Grain Dust and Health: A Competing Risk Analysis for the Grain Workers in Saskatchewan

A Thesis Submitted to the

College of Graduate and Postdoctoral Studies

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

Md Nazmul Hasan

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, S7N 5E6, Canada

Or,

Dean

College of Graduate and Postdoctoral Studies

University of Saskatchewan

116 Thorvaldson Building, 110 Science Place

Saskatoon, Saskatchewan, S7N 5C9, Canada

Abstract

Grain dust industry workers are exposed to a number of work-related hazards, including high levels of endotoxin, microorganisms and dust. Multiple studies have reported immunological, toxicological and clinical effects of occupational exposure to grain dust contaminants. The study aims to determine the effects of various prognostic and demographic factors on health-related outcomes among the grain industry workers in Saskatchewan. Statistical Analyses of the grain dust data can be carried out in a competing risk framework. In this context, competing risk is defined when an individual has a chance of getting one or more events to emulate with event of interest (e.g. death, time to relapse, time o disease type etc.). The competing risk analysis involves fitting the Cox PH model separately for each event type, treating the other (competing) event types as censored in addition to those who are censored from loss to follow-up or withdrawal. One of the assumptions of competing risk analysis is that censoring is independent of events regardless of the different ways that censoring can occur, including failure from competing risks other than the event-type of interest.

We define three competing events for the grain dust industry workers in Saskatchewan: chronic cough or phlegm, shortness of breath and allergy. Each worker can experience any of these events over the follow-up period from 1978 to 2005. We then consider seven covariates to assess their effects on the hazards of each of these three events: age, history of health problem (yes/no), history of asthma (yes/no), body mass index (BMI), forced expiratory volume in one second (FEV1), FEV1/FVC ratio which is the proportion of the amount of air exhaled in the first second (FEV1) to all of the air exhaled during a maximal exhalation (FVC) and smoking.

Our competing risk analyses reveal that FEV1/FVC ratio and smoking are highly significant to the risk of developing chronic cough or phlegm (p-value = 0.0238 and 0.0009 respectively). For shortness of breath, history of asthma and smoking are found significant, with p-values 0.0481 and 0.024, respectively. Results also indicate a high impact of age and FEV1 on allergy.

Our analyses are based on a relatively small sample (n = 226), and therefore caution should be applied to generalize our findings. Nevertheless, our findings could be useful for policy makers to make the environment of grain industries safe and secure for the workers with respect to standards and guidelines. Results could also be useful for human awareness.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Dr. Shahedul Khan for his continuous support, patience, motivation, and immense knowledge in my study and related research. His guidance helped me in all the way during the time of my entire research and writing. I could not have imagined having a better advisor and mentor for my program.

Besides my advisor, I would like to thank the rest of my thesis committee members, Dr. Juxin Liu and Dr. Steven Rayan, for their thoughtful comments and encouragement to my research from various perspectives.

I am also grateful to Dr. Punam Pahwa and Dr. James Dosman, for providing me the data which made it possible to do my research.

I am really thankful to my fellow mates Kangjie Zhang and Naorin Islam for their support and advice during my course works and research. I want to thank Dr. Saima Khan Khosa for enlightening me at the beginning of my program.

I am very obliged to all the professors, graduate students and staffs in the Department of Mathematics and Statistics. My special thanks to administrative coordinator Kyla Denton who has always been there for me starting from my pre-application process for the program. Thanks to Dr. Lawrence Chang and lab co-ordinator Manuela Golban for their support in teaching assistantship duties. I am thankful to the department of Mathematics and Statistics and my supervisor for the financial support provided throughout my Masters program.

Dedication

I would like to dedicate my dissertation work to my family. A special feeling of gratitude to my loving parents, Md. Shahjahan and Umme Nazma whose inspiration has always given me strength and courage. I also dedicate this dissertation to my lovely wife Zahida Sultana Irin who has supported me throughout the process. She is the one who helped to build confidence in me and motivated me all the way. My angels, Zareef Adyan Hasan and Zaheen Ahyan Hasan are always worked as my lifelines. I have never felt to give up during my studies because of their presence. Finally, I also want to dedicate my work to my younger sister Tasfia Jahan Sanjida and my friends who have never left my side during my hard time.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Grain workers are exposed to not only grain dust inhalation but also to other materials, resulting in an increased risk of health problems. In particular, exposure to grain dust may result in a range of acute and chronic respiratory symptoms and reduced lung function (Pahwa et al. 2003). Many studies have reported multiple long-term effects of grain dust, including chronic cough or phlegm, shortness of breath and allergy. When a worker is exposed to grain dust, various prognostic and demographic factors may act differently because of interaction between these factors and grain dust inhalation. Therefore, it is of particular interest to investigate the effects of these factors on the risk of long-term health effects for the grain workers. In this study, we investigate the effects of various factors (covariates) on long-term health hazards for the grain workers in Saskatchewan. Time-to-event data for a cohort of 280 Saskatchewan grain workers followed over a period of 27 years (1978 to 2005) are considered, and survival analysis methodology is used for statistical analysis.

Survival analysis is a set of statistical tools or methods used upon survival data, where the point of interest is time until an event occurs (Kleinbaum and Klein 2010). The event is commonly referred to as survival event, and it can be, e.g., the development of a disease, response to a treatment, relapse or death. The time to the occurrence of the event is referred to as survival time or lifetime or failure time. For example, survival time can be tumor-free time, the time from the start of treatment to response, remission time or time to death. Note that the survival time is considered

to be the response variable in lifetime data analysis, which is a non-negative random variable representing the lifetimes of individuals in some population (Lawless 2011). Typically, data sets on failure times also contain information on explanatory variables or covariates. As a result, it is of particular interest to develop models to characterize the relationship between the response and one or more covariates which are thought to affect some feature of the distribution of the response.

Time-to-event or survival data have a distinctive nature of containing observations with lack of information. For example, the event time is not observed if an individual drops out of the study before the occurrence of the event or does not experience the event by the end of the study. In such a scenario, only partial information is available in that we do not know the exact event time of the individual but we know that the individual is event-free until a certain time point. Incomplete observations of this nature are commonly known as censored observations in survival analysis. It is important to find some acceptable ways of analysis without ignoring censored observations. Skewness in the response variable is another distinguishing feature of time-to-event data (Kartsonaki 2016). For these reasons, standard statistical techniques cannot generally handle an analysis of survival data. Non-parametric methods are usually considered to get basic information on the response variable, which include the Kaplan-Meier (Kaplan and Meier 1958) and Nelson-Aalen (Nelson 1969, Aalen 1978) test to compare survival experiences between two groups (Lawless 2011). Regression methodology is used t understand the relationship between the response and one or more covariates. Both parametric and semi-parametric methods are widely used for regression analyses. For example, parametric regression models can be formulated using the weibull, log-logistic and log-normal distributions (Lawless 2011), and semi-parametric methods can be developed using the Cox-proportional hazards model (Cox 1972).

Another common problem in survival analysis is competing risk, where each subject can experience only one of several different types of events over follow-up, Conventional approaches like the Kaplan-Meier method and the Cox proportional hazards model are not directly compatible for

an analysis of competing risk data (Noordzij et al. 2013). However, the Cox proportional hazards model can be extended for competing risk problems (Prentice et al. 1978)

History of prognostic studies shows us the development and elaboration of statistical analysis for survival data in more than thirty years. Generalization of statistical tools for survival data brings historical changes in medical studies. Nowadays survival analysis is used in a significant number of medical research to obtain more precise knowledge about biomedical history of subjects, including competing risk problems.

1.1 Background of Study

1.1.1 Motivating Examples: Grain Dust and Health

In grain industries, workers are always at risk of hazard from different things. The grain dust is one of the major reasons for exposure in the long run. Canada is one of the largest production home for grains like wheat, barley, oats and rye. Wheat is the largest crop among all these grains. According to Dakers and Fréchette (1998), Canada yields approximately 7% of wheat and barley from world's total production. In 2017, total productions of wheat, barley and oats were approximately 27, 7 and 4 million tones, respectively (Agriculture and Canada 2017). According to the report from Statistics Canada (2017), Saskatchewan is the greatest grain producing province in Canada. Statistics Canada also reported that, in Saskatchewan about 91% of the total cropland was seeded with field grains in 2016. A number of grain industry is running with a huge number of grain workers to process and handle grains. It is now well known to all that, grain workers are facing different health problems due to grain dust. Among these, lung functions and respiratory problems are the most common ones.

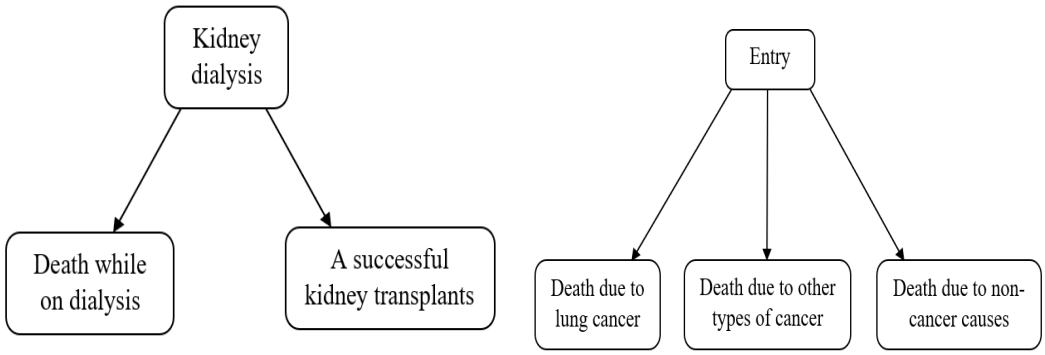
In 1976, Labor Canada took a step to get a complete picture of the grain industry. A national program for inspection of health and monitoring of environment was announced nationally. The program was then started in 1978 as "Grain Dust Medical Surveillance Program(GDMSP)" (Pahwa et al. 2003). After that, they have distributed some guidelines for the program. GDMSP data consist of health information of employees who had been working in the grain industries for more than 90 days in a period of one year or six months irregularly in three years. The test was conducted every three years in eight territories and five different geographical regions. Workers have undergone one or more medical examinations under the guidance of a licensed physician during the program (Pahwa et al. 2003). The purpose of the program was to keep tracking workers respiratory conditions in association with grain dust levels in the industries. We will only use Saskatchewan's data for our research.

A couple of longitudinal analysis has been done using the data set. For example, an article by Pahwa et al. (2003) has consider longitudinal model to see the decline in lung function measurements due to grain dust among grain workers in Saskatchewan. Their result indicates the estimated annual decline for lung function measurements increased with respect to the length of time in the grain industry, considering workers smoking status as an important factor. Note that, we have considered not only observed individual information but also the information of individuals that are censored in the data set. This is the reason our statistical analysis is different from that of their analysis.

1.1.2 Statistical Framework

Statistical Analyses of the grain dust data can be carried out in a competing risk framework. A general overview of the competing risk problems is presented below (for theoretical details, see Chapter 2).

A particular case of time-to-event data analysis is acknowledged as competing risk (Pintilie 2011), where an individual has a chance of getting one or more events to emulate with event of interest is defined as competing risk (Noordzij et al. 2013). According to Feakins et al. (2018), "competing risks are defined as events during follow-up that either preclude the observation of the primary outcome or alter the probability of its occurrence". For example, a patient may get a donor for kidney transplant during the dialysis period, with the event of interest being death while on dialysis. Here, the competing event is "a successful transplant", which is likely to prevent the event of the interest from occurring while on dialysis (Noordzij et al. 2013). The conceptual framework of this problem is presented in figure 1.1a. Similarly, in a study involving cancer patients, the outcome of interest could be death due to a specific type of cancer, whereas death due to other types of cancer and death due to non-cancer causes could be considered competing outcomes (see figure 1.1b). In summary, competing risk problems arise when there are two or more possible ways that a patient can experience the event but practically, the event occurs only for one reason (Kleinbaum and Klein 2010).



