to as a dynamic covariate (Aalen et al., 2004) which represents how the past developments influence the present and the future in the counting process (Aalen et al., 2008). Other examples of dynamic covariate for recurrent events data can be found in Aalen et al. (2008). When incorporating dynamic covariates in Anderson-Gill model, one should beware of following problems. Firstly, the values of some time-independent covariates may be underestimated. The time-independent covariates $\mathbf{x}$ may affect the occurrence of one hospital admission ($dN(t)$). It will obviously also affect the cumulative number of hospital admissions ($N_i(t-)$). Thus, the full effects of $\mathbf{x}$ on $dN(t)$ can be considered as a "sum" of the direct effect on $dN(t)$ and the indirect effect through $N_i(t-)$. This results in a well-organized problem when including dynamic covariates in the model that some of the covariates may be underestimated; see Kalbfleisch and Prentice (2002), and Aalen et al. (2008) for detail. This phenomenon can also be seen in Table 3.3 that the parameter estimate in Model 2 is smaller than those in Model 1. Another challenge of using dynamic covariates is the difficulties in

interpretation. Without sufficient information regarding the processes, it is hard to tell the dynamic effects may represent real effect of the past event occurrences or may represent the unobservable heterogeneity across individuals (Aalen et al., 2008). As far as we know, there is no formal statistical test can be used to compare Model 2 and Model 3. However, the log likelihood value for Model 3 is the biggest one. This suggests that Model 3 is preferred over Model 2. As mentioned earlier in this section that $\hat{\phi} \neq 0$. This means there is variability between individuals. Consequently, the random effects should be used to account for the heterogeneity across individuals. According to the above discussion, Model 3 is better.

Table 3.4: Results from fitting the Model 3 for recurrent hospital admission due to all respiratory diseases.

Variables	EST.	p-value	exp($coef$)	lower 95%	upper 95%
age	-0.0035	0.0000	0.9965	0.9957	0.9973
sex	0.0360	0.0058	1.0366	1.0105	1.0635
Event Counts					
CO	0.0024	0.0000	1.0024	1.0024	1.0025
NO_2	-0.0214	0.0000	0.9788	0.9757	0.9820
O_3	0.0297	0.0000	1.0302	1.0281	1.0322
SO_2	-0.0776	0.0000	0.9254	0.9020	0.9493
$PM_{2.5}$	-0.0073	0.0007	0.9928	0.9886	0.9969
Temp	0.0081	0.0000	1.0081	1.0065	1.0097
RH	-3.0121	0.0000	0.0492	0.0434	0.0558

1. Abbreviations: Temp: Temperature; RH: Relative humidity; EST: Estimated parameters; P: P value.
2. Null log likelihood= -304896.3.
3. I-likelihood is the log partial-likelihood with the frailty terms integrated out. I-likelihood=-297995.7.
4. Fitted log likeihood=-293989.5.
5. Model: $\lambda_i(t|H(t);U_i) = u_i\lambda_0(t)\exp(\mathbf{x}_i(t)'\boldsymbol{\beta})$.
6. Variance of random effect= 0.158.
7. Round to 4 decimal places.

As introduced in Section 2.1.3, the hazard ratio can be used to describe the relative risk

of one covariate level to another. Similarly, in recurrent event analysis, the relative risks (RR) can be used to describe the relative risks of experiencing another event of interest with two different levels of a covariate. Table 3.4 displays the results by applying Model 3. The estimated regression coefficients in Model 3 provide the relative risk for an individual, conditional on the frailty u_i that influence the baseline intensities. In addition, the regression coefficients should be interpreted with other covariates held fixed. The results show that male patients may have a slightly higher risk than female patients ($\widehat{RR}$=1.0366, 95% CI $(1.0105, 1.0635)$). Furthermore, for one year older, the readmission rate decreases moderately by 0.35% ($\widehat{RR}$=0.9965, 95% CI $(0.9975, 0.9973)$). The results in Table 3.4 indicate that all air pollutants have statistically significant effects, but only CO and O_3 can statistically increase the risk of hospital readmission due to all respiratory diseases. For CO, when controlling other covariates, the relative risk of hospital readmission increases by 0.2% for one unit increase in CO concentration, given the random effects ($\widehat{RR}$=1.0024, 95% CI $(1.0024, 1.0025)$). For O_3, the relative risk of hospital readmission increases by 3.0% for one unit increase in O_3 concentration ($\widehat{RR}$=1.0302, 95% CI $(1.0281, 1.0322)$). One thing worth noting is that the point estimates of all air pollutants are close to one and the 95% confidence interval is pretty narrow. This may be due to following reasons. Firstly, it is a large number of patients involved in the study. Secondly, there are some limitations in current work. Specifically, the lag effects of air pollutants are ignored, the potential measurement errors are ignored and a common distribution is assumed for all random effects; see Chapter 4 for detail.

