Additive and Multiplicative hazards Regression Models In Competing Risks Analysis: Application To The Canadian Heart Health Survey

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 Collaborative Biostatistics Program of School of Public Health

University of Saskatchewan

Saskatoon

By

Temitope Adesina

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:

Chair of the Collaborative Biostatistics Program
E Wing Health Sciences
104 Clinic Place
University of Saskatchewan
Saskatoon SK S7N 5E5 Canada
OR
Dean College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 – 110 Science Place Saskatoon SK S7N 5C9

Abstract

Background: In survival analysis, an event whose occurrence influences the occurrence of another event is termed a competing risk event. The Cox hazards model is applicable in standard survival analysis with a single event. To correctly assess covariate effects in competing risks analysis, the Fine & Gray (F-G) subdistribution hazards and the Cox cause-specific hazards models are appropriate. Equally, additive hazards models can be used to examine the covariate effects in a competing risks framework.

Objectives: (i) To examine the additive and multiplicative hazards models in the competing risks setting by applying the said models to the Canadian Heart Health Survey data; (ii) To determine the risk factors for cardiovascular disease using the competing risks approach; (iii) To compare the risk factors identified by the additive and multiplicative hazards models in the context of competing risks.

Methods: The observational Canadian Heart Health Survey database collected between 1986 and 1995 is the baseline data used in this study. Two competing outcomes, cardiovascular disease (CVD) and non-CVD-related deaths, are analyzed with the Cox cause-specific and the F-G multiplicative hazards models. Similarly, the additive hazards models of Aalen and that of Lin & Ying (L-Y) are modeled for the outcomes using the competing risks approach.

Results: There were 13,996 eligible subjects in the data, and 7,071 (50.5%) of them were women. After a median follow-up time of 15 years (interquartile range = 5.52 years), a total of 1,536 deaths were observed, and 549 (35.7%) of these were CVD related deaths. Factors like male gender, old age, and alcohol abstinence significantly increased the risk of CVD mortality in the additive and multiplicative hazards models. Former alcohol users compared to current alcohol users have a 53% (P-value= 0.002) and a 55% (P-value= 0.001) increased risk of CVD mortality in the Cox cause-specific and the F-G models, respectively. In the L-Y additive model, former alcohol users compared to current users increased CVD mortality by adding 16 new cases per 10,000 person-years (P-value = 0.008).

Conclusion: The results from this study suggest that covariate effects in the Cox cause-specific and the F-G subdistribution hazards models may be identical in terms of magnitude and direction. The numerical results from the multiplicative and the additive hazards models give different interpretation of the covariate effects, and using both the additive and multiplicative models together would boost understanding of the data.

Acknowledgements

My profound gratitude goes to my supervisor, Dr. Hyun June Lim for the guidance, assistance, and advises she provided throughout my study. I appreciate her time and mentoring.

My thanks also go to Dr. Bruce Reeder for allowing us access to the data and for his professional feedback and encouragement at all times. Despite being retired, he was never tired of sharing his professional knowledge. I would also like to take this opportunity to thank Dr. Micheal Szafron for his constant assistance and advice. I thank Dr. Cindy Feng for her time and constructive feedback.

I am grateful to the Biostatistics program chair, Dr. Punam Pahwa, for her contributions towards the completion of my degree. I am indebted to the School of Public Health for the financial support, Vicky Duncan, the school's librarian, Heather McWhinney, Dr. Chel Lee, and Dr. Prosanta Mondal for their support.

I would like to extend my deepest appreciation to my awesome parents, Prof and Mrs. A. L. Adesina, I am honored to be one of your offspring. Thank you for everything. You are the best! Thanks to my fabulous siblings for their encouragement over the years. I would also like to thank my extended family members and friends for their support and prayers for the success of this research. Last but not least, I appreciate my soulmate, Jubril Bello and my lovable sons, AbdurRahman and AbdusSalaam, for their understanding, care, and affection.

To the Owner of My Soul, the Best of Planners, the Perfectly Wise, the Most Knowledgeable, the First, and the Last.

Contents

List of Tables

List of Figures

List of Abbreviations

AIDS	Acquired Immune Deficiency Syndrome
APAC	Asia-Pacific
ASCVD	Atherosclerotic Cardiovascular Disease
BMI	Body Mass Index
CHD	Coronary Heart Disease
CHHS	Canadian Heart Health Survey
CIF	Cumulative Incidence Function
CMDB	Canadian Mortality Database
CSH	Cause-Specific Hazard
CVD	Cardiovascular Disease
ESRD	End Stage Renal Disease
F-G	Fine and Gray
HDL	High-Density Lipoprotein
HIV	Human ImmunoDeficiency Virus
ICD	International Classification of Diseases
IPCW	Inverse Probability of Censoring Weighting
K-M	Kaplan-Meier
L-M	Lunn and McNeil
L-Y	Lin and Ying
MI	Myocardial Infarction
M-S	McKeague and Sasieni
PH	Proportional Hazards
PY	Person Years
SCD	Sudden Cardiac Death
SH	Subdistribution Hazard
S-Z	Scheike and Zhang

Chapter 1

Introduction

This chapter introduces the concept of competing risks events, and the analytic procedures relevant to competing risks data and analysis. It also explores the non-parametric approach and modeling methods used for competing risks data. It additionally presents the motivation and objectives of this study, along with the organization of the thesis.

1.1 Background

Survival analysis, also known as failure-time analysis and time-to-event analysis, is an approach used to analyze particular kinds of data characterized by follow-up from a precise start-point to the occasion of a specified event (an event of interest) or study end-point [1]. The event of interest is any event related to the objective of the survival study, for example, death, disease episode, disease relapse, re-hospitalization, or re-arrest. The duration between the start-point and the occurrence of the specified event is the survival time, which can be measured in years, months, weeks, or days [1]. In most survival studies, information concerning the survival time for some of the study subjects is incomplete because they did not experience the specified event during the study. The subject that does not witness the specified event is said to be censored. Once a subject is censored, the particular time that such a subject witnesses the specified event is not known [1]. Usually, censoring can occur as a result of study completion, a competing risk event (event of non-interest) or lost to follow-up [1, 2]. In standard survival analysis, censoring is assumed to be non-informative. This assumption implies that no relationship exists between the censoring time and the survival time of the censored subjects [1].

The standard survival analysis has a single event. However, there may be multiple events in a survival time data [2, 3]. For instance, if a research objective is to analyze the time to cardiac death in a cohort of patients, in reality, some subjects will die of causes that are not cardiac-related. So, in this case, there are more than one possible events (cardiac and non-cardiac deaths) in the study data. Although deaths from non-cardiac causes are not important based on the study objective, the occurrence of a non-cardiac death will prevent the observation of a cardiac death. Thus, non-cardiac death is a competing risk event for cardiac death [3]. Alternatively, a competing risk event may modify the chance of witnessing the event of interest [3]. For example, when a cohort of breast cancer patients are followed for a local recurrence, a managed distant recurrence before the event of interest will alter the probability of experiencing the local recurrence [3].

In summary, the episode of a competing risk event changes the likelihood of the occurrence of the event of interest or prevents the occurrence of the said event [4, 5]. Competing risks data in which subjects are at risk of multiple possible outcomes are a common feature in medical research; the analysis of such data need more than the standard survival analysis [3]. There could be more than one competing risk event in a dataset, or the many competing risk events could be gathered as one event. When competing risk events are present in survival time data, the competing risk events may either be censored or accounted for during the data analysis, depending on the research questions [3].

1.2 Non-Parametric Procedure for Competing Risks Data

In the analysis of survival time data, the non-parametric method for the estimation of survival probability is the Kaplan-Meier (K-M), otherwise known as the product-limit method [1]. The K-M method assumes non-informative censoring; that if follow-up is elongated, those whose observations are incomplete (censored) will eventually witness the event of interest [5, 6]. This approach is intended for use in standard survival studies where only a single failure is present [7]. In such case, the K-M method offers valid solutions with direct interpretation and graphical representation of the survival estimates. The complement of the K-M survival estimate is the probability of occurrence of the event [5]. In the competing risks framework, the complement of the K-M is interpreted as the probability of the event of interest, given the non-existence of the competing risk event [5]. However, such interpretation is misleading in some occasions [4, 8].

As an alternative, an approach that takes the competing event into consideration has been recommended [6]. The cumulative incidence function (CIF) [2, 6, 9], the subdistribution [10], or the cumulative incidence probability [9] are some common names of the non-parametric method used in estimating the failure probability for competing risks data. For the CIF, the probability of an event occurrence is split into that of the event of interest and the (competing) event of non-interest. The CIF allows for valid analysis of the events without assuming independence of the different types of events [6]. When the complement of the K-M survival estimate is employed for competing risks data, the probability of failing from the event of interest is biased upward, since the risk set does not include the competing risk events [5, 6]. However, when there are no competing risk events, the complement of the K-M and the CIF produce the same estimate of the failure probability [5, 6, 11].

1.3 Regression Models for Competing Risks Data

When competing risk events are present in survival time data, regression models that pinpoint and outline the association of treatment or patient characteristics on the multiple competing outcomes are usually employed [6]. These models are used because they consider the "time-until-first-event" and the "type-of-first-event" [12]. Multiplicative hazards regression models and additive hazards regression models are two sets of models utilized in connecting the impact of a covariate to the hazard function [13].

1.3.1 Multiplicative Hazards Regression Models

Multiplicative hazards models are famous for survival analysis due to their accessibility in software programs and simple interpretability. In multiplicative models, the impact of a covariate is multiplicative on the unknown baseline hazard rate. Hence, the hazard ratio is employed to elucidate the covariate effects within the multiplicative hazards models [13].