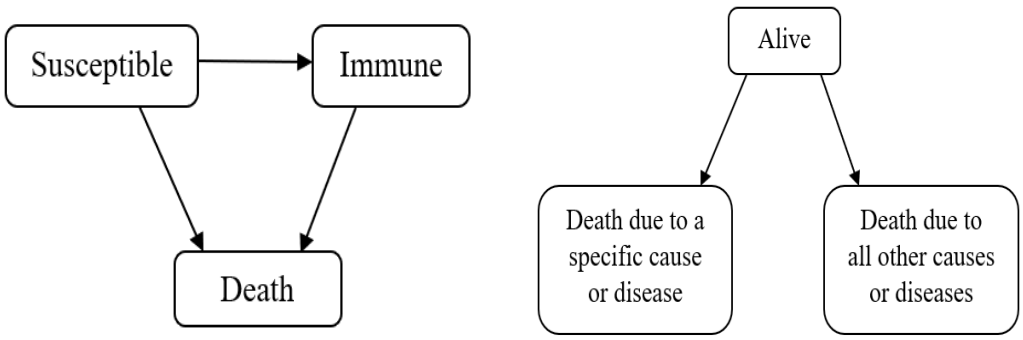
(a) Competing risk of a kidney dialysis patient

(b) Competing risk of a lung cancer patient

Figure 1.1: Competing risk models in studying the effect on population mortality due to different causes

The main idea behind competing risk model was to review the impact of death due to other reason on death due to a specific disease. The competing risk problem was introduced by d'Alembert

and Bernoulli first in 1760s. They noticed evidence of competing event effects in a cohort study involving smallpox patients. While studying the impact of immunization for smallpox, discovered a relationship between long and short term mortality. For a preliminary analysis, Bernoulli defined two disease states: susceptible and immune. In this model, death was considered as an absorbing state, with transition defined in terms of infection rate. Moreover, no interim state was assumed due to relatively short duration of smallpox (figure 1.2a). d'Alembert considered this problem from a different point of view to give it a more simpler look. He defined two independent absorbing states for death along with a single state of alive (see figure 1.2b). The two absorbing states were defined as death due to the disease of interest and death due to all other causes (Feakins et al. 2018). Note that Bernoulli's model involves a multi-state modelling framework, whereas d'Alembert's model considers a competing risk problem. These are the very initial steps towards analysis of time-to-event data for competing events. Competing risk models have been extensively used in medical, health and epidemiological research ever since the first discussion by Bernoulli of the effects on population mortality of removing smallpox through vaccination.



(a) Bernoulli's Model **(b)** d'Alembert's Model

Figure 1.2: Conceptual models in studying the effect on population mortality of removing smallpox through vaccination; Bernoulli assumed a multi-state modelling framework, whereas d'Alembert considered a competing risk analysis

Competing risk analysis is commonly used for elderly population who are struggling to get fit for a longer period of time. It is very crucial for a patient to know about risk for a certain condition

and get proper treatment or available options. It also helps physicians to get optimal information about a patient to provide the best possible treatment for the conditions. For the time being, this age group represents a higher portion of competing events. In a review, 70% studies out of 50 clinical studies have been reported competing risks as an important issue for elderly population in high-impact journals (Tan et al. 2018).

The objective of the study has always been an important part of deciding methods of analysis. Simple Kaplan-Meier and Competing risk processes are analogous, where censoring events assumed to have no information in absence of competing events (Tan et al. 2018). However, informative censoring can come up with biased estimates in the presence of competing events (Tan et al. 2018). In accordance with Tan et al. (2018), "A recent review of 100 studies from prominent medical journals found that 46% of studies that used Kaplan-Meier estimates ignored potential competing risks and Kaplan-Meier estimates were biased by at least 10%". The information provide an idea about dominance of competing risk analysis over other accredited statistical analysis in recent years. It has been acknowledged in many studies that competing risks analysis is more efficient to give admissible results than ever before.

1.1.3 Motivation

Health related outcomes of grain industry workers can be conveniently represented by a competing risk framework. Grain industry workers may suffer from different types of complications, including chronic cough, chronic phlegm, chronic bronchitis and so on. An intuitive modeling framework would be to consider transitions from the healthy state to one of such complications with competing risks from other health outcomes. Although many authors reported strong statistical association between exposures to grain dust and adverse health outcomes (Pahwa et al. 2003, Swan et al. 2007), competing risk analyses have never been considered in these studies. Strong evidence of biased estimates of conventional approaches motivate us to think beyond conventional

ways of analysis when competing events are the main focus.

1.2 Objectives

The primary objective of this study is to investigate the effects of various prognostic and demographic factors on health-related outcomes among the grain industry workers in Saskatchewan. To achieve our goal, we consider a competing risk framework for statistical analyses with three competing events: chronic cough, chronic shortness of breath and allergy. Overall, our findings could be useful not only to policy makers with respect to standards and guidelines for grain industry workers, but also for human awareness.

In chapter 2, we describe the background and theoretical development of the competing risk models. Analysis of the grain dust data in a framework is presented in Chapter 3. Our findings are summarized in chapter 4, which include limitations of our study and directions for future work.

Chapter 2

Competing Risk Modelling

In chapter 1, we have mentioned competing risk model to analyze the exposures to grain dust among grain workers in Saskatchewan. This chapter provides a complete overview of survival analysis and competing risk model including definitions, relations between functions and methods of estimation.

2.1 Basic Concepts of Survival Analysis

Some important functions in the analysis of survival data are cumulative density function, survival function and hazard function. All these functions will be introduced here with mathematical definitions as well as their relations with other functions. We are using the same terminology and notation from Kleinbaum and Klein (2010) book.

2.1.1 Definitions and Relations

Let T be a non-negative ($T \geqslant 0$) continuous random variable commonly known as survival time. Time until an event occurs is the survival time of an individual. Time can be expressed as days, month, years or weeks (Kleinbaum and Klein 2010). Specific value of random variable T is denoted as t. Now, survival function of an individual can be defined as-

$$S(t) = P(T > t), \tag{2.1}$$

which means probability of survival of an individual longer than a specified time t where t lies between 0 to infinity range. Survival function is a monotone decreasing function where the probability of survival is 1 at the beginning of a study when time $t = 0$ and the probability tends to 0 when survival time t approaches infinity.

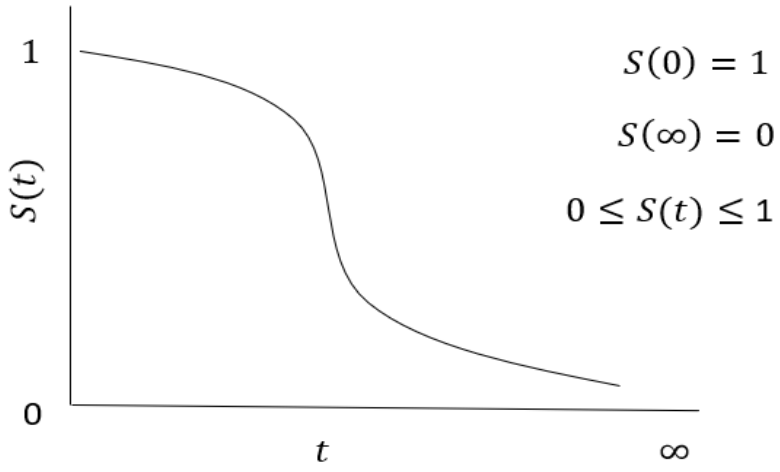


Figure 2.1: An example of survival function – indicating a monotone decreasing trend with probability of survival is 1 when time t is 0 and probability of survival is 0 when time t approaches infinity.

Mathematically, we can write- $S(t) = 1$ when $t = 0$ and $S(t) = 0$ when $t = \infty$. Hazard function is denoted by $h(t)$ and defined as-

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leqslant T < t + \Delta t | T \geqslant t)}{\Delta t}, \tag{2.2}$$

this mathematical hazard function (2.2) can be expressed in words. Hazard function $h(t)$ describes instantaneous rate of experiencing an event by a subject in a study who has survived till time t. Hazard function is also known as conditional failure rate as it has a conditional format of expression (Kleinbaum and Klein 2010). Cumulative hazard function can be expressed in terms of hazard

function and it has a relationship with survival function-

$$H(t) = \int_0^\infty h(u)du, \tag{2.3}$$

$$S(t) = e^{-H(t)}, \tag{2.4}$$

which means survival function can also be expressed as- a exponential of negative cumulative hazard function. Similarly, we can express hazard function with respect to survival function-

$$S(t) = \exp\left(-\int_0^\infty h(u)du\right), \tag{2.5}$$

$$h(t) = -\frac{1}{S(t)}\frac{dS(t)}{dt}, \tag{2.6}$$

we know that the distinguished feature of survival data is to contain censoring information. Missing observations are generally known as censored in survival analysis. In most research, prior to the analysis those rows have been removed which contain missing observations. However, censored observations are considered in survival analysis which has given the analysis a distinctive feature. Therefore, an indicator variable is required to identify censored observations in a data set. We assume δ as the censoring indicator where, 1 for the occurrence of an event and 0 for censoring. If T and C be the time to an event and censoring respectively then-

$$\delta = \begin{cases} 1 & \text{if } T \leq C \\ 0 & \text{if } T > C \end{cases}$$

let δ_i is the censoring status of i^{th} individual in a study. Likelihood function for a survival model

can be written as-

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}, \tag{2.7}$$

we can maximize the likelihood function to get estimates. Non-parametric and parametric methods are used in the analysis of survival data.

2.1.2 Non-Parametric Methods

Different non-parametric methods are used to estimate survival and hazard functions. Some non-parametric methods also provide graph to demonstrate survival and hazard functions. Kaplan-Meier and Nelson-Aalen are two commonly used non-parametric methods in survival analysis. Descriptions are provided for these methods in this section.

Kaplan-Meier Estimator

Kaplan-Meier estimator is often known as product limit estimator. Kaplan-Meier estimator is an effective process of computing and portraying survival functions from time-to-event data. In the absence of censored observations in a sample with size n, the empirical survivor function is defined as-

$$\hat{S}(t) = \frac{\text{Number of observations} \geqslant t}{n}, \tag{2.8}$$

this function (2.8) is a step function, where $t \geq 0$. It decreases by $1/n$ immediately after each observed survival time where all observations are unique. In other words, empirical survivor function can drop by d/n if number of survival time d is equal to t (Lawless 2011). An extension of the equation 2.8 has been introduced by Kaplan and Meier in 1958 that is capable of dealing survival data with censored observations. Kaplan and Meier (1958) described the estimate and its properties first time in the article.

Suppose we have a data set contains n individuals with k different times $t_1 \leq t_2 \ldots \leq t_k$ at which an event of interest occur. If our event of interest is death then it is possible to get more than one death at time t_j and where $j = 1, 2, \ldots, k$. Let d_j and n_j respectively represent the number of death and the number of individual at risk at time t_j. Now, censoring times can be denoted as λ_j that are from the interval $[t_{(j-1)}, t_j)$, whereas observed censoring times is L_i^j $(i = 1, \ldots, \lambda_j)$. We can keep censoring times that occur before first or after last observed individuals' lifetime by considering $t_0 = 0$, $t_{(k+1)} = \infty$ and $j = 1, \ldots, k+1$. If underlying survivor function is $S(t)$ and the probability of dying of an individual is $S(t_j) - S(t_j + 0)$ where $S(t)$ is a non-increasing left continuous function (Lawless 2011) then the observed likelihood function can be written as-

$$L = \prod_{j=1}^{k} \left[\left(\prod_{i=1}^{\lambda_j} S(L_i^j) \right) [S(t_j) - S(t_j + 0)]^{d_j} \right] \prod_{i=1}^{\lambda_{k+1}} S(L_i^{k+1}) \tag{2.9}$$

to maximize equation 2.9 with respect to $S(t)$, define $P_0 = 1$ and $S(t_j + 0) = P_j (j = 1, \ldots, k)$ and consider only-

$$L_1 = \prod_{j=1}^{k} (P_{j-1} - P_j)^{d_j} P_j^{\lambda_{j+1}}, \tag{2.10}$$

now maximize L_1 with respect to $P_1, \ldots, P_k$ to get the estimated survivor functions where $p_j =$ probability of an individual survives beyond the interval $[t_{j-1}, t_j) = P_j / P_{j-1}$ and $q_j = 1 - p_j$ $(j = 1, \ldots, k)$ (Lawless 2011). Here, We have-

$$L_1 = \prod_{j=1}^{k} (p_1 \ldots p_{j-1} q_j)^{d_j} (p_1 \ldots p_j)^{\lambda_{j+1}} = \prod_{j=1}^{k} q_j^{d_j} p_j^{n_j - d_j}, \tag{2.11}$$

after that equation 2.11 can be maximized considering $p_j = (n_j - d_j)/n_j$ and $S(t_j + 0) = S(t_{j-1} + 0) \frac{n_j - d_j}{n_j}$. Finally, the equation becomes-

$$\hat{S}(t) = \prod_{j : t_j \leq t} \frac{n_j - d_j}{n_j}, \tag{2.12}$$

We can draw survival curve based on estimated survival functions. This curve is also called as Kaplan-Meier curve.