In conclusion, the results show that CO and O_3 significant effects on recurrent hospital admission due to all respiratory diseases among patients who were admitted into the hospitals governed by Regina Qu'Appelle Regional Health Authority and Saskatoon Regional Health Authority, Saskatchewan.

3.2.2 Statistical Analysis for Asthma

Asthma is a serious chronic lung disease that caused by the inflammation of the airways to the lungs. It can cause shortness of breath, chest tightness, coughing and wheezing (Statistics Canada, 2013). In Canada, asthma is one of the most prevalent chronic respiratory diseases and is a leading cause of hospital admissions (Asthma Society of Canada, 2013). Thus, in

this section, we investigate the effects of air pollutants on recurrent hospital admissions due to asthma.

The study period ranges from January 1, 2005 to December 30, 2011, a total of 478 patients (Male n=308, Female n=170) age 40 years and older admitted into the hospitals governed by Regina Qu'Appelle Reginal Health Authority and Saskatoon Reginal Health Authority with primary diagnoses of respiratory disease (ICD-9 codes 493, ICD-10-CA codes J45). Table 3.5 shows the summary of hospital admission data. The total number of hospital admissions was 1,086, including 609 readmissions. Most patients (85.15%) had only one hospital admission.

Table 3.5: Number of admissions due to asthma (ICD1-10-CA codes J45) in Regina Qu'Appelle Reginal Health Authority and Saskatoon Reginal Health Authority, January 1, 2005 to December 30, 2011

Number of admissions	Number of individuals	percent
0	1	0.21
1	407	85.15
2	46	9.62
3	12	2.51
4	4	0.84
>4	8	1.67

Table 3.6 displays the results of fitting three models to asthma data. In Model 2, the number of previous events ($N_i(t-)$) (p-value=0.1677) has no significant effects on recurrent hospital admissions due to asthma. In addition, a likelihood ratio test of Model 1 versus Model 2 is -2(-3605.062.8+3604.019)=2.086 on one degree of freedom and p-value is 0.1487. This result also indicates that ($N_i(t-)$) has no significant effects. Thus, Model 1 is better than Model 2.

In Model 3, the estimated variance of the random effects is only $\hat{\phi}$=5e-07, suggesting there may be no individual to individual variability that cannot be explained by fixed covariates. The likelihood ratio test of Model 1 versus Model 3 under the null hypothesis $H_0 : \phi = 0$ is -2(-3605.1+3605.062)=0.076 on 1 degree of freedom with p-values 0.7828. This result also

indicates a little need to model the heterogeneity across individuals. According to what have been mentioned above, Model 1 is better than Model 3.

Table 3.7 displays the results by applying Model 1. The results show that the gender of patients may have no significant effects of hospital admission due to asthma ($\widehat{RR}$=1.0225, 95% CI $(0.8612, 1.2141)$). Furthermore, for one year older, the readmission rate decreases moderately by 1.2% (RR=0.9874, 95% CI $(0.9816, 0.9933)$). According to Table 3.7, CO has a significant effect on hospital admissions ($\widehat{RR}$=1.0024, 95% CI $(1.0019, 1.0028)$). This means when controlling other covariates, the relative risk of hospital readmission due to asthma increases by 0.2% when CO concentration increase by one unit. The findings also show significant effect of O_3 on the recurrent respiratory hospital admissions. The relative risk of hospital readmission increases by 2.3% for an one-unit increase in O_3 concentration ($\widehat{RR}$=1.0232, 95% CI $(1.0108, 1.0359)$). The estimated coefficients of NO_2, SO_2 and fine particulate matters $PM_{2.5}$ are not statistically significant, which means they may not have significant effects on hospital admission based on asthma data set.

In conclusion, the results have illustrated that CO and O_3 have significant effects on recurrent hospital admission due to asthma among patients who were admitted into the hospitals governed by Regina Qu'Appelle Regional Health Authority and Saskatoon Regional Health Authority, Saskatchewan.