Two important types of hazards are often modeled for competing risks data. These are: the cause-specific hazard and the subdistribution hazard [3]. The cause-specific hazard (CSH) is the characterization of the instantaneous failure rate when the competing risk event is censored [2]. On the other hand, the subdistribution hazard (SH) is the failure rate when the competing risk event is not censored [6]. Research which aims at learning the biological influence of the covariates on the outcome of interest offers appropriate results by censoring the events of non-interest through modeling of the CSH [3]. On the other hand, a prognostic analysis that predicts a subject's risk for the outcome of interest or resource allocation requires the non-censoring of the competing events. This type of analysis can be achieved by modeling the SH [3, 6, 14–16].

The Cox multiplicative hazards model projects itself as a convenient methodology for survival analysis [13]. In the context of competing risks, two novel versions of the Cox hazards model are often used: the Cox cause-specific hazards (CSH) model and the Fine & Gray (F-G) subdistribution hazards (SH) model [10]. Even though the Cox CSH and the F-G SH models are both proportional hazards models, the F-G model distinguishes itself from the Cox CSH model by modeling the covariate impact directly on the CIF [10, 17].

When the Cox model is employed for the cause-specific hazards and the all-cause hazard, the solutions are inconsistent, because covariate effect on the cause-specific hazards does not add up to the covariate effect on the all-cause hazard. Alternatively, the additive hazards regression models were suggested for competing risks data analysis to resolve the problem of inconsistent solutions and produce coherent results [17, 18].

1.3.2 Additive Hazards Regression Models

The class of additive hazards models provide another form of association between the regression function of covariates and the baseline hazard function [19]. Contrary to the multiplicative hazards models, the additive hazards models fit the hazard function as a sum of the covariates function and the baseline hazard function [19]. The additive hazards models estimate the absolute effect of a covariate on the hazard, instead of its relative effect [13].

Researchers rarely use the additive hazards models for survival analysis due to their restricted analytical computer programs and solutions [17]. Regardless, the idea of risk additivity has been favored for modeling the competing risks data because the sum of a covariate effect on the event of interest and the competing risk event equals the covariate effect on the all-cause hazard in the additive hazards models [17]. Several authors have advocated the different versions of the additive hazards models for analyzing competing risks data [17, 18, 20, 21].

1.4 Motivation and Objectives of Study

The additive hazards models separate the overall effect of a covariate into that from the competing events. Thus, they are the preferred choice for the analysis of competing risks data [17]. Besides, from the public health point of view, the estimate of the regression coefficient from the additive hazards model is more meaningful in describing the relationship between risk factors and disease outcomes [22, 23].

To my knowledge, no study has investigated the risk factors for cardiovascular disease (CVD) through the competing risks framework within the Canadian population. Moreover, no study has used the additive hazards models for the competing risks analysis of CVD, hence my interest in this research. The objectives of my study are as follows:

(i) To examine the additive and multiplicative hazards models in a competing risks setting by applying said models to the Canadian Heart Health Survey data
(ii) To determine the risk factors for cardiovascular disease using the competing risks approach
(iii) To compare the risk factors identified by the additive and multiplicative hazards models in the context of competing risks.

To address these research objectives, I examined multiplicative and additive hazards regression models. For the multiplicative hazards regression models, I considered the Cox cause-specific hazards (CSH) model and the Fine & Gray (F-G) subdistribution hazards model. Furthermore, under the additive hazards regression models, I examined Aalen additive model and the Lin & Ying (L-Y) additive model. I applied the proposed four hazards regression models (Cox CSH, F-G, Aalen, and L-Y) to real cardiovascular disease dataset and examined the risks factors for the disease. The dataset included causes of death, and all deaths that were not related to CVD were grouped together as non-CVD deaths. When CVD death was modeled as the event of interest, I consider non-CVD death as a competing risk event, and vice versa.

1.5 Thesis Structure

This thesis contains six chapters. In Chapter 1, I introduce the statistical and biomedical basis of this research work. Chapter 2 includes literature reviews on the Cox multiplicative model, F-G multiplicative model, Aalen additive model, and the Lin & Ying additive model. Also in Chapter 2, I review the epidemiology of cardiovascular disease. In Chapter 3, I discuss the multiplicative and the additive hazards models' techniques within the competing risks setting. In Chapter 4, the techniques mentioned in Chapter 3 are applied to the Canadian Heart Health follow-up study. The discussion of the results is presented in Chapter 5, together with the study limitations and strengths. Finally, in Chapter 6, I include comments and explain the significance of this study to medical research, and suggest future research.

Chapter 2
Literature Review

This chapter contains a review of popular multiplicative hazards regression models used in survival analysis. I describe the extension of the Cox hazards regression model for competing risks analysis. In addition, I review the additive hazards models proposed by Aalen and the Lin & Ying, and their application to competing risks data. This chapter concludes with a brief discussion of the epidemiology and risks factors associated with CVD.

2.1 Multiplicative Hazards Regression Models

A notable feature of multiplicative hazards models is that when covariates are fixed at time zero, the hazard rate of two subjects with separate covariate values is assumed to be independent of time translating to an unchanging covariate effect for the entire follow-up time [24]. The Cox hazards model and its extensions: the Cox cause-specific hazards model and the F-G subdistribution hazards model are discussed in this section.

2.1.1 Cox Hazards Model

The Cox hazards model is the most popular model used to express the impact of covariates on the survival function [9]. In 1972, Sir David Cox proposed the regression model, which allows constant covariates effect to act in a multiplicative manner on an unspecified baseline hazard function [25]. In the Cox model, no assumption is made about the distribution of the baseline hazard function, but the effect of the predictor variables are assumed to be constant. Hence, the Cox model is called a semi-parametric model [13]. David Cox maximized the partial-likelihood approach to estimate his model's regression coefficients, whose exponent is interpreted as hazard ratio or relative risk. The hazard ratio in the Cox model is assumed to be independent of time, so, the model is called proportional hazards (PH) model [1, 9, 13].

2.1.1.1 Cox Cause-Specific Hazards Model

Regression models based on the cause-specific hazards (CSH) are one of the familiar hazards-based approaches to modeling covariate impact in survival data when competing risk events are present [24]. The cause-specific hazard measures the probability of experiencing an event given that no prior event has occurred [2, 24]. The Cox CSH model, akin to the Cox PH model assumes constant covariate effect through the follow-up time, and also employs the partial-likelihood method for the parameter estimation [1].

The Cox cause-specific model has been implemented in series of competing risks studies [26–30]. Particularly in cardiovascular study, research has assessed how cancer and cardiac events interact in patients who received percutaneous coronary intervention (PCI) [31]. In the PCI study, the event of interest was any cardiovascular event that occurred after PCI initiation among the patients. Non-cardiac death was censored in the Cox cardiac-specific hazards model [31]. In another CVD study, Cox cardiac-specific model was employed to examine the effect of cardiorespiratory wellness on the lifetime risk of CVD mortality. The authors conducted a separate analysis where non-cardiac-specific death was modeled with the Cox CSH approach. For both analyses, the event not being modeled was treated as a censored observation [32].

Conventional statistical softwares like R [27], SAS [26–29, 31], and STATA [30, 31] were used in fitting the Cox cause-specific hazards model.

2.1.1.2 Fine & Gray Subdistribution Hazards Model

When the CSH is modeled, the influence of the covariate does not automatically mirror its impact on the CIF. In extreme cases, covariates have an effect on the CSH and no effect on the CIF [33]. A multiplicative competing risks regression model was proposed, which focuses on examining the effect of covariates directly on the CIF [10]. Initially, Gray (1988) developed nonparametric tests to differentiate the CIF of a specific failure among diverse groups [33]. The Gray's test was further studied in the subdistribution model, where covariate effects are linked to the CIF under the proportional hazard assumption [10]. F-G showed that the partial likelihood approach could be used to estimate the regression parameters when all subjects in the data experience an event (complete data). However, with incomplete data when individuals are right-censored, F-G used the inverse probability censoring weighting (IPCW) approach to derive consistent estimates of the regression parameters from the weighted score function [10]. Furthermore, F-G compared the results from their subdistribution model and the Cox cause-specific model using a real dataset, but found that the parameter estimates in both models were similar. Despite obtaining similar covariate effects on the Cox CSH and the subdistribution models, F-G insisted that it is unconventional to test covariate effects directly on the CIF within the cause-specific analysis [10].

With modifications to allow for the competing events to partake in the subdistribution analysis, the F-G model has been implemented for numerous diseases [34–43]. Specifically, in CVD studies, the F-G model was used in the Rotterdam cohort to investigate the lifetime risk of CVD morbidity and mortality in both sexes. In order to determine the risk of the event of interest (CVD), non-CVD death was recognized as a competing risk event [41]. In a Norwegian cohort of patients treated for coronary artery disease (CAD), some researchers evaluated the impact of the competing non-cardiac event when analyzing treatment effects on cardiovascular outcomes [42]. Those researchers reiterated that the Cox CSH model is not appropriate for predicting the chance of cardiac events that may be precluded by non-cardiac outcomes; therefore, they used the F-G model to avoid misinterpreting their study results [42]. Another cohort study in Beijing examined the predictors of CVD mortality in the competing risks analysis of older adults [43]. In the study, death from cerebrovascular disease, cancer, and other diseases were considered as competing risk events using the subdistribution approach proposed by F-G [43].

Generally, multiplicative hazards models rely on the proportional hazard assumption [44]. This proportionality assumption does not always hold for the survival time data. In fact, when covariates in the survival data have time-dependent effects on the hazard rate, the underlying assumption of proportional hazard for the Cox-type model is violated [44].