Nelson-Aaelan Estimator

Another familiar non-parametric estimator is Nelson-Aalen estimator. The estimator was first introduced by Nelson (1969), Aalen (1978) which can compute cumulative hazard functions from survival model. Nelson also provided the graphical presentation of cumulative hazard function to visualize the pattern. Aalen then extended the method for small and large sets of data beyond the survival and competing risks problems. Relationship between cumulative hazard function and survival function can be expressed in a different way, such as- $H(t) = -logS(t)$ and its natural estimate is-

$$\hat{H}(t) = -\log \hat{S}(t), \tag{2.13}$$

where $\hat{S}(t)$ is the estimated Kaplan-Meier survivor function. Derivation of Kaplan-Meier estimate has been discussed in the previous section. More generally we can rewrite the expression (2.13) with respect to number of death and number of individuals at a specific time t_j,

$$\hat{H}(t) = -\sum_{j:t_j \leq t} \log\left(1 - \frac{d_j}{n_j}\right) \tag{2.14}$$

Estimated hazard functions $\hat{H}(t)$ are known as Nelson-Aalen estimates. Estimates are used in making cumulative hazard curve.

2.2 Cox Proportional Hazard Model

Most popular estimation process in survival analysis is called Cox proportional hazard model. It was first introduced by Cox (1972) in the year of 1972. Cox proportional hazard model is a robust

model, although the baseline hazard function of the model is unspecified during the estimation. Cox proportional hazard regression model can provide estimates of regression coefficients, hazard ratios and adjusted survival curve from wide range of problems (Kleinbaum and Klein 2010). Another characteristics of the model is - estimates of regression coefficients can be obtained despite the fact of unknown baseline hazard function. The Cox proportional hazard model is always preferred in survival analysis over logistic regression model in the presence of censored observations because the Cox model uses maximum information about survival times and censoring times (Kleinbaum and Klein 2010). Whereas, censoring times are ignored in logistic model. For these reasons Cox proportional hazard model is very common in clinical studies.

2.2.1 Statistical Inference

The Cox regression model is formed with baseline hazard function, regression coefficients and covariates. It is written as-

$$h(t, x) = h_0(t) \exp \sum_{i=1}^{p} \beta_i x_i, \tag{2.15}$$

where $h_0(t)$ is the baseline hazard function, β_i $(i = 1, 2, \ldots, p)$ is a vector of parameters and x_i is a vector of explanatory variables. Explanatory variables are time-independent for the model (Kleinbaum and Klein 2010).

Regression coefficients ($\beta' s$) can be estimated using maximum likelihood method from the Cox proportional hazard model. Likelihood of the Cox proportional hazard model is formulated based on the distribution of the outcomes. One of the important features of the Cox model is that there is no distribution for time to event or outcome variable. Therefore, a full likelihood cannot be constructed based on the outcome distribution. It can be formulated based on the observed order of events for which Cox likelihood is called partial likelihood (Kleinbaum and Klein 2010).

Likelihood of individual i at time t_i is written as-

$$L_i(\beta) = \frac{h(t_i|x_i)}{\sum_{j:t_j \geq t_i} h(t_i|x_j)}, \tag{2.16}$$

summation is over the set of subjects j where the event did not occur before time t_i including i^{th} subject. The effect of the covariates can be estimated without changing the hazard over time. The joint probability of all events is-

$$L(\beta) = \prod_i L_i(\beta), \tag{2.17}$$

Wald test statistic is used to get significant factors in a model. It is also known as Z statistics. The statistic can be defined using estimates of coefficients and their standard errors. Divide each estimate by their corresponding standard error will give values that are approximately equal to the standard normal quantity. The statistic from the above calculation is known as wald statistics.

$$z = \frac{\text{Estimate of regression coefficient}}{\text{standard error of estimate}} = \frac{\hat{\beta}}{se(\hat{\beta})}, \tag{2.18}$$

estimated value of z is compared with tabulated value to see the presence of significant factors.

2.2.2 Checking PH Assumption

One of the way of checking proportional hazard assumption is to plot Schoenfeld residuals. Schoenfeld residuals is defined for every individuals considering each covariates in the model. Let, i^{th} subjects and k^{th} covariates have the estimated Schoenfeld residual r_{ik} which is defined by (notation from article of Hosmer Jr and Lemeshow (1999))

$$\hat{r}_{ik} = x_{ik} - \hat{\bar{x}}_{w_i k}; \tag{2.19}$$

here x_{ik} is the value of the k^{th} covariate for individual i, and weighted mean of covariate values are denoted as $\hat{\bar{x}}_{w_i k}$ in which the individuals in the risk set at the given event time are considered. Sum of Schoenfeld residuals is equal to zero but Schoenfeld residuals will lie between –1 and 1 for a dichotomous variable with 0 and 1 coding. A positive r_{ik} value means the higher expectation at a specific event time (Gillespie 2006).

2.3 Competing Risk Model

Competing risk analysis is an important part of survival analysis. In the presence of competing events, different statistical methods are used in the analysis of survival data. As we mentioned in our chapter 1 the competing risk was established to see the effect of different reasons of death on event (death due to a certain cause) of interest. Later on, the method has been modified for other types of competing models in which the event of interest is different (e.g affected by any kind of chronic diseases) than death. In other words, model was modified to work with adverse health outcomes along with death. Competing events develop when subjects have a chance of experiencing the events by more than one possible ways. For example, a patient could die from lung cancer or cardiovascular disease or any other chronic diseases. Types of failure could be more than one but event can only occur for one reason (Kleinbaum and Klein 2010). Therefore, every other event is censored by another event in the model (Fermanian 2003). We can have a clear picture from the Figure 2.2.

2.3.1 Cumulative Incidence Curve

One of the common estimation procedure for competing risk model is to estimate and plot cumulative incidence functions for every event in a study of time-to-event data. Cumulative incidence function provides a proportion of individual who had experienced the event till time t from any cause k considering the fact that subjects can experience event due to other causes (Hinchliffe

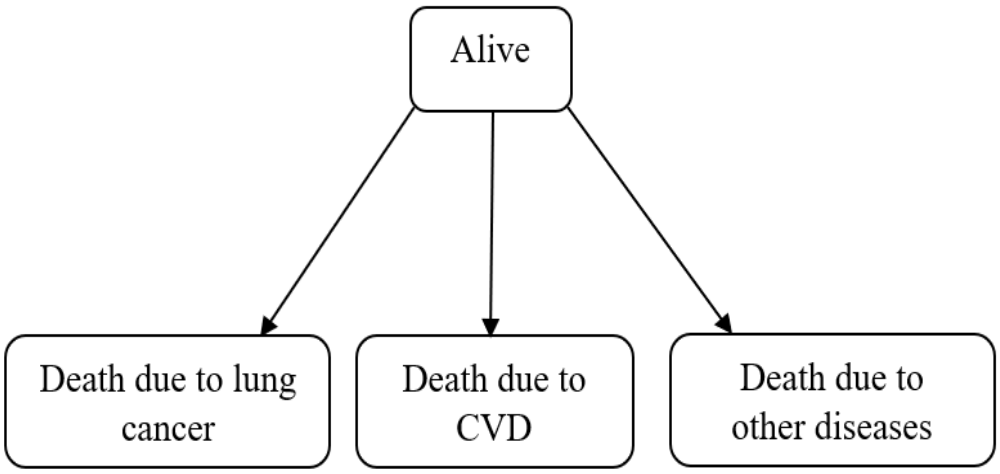


Figure 2.2: An example of conceptual competing risk model indicating three independent events: death due to lung cancer, cardiovascular disease and any other chronic diseases

and Lambert 2013), where k be the number of failure types in the model. Cumulative incidence curve is plotted separately for each of the events in the model. Cumulative incidence function can be defined as-

$$C_k(t) = \int_0^t h_k(u|x)S(u)du, \tag{2.20}$$

here $h_k(u|x)$ is the cause specific hazard with $k = 1, 2, \ldots, m$ and x is the vector of covariates. $S(u)$ is overall survival function.

2.3.2 Cause-Specific Hazard Model

Cause specific hazard model is based on the functions where risk of getting an event from a certain cause will be specified. Let $h_k(t)$ be the cause specific hazard where momentary risk of failure from a specific reason or cause given the information that the individual is still alive at t (Prentice et al. 1978). Survival and hazard functions are different in notations for the model. By

definition, these functions can be written in following form-

$$S_k(t) = \exp\left(-\int_0^t h_k(u)du\right) \tag{2.21}$$

$$h_k(t) = \lim_{\delta t \to 0} \frac{P(t \leqslant T < t + \delta t | T \geqslant t)}{\delta t}; \tag{2.22}$$

where $S_k(t)$ and $h_k(t)$ are the cause specific survival and hazard functions of the model. Cause-specific cox proportional hazard model looks like-

$$h_k(t) = h_0(t) \exp(x^T \beta), \tag{2.23}$$

here $h_0(t)$, x and β are presenting baseline hazard function, a vector of covariates and corresponding covariate coefficient vector.

2.4 Computational Functions

Various statistical software can be used in the analysis of competing risk model. The analysis of the study has been operated using *R* programing language created by R Core Team (2013). In the study of competing risk model using *R* program is very effective due to some important packages. *survival*, *survminer*, *survMisc*, *cmprsk* and *lubridate* (Terry M. Therneau and Patricia M. Grambsch 2000, Kassambara et al. 2019, Dardis 2018, Gray 2014, Grolemund and Wickham 2011) packages have been used in the study. These packages include important R-functions such as- $survfit$, $Surv$, $coxph$, $cox.zph$, $cuminc$, $ggcoxzph$, $ggcompetingrisks$, $ggforest$ that are used in finding the estimates and plotting curves of hazard function or survival functions of study population.

2.5 Summary

In this chapter, we have described basic components of survival analysis and competing risk model. Commonly used parametric and non-parametric methods are described for time-to-event data. Important computational functions and their packages from statistical software are also mentioned in this chapter.

Our results are shown in chapter 3 which includes summary table and graphs as well as their interpretations.

Chapter 3

Analysis

In this chapter we present a description of the data set, the conceptual framework for statistical analysis, and the fitting of the proposed models. We then present our hypothesis, and summarize our findings.

3.1 The Grain Dust Medical Surveillance Program

As mentioned in Chapter 1, the Grain Dust Medical Surveillance Program (GDMSP) started in 1978 and had run for fifteen years (Pahwa et al. 2003). The participated provinces and territories were divided into five regions: Atlantic (east of Quebec), St Lawrence (Quebec only), Great Lakes (Ontario - east of Thunder Bay), Central (Ontario - Thunder Bay and westward, Manitoba and Saskatchewan), and Mountain (Alberta, British Columbia, Yukon and Northwest Territories). Data were collected from each region in every three years in five cycles, including information on company, province, region, type of elevator, age, height, weight, smoking information, lung function measurements, respiratory symptoms, grade change and physician. The first two cycles' data were coded by Labor Canada, and the subsequent cycles' data were coded and entered in a computer system by the staff at the Environmental Epidemiology Unit, Centre for Agriculture Medicine, University of Saskatchewan (Pahwa et al. 2003). Note that most of the variables were coded as categorical or binary, and there were only a few continuous variables, measuring certain functions of the lung and the respiratory system.