As shown in Table 3.5, only 8 patients have more than 4 hospital admissions. It is only 1.67% of the total population and may be treated as outliers. The results by ignoring these 8 patients could be different from what we had obtained using the whole population.

Table 3.6: Results for recurrent hospital admission due to asthma

	Model 1		Model 2		Model 3	
Covariate	EST.	P	EST.	p	EST.	p
age	-0.0127	0.0000	-0.0125	0.0000	-0.0127	0.0000
sex	0.0223	0.7993	0.0251	0.7744	0.0223	0.8000
Event Counts						
$(N(t-))$	-	-	-0.0831	0.1678	-	-
CO	0.0024	0.0000	0.0024	0.0000	0.0024	0.0000
NO_2	-0.0038	0.7187	-0.0020	0.8490	-0.0038	0.7200
O_3	0.0230	0.0002	0.0219	0.0005	0.0230	0.0002
SO_2	0.1460	0.0630	0.1388	0.0771	0.1460	0.0630
$PM_{2.5}$	-0.0072	0.6381	-0.0077	0.6175	-0.0072	0.6400
Temp	0.0150	0.0092	0.0152	0.0083	0.0150	0.0092
RH	-2.8961	0.0000	-2.8481	0.0000	-2.8962	0.0000
Variance (ϕ)	-		-		5e-07	
Fitted Loglik	-3605.062		-3604.019		-3605.062	
I-likelihood	-		-		-3604	

1. Abbreviations: Temp: Temperature; RH: Relative humidity; EST: Estimated parameters; P: P value.
2. Null log likelihood= -3757.113.
3. I-likelihood is the log partial-likelihood with the frailty terms integrated out.
4. Model 1: $\lambda_i(t|H(t)) = \lambda_0(t)\exp(\mathbf{x}_i(t)'\boldsymbol{\beta})$.
5. Model 2: $\lambda_i(t|H(t)) = \lambda_0(t)\exp(\mathbf{x}_i(t)'\boldsymbol{\beta} + \gamma N_i(t-))$.
6. Model 3: $\lambda_i(t|H(t); u_i) = u_i\lambda_0(t)\exp(\mathbf{x}_i(t)'\boldsymbol{\beta})$.
7. Round to 4 decimal places.

Table 3.7: Results from fitting the Model 1 for recurrent hospital admission due to asthma.

Variables	EST.	p-value	$\exp(coef)$	lower 95%	upper 95%
age	-0.0127	0.0000	0.9874	0.9816	0.9933
sex	0.0223	0.7994	1.0225	0.8612	1.2141
CO	0.0024	0.0000	1.0024	1.0019	1.0028
NO_2	-0.0038	0.7187	0.9962	0.9759	1.0170
O_3	0.0230	0.0002	1.0232	1.0101	1.0359
SO_2	0.1460	0.0630	1.1572	0.9921	1.3498
$PM_{2.5}$	-0.0072	0.6381	0.9928	0.9633	1.0232
Temp	0.0150	0.0092	1.0151	1.0037	1.0267
RH	-2.896	0.0000	0.0552	0.0235	0.1301

1. Abbreviations: Temp: Temperature; RH: Relative humidity; EST: Estimated parameters; P: P value.
2. Null log likelihood= -3757.113.
3. Fitted log likeihood=-3605.062.
4. Model: $\lambda_i(t|H(t);U_i) = u_i\lambda_0(t)\exp(\mathbf{x}_i(t)'\boldsymbol{\beta})$.
5. Round to 4 decimal places.

Chapter 4

Concluding Remarks and Future Work

In contemporary society, adverse influence of poor air quality on public health has been an increasingly disturbing issue. Asthma, allergies, lung cancer and some other respiratory diseases have been linked to poor environmental quality (Health Canada, 2006).

Many researchers have reported that the air pollutants have adverse influence on public health in some Canadian cities such as Windsor, Vancouver and northern Alberta (Luginaah et al., 2005; Fung et al., 2006; Villeneuve, 2007). However, no such research has been done for the Province of Saskatchewan, Canada. The primary purpose of this study is to investigate the effects of air pollutants on recurrent hospital admissions due to respiratory diseases in the two major cities of Saskatchewan, namely, Regina and Saskatoon.