Similar to the Cox cause-specific hazards model, popular computer programs including R [38–43], SAS [34], STATA [35–37], and SPSS [41] have been employed for the F-G subdistribution hazards model.

2.1.2 Comparison of Cox CSH and F-G SH Models

The Cox CSH and the F-G SH models are popularly used for competing risks data analysis [45–53]. This duo assumes proportionality of hazard, yet, they specialize in distinct measures of the competing risks data [3, 24]. In the SH approach, individuals who witness the event of non-interest stay within the risk set. In distinction, such persons exit the risk set in the cause-specific analysis [3].

Results of the Cox CSH and the F-G SH models were compared in series of competing risks studies using a single dataset [45–51]. For CVD, Rotterdam cohort was analyzed with the standard Cox, the Cox CSH, and also the F-G SH models [52]. The study compared estimates from the three models in the prediction of coronary heart disease (CHD) outcome in the presence of non-CHD deaths [52]. A prospective study among Americans forecasted the chance of sudden cardiac death (SCD) among hemodialysis patients using the subdistribution mode [53]. Covariates associated with SCD, non-SCD, and non-cardiac deaths were identified in separate analysis censoring the events of non-interest in each of the analysis through the CSH model of Cox [53]. In a few of the studies, the results from the Cox CSH and the F-G SH models were similar [47, 50, 52]. On the other hand, other authors concluded that estimates from the two models disagree in direction, and in certain cases, the covariates selected in each model were different [45, 46, 51, 53]. The preference for the CSH model or the SH model relies upon the study objectives and the researcher's interest since each model performs a definite task and reaches a peculiar conclusion [37, 54, 55].

2.2 Additive Hazards Regression Models

Additive hazards models relate the connection between the failure-time and the impact of the covariates in terms of risk difference [19]. Although these models offer a different interpretation of the covariate effects on the hazard function, a few survival data has been analyzed using additive hazards models. However, the additive hazards models are approved for modeling competing risks data because they give consistent solution for such data [17].

One reason why additive hazards models are seldom used is the limited computer programs for their analysis [17]. Only recently have researchers created packages within the R software system or written SAS macro language for the additive hazards models [18]. Another weakness of the additive model is that the estimated hazard rate may not be positive; if a person has extreme covariate effects, a negative hazard rate may be obtained [17]. In this section, I focus on the nonparametric Aalen and the semiparametric Lin & Ying additive hazards models.

2.2.1 Aalen Additive Hazards Model

Aalen introduced a nonparametric additive hazards model, the aftermath of the multiplicative model he studied in 1978. In the Aalen's additive hazards model, both the baseline hazards function and the regression functions depend on time [56–58]. Although the direct estimation of the regression function in Aalen's model is onerous, the cumulative regression coefficient is easily obtained through the least squares approach. Thus, valuable information regarding the covariate effect on survival is deduced from the slope of the plots of the cumulative regression functions against time. An increasing slope implies an increasing additive covariate impact whereas, a decreasing slope suggests a decreasing additive impact [13].

Aalen (1993) developed his nonparametric regression model further to permit the fitness of the model to be checked by martingale residuals [58]. He also used bootstrap replications to examine the features of the regression plots that were reflecting real phenomena. Additionally, Aalen studied the kernel estimation of the regression functions so that facts regarding the covariates may be inferred directly from the regression functions instead of the cumulative regression plots [58]. Since Aalen's additive model is nonparametric, it is not fully developed for the purpose of making an inference. Consequently, the influence of the covariates in a dataset become unnecessarily complicated to report. [9, 12, 59]. Another limitation of the Aalen model is that it handles only a few number of covariates [59]. Aalen's model cannot be directly analyzed in SAS but SAS macros and R-package *timereg* can be used to fit the model.

2.2.2 Lin & Ying Additive Hazards Model

Even though a variety of the additive hazards models were developed and recommended by many authors [56–60], none of these authors employed the semiparametric method for the estimation of the regression coefficients [19]. Lin & Ying (L-Y) (1994) observed this gap, and suggested an additive hazards model that is identical to the semiparametric multiplicative hazards model of Cox [19]. L-Y substituted the time-varying regression functions in the Aalen model with parameters that do not depend on time. In that paper, L-Y provided a precise semiparametric estimator for the regression coefficients that imitates the approach employed within the Cox model. Their approach resulted in consistent and asymptotically normal solutions with a simple covariance estimator [19]. Also, L-Y estimated the cumulative baseline hazard function for their semiparametric additive model, which mimicked the Breslow's method in the multiplicative Cox model, but the regression estimates and variances in the L-Y model have an explicit formula. The L-Y additive model, just like the Cox model, contains covariates that have a constant effect on the hazard function [19]. SAS macros and R-package *ahaz* can be used to fit the L-Y model.

2.3 Additive vs Multiplicative Hazards Model

The additive and multiplicative hazards models present a diverse relationship between the hazard function and the covariates [19, 61]. Although multiplicative hazards models are very popular, additive hazards models provide more flexibility in dealing with survival-time data, and in public health, it is of high importance to understand the additive effect of a covariate well-above its relative effect [23]. Some authors suggested that both models should be examined complementarily so as to have a better understanding of the covariates impact on the hazard function [61–63]. Studies are emerging in which both the additive and multiplicative models are likened in the competing risks framework [17, 20]. Klein investigated the effect of covariates on standard competing events in cancer study using Cox multiplicative CSH model, Aalen, and L-Y additive hazards models [17]. He maintained that even though the multiplicative hazards model is theoretically sound, it may be absurd in real life competing risks setting. Hence, he recommended the additive hazards model for competing risks analysis due to their theoretical and practical credibility [17]. Researchers Shen and Cheng used the additive hazards model of L-Y to study prognostic factors responsible for failing from melanoma and other causes amongst survivors of the disease [20]. In that paper, they compared the regression estimates from the additive model with that from the Cox cause-specific model [20].

Successful efforts have been made to join the additive and multiplicative hazards models into a single hazards model [22, 64, 65]. One of such attempts was by L-Y (1995) in which they studied a general joint hazards model that contains a semiparametric additive component and a semiparametric multiplicative component [64].

2.4 Cardiovascular Disease

Cardiovascular disease (CVD) is a notable human disease affecting developed and developing countries [66]. CVD is a set of diseases and injuries that affect the proper functioning of any of the blood circulatory system, which comprises of the heart and the blood vessels. These diseases are noncommunicable, chronic, and terminal but mostly preventable including congenital heart disease, coronary heart disease, cerebrovascular disease, etc [66].

2.4.1 Epidemiology of CVD

CVD is acknowledged as a severe disease, liable for more deaths than any other known disease, which accounted for 31% of all global deaths in 2015 [66]. The rising human and financial burden of this disease lead to a recent effort around the world, towards decreasing untimely CVD mortality (defined by World Health Organization (WHO) as occurring among individuals aged 30-70 years) by 25% in 2025 [67].

In Canada, cardiovascular mortality has declined since the 1960's due to better disease management and lower CVD incidence rate, but the economic burden of the disease remains a concern [68, 69]. In the last decade, CVD accounted for 32.1% of all deaths and 16.9% of all hospital admissions in Canada [68]. In addition, CVD accounted for over $20 billion of Canada's total health costs, with coronary heart disease (CHD) accounting for more than half of all CVD hospital costs [68]. Canadians are still at a high risk of developing CVD and mortality rate may increase in the future because more than one behavioral risk factor for the disease is present among most Canadians [69].

CVD is a complex disease attributed to multiple factors including environmental and genetic [70]. These factors have mutual effect on other multifactorial diseases; resulting in the high prevalence of CVD comorbidities. The presence of comorbid diseases influence mortality outcome and disease management among others [71]. Understanding the interaction between CVD and co-occurring diseases is critical in the prevention and control of CVD.

2.4.2 Risk Factors for CVD

Various factors can predispose a subject to CVD events. While some of these factors cannot be modified, the majority of CVD cases result from modifiable risk factors [68]. Age, sex, and family history of the disease are some non-modifiable risk factors for the cardiovascular disease [68]. Typical behavioral risk factors for CVD include tobacco use, alcohol consumption, sedentary lifestyle, high cholesterol level, and an unhealthy diet [68]. Socio-demographic factors like marital status [72–74], level of education, and income have also been shown to affect the risks of CVD outcomes [75]. In addition, clinical factors such as high blood pressure [76, 77], diabetes [78–80], and stroke [81, 82] have an impact on CVD occurrence. When a subject has a single risk factor, it may not cause cardiac disease, but the chance of developing the disease increases as the risk factor increases [66, 83]. The Framingham Heart Study, which was launched in the United States of America by the then National Heart Institute and Boston University was instrumental in identifying the major predisposing risk factors for CVD. This long-term cohort study started in 1948, and it continues to give meaningful insights into the factors that contribute to the development of CVD [84].