Saskatchewan was one of the places where the GDMSP was implemented. Although the program ended its activities in all the regions in 1993, Labor Canada continued the survey in Saskatchewan for another twelve years. Thus, the survey in Saskatchewan includes 27 years of data (1978 to 2005), collected in nine cycles three years apart (see Table 3.1). Data are available for 280 Saskatchewan workers (244 males and 36 females).

Table 3.1: Nine cycles of the grain dust program in Saskatchewan.

Cycle	Interval
Cycle 1	(October 1978 to September 1981)
Cycle 2	(October 1981 to September 1984)
Cycle 3	(October 1984 to September 1987)
Cycle 4	(October 1987 to September 1990)
Cycle 5	(October 1990 to September 1993)
Cycle 6	(October 1993 to September 1996)
Cycle 7	(October 1996 to September 1999)
Cycle 8	(October 1999 to September 2002)
Cycle 9	(October 2002 to September 2005)

3.2 Data and Variables

There are 280 individuals and 144 variables in the data set, which include information on workers' date of birth, date of examination, demographic and prognostic factors, and some health-related outcomes. Note that cases with missing observations are discarded, leading to a sample of size $n = 226$. Ten considered for analyses: three health outcomes and seven covariates. The seven covariates are age, history of health problem (yes/no), history of asthma (yes/no), body mass index (BMI), forced expiratory volume in one second (FEV1), FEV1/FVC ratio which is the proportion of the amount of air exhaled in the first second (FEV1) to all of the air exhaled during a maximal

exhalation (FVC), expressed as FEV1%, and smoking status. The health outcomes are considered to be the competing risk events, and are defined as follows.

- **Event 1:** Chronic cough or phlegm if any of the following four conditions hold.
  - coughing on most days for as much as three months of the year, or
  - suffering from cough for at least a year, or
  - bringing up phlegm from the chest on most days for as much as three months of the year, or
  - suffering from phlegm from the chest for at least a year.
- **Event 2:** Shortness of breath if any of the following two conditions hold.
  - wheeze occasionally apart from cold, or
  - suffering from shortness of breath.
- **Event 3:** Occupational allergy.

A summary of the competing events are displayed in Table 3.2. Thirty three events are observed for chronic cough or phlegm (14.6%), 34 events are observed for shortness of breath (15.0%) and 43 events are observed for allergy (19.0%). Thus, there are 193 censored observations for chronic cough or phlegm (85.4%), 192 for shortness of breath (85.0%), and 183 for allergy (81.0%).

Table 3.2: A summary of the competing events under study.

Event	Count (%)
Chronic cough or phlegm	33 (14.6%)
Shortness of breath	34 (15.0%)
Allergy	43 (19.0%)
Total	110 (48.6%)

Table 3.3: A summary of the baseline measurements of the continuous covariates.

Variable	Definition	Mean	Standard deviation
Age	Age of workers	29.1	6.9
BMI	Body mass index in kg/m^2, weight in kg/(height in m)^2	26.9	4.2
FEV1	Forced expiratory volume in 1 second (volume of air that can be forced out in one second after taking a deep breath)	4.4	0.8
FVC	Forced vital capacity (volume of air in the lungs that can be exhaled following a deep inhalation)	5.6	0.9
Ratio	FEV1/FVC in percentage (percent of the lung size (FVC) that can be exhaled in one second)	79.9	5.6

Table 3.4: A summary of the baseline measurements of the binary covariates.

Variable	Definition	Count (yes)	Percentage
History of health problem	Any health issue at the time of diagnosis, 1 if yes, 0 Otherwise	16	7.1%
History of asthma	History of asthma in the family, 1 if yes, 0 Otherwise	7	3.1%
Smoking status	Current smoker 1 if yes, 0 Otherwise	71	31.4%

Note that we are interested to estimate the effects of the covariates for each failure type/event, allowing for competing risks from the other two failure types. A summary of the covariates at baseline are displayed in Tables 3.3 and 3.4. There are four continuous covariates: age, BMI, FEV1, and FEV1/FVC ratio (expressed in percentage). We see that (Table 3.3) average age of the

workers at baseline is 29.1 (standard deviation = 6.9), average BMI is 26.9, (standard deviation = 4.2), average FEV1 is 4.4 (standard deviation = 0.8), and average FEV1/FVC ratio is 79.9 (standard deviation = 5.6). We also have three binary covariates: history of health problem, history of asthma, and smoking status. We see that (Table 3.4) 16 workers had a history of health problem (7.1%), 7 workers had a history of asthma (3.1%), and 71 workers were smokers (31.4%).

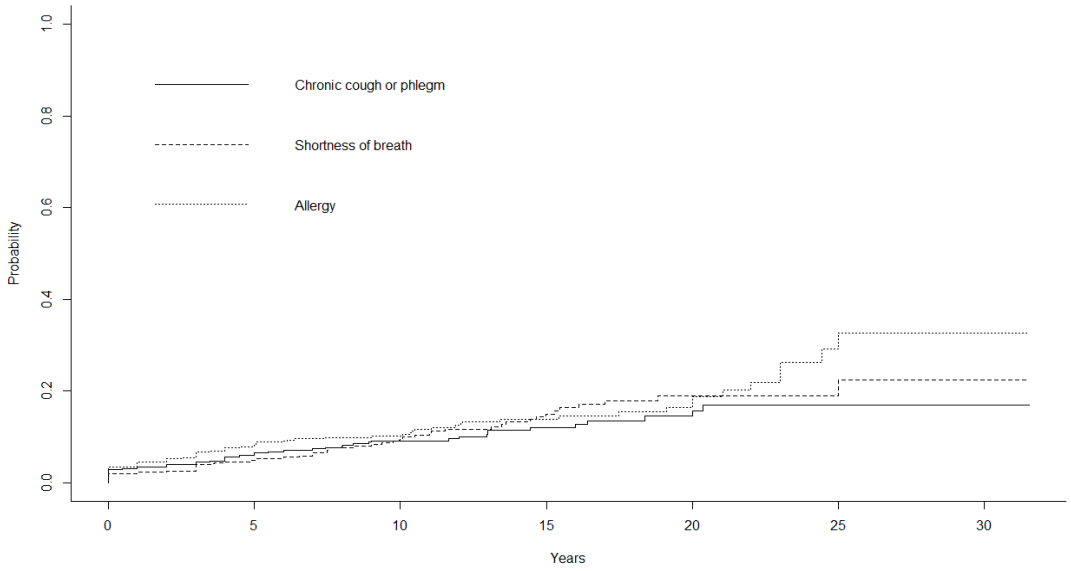


Figure 3.1: Non-parametric estimation of the cumulative incidence curves for the Saskatchewan grain dust industry workers from 1978 to 2005- solid, dashed and dotted lines are representing chronic cough and phlegm, shortness of breath and allergy respectively.

3.3 Trends of the Competing Events

In particular, a cumulative incidence curve is useful to understand the trends of the event hazards over time. The 27-year cumulative incidence rates (1978-2005) for the Saskatchewan grain industry workers are displayed in Figure 3.1 indicates three events: chronic cough and phlegm, shortness of breath and allergy with solid, dashed and dotted lines respectively. We see that the hazards for the three events are similar until around 1998 (20 years from the beginning of the

study). After about 20 years, a worker has a higher chance to suffer from occupational allergy, whereas a relatively lower chance to suffer from chronic cough or phlegm. We also see a rapid increase in hazards for shortness of breath after about 15 years, and it remains to be the highest risk event from year 15 to year 20. As mentioned above, allergy takes over the other two competing risks in terms of hazards after about 20 years. In summary, (a) there is no substantial difference during the first 15 years with respect to the occurrence of the events, (b) there is relatively a higher risk for shortness of breath between year 15 and year 20, and (c) there is relatively a higher risk for allergy after about 20 years of working in the industries. Together these lead to the conclusion that there is a long-term health risk for the grain dust industry workers in Saskatchewan.

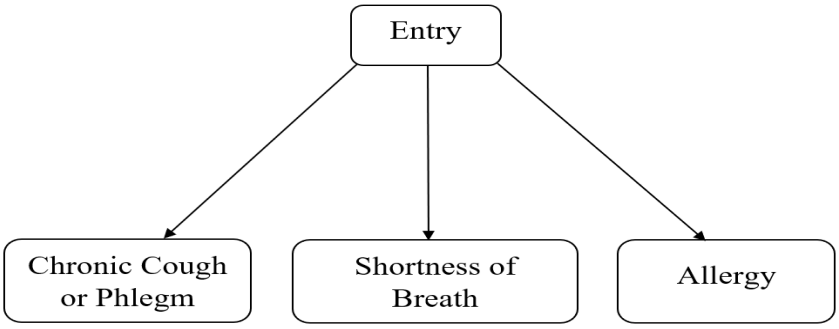


Figure 3.2: Conceptual framework of competing risk analyses with three events.

3.4 Competing Risk Analyses

Recall that the three competing events under study are chronic cough or phlegm, shortness of breath and allergy. The conceptual framework of the competing risk model is shown in Figure 3.2. The data contain 33 uncensored and 193 censored observations for chronic cough or phlegm, 34 uncensored and 192 censored observations for shortness of breath, and 43 uncensored and 183 censored observations for allergy.

The proportional hazards model for event type j ($j = 1, 2, 3$ for chronic cough or phlegm,

shortness of breath and allergy, respectively) can be written as

$$h_j(t) = h_{0j}(t)\exp(\beta_{1j}x_1 + \beta_{2j}x_2 + \beta_{3j}x_3 + \beta_{4j}x_4 + \beta_{5j}x_5 + \beta_{6j}x_6 + \beta_{7j}x_7), \tag{3.1}$$

where $h_{0j}(t)$ is the baseline hazard function for event type j, x_1 = age, x_2 = I(history of health problem), x_3 = I(history of asthma), x_4 = body mass index, x_5 = FEV1, x_6 = FEV1/FVC ratio and x_7 = I(current smoker), where I(A) is an indicator function that equals 1 if A is true and 0 otherwise. Since, there are very few female grain workers in the data set therefore inclusion of gender leads to a convergence issue. For this reason we drop gender from the model for the analyses.

For the competing risk model (3.1), we assume that censoring is independent, that is, a subject in the risk set at time t is as likely to experience any competing event as to be lost to follow-up. Under this assumption, the typical approach for analyzing competing risk data using the Cox proportional hazards model (3.1) involves fitting separate models for each competing risk while treating the other competing risks as censored observations. Note that, this assumption may not be reasonable for all cases, which is one of our limitations in this study. Our R codes to fit these models are given in the Appendix.

3.4.1 Analyis for Chronic Cough or Phlegm

We first fit the competing risk model for chronic cough or phlegm with shortness of breath and allergy censored. Numerical results are summarized in Table 3.5. We see that FEV1/FVC ratio and smoking are highly significant for chronic cough or phlegm (p-value = 0.0238 and 0.0009, respectively), whereas age is marginally significant (p-value = 0.0913). The estimates of the hazard ratios are given in Table 3.5, and the hazard ratio plots along with 95% confidence intervals are displayed in Figure 3.3. We see that the estimates of the hazard ratios for FEV1/FVC ratio, smoking and age are 1.0811, 3.3517 and 0.9348, respectively, suggesting

- the hazard risk for chronic cough or phlegm will increase about 8.1% for one unit increase of the FEV1/FVC ratio controlling for the other factors (i.e., a positive association between

FEV1/FVC ratio and chronic cough or phlegm);

- smoking will increase the hazard risk for chronic cough or phlegm about 235% controlling for the other factors; and
- one year increase in age will reduce the hazard for chronic cough or phlegm about 7% controlling for the other factors (i.e., a negative association between age and chronic cough or phlegm).

Table 3.5: Competing risk analysis for chronic cough or phlegm with shortness of breath and allergy censored – estimates of the coefficients, hazard ratios, standard errors of the estimates, Wald statistic (z), and p values.