In this study, we use intensity-based model to analyze the recurrent hospital admission data (Fleming et al., 1991; Andersen et al., 1993; Therneau and Grambsch, 2000; Cook and Lawless, 2007). In addition, the subject-specific random effects are included in the model to account for the heterogeneity across individuals (Therneau and Grambsch, 2000; Cook and Lawless, 2007; Duchateau and Janssen, 2008). The penalized partial likelihood method (Therneau and Grambsch, 2000) is used to estimate the unknown parameters and the variance of the random effects.

Our analysis indicates that CO and O_3 have significant effects on recurrent hospital admission due to respiratory diseases in Regina and Saskatoon. In practice, there is a tendency of strong correlation between CO and other pollutants (Burnett et al., 1999), thereby resulting in difficulties in assessing the effects of CO independently. As a result, the extent of the association between CO and respiratory hospitalization differs from one study to another. For instance, Cho et al. (2000) found significant association between CO and respiratory hospitalization in Korea controlling for temperature and seasonal effects. In contrast, Lugi-

naah et al. (2005) found no significant effects of CO on respiratory hospitalization for women 65 years old above. For O_3, the results are comparable with those of Burnett et al. (1997) who reported results of the effects of air pollutants on respiratory hospital admissions for 16 Canadian cities.

There is a scope of further research to investigate some related points to this research.

1. The lag effects of air pollutants on recurrent hospital admission due to respiratory diseases are not considered. In this study, we only consider daily average of air pollutant levels, and model the time to occurrence of an event (i.e., date of an admission) as a function of the pollutant levels measured on the same date. Instead of having immediate effects on the hospital admission due to respiratory diseases, the air pollutants might have significant lag effects (e.g., 1-day or 3-day or event 7-dat lag) on respiratory hospital admission. For example, Burnett et al. (1997) investigated the effects of ozone on hospitalization due to respiratory diseases in 16 Canadian cities. They reported that the concentration of ozone measured one day prior to the admission have stronger association with the number of respiratory hospitalizations than that for the concentration of ozone measured on the day of admission or two days before admission. Thus, it might be worthwhile to investigate the lag pollutant effects on respiratory hospital admissions in Saskatchewan.

2. The potential measurement errors in the air pollution and weather data are ignored in this study. The air pollutant data are obtained from two fixed-monitors in Regina and Saskatoon. We use these data to represent the same exposure for all patients, even though they came from different monitoring stations. So, further investigation is necessary to deal with this problem for our data Goldberg et al., (2001) discussed this problem in their environmental exposure study.

3. In this study, we assume a common distribution for all the random effects. As shown in Tables 3.2 and 3.5, about 76% of the individuals had only one hospital admission, whereas about 24% had two or more hospital admissions. So, it would be more logical to treat the individuals with two or more hospital admissions differently from those with only one admission (Xu et al., 2012). With this respect, individuals could be

considered to come from different subgroups, having different distributions of the random effects. The misspecified distribution of random effects might have an influence on statistical inference. No statistical analysis has been conducted so far to analyze hospital admission data by addressing this issue. So, it might be worthwhile to carry out further investigation in analyzing this type of data by modeling the random effects using a mixture distribution.

References

[1] O.O. Aalen. *Statistical inference for a family of counting processes*. PhD thesis, Univerisy of California,Berkeley, 1976.

[2] O.O. Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4):701–726, 1978b.

[3] O.O. Aalen. Heterogeneity in survival analysis. *Statistics in medicine*, 7(11):1121–1137, 1988.

[4] O.O. Aalen, Ørnulf Borgan, Håkon K Gjessing, and Stein Gjessing. *Survival and event history analysis: a process point of view*. Springer, 2008.

[5] O.O. Aalen, Johan Fosen, Harald Weedon-Fekjær, Ørnulf Borgan, and Einar Husebye. Dynamic analysis of multivariate failure time data. *Biometrics*, 60(3):764–773, 2004.

[6] O.O. Aalen and S Johansen. An empirical transition matrix for nonhomogeneous markov chains based on censored observation. *Scandinavian Journal of Statistics*, 5:141–150, 1978.

[7] Paul D Allison. *Event history analysis: Regression for longitudinal event data*. SAGE Publications, Incorporated, 1984.

[8] Per Kragh Andersen. *Statistical models based on counting processes*. Springer Verlag, 1993.

[9] Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

[10] Suresh H. Moolgavkar Anup Dewanji. A poisson process approach for recurrent event data with environmental covariates. *Environmetrics*, 11:665–673, 2000.