Age and gender are shown to affect the risk of CD in many epidemiological studies [41, 85–87]. During life phases, men and women exhibit distinct risk of CVD outcomes. Men above 45 years of age have higher chances of witnessing CVD, however, the risk of developing the disease is delayed in women post menopause or until they are above 55 years of age [68]. The protective contribution of the female hormones has been identified for the delayed occasion of cardiovascular events among women [85, 86]. The risk of death from cardiovascular and non-cardiovascular disease among adults was investigated in a Rotterdam study,. Overall, the lifetime risk of CVD death and non-CVD death was lower in females compared with males [41]. Another study among Finland residents examined how gender and age affect the onset and death from coronary heart disease (CHD) [85]. In that study, the risk of CHD outcomes increased with age in both sexes. However, men compared to women have three times hazards of CHD occurrence and five times hazards of CHD death [85]. In an American cohort study, the age-specific CHD mortality rates increased in both sexes and races, and females have lower mortality rates [86]. Among whites, the adjusted risk of CHD mortality was fivefold for men compared to women at age 45 but by age 95, the gender differences in CHD mortality risk vanished. The corresponding man versus woman risk for blacks at age 45 was doubled, and this did not differ with increasing age. Black women at 45 years have fourfold increased risk relative to white women of the same age but the racial dissimilarity reduced with age. Contrarily, there was no significant gap in CHD mortality risk between black and white men in all age groups [86]. Another non-modifiable risk factor for CVD is the history of the illness in familial generations, which suggests that genetic and shared environmental factors can increase the risk of developing CVD [87]. To investigate how the presence of myocardial infarction (MI) in first-degree relatives can increase the risk of CHD, a cohort of males and females residing in Iceland was studied [87]. Those with a positive family history of MI had a significant adjusted risk of CHD compared to those with a negative family history of the disease [87]. Similarly, the impact of sibling cardiac disease on the progression of the CVD was reviewed using data from the Framingham study. The result showed that sibling cardiac disease versus no sibling cardiac disease raised the risk of CVD death significantly [88]. In a review of articles from prominent databases, it was found that progeny with an existing record of CVD in both parents was at a greater risk of having CVD relative to those with a negative history of CVD in both parents [89].

Meaningful association between cardiovascular mortality and marital status has been reported in various studies [72–74]. In general, research showed that married people are protected against various health outcomes and CVD was no exception. A Scottish study reported that compared with married subjects, the risk of CVD mortality was high among unmarried people [72], and in another national study, it was found that the odds of dying from vascular disease increased substantially among never married compared with married subjects [73]. A cohort of Japanese adults stratified by sex showed that relative to married men, the risk of CVD was higher among never-married men and widowed men. However, the study did not show any relevant link between marital status and CVD morbidity risk among women [74]. Using data from the Finnish cardiovascular risk factor survey, researchers found that men with below high school education had a greater risk of CVD mortality compared to men with above high school education [75]. Also, the Finnish study found that the adjusted risk of CVD mortality for low-income men almost doubled that of high-income men. In addition, they found that the risk of CVD mortality among manual workers was found to increase by twofold compared to upper-level workers. However, they did not find a meaningful association between the socioeconomic factors and CVD mortality in women [75]. In a related study in Australia, there was no significant difference in the hazards of CVD failure for men and women who completed primary education compared to those who completed tertiary education [90].

The effect of the voluntary and involuntary use of tobacco products on the heart and the blood vessels has been the focus of many research [91–93]. In all these studies, a positive association was established between cardiovascular outcomes and active use of tobacco products [91, 92] or exposure to secondary smoke [93]. Several studies also reported that light to moderate alcohol consumers compared to abstainers have the lowest CVD mortality and morbidity risk while heavy and binge consumers have the highest risk [94–96]. Lately, research has found no difference in the CVD risk among binge and non-binge alcohol consumers [97]. Another modifiable risk factor for cardiovascular disease is physical activity. Physical activities including active commuting and hiking have been shown to decrease the risk of CVD outcomes among men and women [98, 99]. A meta-analysis procedure showed that active travel to work versus non-active travel protected against cardiovascular endpoints [98], and a prospective study of adults reported that trekking more than four hours in a week versus less than one hour in a week lowered the hazards of CVD hospitalization and mortality significantly in both sexes [99].

Studies have shown that the presence of an ailment can influence the risk of CVD [76–81]. In a Chinese study, the different levels of hypertension was compared to normotension.

The result showed that being hypertensive significantly increased the risk of CVD outcomes in both genders [76]. Similarly, the Framingham Heart Study found that non-optimal blood pressure raised the risk of CVD incidence and death among the study subjects [77]. A population study in the United Kingdom examined how the presence of type 1 diabetes influence morbidity and mortality from major CVD outcomes. The study result showed that the risk of CVD outcomes in type 1 diabetes patients is higher compared to non-diabetic control group [78]. In another diabetes cohort study, type 1 and type 2 diabetic participants have a fivefold hazard for CVD mortality, a twofold hazard for non-CVD mortality, and a threefold hazard for total mortality in comparison to non-diabetic people [79]. The Asia-Pacific (APAC) cohort study reported that in comparison to non-diabetic subjects, diabetes patients have a twofold hazard for CVD mortality, non-CVD mortality, and total mortality [80]. Following a stroke, research has shown that CVD is the chief cause of death [81]. In a population study, the cause of mortality in a group of Minnesota residents who have experienced a stroke was investigated. After following the cohort for a decade, the result showed that those with stroke compared to controls have a higher cardiac mortality rate [81]. Similarly, a meta-analysis of studies from the APAC region reported a significant increased risk of cardiovascular mortality and morbidity for every one unit increase in the level of total blood cholesterol [100].

Chapter 3
Methodology

In this chapter, I review the modeling of survival data with more than one event, beginning with a general methodology of survival analysis. Also, I review multiplicative and additive hazards models for survival analysis. Then, I discuss how the multiplicative and the additive hazards models are extended for use in competing risks settings, and describe the methods for assessing the goodness of fit of the models.

3.1 Key Concepts in Survival Analysis

3.1.1 Survivor and Hazard Functions

The actual survival time of an individual is considered as the non-negative value of a random variable T. The set of values that T can take is associated with a probability distribution with underlying probability density function (pdf), *f(t)* [1]. The cumulative distribution function (cdf) of *T*, *F(t)*, which denotes the chance that the survival time is less or equal to some value t is given as [6]

$$F(t) = P(T \leq t) = \int_0^t f(x)dx.$$

The survivor function *S(t)*, is the chance that the survival time exceeds some given time t expressed as [6]:

$$S(t) = P(T > t) = 1 - F(t). \tag{3.1.1}$$

The hazard function *h(t)*, also known as the intensity rate, the hazard rate, or the force of mortality is defined as the instantaneous event rate for an individual who survived the event to time t [1, 6]. It is obtained from the conditional probability of an event occurring within a small time interval, t and (t + δt) given the non-occurrence of the event up to time t. The hazard function *h(t)* is given as [6]:

$$\begin{aligned} h(t) &= \lim_{\delta t \to 0} \left\{ \frac{P(t < T \leq t + \delta t | T > t)}{\delta t} \right\} \\ &= \frac{f(t)}{S(t)} \end{aligned} \tag{3.1.2}$$

When the event of interest is death, then *h(t)* is interpreted as the hazard of death at time t [1]. Useful associations between the probability density function *f(t)*, the cumulative distribution function *F(t)*, the survivor function *S(t)*, and the hazard function *h(t)* are specified below [6]:

$$f(t) = \frac{dF(t)}{d(t)} = -\frac{dS(t)}{dt} \tag{3.1.3}$$

$$\frac{d}{dt}\left\{logS(t)\right\} = \frac{1}{S(t)}S(t)' = -\frac{f(t)}{S(t)} = -h(t) \tag{3.1.4}$$

$$S(t) = exp\left\{ -\int_0^t h(x)dx \right\} = exp\left\{-H(t)\right\}. \tag{3.1.5}$$

$$H(t) = -log[S(t)]. \tag{3.1.6}$$

3.1.2 Estimating the Survivor and Hazard Functions

The survivor and hazard functions are estimated from the observed event times in the survival data. The *Kaplan-Meier* estimate of the survivor and hazard functions is the most widely used [1], and it is described in details here. Assume that there are r observed death times and let the j^{th} death time be denoted as t_j for j=*1,2,...r* and $t_1 \leq t_2 \leq ... \leq t_r$. Those at risk of death at t_j is n_j, and the death that occurred at t_j is d_j. Then, the estimated probability that a death occurs within a small time interval, which accommodates a single death time is

$\frac{d_j}{n_j}$, and its complement of surviving is $1 - \frac{d_j}{n_j}$ [1]. Then,

$$\hat{S}(t) = \prod_{t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right) \tag{3.1.7}$$

is the *Kaplan-Meier* or product-limit estimate of the survivor function [1, 6]. Based on the Greenwood's formula, the variance of the *Kaplan-Meier* survivor estimate is

$$\widehat{Var}\Big(\hat{S}(t)\Big) = \Big\{\hat{S}(t)\Big\}^2 \sum_{t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}, \tag{3.1.8}$$

and the 100(1 - α)% confidence interval of the *Kaplan-Meier* survival estimate is

$$\hat{S}(t) \pm z_{1-\alpha/2} \sqrt{\widehat{Var}\Big(\hat{S}(t)\Big)}, \tag{3.1.9}$$

where z_α is the α quantile of the standard normal distribution [6]. Also, the hazard function in the time interval $(t_j\ ,\ t_{j+1})$ is obtained as [6]:

$$\hat{h}(t) = \frac{d_j}{n_j \tau_j}, \tag{3.1.10}$$

for $t_j \leq t < t_{j+1}$, where d_j and n_j are the number of death and the number at risk at t_j respectively, and $\tau_j = t_{j+1} - t_j$ [6]. An approximate variance of $\hat{h}(t)$ is [6]:

$$\widehat{Var}\Big(\hat{h}(t)\Big) = \Big\{\hat{h}(t)\Big\}^2 \left(\frac{n_j - d_j}{n_j d_j} \right). \tag{3.1.11}$$

An estimate of the cumulative hazards to time t, denoted as $\hat{H}(t)$ can be computed using the relation in (3.1.6) [6]. Differentiating the cumulative hazard function gives the hazard function $h(t)$ itself. So, useful information about the shape of the hazard function can be acquired from the slope of the cumulative hazard function [1].