	Estimate	Hazard ratio	Standard error (SE) of the estimate	$z = \frac{\text{estimate}}{\text{SE}}$	$\Pr(>\|z\|)$
Age (β_{11})	-0.0675	0.9348	0.0400	-1.6890	0.0913
History of health problem (β_{21})	0.7287	2.0724	0.5239	1.3910	0.1642
History of asthma (β_{31})	0.7786	2.1785	1.1163	0.6980	0.4855
BMI (β_{41})	-0.0738	0.9288	0.0463	-1.5970	0.1103
FEV1 (β_{51})	0.0120	1.0120	0.2922	0.0410	0.9673
FEV1/FVC ratio (β_{61})	0.0779	1.0811	0.0345	2.2600	0.0238
Smoking (β_{71})	1.2095	3.3517	0.3635	3.3270	0.0009

The hazard plot (3.3) for chronic cough model consisted with hazard ratio and their 95% confidence interval for all the variables in the model. This graph (3.3) also contains the information of global long-rank test p-value, Akaike Information Criterion (AIC) and concordance index of the model. From the plot (3.3), the variables that are significant have confidence intervals before or after the marginal line which is belongs to 1. Confidence interval for ratio and smoking status indicates a strong evidence of being highly significant.

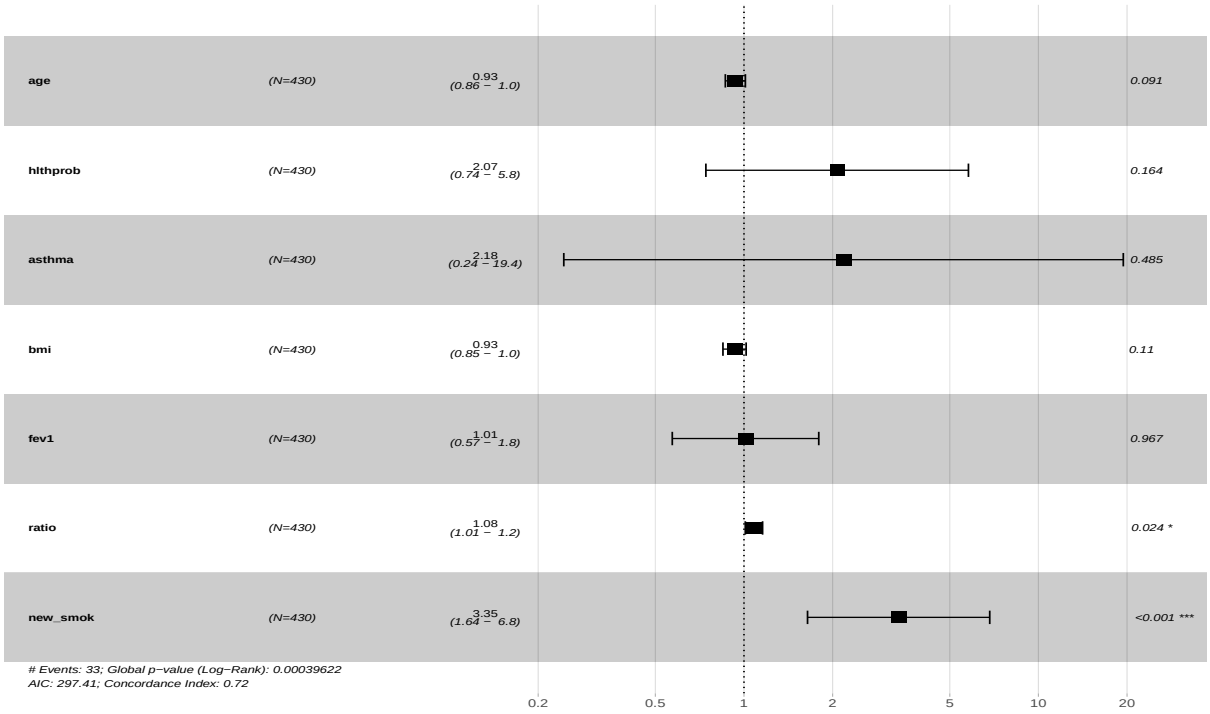


Figure 3.3: Competing risk analysis for chronic cough or phlegm – hazard ratio along with a 95% confidence interval for each of the covariates, indicating (a) FEV1/FVC ratio and smoking are highly significant, (b) age is marginally significant, and (c) history of health problem, history of asthma, BMI and FEV1 are not statistically significant.

Table 3.6: Competing event chronic cough or phlegm – test for the PH assumption for each covariate, along with a global test of the model as a whole.

	rho	chisq	p
Age (β_{11})	0.01632	0.01486	0.903
History of health problem (β_{21})	0.19953	1.45291	0.228
History of asthma (β_{31})	-0.11417	0.52047	0.471
BMI (β_{41})	-0.10371	0.40634	0.524
FEV1 (β_{51})	-0.00997	0.00378	0.951
FEV1/FVC ratio (β_{61})	0.09800	0.36173	0.548
Smoking (β_{71})	-0.14648	0.66650	0.414
GLOBAL	NA	3.02113	0.883

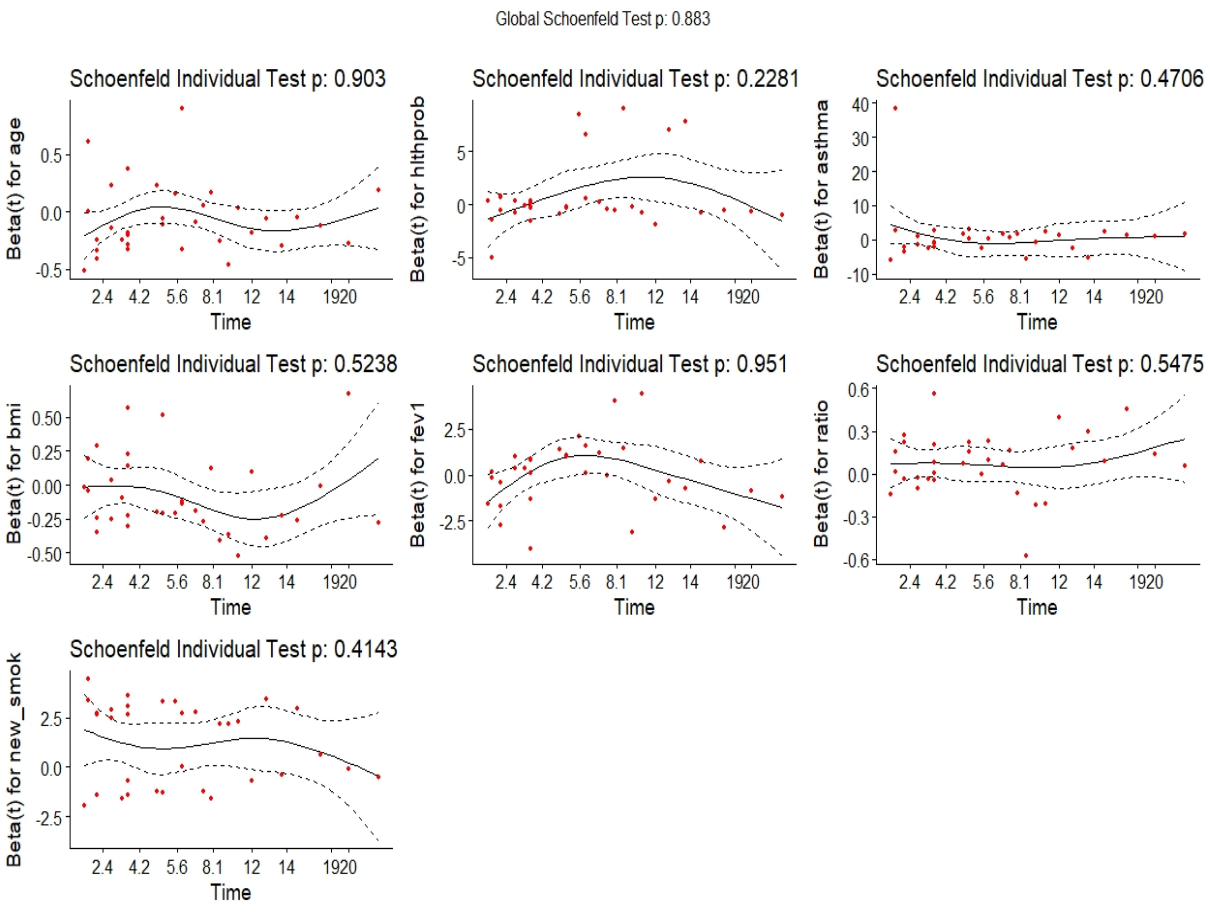


Figure 3.4: Schoenfeld residual plots of all the covariates for the competing event chronic cough or phlegm; dots represent Schoenfeld residuals, the solid line is a smoothing spline fit to the plot, with the dashed lines representing a ± 2-standard-error band around the fit. The assumption of proportional hazards appears to be supported for all the covariates.

To check the validity of the proportional hazards assumption, we consider tests and graphical diagnostics based on the Schoenfeld residuals. A test for each covariate along with a global test for the model as a whole is summarized in Table 3.6. Note that the proportional hazards assumption is supported by a non-significant relationship between residuals and time. We see that (Table 3.6) the test is not statistically significant for any of the covariates, and the global test is also not statistically significant. Therefore, there is no evidence against the proportional hazards assumption for any of the covariates. A graphical diagnostic of the proportional hazards assumption is displayed in Figure 3.4 (the solid line is a smoothing spline fit to the plot, with the dashed lines represent-

ing a ± 2-standard-error band around the fit). Here, systematic departures from a horizontal line are indicative of non-proportional hazards. From the graphical inspection, we see no obvious pattern with time. Thus, the assumption of proportional hazards appears to be supported for all the covariates.

3.4.2 Analysis for Shortness of Breath

Furthermore, we fit the competing risk model for shortness of breath with chronic cough or phlegm and allergy censored. Numerical results are summarized in Table 3.7. We see that history of asthma and smoking are highly significant for shortness of breath (p-value = 0.0481 and 0.024, respectively), whereas age and FEV1/FVC ratio are marginally significant (p-value = 0.0913 and 0.0721, respectively). The estimates of the hazard ratios are given in Table 3.7, and the hazard ratio plots along with 95% confidence intervals are displayed in Figure 3.5. We see that the estimates of the hazard ratios for history of asthma, smoking, age and FEV1/FVC ratio are 5.3932, 2.2289, 0.9363 and 0.9375, respectively, suggesting

- the hazard risk for shortness of breath will increase about 439% for one unit increase of the history of asthma controlling for the other factors (i.e., a positive association between history of asthma and shortness of breath);
- smoking will increase the hazard risk for shortness of breath about 123% controlling for the other factors;
- one year increase in age will reduce the hazard for shortness of breath about 6.4% controlling for the other factors (i.e., a negative association between age and shortness of breath); and
- the hazard risk for shortness of breath will decrease about 6.3% for one unit increase of the FEV1/FVC ratio controlling for the other factors (i.e., a negative association between FEV1/FVC ratio and shortness of breath).

Table 3.7: Competing risk analysis for shortness of breath with chronic cough or phlegm and allergy censored – estimates of the coefficients, hazard ratios, standard errors of the estimates, Wald statistic (z), and p values.

	Estimate	Hazard ratio	Standard error (SE) of the estimate	$z = \frac{\text{estimate}}{\text{SE}}$	Pr(> |z|)
Age (β_{12})	-0.0658	0.9363	0.0390	-1.6880	0.0913
History of health problem (β_{22})	-0.2891	0.7489	0.6581	-0.4390	0.6605
History of asthma (β_{32})	1.6851	5.3932	0.8528	1.9760	0.0481
BMI (β_{42})	0.0481	1.0492	0.0382	1.2600	0.2077
FEV1 (β_{52})	0.3123	1.3665	0.2837	1.1010	0.2710
FEV1/FVC ratio (β_{62})	-0.0646	0.9375	0.0359	-1.7990	0.0721
Smoking (β_{72})	0.8015	2.2289	0.3552	2.2570	0.0240

The hazard plot (3.5) for shortness of breath model consisted with hazard ratio and their 95% confidence interval for all the variables in the model. This graph (3.5) also contains the information of global long-rank test p-value, Akaike Information Criterion (AIC) and concordance index of the model. From the plot (3.5), the variables that are significant have confidence intervals before or after the marginal line which is belongs to 1. Confidence interval for history of asthma and smoking status indicates a strong evidence of being highly significant.