[11] RW Atkinson, HR Anderson, DP Strachan, JM Bland, SA Bremmer, and A Ponce de Leon. Short-term associations between outdoor air pollution and visits to accident and emergency departments in london for respiratory complaints. *European Respiratory Journal*, 13(2):257–265, 1999.

[12] Janet M Box-Steffensmeier and Bradford S Jones. *Event history modeling: A guide for social scientists*. Cambridge University Press, 2004.

[13] Charlotte Braun-Fahrländer, Ursula Ackermann-Liebrich, Joel Schwartz, Hans Peter Gnehm, Markus Rutishauser, and Hans Urs Wanner. Air pollution and respiratory symptoms in preschool children. *American Review of Respiratory Disease*, 145(1):42–47, 1992.

[14] Richard T Burnett, Jeffrey R Brook, Wesley T Yung, Robert E Dales, and Daniel Krewski. Association between ozone and hospitalization for respiratory diseases in 16 canadian cities. *Environmental research*, 72(1):24–31, 1997.

[15] Richard T Burnett, Marc Smith-Doiron, Dave Stieb, Sabit Cakmak, and Jeffrey R Brook. Effects of particulate and gaseous air pollution on cardiorespiratory hospitalizations. *Archives of Environmental Health: An International Journal*, 54(2):130–139, 1999.

[16] Environment Canada. National air pollution surveillance program (naps), 2013.

[17] Health Canada. Health effects of air pollution. `http://www.hc-sc.gc.ca/ewh-semt/air/out-ext/effe/health_effects-effets_sante-eng.php`, 2006.

[18] Statistics Canada. `http://www.statcan.gc.ca/pub/82-625-x/2012001/article/11658-eng.htm`, 2012.

[19] Belong Cho, Jaewook Choi, and Yong-Tae Yum. Air pollution and hospital admissions for respiratory disease in certain areas of korea. *J Occup Health*, 42(4):185–191, 2000.

[20] Antonio Ciocco and Donovan J Thompson. A follow-up of donora ten years after: methodology and findings. *American Journal of Public Health and the Nations Health*, 51(2):155–164, 1961.

[21] David Collett. *Modelling survival data in medical research*. CRC press, 2003.

[22] Richard J Cook and Jerald F Lawless. *The statistical analysis of recurrent events*. Springer, 2007.

[23] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[24] David R Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[25] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[26] F Derriennic, S Richardson, A Mollie, and J Lellouch. Short-term effects of sulphur dioxide pollution on mortality in two french cities. *International journal of epidemiology*, 18(1):186–197, 1989.

[27] Luc Duchateau and Palul Janssen. *The frailty model*. New York : Springer Verlag, 2008.

[28] J Firket. Fog along the meuse valley. *Transactions of the Faraday Society*, 32:1192–1196, 1936.

[29] Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware. *Applied longitudinal analysis*. Hoboken, N.J. : Wiley, 2011.

[30] Thomas R Fleming and David P Harrington. *Counting Processes and Survival Analysis*. New York : John Wiley & Sons, 1991.

[31] Canadian Insitute for Health Information. `http://www.cihi.ca`, 2011.

[32] Karen Y Fung, Shahedul Khan, Daniel Krewski, and Yue Chen. Association between air pollution and multiple respiratory hospitalizations among the elderly in vancouver, canada. *Inhalation toxicology*, 18(13):1005–1011, 2006.

[33] MH Gail, TJ Santner, and CC Brown. An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics*, pages 255–266, 1980.

[34] RD Gill. Discussion of the paper by d. clayton and j. cuzick. *Journal of the Royal Statistical Society A*, 148:108–109, 1985.

[35] Mark S Goldberg, Richard T Burnett, John C Bailar III, Jeffrey Brook, Yvette Bonvalot, Robyn Tamblyn, Ravinder Singh, and Marie-France Valois. The association between daily mortality and ambient air particle pollution in montreal, quebec: 1. nonaccidental mortality. *Environmental Research*, 86(1):12–25, 2001.

[36] IJ Goodd and RA Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58(2):255–277, 1971.

[37] Nelson Gouveia and Tony Fletcher. Respiratory diseases in children and outdoor air pollution in sao paulo, brazil: a time series analysis. *Occupational and environmental medicine*, 57(7):477–483, 2000.

[38] John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data, Second Edition*. John Wiley & Sons, 2002.

[39] Patrick J Kelly and Lynette L-Y Lim. Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in medicine*, 19(1):13–33, 2000.