3.1.3 Cox Hazards Model

In survival analysis, it is a routine to model the effect of covariates so as to foretell which of the covariate combinations affect the form of the hazard function and to what extent. Modeling is also important because, through it, one can have estimates of the hazard function for subjects in the study [1]. The hazards regression model introduced by Sir David Cox in 1972 is one of the most popular approaches used in survival analysis [1]. Cox hazards model is a multiplicative one, which is also known as the Cox proportional hazards (PH) model. The model is so called because it is based on the proportional hazard assumption. The assumption implies that the expected hazard of death for a subject in a particular group compared to the expected hazard for a similar subject in the group is constant [1, 13]. The hazard ratio under proportionality assumption does not depend on time. Also, the Cox model is referred to as a semi-parametric model in the sense that a parametric form is considered for the covariate effect but no probability distribution is assumed for the baseline hazard function [1, 13]. The Cox multiplicative hazards model is expressed as [1]:

$$h(t) = h_0(t)exp(\boldsymbol{\beta}^T\mathbf{X}), \tag{3.1.12}$$

where $\mathrm{h}_0(\mathrm{t})$ is the unspecified baseline hazard function, $\mathbf{X}^T = (\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_q)$ represents the vector of q explanatory variables or covariates, and $\boldsymbol{\beta}$ is the vector of regression coefficients corresponding to the covariates [1]. Assuming we observed r distinct death times, $j = 1, 2, ..., r$, and the death times are ordered such that $t_1 \leq t_2 \leq ... \leq t_r$. Then, the partial likelihood function for the Cox hazards model is given by [1]:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \left(\frac{exp(\boldsymbol{\beta}^T\mathbf{X}_j)}{\sum_{l \in R_j} exp(\boldsymbol{\beta}^T\mathbf{X}_l)} \right), \tag{3.1.13}$$

where $\mathbf{X}_j$ is the vector of covariates for those subjects who have had the event at time t_j and R_j is the risk set of subjects who are yet to experience any event at t_j. In other words, the numerator in the equation (3.1.13) strictly depends on the contribution of those who have experienced the event and the summation in the denominator covers all those who are at risk of the event at time t_j [1]. By maximizing the likelihood function using an iterative process, the estimate of the regression coefficients in the Cox hazards model are obtained [1].

3.1.4 Additive Hazards Models

Additive hazards model was first introduced by Odd Aalen in 1980 after he studied a multiplicative intensity model two years earlier [56]. Aalen additive hazards model is a non-parametric model, which allows a covariate effect to change with time. Constant covariates are also allowed to have time-varying effects in his model [13]. Let $\boldsymbol{Z(t)} = (Z_1(t), ..., Z_q(t))$ be a set of covariates, then the Aalen model takes the form [13]:

$$h(t|\mathbf{Z}(t)) = \beta_0(t) + \sum_{a=1}^{q} \beta_a(t) Z_a(t), \tag{3.1.14}$$

for $\beta_a(t), a = 1, ..., q$, being the regression functions to be estimated and $\beta_0(t)$ denoting the baseline hazard function. The cumulative regression functions in the Aalen model are obtained through the least squares method as $B_a(t) = \int_0^t \beta_d(u)du, a = 0, ..., q$ [13]. The non-parametric terms in the Aalen model are not well developed, and reporting covariate effect numerically in the model is difficult [9].

A semiparametric additive hazards model was proposed by Lin and Ying (L-Y) in 1994 [19]. In the L-Y model, the covariates effect are assumed to be constant. For a q-dimensional vector of covariate $\mathbf{Z}(.)$ that could depend on time, the L-Y additive hazards model is specified as [13]:

$$h(t|\mathbf{Z}(t)) = \alpha_0(t) + \sum_{a=1}^{q} \alpha_a Z_a(t), \tag{3.1.15}$$

where $\alpha_0(t)$ is baseline hazard function and $\alpha_a, a = 1, ..., q$ are the regression coefficients corresponding to $\mathbf{Z}$. The estimation of the regression coefficients in the L-Y model is done explicitly from the score equation [13].

3.2 Competing Risks Models

In order to model competing risks data, the joint probability distribution of the bivariate random variables T and C is important. The variable C is the cause of event (failure), and it takes the value *0* when the observation is censored, or it takes one value from the set of p unique events (k=1, 2..., p), if otherwise. T is the failure time in case an event occurs, and when the observation is censored, it is the censoring time [6]. In my study, I categorized events as death from cardiovascular disease (CVD) ($k = 1$) and non-CVD related death ($k = 2$). T is defined as the time from survey to the first of the two failures or censoring. Figure 3.1 depicts this kind of competing risks model, where subjects can die from two distinct causes [2].

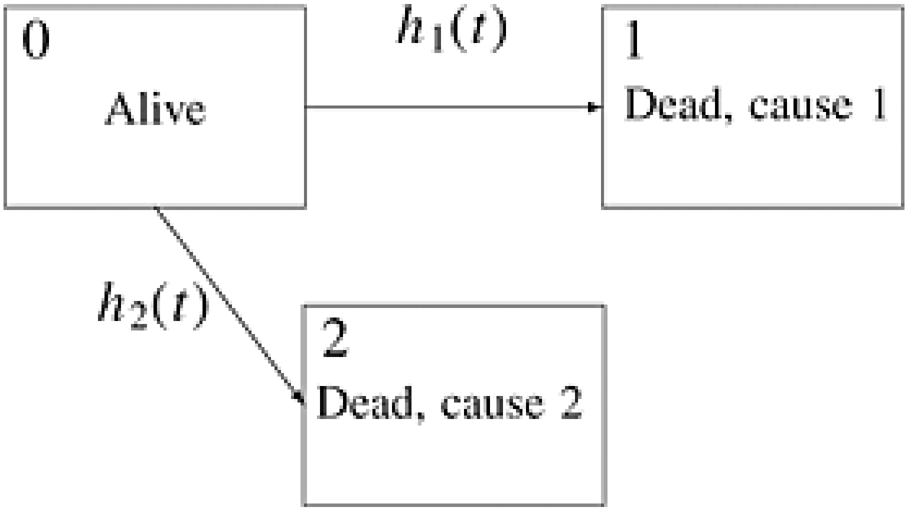


Figure 3.1: The competing risks model with two failure causes

The joint probability distribution of the event time (T) and the cause of event (C) can be outlined with respect to the cumulative incidence function (CIF). The CIF is the probability of occurrence of a particular event, say k at time t or a time before t and it is specified as [6]:

$$F_k(t) = P(T \leq t, C = k). \tag{3.2.1}$$

The distribution of T and C can also be outlined by the subsurvivor function, which is the probability of the non-occurrence of an event of type k by time t, given by $S_k(t) = P(T > t, C = k)$ [6]. The CIF and the subsurvivor function for the k^{th} event are related by:

$$F_k(t) + S_k(t) = P(C = k),$$

where

$$P(C = k) = \lim_{t \to \infty} F_k(t),$$

is the marginal probability distribution of the type k event [6]. The CIF is an improper probability distribution since it only takes values up to $P(C = k)$ instead of 1 at $t = \infty$. Hence, it is called a subdistribution function [2, 6]. The cumulative distribution function, *F(t)*, which is the probability of the occurrence of any event before or at time t takes values between 0 and 1. *F(t)* is the sum of all the CIFs for all the event types and it is specified as [6]:

$$\begin{aligned} F(t) &= P(T \leq t) = 1 - S(t) \\ &= \sum_{k=1}^{p} P(T \leq t, C = k) = \sum_{k=1}^{p} F_k(t). \end{aligned}$$

In the competing risks framework, the hazard function of failure from the event of type k when the other events are censored is defined as [6]:

$$\begin{aligned} \tilde{h}_k(t) &= \lim_{\delta t \to 0} \left\{ \frac{P(t < T \leq t + \delta t, C = k | T > t)}{\delta t} \right\} \qquad (3.2.2) \\ &= \{P(T > t)\}^{-1} \lim_{\delta t \to 0} \left\{ \frac{P(t < T \leq t + \delta t, C = k}{\delta t} \right\} = \frac{f_k(t)}{S(t)}, \end{aligned}$$

where $\mathrm{f}_k(\mathrm{t})$ is the cause-specific density function and S(t) is the overall survivor function [1]. Equation (3.2.2) is the subhazard function. It is also known as the cause-specific hazard

(CSH) function, which is the instantaneous rate of failing from the k^{th} event among subjects who are yet to fail from any of the events [6]. Summing over the CSH rates gives the overall hazard rate of failure, regardless of the type of event [6]. Hence, for a set of competing risk events, the cause-specific hazard rate, $\tilde{h}_k(t)$, and the all-cause hazards rate, *h(t)* , are related by [17]:

$$h(t) = \sum_{k=1}^{p} \tilde{h}_k(t). \tag{3.2.3}$$

3.2.1 Multiplicative Hazards Models

For competing risks data, the covariate impact can be estimated from the CIF directly by modeling the subdistribution hazards (SH) but not the cause-specific hazards (CSH) [10]. While there are many regression models that can be employed to model the CSH, the model proposed by F-G is mainly used for modeling the hazard of subdistribution. The Cox cause-specific and F-G subdistribution hazards models are described below.