To check the validity of the proportional hazards assumption, we consider tests and graphical diagnostics based on the Schoenfeld residuals. A test for each covariate along with a global test for the model as a whole is summarized in Table 3.8. Note that the proportional hazards assumption is supported by a non-significant relationship between residuals and time. We see that (Table 3.8) the test is not statistically significant for any of the covariates, and the global test is also not statistically significant.

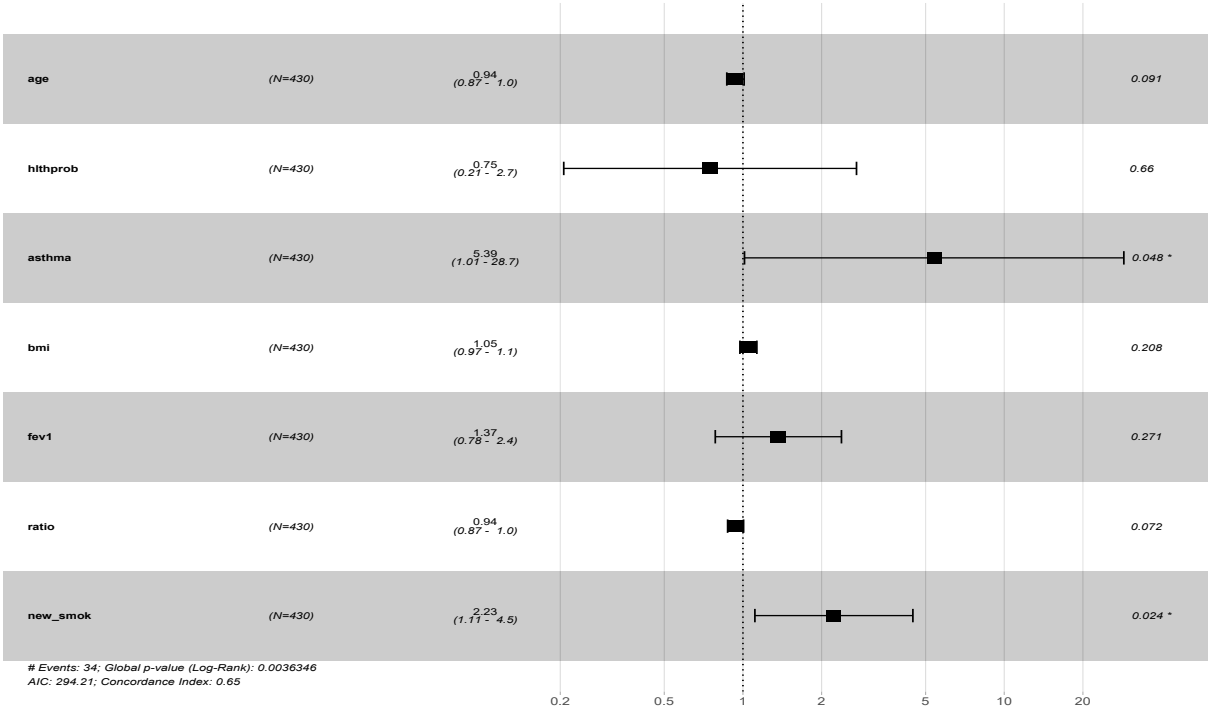


Figure 3.5: Competing risk analysis for shortness of breath – hazard ratio along with a 95% confidence interval for each of the covariates, indicating (a) history of asthma and smoking are highly significant, (b) FEV1/FVC ratio and age are marginally significant, and (c) history of health problem, BMI and FEV1 are not statistically significant.

Table 3.8: Competing event shortness of breath – test for the PH assumption for each covariate, along with a global test of the model as a whole.

	rho	chisq	p
Age (β_{12})	-0.0327	0.0616	0.804
History of health problem (β_{22})	0.1523	0.7492	0.387
History of asthma (β_{32})	-0.0464	0.0597	0.807
BMI (β_{42})	-0.1786	0.8955	0.344
FEV1 (β_{52})	0.1846	1.2647	0.261
FEV1/FVC ratio (β_{62})	-0.0568	0.1387	0.71
Smoking (β_{72})	0.1187	0.4906	0.484
GLOBAL	NA	3.5235	0.833

Therefore, there is no evidence against the proportional hazards assumption for any of the covariates. A graphical diagnostic of the proportional hazards assumption is displayed in Figure 3.6 (the solid line is a smoothing spline fit to the plot, with the dashed lines representing a ± 2-standard-error band around the fit). Here, systematic departures from a horizontal line are indicative of non-proportional hazards. From the graphical inspection, we see no obvious pattern with time. Thus, the assumption of proportional hazards appears to be supported for all the covariates.

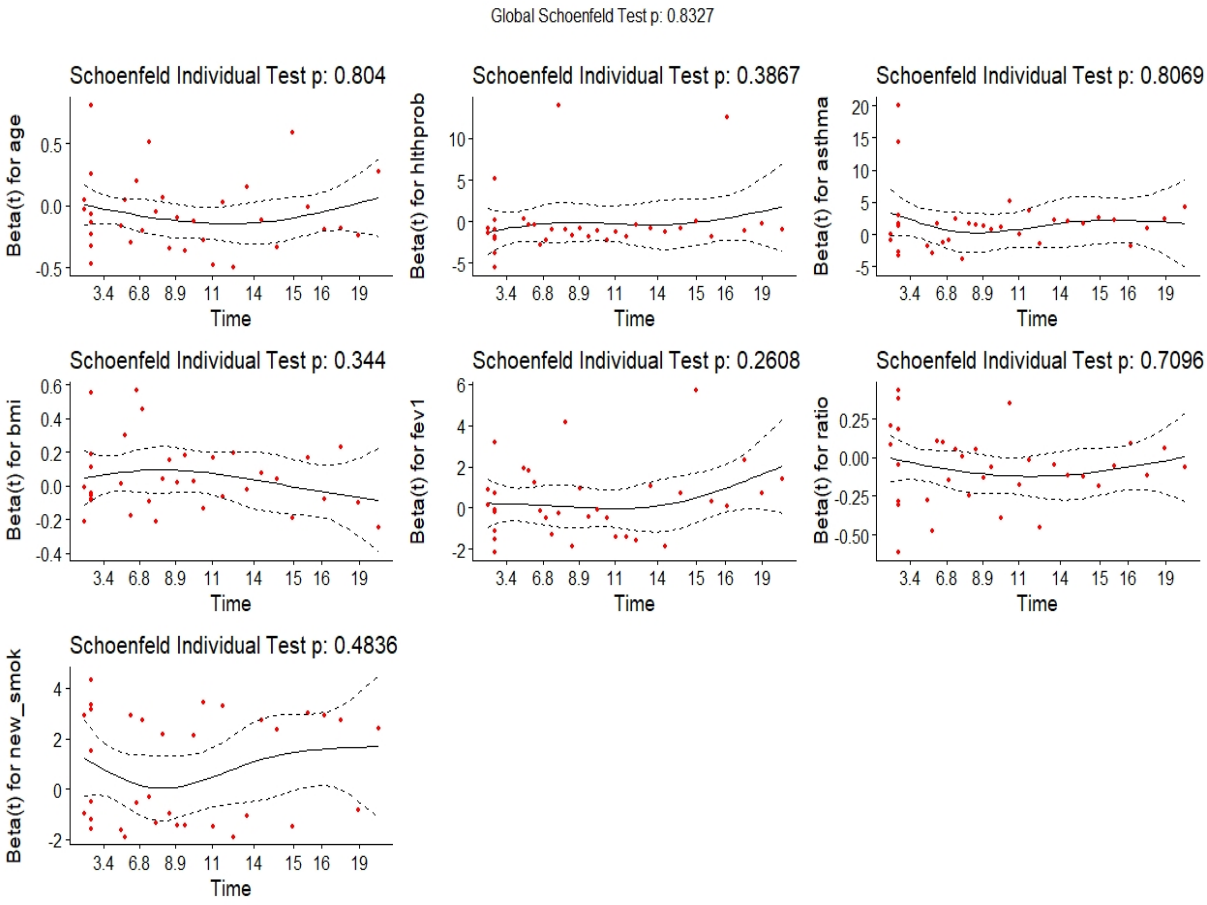


Figure 3.6: Schoenfeld residual plots of all the covariates for the competing event shortness of breath; dots represent Schoenfeld residuals, the solid line is a smoothing spline fit to the plot, with the dashed lines representing a ± 2-standard-error band around the fit. The assumption of proportional hazards appears to be supported for all the covariates.

3.4.3 Analysis of Allergy

Finally, we fit the competing risk model for allergy with chronic cough or phlegm and shortness of breath censored. Numerical results are summarized in Table 3.9. We see that age and FEV1 are highly significant for allergy (p-value = 0.0008 and 0.001, respectively), whereas BMI is marginally significant (p-value = 0.0842). The estimates of the hazard ratios are given in Table 3.9, and the hazard ratio plots along with 95% confidence intervals are displayed in Figure 3.7.

Table 3.9: Competing risk analysis for allergy with chronic cough or phlegm and shortness of breath censored – estimates of the coefficients, hazard ratios, standard errors of the estimates, Wald statistic (z), and p values.

	Estimate	Hazard ratio	Standard error (SE) of the estimate	$z = \frac{\text{estimate}}{\text{SE}}$	$\Pr(>\|z\|)$
Age (β_{13})	-0.1204	0.8865	0.0358	-3.3660	0.0008
History of health problem (β_{23})	0.4318	1.5400	0.4478	0.9640	0.3349
History of asthma (β_{33})	0.6644	1.9434	0.7702	0.8630	0.3883
BMI (β_{43})	0.0565	1.0582	0.0327	1.7270	0.0842
FEV1 (β_{53})	-0.8589	0.4236	0.2610	-3.2910	0.0010
FEV1/FVC ratio (β_{63})	-0.0284	0.9720	0.0302	-0.9410	0.3467
Smoking (β_{73})	-0.4079	0.6650	0.3842	-1.0620	0.2883

We see that the estimates of the hazard ratios for age, FEV1 and BMI are 0.8865, 0.4236 and 1.0582, respectively, suggesting

- one year increase in age will reduce the hazard for allergy about 11.4% controlling for the other factors (i.e., a negative association between age and allergy);
- the hazard risk for allergy will decrease about 58.6% for one unit increase of the FEV1 controlling for the other factors (i.e., a negative association between FEV1 and allergy); and

- body mass index will increase the hazard risk for allergy about 5.8% controlling for the other factors.

The hazard plot (3.7) for allergy model consisted with hazard ratio and their 95% confidence interval for all the variables in the model. This graph (3.7) also contains the information of global long-rank test p-value, Akaike Information Criterion (AIC) and concordance index of the model. From the plot (3.7), the variables that are significant have confidence intervals before or after the marginal line which is belongs to 1. Confidence interval for age and FEV1 indicates a strong evidence of being highly significant.