[40] John P Klein. Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, pages 795–806, 1992.

[41] David G Kleinbaum and Mitchel Klein. *Survival analysis: A Self-Learning Text*. Springer, 2012.

[42] Jerry Lawless, Joan Hu, and Jin Cao. Methods for the estimation of failure distributions and rates from automobile warranty data. *Lifetime Data Analysis*, 1(3):227–240, 1995.

[43] Hyun Ja Lim, Jingxia Liu, and Marlene Melzer-Lange. Comparison of methods for analyzing recurrent events data: application to the emergency department visits of pediatric firearm victims. *Accident Analysis & Prevention*, 39(2):290–299, 2007.

[44] Danyu Y Lin, Lee-Jen Wei, and Zhiliang Ying. Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80(3):557–572, 1993.

[45] DY Lin and LJ Wei. Goodness-of-fit tests for the general regression model. *Statistica Sinica*, 1:1–17, 1991.

[46] DY Lin, LJ Wei, I Yang, and Z Ying. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):711–730, 2000.

[47] WP Logan. Mortality in the london fog incident, 1952. *Lancet*, 1(6755):336, 1953.

[48] Isaac N Luginaah, Karen Y Fung, Kevin M Gorey, Greg Webster, and Chris Wills. Association of ambient air pollution with respiratory hospitalization in a government-designated area of concern: the case of windsor, ontario. *Environmental health perspectives*, 113(3):290, 2005.

[49] CA McGilchrist and CW Aisbett. Regression with frailty in survival analysis. *Biometrics*, pages 461–466, 1991.

[50] Clyde A McGilchrist. Reml estimation for survival models with frailty. *Biometrics*, pages 221–225, 1993.

[51] Melvin L Moeschberger and John P Klein. *Survival analysis: Techniques for censored and truncated data*. Springer, 2003.

[52] Gert G Nielsen, Richard D Gill, Per Kragh Andersen, and Thorkild IA Sørensen. A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, pages 25–43, 1992.

[53] Asthma Scoiety of Canada. Asthma facts & statistics. `http://www.asthma.ca/corp/newsroom/pdf/asthmastats.pdf`, 2013.

[54] Saskatchewan Ministry of Environment. Current and historical saskatchewan air quality data. `http://www.environment.gov.sk.ca/airqualityindex`, 2013.

[55] Ross L Prentice, Benjamin J Williams, and Arthur V Peterson. On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379, 1981.

[56] R. www.r-project.org, 2013.

[57] Willem Roemer, Gerard Hoek, and Bert Brunekreef. Effect of ambient winter air pollution on respiratory health of children with chronic respiratory symptoms. *American review of respiratory disease*, 147(1):118–124, 1993.

[58] Steven G Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

[59] Terry Therneau and Patricia M. Grambsch. *Modeling survival data : Extending the Cox model*. New York : Springer, 2000.

[60] Terry M Therneau, Patricia M Grambsch, and Thomas R Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

[61] Terry M Therneau, Patricia M Grambsch, and V Shane Pankratz. Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, 12(1):156–175, 2003.

[62] James W Vaupel, Kenneth G Manton, and Eric Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454, 1979.

[63] Paul J Villeneuve, Li Chen, Brian H Rowe, Frances Coates, et al. Outdoor air pollution and emergency department visits for asthma among children and adults: a case-crossover study in northern alberta, canada. *Environ Health*, 6(1):40, 2007.

[64] Lee-Jen Wei, Danyu Y Lin, and L Weissfeld. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84(408):1065–1073, 1989.

[65] WHO. International classification of diseases (icd). `http://www.who.int/classifications/icd/en/`, 2013.

[66] Tze Wai Wong, Tai Shing Lau, Tak Sun Yu, Anne Neller, Siu Lan Wong, Wilson Tam, and Sik Wing Pang. Air pollution and hospital admissions for respiratory and cardiovascular diseases in hong kong. *Occupational and environmental medicine*, 56(10):679–683, 1999.

[67] Xiping Xu, Jun Gao, Jun Gao, and Yude Chen. Air pollution and daily mortality in residential areas of beijing, china. *Archives of Environmental Health: An International Journal*, 49(4):216–222, 1994.

[68] Ying Xu, Yin Bun Cheung, KF Lam, and Paul Milligan. Estimation of summary protective efficacy using a frailty mixture model for recurrent event time data. *Statistics in Medicine*, 31(29):4023–4039, 2012.