3.2.1.1 Cox Cause-Specific Hazards Model

Let $\mathbf{X}^T$= (X$_1$, X$_2$,...,X$_q$) be a vector of covariates and let $\boldsymbol{\beta}_k$ be the vector of regression coefficients corresponding to $\mathbf{X}$. Let $\tilde{h}_{0k}(t)$ be the baseline cause-specific hazard function relevant to the k^{th} event, for $k = 1, 2, ..., p$, then, the Cox cause-specific hazards model for the event of type k is given as [1]:

$$\tilde{h}_k(t) = \tilde{h}_{0k}(t) exp(\boldsymbol{\beta}_k^T \mathbf{X}). \tag{3.2.4}$$

The vector $\boldsymbol{\beta}_k$ of regression coefficients purely reflect the covariate effect for event k since the competing events are treated as censored observations [9]. The estimate of the regression coefficients for the k^{th} event is obtained through the partial likelihood approach outlined in Section 3.1.3 [1].

3.2.1.2 Fine and Gray Subdistribution Hazards Model

Fine and Gray proposed a hazard-based approach to evaluate covariates influence directly on the CIF when the competing events are considered [10]. The hazards function of the subdistribution for the k^{th} event is the probability of failing from event k given survival up to time t without any event or with the occurrence of a competing risks event before time t. The subdistribution hazard function is of the form [33]:

$$\begin{aligned}\gamma_k(t) &= \lim_{\delta t \to 0} \left\{ \frac{P(t < T \leq t + \delta t, C = k | (T > t) \cup (T \leq t \cap C \neq k))}{\delta t} \right\} \qquad (3.2.5)\\ &= \frac{f_k(t)}{1 - F_k(t)}.\end{aligned}$$

The so-called subdistribution hazards model is based on the Cox model [10], and so it follows that the subdistribution hazard for event k is modeled as [6]:

$$\gamma_k(t) = \gamma_{0k}(t) exp(\boldsymbol{\beta}_k^T \mathbf{X}). \qquad (3.2.6)$$

where $\gamma_k(t)$ is the subdistribution hazard function for cause k, $\gamma_{0k}(t)$ is the baseline subdistribution hazard function for the k^{th} event, $\boldsymbol{\beta}_k$ is the vector of regression coefficients for the incidence of the k^{th} event, and $\mathbf{X}$ is the vector of covariates [6]. For a collection of r ordered death times in which an event of interest occurs, $j = 1, 2, ..., r$, the partial likelihood for the hazard of the subdistribution for a single covariate X is [6]:

$$L(\beta) = \prod_{j=1}^{r} \frac{exp(\beta X_j)}{\sum_{l \in R_j} w_{jl} exp(\beta X_l)}. \qquad (3.2.7)$$

The risk set (R_j) in equation (3.2.7) includes those who are yet to experience the event of interest by time t_j and those who had failed from the competing event by that time [6].

Hence, R_j for the subdistribution model is specified as [6]:

$$R_j(t) = [l; T_l \geq t \text{ or } (T_l \leq t \text{ and } C \neq k)], \tag{3.2.8}$$

and the weight, w_{jl} in equation (3.2.7) is given as

$$w_{jl} = \frac{\hat{G}(t_j)}{\hat{G}(min(t_j, t_l))}, \tag{3.2.9}$$

where $\hat{G}(.)$ is the Kaplan-Meier estimate of the survival function of the censoring distribution (T_l, C_l) [6]. T_l is the survival time to the first censoring event and C_l is the censoring variable. C_l equals one when event of non-interest occurs and zero if censoring occurs. At each time t_j that an event of interest happens, the index l represents the individuals who are at risk of any event and those who have had the competing risk event at a time before t_j [6]. The definition of the risk set and the weights (w_{jl}) included in equation (3.2.7) are the two distinctions between the cause-specific hazards and the subdistribution hazards models [6, 15]. Estimates of the regression coefficients in the subdistribution model are obtained from the partial likelihood function [6].

3.2.1.3 Goodness of Fit

Various facets of the multiplicative hazards model are examined to diagnose the fitness of the model, including the adequacy of the proportional hazards assumption [1, 13].

3.2.1.3.1 Cox-Snell Residuals

Cox-Snell residual plot is widely used in the evaluation of the general fitness of the Cox hazards model [101]. For the i^{th} subject having covariate vector $\boldsymbol{X}$, the Cox-Snell residual is given by [1]:

$$r_{ci} = \hat{H}_0(t_i) exp(\hat{\boldsymbol{\beta}}^T \mathbf{X}_i)$$

where $\hat{H}_0(t_i)$ is the Breslow's estimator of the aggregate baseline function at time t. If the model fits the data satisfactorily, I expect the plot of the estimated cumulative hazards of the Cox-Snell residual (or its log) versus the Cox-Snell residual (or its log) to have a unit slope and an intercept of zero [13].

3.2.1.3.2 Martingale Residuals

Martingale residual is obtained through the transformation of the Cox-Snell residuals. They are however not symmetrically distributed even in models with acceptable fit [13]. Suppose that for the i^{th} subject in the sample, there is a vector $\boldsymbol{X}_i(t)$ of possible time-dependent covariates. If the subject has witnessed the event of interest, then $N_i(t)$ has a unit value and zero if not [13]. Let $Y_i(t)$ indicate that subject i is in the study at a time prior to t when $\boldsymbol{\beta}$ is the vector of regression coefficients. $\hat{H}_0(\mathrm{t}_i)$ is the Breslow estimator of the aggregate baseline hazard, and martingale residual is defined as [13]:

$$\hat{M}_i = N_i(\infty) - \int_0^\infty Y_i(t) exp[\boldsymbol{\beta}^T \mathbf{X}_i(t)] d\hat{H}_0(t_i), i = 1, ..., n. \tag{3.2.10}$$

3.2.1.3.3 Deviance Residuals

The deviance is a statistic used to give a concise report of the abnormal behavior of a reduced model; the smaller its value, the better [13]. The deviance is given by:

$$D = -2\{\log \hat{L}_r - \log \hat{L}_s\} \tag{3.2.11}$$

where $\hat{L}_r$ and $\hat{L}_s$ are the maximized likelihood under the reduced model and saturated model respectively [1]. In a reduced model, the hazard of death for a subject is determined by the subject's covariates values through the risk score, $\boldsymbol{\beta}^T\mathbf{X}$. Negative risk score implies that the hazard of the event is lower than the anticipated hazard while those with high positive risk score have hazard that is greater than the average risk [1]. When censoring is light, the deviance residual is normally distributed about zero for a good model. Observations with fairly large deviance residuals are considered as outliers [13].

3.2.1.3.4 Graphical Checks

Plots are usually used to inspect the proportional hazard assumption for a given covariate [1]. This is achieved by examining the plot of the log-cumulative hazard against the log of time; roughly parallel curves validate the proportional hazard assumption [1].

For the proportionality of the subdistribution hazards, the plot of log (- log (1 - F)) against the log of time can also be assessed, where F is the CIF for the event of interest [6]. The curves for the levels of a factor are inspected, and their divergence supports the proportionality assumption [6].

3.2.2 Additive Hazards Models

Hazards models in which the regression functions relate in an additive manner on the baseline hazard have been around for a while. However, they are not the preferred choice in survival analysis, mainly because of the scarcity of computer software [17]. Notwithstanding, in competing risks analysis, additive hazards models are embraced as substitutes for the proportional hazards model when modeling the cause-specific hazard rates. The benefit of using the additive hazards models is that equation (3.2.3) holds concurrently for the entire competing events. As such, the covariate effects on the cause-specific hazards amount to the overall covariate effect on the all-cause hazard. Hence, the additive hazards models give non-conflicting solutions for competing risks data [17, 18].

3.2.2.1 Aalen Additive Hazards Model

The Aalen model is a nonparametric additive hazards model, where the impact of time-varying covariates act absolutely on some unspecified baseline hazard function [56]. Given a set of possibly time-varying covariates $\boldsymbol{Z(t)} = (Z_1(t), ..., Z_q(t))$ and time-varying regression coefficients $\beta_{ak}(t)$ for $a = 1, ..., q$, the hazard rate for the event of type k is [13]:

$$h_k(t|\mathbf{Z}(t)) = \beta_0(t) + \sum_{a=1}^{q} \beta_{ak}(t) Z_a(t). \tag{3.2.12}$$

The nonparametric estimation of the regression functions $\beta_a(t)$ in Aalen's model is challenging, but it is easy to estimate the cumulative regression function as [13]:

$$B_a(t) = \int_0^t \beta_a(u)du, \qquad a = 0, ..., q. \tag{3.2.13}$$

Crude estimates of the regression function can be found from the slope of the estimates of $B_a(t)$. However, smoothing the estimates of $B_a(t)$ will give better estimates of the regression functions [13]. The least-squares method is used to obtain the estimates of the cumulative functions $B_a(t)$. To get the estimates, a design matrix $\boldsymbol{X}$(t) having n by ($q+1$) dimension is specified as follows; if the i^{th} subject is at risk at time t then the i^{th} row of $\boldsymbol{X}$(t) is configured as $\boldsymbol{X}_i(t)$ = (1, Z$_1$(t)),...,Z$_q$(t)) [13]. However, if the i^{th} subject is no more at risk at time t then the corresponding $q+1$ vector contains all zeros. Say $\boldsymbol{I}$(t) is the $n \times 1$ column vector having its i^{th} element as one if subject i dies at t and zero otherwise. Then, the least-squares estimate of the vector $\boldsymbol{B}(t)$ is outlined by [13]:

$$\hat{\mathbf{B}}(\mathbf{t}) = \sum_{T_i \le t} [\mathbf{X}^T(T_i)\mathbf{X}(T_i)]^{-1}\mathbf{X}^T(T_i)\mathbf{I}(T_i), \tag{3.2.14}$$

and the covariance matrix of $\mathbf{B}$(t) is given as [13]:

$$\widehat{Var}[\hat{\mathbf{B}}(t)] = \sum_{T_i \le t} [\mathbf{X}^T(T_i)\mathbf{X}(T_i)]^{-1}[\mathbf{X}^T(T_i)\mathbf{I^D}(T_i)\mathbf{X}(T_i)]\{[\mathbf{X}^T(T_i)\mathbf{X}(T_i)]^{-1}\}^T. \tag{3.2.15}$$