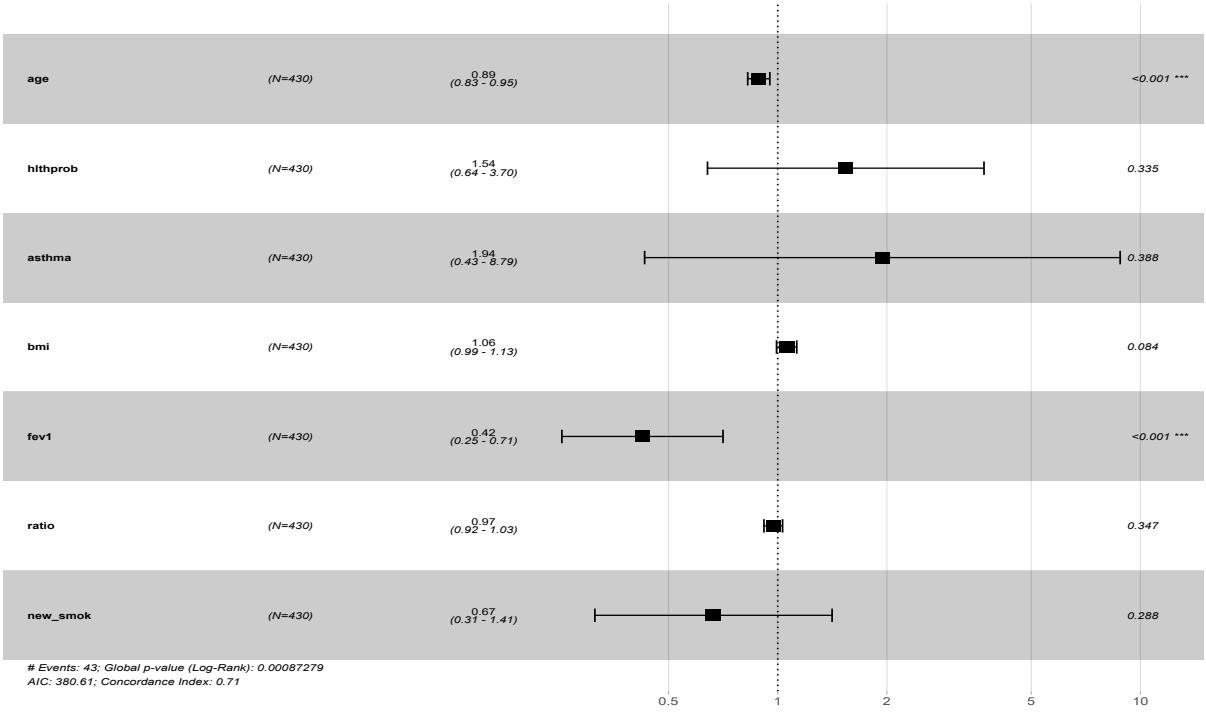


Figure 3.7: Competing risk analysis for allergy – hazard ratio along with a 95% confidence interval for each of the covariates, indicating (a) age and FEV1 are highly significant, (b) BMI is marginally significant, and (c) history of health problem, history of asthma, FEV1/FVC ratio and smoking are not statistically significant.

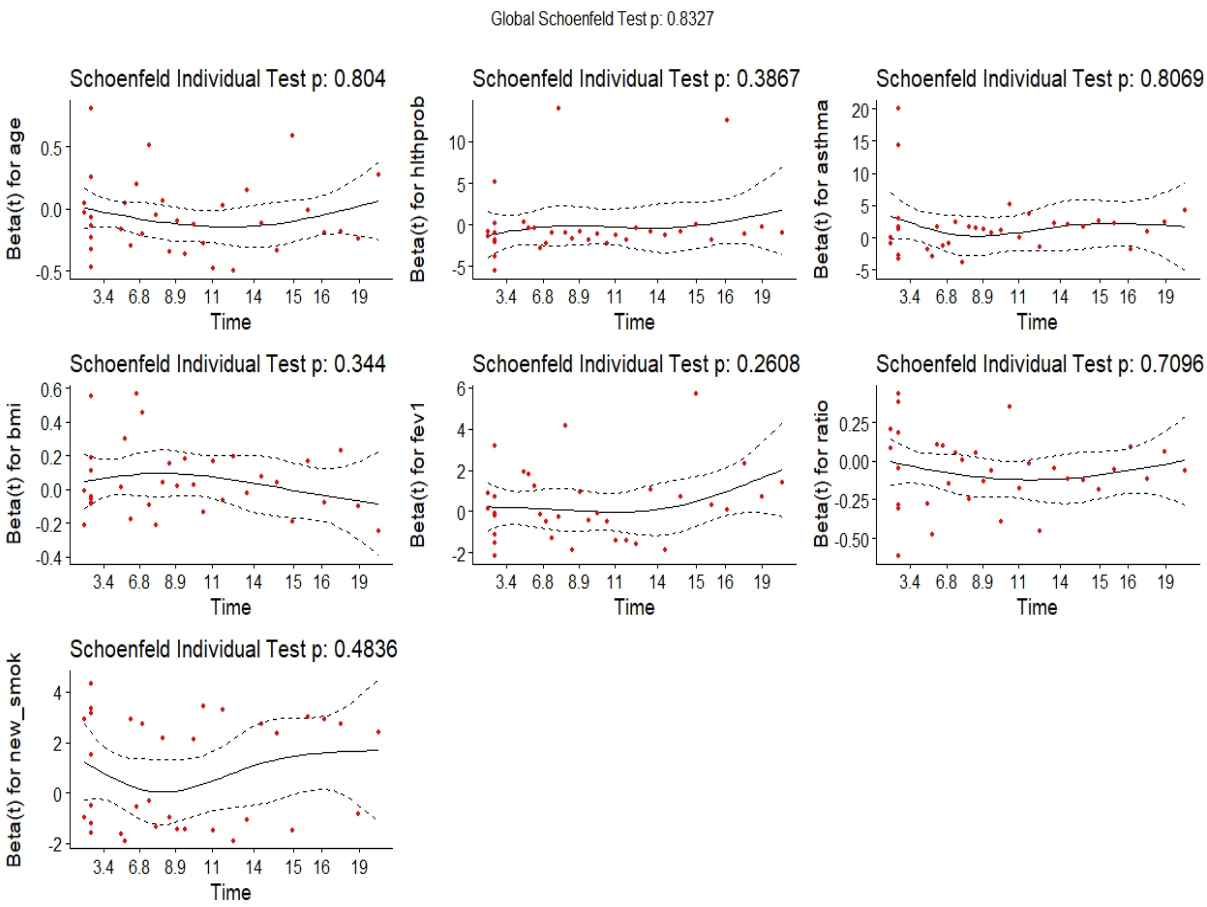


Figure 3.8: Schoenfeld residual plots of all the covariates for the competing event allergy; dots represent Schoenfeld residuals, the solid line is a smoothing spline fit to the plot, with the dashed lines representing a ± 2-standard-error band around the fit. The assumption of proportional hazards appears to be supported for all the covariates.

To check the validity of the proportional hazards assumption, we consider tests and graphical diagnostics based on the Schoenfeld residuals. A test for each covariate along with a global test for the model as a whole is summarized in Table 3.10. Note that the proportional hazards assumption is supported by a non-significant relationship between residuals and time. We see that (Table 3.10) the test is not statistically significant for any of the covariates, and the global test is also not statistically significant. Therefore, there is no evidence against the proportional hazards assumption for any of the covariates. A graphical diagnostic of the proportional hazards assumption is displayed in Figure 3.8 (the solid line is a smoothing spline fit to the plot, with the dashed lines represent-

ing a ± 2-standard-error band around the fit). Here, systematic departures from a horizontal line are indicative of non-proportional hazards. From the graphical inspection, we see no obvious pattern with time. Thus, the assumption of proportional hazards appears to be supported for all the covariates.

Table 3.10: Competing event allergy – test for the PH assumption for each covariate, along with a global test of the model as a whole.

	rho	chisq	p
Age (β_{13})	-0.1878	2.07724	0.15
History of health problem (β_{23})	0.04457	0.08256	0.774
History of asthma (β_{33})	-0.0052	0.0013	0.971
BMI (β_{43})	-0.1392	1.20568	0.272
FEV1 (β_{53})	0.01221	0.00735	0.932
Ratio (β_{63})	0.15181	0.99136	0.319
Smoking (β_{73})	-0.0297	0.03787	0.846
GLOBAL	NA	5.51322	0.598

3.5 Summary

A summary of our competing risk analysis is presented in Table 3.11 Here, we summarize our findings for the grain dust industry workers in Saskatchewan.

- Age is marginally significant for chronic cough or phlegm and shortness of breath, whereas it is highly significant for allergy.It is interesting to see a negative association for age, suggesting that the higher the age, the less hazard to get an event (i.e., younger workers are at high risk compared to older workers). Young workers are exposed to more severe condition compared to older workers. We suspect that young workers may be involve more in field works or severe workplace environment leading to a higher risk of health related hazard.
- History of health problem is not significantly associated with the occurrence of these events.

- History of asthma is found highly significant for shortness of breath. We see that a history of asthma can significantly increase the hazard of shortness of breath.
- BMI is found marginally significant for allergy: the higher the BMI, the more hazard to get allergy.
- FEV1 is highly significant for allergy: the lower the FEV1, the more hazard to get allergy.
- FEV1/FVC ratio is significant for chronic cough or phlegm and shortness of breath. Note that this ratio is positively associated with chronic cough or phlegm, whereas negatively associated with shortness of breath.
- Smoking is significantly associated with the risk of chronic cough or phlegm and shortness of breath. We see a positive association between smoking and the hazards of these two events.

Table 3.11: A summary of the competing risk analysis with competing events chronic cough or phlegm, shortness of breath and allergy – estimates of the coefficients and p-values.

	Chronic cough or phlegm Estimate (p-value)	Shortness of breath Estimate (p-value)	Allergy Estimate (p-value)
Age	−0.067 (0.091)	−0.066 (0.091)	−0.120 (<0.001)
History of health problem	0.729 (0.164)	−0.289 (0.661)	0.432 (0.335)
History of asthma	0.779 (0.485)	1.685 (0.048)	0.664 (0.388)
BMI	−0.074 (0.110)	0.048 (0.208)	0.057 (0.084)
FEV1	0.012 (0.967)	0.312 (0.271)	−0.859 (<0.001)
FEV1/FVC ratio	0.078 (0.024)	−0.065 (0.072)	−0.028 (0.347)
Smoking	1.209 (<0.001)	0.802 (0.024)	−0.408 (0.288)

Chapter 4

Conclusion

Grain dust industry workers are exposed to a number of work-related hazards, including high levels of endotoxin, microorganisms and dust. Multiple studies have reported immunological, toxicological and clinical effects of occupational exposure to grain dust contaminants (e.g., Pahwa et al. (2003), Swan et al. (2007)). The main goal of this study is to investigate the effects of demographic and biological factors on multiple health outcomes for workers who are exposed to grain dust contaminants for a long period of time. The Saskatchewan data under the Grain Dust Medical Surveillance Program are considered, which include demographic and health information of employees from 1978 to 2005. In particular, seven covariates (age, history of health problem, history of asthma, BMI, FEV1, FEV1/FVC ratio and smoking) are considered to investigate their effects on three health outcomes (chronic cough or phlegm, shortness of breath and allergy). Under this setup, a worker can suffer an adverse health outcome in three possible ways, leading to a competing risk problem as illustrated in Figure 3.2. Thus, competing risk survival analysis is carried out to achieve our goal.

Based on our analyses, FEV1/FVC ratio and smoking are highly responsible for chronic cough or phlegm among grain industry workers which also indicate the positive association between FEV1/FVC ratio and smoking with chronic cough or phlegm (Table 3.5). The estimates of the hazard ratios indicate increasing hazard rates for a single unit increase of the FEV1/FVC ratio and smoking. However, estimate of age suggesting a negative association with chronic cough or

phlegm reporting marginally significant. The estimate of hazard ratio for age indicates a low risk of chronic cough or phlegm among young workers compare to elder ones. The hazard ratio plot along with 95% confidence intervals are displayed in Figure 3.3. Validity of the proportional hazard assumption is checked with a test for every covariates as well as a global test and summarized in Table 3.6. It is reported that a non-significant relationship between residuals and time supported the proportional hazard assumption.

Results show us that history of asthma and smoking are highly significant for shortness of breath as per our expectation before the analyses. That is, history of asthma and smoking will rapidly increase the hazard of shortness of breath. The estimates of history of asthma and smoking premise the positive correlation with shortness of breath among workers from grain industries in Saskatchewan. The results also provide negative estimates for age and FEV1/FVC ratio suggesting a higher rate of hazard of getting shortness of breath among young workers with chronic cough or phlegm and allergy as censored. The table 3.8 represented to check the validity of proportional hazard assumption with a test for each covariates considering global test as well. It is shown that a insignificant association between residuals and time supported proportional hazard assumption.

Our final model for allergy suggests two highly significant covariates: age and FEV1 are strongly associated (negative association) with the allegy (Table 3.9). Although, BMI is slightly associated with allegy and has a positive estimated value. The estimates of hazard ratio for age and FEV1 also indicate the correlation with allergy that reduce the rate of hazard with one unit change in age and FEV1 among workers in grain industries in Saskatchewan.