Appendix A

A.1 Statistical Computation

In this section, we introduce the **R** functions that used in the data analysis.The **R** packages we used in this study is called **survival**, which is available from the Comprehensive **R** Archive Network at `http://cran.r-project.org/web/packages/survival` (Therneau, 2013).

In Section 2.1.3, we presented that the Cox proportional hazards model can be used to fit univariate survival data by treating the events independent as they come from different individuals. In Section 2.2, we introduced that the Cox proportional hazards model can be reformulated based on the counting process technique. This reformulation extends the Cox hazards model to fit recurrent event data. The Anderson-Gill model used in this study for fitting recurrent hospital admission data is one extension of the Cox hazards model. It assumes that the observations in nonoverlapping intervals are independent. Cook and Lawless (2007) pointed out that software for fitting the Cox hazards model has been extended to fit the Anderson-Gill model for recurrent event data. Additionally, Therneau (2013) pointed out that the **coxph** function can be used to fit recurrent event data by using the counting process formulation (`http://cran.r-project.org/web/packages/survival/survival.pdf`). Furthermore, shared gamma frailty model can be fitted by using the **frailty (id)** option in **coxph** function (Cook and Lawless, 2007; Therneau, 2013).

Before we introduce the **R** function, it is important to present several necessary variables when fitting the recurrent event data. As mentioned in Section 2.2.1, we use at-risk process to indicate whether the individual is under observation. The term "status" is used to denote the at-risk process which remains one when an individual is under observation and becomes zero otherwise. Additionally, Start Time and Stop Time are the other two necessary variables. Start Time shows the time at which the individual entered study or the time at which the last event occurred if it is not the first event. Stop Time shows the time at which the event

occurred or the end of follow-up if there is no more event occurrence. In this study, the origin time is January 1, 2005, and the corresponding Start Time is 0, while the end of study time is December 30, 2011 with the corresponding Stop Time 2554 days. Table A.1 shows three patients' data as an example of the data layout. For example, patient 2 entered study at Start Time (t.start=0) and experienced the first event at Stop Time (t.stop=2038) and then censored at Stop Time (t.stop=2554). The status equals to 1 until censoring. The counts denotes the number of previous events. For instance, patient 2 experienced the first event at time t.stop=2083, at this time the number of event happened before t.stop=2083 is 0.

Table A.1: Extract from the datasets

id	start	stop	status	age	sex	counts	CO_2	NO_2	O_3
2	0	2083	1	90	0	0	0.2	8	18
2	2083	2554	0	90	0	1	0.2	8	16
39	0	980	1	70	0	0	0.1	9	9
39	980	1002	1	70	0	1	0.1	9	12
39	1002	2554	0	70	0	2	0.3	19	10
51	0	1579	1	80	1	0	0.2	6	32
51	1579	1599	1	80	1	1	0.2	6	25
51	1599	2005	1	80	1	2	0.3	6	26
51	2005	2394	1	80	1	3	0.2	4	22
51	2394	2554	0	80	1	4	0.3	19	10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

A.2 R Code

In section 3.2, we have fitted Poisson process model (Anderson-Gill model), Poisson process model with $N_i(t-)$ and mixed Poisson process model for the recurrent hospital admission data. Before displaying the **R** function for each model, we need to load the package **survival**.

```
> library(survival)
```

Poisson Process Model (Anderson-Gill Model)

```
> Coxph(Surv(start,stop,status)~ age+sex+p1+p2+p3+p4+p5+w1+w2,data)
```

- Start and Stop: Start Time and Stop time described in Section A.1;
- Status: When the individual is under observation equals to 1 and equals to 0 otherwise;
- p1: CO; p2: NO_2; p3: SO_2; p4: O_3 and p5: $PM_{2.5}$;
- w1: Temperature and w2: relative humidity;
- Data: The dataset used in the statistical analysis.

Poisson Process Model with $N_i(t-)$

```
> Coxph(Surv(start,stop,status)~ age+sex+pre.counts+p1+p2+p3+p4+p5+w1
                                 +w2,data)
```

- Pre.counts: $N_i(t-)$ records the number of events before time t.

Mixed Poisson Process Model

```
> Coxph(Surv(start,stop,status)~ age+sex+p1+p2+p3+p4+p5+w1+w2+frailty(id)
                                 ,data)
```

Extensive discussion about using the frailty option **coxph** function can be found in Therneau and Grambsch (2000).