$\boldsymbol{I}^D(t)$ is the diagonal matrix of $\boldsymbol{I}(t)$, and the estimator of the cumulative function $\boldsymbol{B}(t)$ is obtained as long as $\boldsymbol{X}^T(T_i)\boldsymbol{X}(T_i)$ has an inverse [13]. The plot of the cumulative regression function against time depicts the influence of a covariate on the survival probability over time. When the covariate has no impact on the hazards of the outcome, the slope will be approximately zero. A positive slope is seen when the covariate effect enhances the risk of the outcome while a negative slope indicates that the covariate decreases the hazard of the event [13]. The confidence interval for the cumulative function B(t) is constructed as [13]

$$\hat{B}_q(t) \pm z_{1-\alpha/2}[\widehat{Var}(\hat{B}_q(t))]^{1/2}.$$

Two hypotheses about the functional form of the regression coefficients in the Aalen model are usually of interest. These are the hypothesis of no effect of the q^{th} covariate and the hypothesis of a time-independent effect of the q^{th} covariate [9]. When these hypotheses are based on the cumulative regression function, they are specified respectively as:

$$H_{01} \quad : \quad B_q(t) = 0$$

and

$$H_{02} \quad : \quad B_q(t) = bt$$

for all $t \; \epsilon \; [0, \; \tau]$ and τ being the largest observed time point [9]. Valid test-statistics are constructed in order to test the above hypotheses. For the q^{th} cumulative regression function, H_{01} is based on the supremum test statistic [9]:

$$T_{sup} = \underset{t \, \epsilon \, [0,\tau]}{sup} |\frac{\hat{B}_q(t)}{\hat{\sigma}_q(t)}|,$$

for $\hat{\sigma}_q$(t) $= [Var(\hat{B}_q(t))]^{1/2}$, being the estimated standard deviation of $\hat{B}_q(t)$. The Kolmogorov-Smirnov-type test statistic can be used to test the hypothesis of a time-independent effect of the q^{th} covariate as [9]

$$T_{KS} = \underset{t \, \epsilon \, [0,\tau]}{sup} |B_q(t) - \frac{t}{\tau} B_q(\tau)|.$$

3.2.2.2 Lin and Ying Additive Hazards Model

Lin and Ying (1994) proposed the additive analog of the traditional Cox multiplicative hazards model, but estimates in their models are deduced from explicit formulas [13]. Similar to the Aalen model, covariate effect relates with the baseline hazard in a linear form, though the regression coefficients in the L-Y model can be estimated easily [19, 23]. When there is a q-dimensional vector of covariate $\mathbf{Z}(.)$ that could depend on time, the hazards model for the event of type k is given as:

$$h_k(t|\mathbf{Z}(t)) = \alpha_0(t) + \sum_{a=1}^{q} \alpha_{ak} Z_a(t), \tag{3.2.16}$$

The estimator of the regressors and the variances for the L-Y model have an explicit form with consistent and asymptotic characters, and the model also has to its credit natural result interpretation [19]. To estimate α_a, assume δ_i is the event indicator and $\boldsymbol{Z}_i(t) = (Z_{i1}(t),\ldots,\ Z_{iq}(t))$ is the vector of covariates for i=$1,\ldots,n$. When the i^{th} subject is at risk at time t, Y_i(t) takes the value of *1*, and *0*, if otherwise. Then, we define the vector $\bar{\mathbf{Z}}$(t), which is the mean value of the covariates at time t as [13]:

$$\bar{\mathbf{Z}}(t) = \frac{\sum_{g=1}^{n} \mathbf{Z}_g Y_g(t)}{\sum_{g=1}^{n} Y_g(t)}, \tag{3.2.17}$$

where the numerator is the covariate sum for subjects in the risk set at time t and the divisor represents those in the risk set at that time [13]. The square matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ with dimension q and the q-vector $\boldsymbol{B}$ are given as follows [13]:

$$\boldsymbol{A} = \sum_{g=1}^{n} \sum_{i=1}^{g} (T_i - T_{i-1}) \left(\mathbf{Z}_g - \bar{\mathbf{Z}}(T_i)\right)' \left(\mathbf{Z}_g - \bar{\mathbf{Z}}(T_i)\right) \tag{3.2.18}$$

$$\boldsymbol{B}' = \sum_{g=1}^{n} \delta_g \left[\mathbf{Z}_g - \bar{\mathbf{Z}}(T_g)\right] \tag{3.2.19}$$

$$\boldsymbol{C} = \sum_{g=1}^{n} \delta_g \left[\mathbf{Z}_g - \bar{\mathbf{Z}}(T_i)\right]' \left[\mathbf{Z}_g - \bar{\mathbf{Z}}(T_i)\right] \tag{3.2.20}$$

Thus, the estimate of the regression coefficient is given as [13]:

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{A}^{-1}\boldsymbol{B}' \tag{3.2.21}$$

and the estimated variance is:

$$\hat{\mathbf{V}} = \widehat{Var}(\hat{\boldsymbol{\alpha}}) = \boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} \tag{3.2.22}$$

3.2.2.3 Goodness of Fit

Checking how well the additive models fit the data is important because the model could be misspecified and the effect of covariates could also be misrepresented. To assess the adequacy of the Aalen additive model and that of the L-Y additive model, martingale residual can be utilized [9].

3.2.2.3.1 Martingale Residuals

Martingale residuals give the difference in the actual and predicted number of events. Being a function of time, martingale residual gives a precise indication of the point of the problem in the fitted model [13, 102]. For the i^{th} subject at time t, the martingale residual is given as

$$\hat{M}_i(t) = N_i(t) - \hat{H}[t|\mathbf{Z}_i(t)], i = 1, 2, ..., n,$$

where $N_i(t)$ represent the actual number of events and $\hat{H}$[t—$\mathbf{Z}_i$(t)] is the predicted number of events [102]. To check the model fitness, the martingale residuals are sum based on the levels of the covariate, and for a good fit, the sums are near zero when plotted against time. However, when the residuals are many, P-values are used to quantify the departure of the residuals from the null [9].

3.2.3 Software

Data cleaning and data preparation for this study were conducted using the SAS software, version 9.4 (SAS Institute Inc., Cary NC, USA). R-Studio, version 3.1.3 [103] was used for the competing risks data analysis using *cmprsk*, *survival*, *ahaz*, and the *timereg* packages. The appendices contain the implementation codes used in generating the results. Microsoft word and LaTex were used in typing and producing the figures and tables.

Chapter 4
Application

Here, I begin with a description of the data, the data coding, and I present the results from the additive and multiplicative hazards models with their interpretation. In this thesis, a 5% significance level is adopted for covariates in all the models. This study is approved by the Biomedical Research Ethics Board of the University of Saskatchewan (Bio # 14-123).

4.1 Source of Data

The Canadian Heart Health Survey (CHHS) was launched by the provincial and federal ministries of health and collaborating academic institutions. The purpose of this collaborative effort was to examine the prevalence of CVD risk factors and the level of awareness regarding the causes, effects, and prevention of CVD in the nation [104]. The CHHS was conducted between 1986 and 1992 in all Canadian provinces. Furthermore, a supplementary survey on new subjects was conducted in Nova Scotia in 1995. Details of the survey framework have been described elsewhere [104, 105]. CHHS is a complex survey of the inhabitants of Canadian provinces who are between 18 and 74 years old. A stratified, 2-stage probability design was used to draw samples from the provincial health insurance registries that provided approximately 2,000 responses in each province. Residents of Indian Reserves, military camps, and prisons were excluded from the survey. The potential survey participants sampled from the health registries were contacted by phone or letter for the home and clinic interviews [104, 105]. All ten provinces provided the core information, but family history information was only available for residents in Alberta, Ontario, Saskatchewan, and Quebec. In the home interview, participants provided their demographics, risk behaviors, knowledge, and beliefs about CVD. Within two weeks of the home interview, survey participants were invited to a clinic for the second round of the interview. In the clinic visit, anthropometric measurements and other clinical data were collected. In each province, two probability weights were calculated for each individual to adjust for the unequal probabilities of selection and non-response at the home and clinic interviews [104, 105].

In order to determine the mortality status of the CHHS participants, a Follow-up study was carried out. In the Follow-up study, the CHHS dataset was linked with the Canadian Mortality Database (CMDB) at Statistics Canada through computerized probabilistic record linkage system using participants' unique identifier [106, 107]. Statistics Canada calculated bootstrap weights for participants using the probability sampling unit (PSU) and strata because the CHHS was based on a complex survey design. However, bootstrap weights were only available for Alberta, Manitoba, and Saskatchewan participants. The CMDB has death information of Canada residents' since 1950 with an high level of accuracy. The database is updated regularly based on information from the provincial and territorial death registries [108]. The cause of death is recorded in the CMDB based on the version of the International Classification of Diseases (ICD) code in use at death time [108]. The ICD 9^{th} revision was used till December 31, 1999, and the $10th$ revision is in use till now [109].