The competing risk analysis involves fitting the Cox PH model separately for each event type, treating the other (competing) event types as censored in addition to those who are censored from loss to follow-up or withdrawal. One of the assumptions of competing risk analysis is that censoring is independent of events regardless of the different ways that censoring can occur, including

failure from competing risks other than the event-type of interest. This assumption can never be explicitly proved for given data. Therefore caution is warranted, given that violation of this assumption may lead to biased estimates. Moreover, our findings are based on a relatively small sample (n=226), and therefore caution should be applied to generalize or extrapolate our findings.

In this analysis, we did not compare our outcomes with any previous research due to some inconsistency between Saskatchewan data and the data set for all five regions. Moreover, we have only used information about grain workers in Saskatchewan but did not consider grain workers and general population from all over Canada for the analysis. General population may also be exposed to grain dust in which a similar competing risk setting could be used. In our future work we could try a multi-state model for the same data set and compare both models and their findings.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726.

Agriculture and Canada, A.-F. (2017). Canadian grains.

Canada, S. (2017). Saskatchewan remains the breadbasket of canada.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Dakers, S. and Fréchette, J.-D. (1998). *The grain industry in Canada*. Parliamentary Research Branch.

Dardis, C. (2018). *survMisc: Miscellaneous Functions for Survival Data*. R package version 0.5.5.

Feakins, B. G., McFadden, E. C., Farmer, A. J., and Stevens, R. J. (2018). Standard and competing risk analysis of the effect of albuminuria on cardiovascular and cancer mortality in patients with type 2 diabetes mellitus. *Diagnostic and prognostic research*, 2(1):13.

Fermanian, J.-D. (2003). Nonparametric estimation of competing risks models with covariates. *Journal of Multivariate Analysis*, 85(1):156–191.

Gillespie, B. (2006). Checking assumptions in the cox proportional hazards regression model. *Midwest SAS Users Group (MWSUG)*.

Gray, B. (2014). *cmprsk: Subdistribution Analysis of Competing Risks*. R package version 2.2-7.

Grolemund, G. and Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25.

Hinchliffe, S. R. and Lambert, P. C. (2013). Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. *BMC medical research methodology*, 13(1):13.

Hosmer Jr, D. W. and Lemeshow, S. (1999). Applied survival analysis: regression modelling of time to event data (1999). *Eur Orthodontic Soc*, pages 561–2.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Kartsonaki, C. (2016). Survival analysis. *Diagnostic Histopathology*, 22(7):263–270.

Kassambara, A., Kosinski, M., and Biecek, P. (2019). *survminer: Drawing Survival Curves using 'ggplot2'*. R package version 0.4.6.

Kleinbaum, D. G. and Klein, M. (2010). *Survival analysis*, volume 3. Springer.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1):27–52.

Noordzij, M., Leffondré, K., van Stralen, K. J., Zoccali, C., Dekker, F. W., and Jager, K. J. (2013). When do we need competing risks methods for survival analysis in nephrology? *Nephrology Dialysis Transplantation*, 28(11):2670–2677.

Pahwa, P., Senthilselvan, A., McDuffie, H. H., and Dosman, J. A. (2003). Longitudinal decline in lung function measurements among saskatchewan grain workers. *Canadian respiratory journal*, 10(3):135–141.

Pintilie, M. (2011). An introduction to competing risks analysis. *Revista Española de Cardiología (English Edition)*, 64(7):599–605.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, pages 541–554.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Swan, J., Blainey, D., and Crook, B. (2007). The hse grain dust study—workers' exposure to grain dust contaminants, immunological and clinical response.(rr540). *Health and Safety Laboratory. Buxton, Derbyshire*.

Tan, K. S., Eguchi, T., and Adusumilli, P. S. (2018). Competing risks and cancer-specific mortality: why it matters. *Oncotarget*, 9(7):7272.

Terry M. Therneau and Patricia M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.

Appendix

This chapter includes R programming codes used for competing risk modeling for health of Saskatchewan.

```
####################For Competing risk model#################
rm(list=ls())
library(survival)
library(survminer)
library(survMisc)
library(cmprsk)
library(lubridate)
#Load data set and remove unneccessary columns and rows#
##########################################################
##########################################################

data<-read.csv("d:/1 MSc thesis/Ms thesis/Thesis Data/Data_Sets/grain-dust.csv",
    header=TRUE)
data<-data[,-4]
data[1:20,]
n.org<-length(unique(data$new_id))
n.org

cou<-as.numeric(data$cge3mo==1 | data$coughyrs>=1 | data$pge3mo==1 |
```

```
    data$phlegyrs>=1)
sum(cou,na.rm=TRUE)

sofb<-as.numeric(data$wwocold==1 | data$sob==1)
sum(sofb,na.rm=TRUE)

aller<-as.numeric(data$allergy==1)
sum(aller,na.rm=TRUE)

event.mat<-cbind(cou,sofb,aller)

n.event<-NULL
for(i in 1:nrow(event.mat)){
if(all(is.na(event.mat[i,]))){
n.event[i]<-NA
} else{
n.event[i]<-sum(event.mat[i,],na.rm=TRUE)
}
}

cbind(event.mat,n.event)

dat1<-data.frame(data,cou=cou,sofb=sofb,aller=aller,n.event=n.event)
dat1[100:150,]

#cbind(dat1$n.event,dat1$new_id)
```

```
# remove unnecessary rows
dat11<-split(dat1,dat1$new_id)
dat12<-NULL
for (i in 1:length(dat11)){
if(nrow(dat11[[i]])==1){
if(!is.na(dat11[[i]]$n.event)){
dat12<-rbind(dat12,dat11[[i]])
}
}
if(nrow(dat11[[i]])>1){
n0.event<-dat11[[i]]$n.event
obs0<-which(n0.event>0)
n0<-length(obs0)
if(n0==0){
dat12<-rbind(dat12,dat11[[i]])
}
if(n0>0){
obs1<-1:obs0[1]
ddat<-dat11[[i]][obs1,]
dat12<-rbind(dat12,ddat)
}
}
}
dat12[1:20,]

sum(dat12$cou,na.rm=TRUE)
```

```
sum(dat12$sofb,na.rm=TRUE)
sum(dat12$aller,na.rm=TRUE)


# identify id's with overlapping events
id0<-dat12$new_id[dat12$n.event>1]
# remove these id's from the data set
dat13<-dat12[!(dat12$new_id %in% id0),]
dat13[1:20,]
n<-length(unique(dat13$new_id))
n


dat13$cou[is.na(dat13$cou)]<-0
dat13$sofb[is.na(dat13$sofb)]<-0
dat13$aller[is.na(dat13$aller)]<-0

sum(dat13$cou)
sum(dat13$sofb)
sum(dat13$aller)

# Find survival times and create a data set
# with a single row for each subject
dat21<-split(dat13,dat13$new_id)
dat22<-NULL
for(i in 1:length(dat21)){
```

```
if(nrow(dat21[[i]])==1){
stop<-dat21[[i]]$yrsind
start<-0
id<-i
}
if(nrow(dat21[[i]])>1){
dd<-ISOdate(dat21[[i]]$ydot,dat21[[i]]$mdot,dat21[[i]]$ddot)
st0<-time_length(diff(dd),"years")
stop<-cumsum(c(dat21[[i]]$yrsind[1],st0))
start<-c(0,stop[-length(stop)])
id<-rep(i,nrow(dat21[[i]]))
}
dat22<-rbind(dat22,cbind(dat21[[i]],start=start,stop=stop,id=id))
}
dat22[1:100,]
rownames(dat22)<-NULL
sum(dat22$cou)
sum(dat22$sofb)
sum(dat22$aller)
length(unique(dat22$id))

# Check missing values in the covariates
dat3<-dat22
row1<-which(is.na(dat3$sex))
row2<-which(is.na(dat3$age))
row3<-which(is.na(dat3$hlthprob))
```

```
row4<-which(is.na(dat3$asthma))
row5<-which(is.na(dat3$height))
row6<-which(is.na(dat3$weight))
row7<-which(is.na(dat3$fev1))
row8<-which(is.na(dat3$fvc))
row9<-which(is.na(dat3$new_smok))

missing.row<-unique(c(row1,row2,row3,row4,row5,row6,row7,row8,row9))
# Find the corresponding ID's
missing.id<-sort(unique(dat3$id[missing.row]))
# Remove these from the data
dat4<-dat3[which(as.numeric(dat3$id %in% missing.id)==0),]
rownames(dat4)<-NULL
# Double check
row1<-which(is.na(dat4$sex))
row2<-which(is.na(dat4$age))
row3<-which(is.na(dat4$hlthprob))
row4<-which(is.na(dat4$asthma))
row5<-which(is.na(dat4$height))
row6<-which(is.na(dat4$weight))
row7<-which(is.na(dat4$fev1))
row8<-which(is.na(dat4$fvc))
row9<-which(is.na(dat4$new_smok))
missing.row<-unique(c(row1,row2,row3,row4,row5,row6,row7,row8,row9))
#missing.row<-unique(c(row2,row3,row4,row5,row6,row7,row8,row9))
```

```
sum(dat4$cou)
sum(dat4$sofb)
sum(dat4$aller)
length(unique(dat4$id))

#####################################################
#####################################################
#####################################################
final.dat<-dat4
final.dat$sex<-ifelse(final.dat$sex==1,1,0)
final.dat$hlthprob<-ifelse(final.dat$hlthprob==1,1,0)
final.dat$asthma<-ifelse(final.dat$asthma==1,1,0)
final.dat$new_smok<-ifelse(final.dat$new_smok==1,1,0)
final.dat$bmi<-dat4$weight/((dat4$height/100)^2)

#####################################################
### Descrptuve statisics mean and standard deviation ###
library(dplyr)
a<-final.dat %>% group_by(id) %>% slice((1))

mean(a$age)
sd(a$age)
mean(a$bmi)
sd(a$bmi)
mean(a$ratio)
sd(a$ratio)
```

```
mean(a$fev1)
sd(a$fev1)
mean(a$fvc)
sd(a$fvc)
table(a$hlthprob)
table(a$asthma)
table(a$new_smok)
################################################
########### Analysis Of Chronic Cough ############
fit.cou<-coxph(Surv(start,stop,cou)~age+hlthprob+asthma+bmi+fev1+
        ratio+new_smok,data=final.dat)
fit.cou
summary(fit.cou)
cox.zph(fit.cou)
ggcoxzph(cox.zph(fit.cou))
ggforest(fit.cou, final.dat)

########### Analysis Of Chronic Shortness Of Breath ############
fit.sofb<-coxph(Surv(start,stop,sofb)~age+hlthprob+asthma+bmi+fev1+
        ratio+new_smok,data=final.dat)
fit.sofb
summary(fit.sofb)
cox.zph(fit.sofb)
ggcoxzph(cox.zph(fit.sofb))
ggforest(fit.sofb, final.dat)
```

```
########### Analysis Of Allergy ############
fit.aller<-coxph(Surv(start,stop,aller)~age+hlthprob+asthma+bmi+fev1+
        ratio+new_smok,data=final.dat)
fit.aller
summary(fit.aller)
cox.zph(fit.aller)
ggcoxzph(cox.zph(fit.aller))
ggforest(fit.aller, final.dat)

#############################################################
#############################################################
#############################################################
final.dat$status[final.dat$cou==0 & final.dat$sofb==0
                & final.dat$aller==0]<-0
final.dat$status[final.dat$cou==1 & final.dat$sofb==0
                & final.dat$aller==0]<-1
final.dat$status[final.dat$cou==0 & final.dat$sofb==1
                & final.dat$aller==0]<-2
final.dat$status[final.dat$cou==0 & final.dat$sofb==0
                & final.dat$aller==1]<-3

############Cumulative IncidenCe Curves############
fit<-cuminc(ftime = c(final.dat$start,final.dat$stop),
            fstatus = final.dat$status,cencode=0)

plot(fit,main="Cumulative Incidence Curve", curvlab
```

```
=c("Chronic Cough", "Chronic Shortness of Breath",
 "Allergy"), ylim=c(0, 1), wh=2, xlab="Years",
  ylab="Probability", lty=1:length(fit),
  color=1, lwd=par('lwd'))
```