My study uses data from the CHHS Follow-up study, which contains information for six of the ten provinces: Alberta, British Columbia, Manitoba, Newfoundland, Nova Scotia, and Saskatchewan (N = 14,018 participants). The ICD code was used to identify the causes of death. Death due to CVD (ICD-9: 390-448; ICD-10: 100-178) is the event of interest in this study, and all deaths caused by non-CVD constitute the competing risk. Participants with no death information were considered alive. In my study, I excluded subjects who were on the survey register but who had experienced death by the time they were visited for the survey, and subjects whose sex in the CMDB and the CHHS differed. Figure 4.1 below depicts the extraction of the eligible study subjects. After applying the exclusion criteria, a total of 13,996 subjects were eligible and thus considered in my data analysis. At the end of the study linkage period on December 31, 2004, 1,536 (11%) deaths were recorded and 549 (4%) of the survey subjects experienced CVD death. 12,460 (89%) of the survey subjects were censored (Figure 4.1).

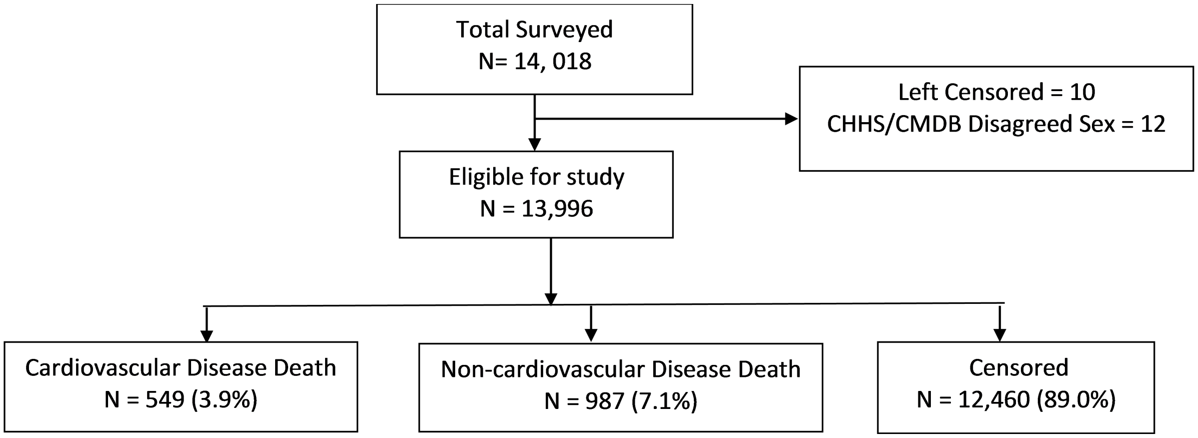
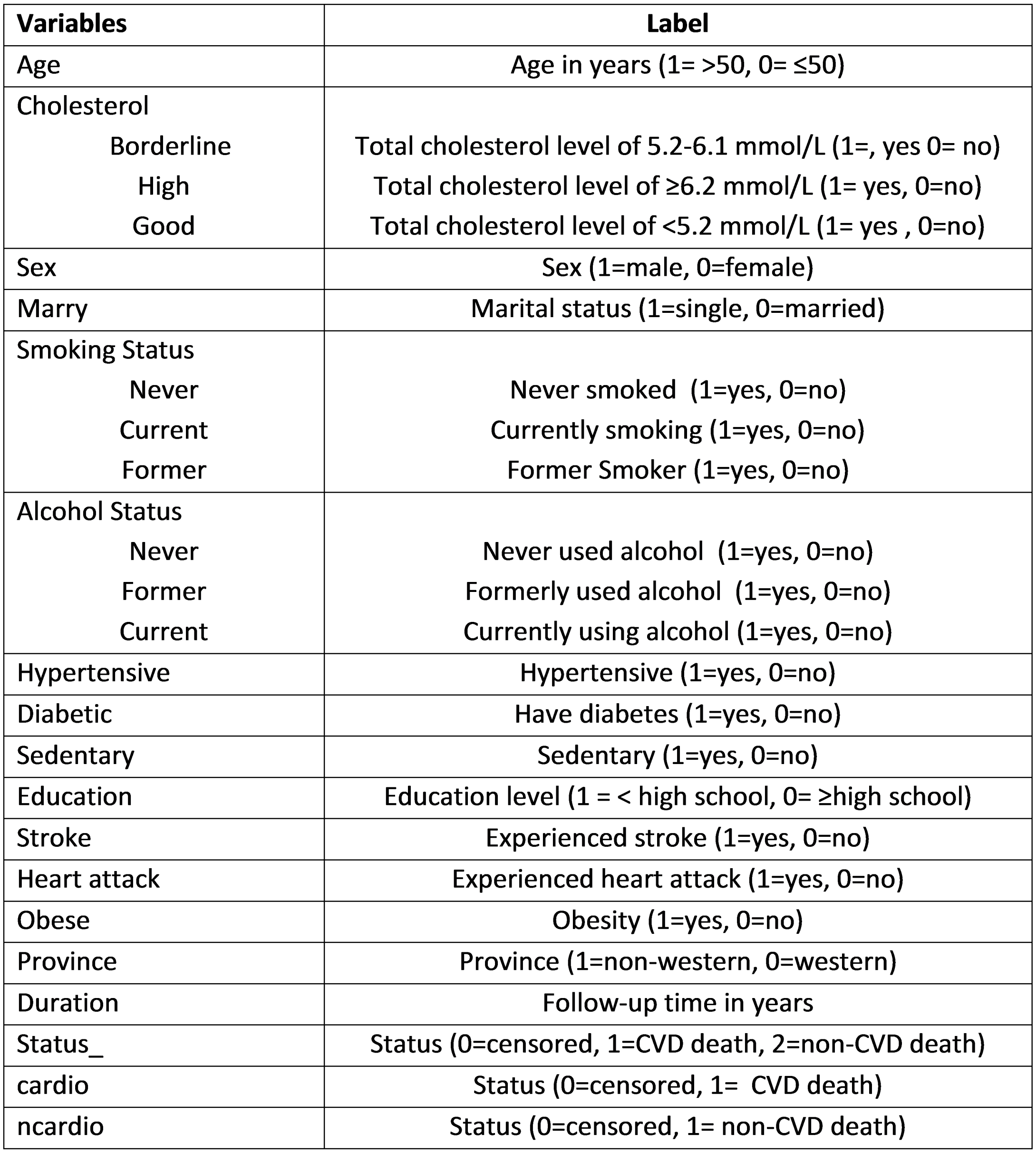


Figure 4.1: Data extraction chart

4.2 Coding of the Data

The explanatory variables (covariates) considered in my data analysis are listed in Table 4.1. CVD death is common in men over 45 years of age and women over 55 years of age [68]. Based on this a priori information, I categorized the covariate Age into two distinct groups in relation to CVD: ≤50 and >50 years. I chose age ≤50 years as the reference category. In my data, the fasting total plasma cholesterol has three levels: good cholesterol level (< 5.2 mmol/L), borderline cholesterol level (5.2-6.1 mmol/L), and high cholesterol level (≥6.2 mmol/L). The good cholesterol level was chosen as the reference level. Also, smoking status was categorized into three: those who never smoked, current smoker (smoked a cigar, cigarette or pipe regularly or occasionally at the time of survey), and former smoker. Those who never smoked was used as the referent group. Alcohol status was also grouped into three: those who never used alcohol, current alcohol users, and former alcohol users. Current alcohol user category was my reference because of the small proportion of those who never used alcohol in the data. Female was referenced for sex, and active subjects, who exercised at least once a week in the months that preceded their survey were referenced. Marital status was compressed to single (never married, divorced and widow/widower) and married (married/common law), and the latter was the reference group. In addition, educational status was categorized as below high school (elementary and some secondary education) and above high school (secondary school completed and university degree). Above high school category was used as the reference group. This dichotomization was done to ensure reasonable representation in each category. Standard body mass index (BMI) was used to categorize subjects as obese (BMI $\geq$ 30) and not obese (BMI $<$ 30). The latter was the reference group. Study subjects were categorized as hypertensive if they were using a pharmaceutical or non-pharmaceutical drug for hypertension or their average diastolic blood pressure was $\geq$ 90 mmHg or their average systolic blood pressure was $\geq$ 140 mmHg at the time of the survey. Otherwise, they were categorized as non-hypertensive, which was the reference category. Diabetes status was based on subjects' report of the disease and no-diabetes was referenced. If study subjects had experienced a heart attack in the past, they were categorized as having a heart attack. Those who had no heart attack was used as the reference group. Subjects were asked if they had experienced a stroke and their response was used to group them into two. Those with no stroke was used as the reference. Province was categorized as a western province (Alberta, British Columbia, Manitoba, and Saskatchewan) and a non-western province (Nova-Scotia and Newfoundland). Western province was used as the reference group. The covariate "Province" was included in my data analysis to investigate the impact of a subject's province of residence on the study outcome.

Table 4.1: Description of variables



Variables	Label
Age	Age in years (1= >50, 0= ≤50)
Cholesterol	
Borderline	Total cholesterol level of 5.2-6.1 mmol/L (1=, yes 0= no)
High	Total cholesterol level of ≥6.2 mmol/L (1= yes, 0=no)
Good	Total cholesterol level of <5.2 mmol/L (1= yes , 0=no)
Sex	Sex (1=male, 0=female)
Marry	Marital status (1=single, 0=married)
Smoking Status	
Never	Never smoked (1=yes, 0=no)
Current	Currently smoking (1=yes, 0=no)
Former	Former Smoker (1=yes, 0=no)
Alcohol Status	
Never	Never used alcohol (1=yes, 0=no)
Former	Formerly used alcohol (1=yes, 0=no)
Current	Currently using alcohol (1=yes, 0=no)
Hypertensive	Hypertensive (1=yes, 0=no)
Diabetic	Have diabetes (1=yes, 0=no)
Sedentary	Sedentary (1=yes, 0=no)
Education	Education level (1 = < high school, 0= ≥high school)
Stroke	Experienced stroke (1=yes, 0=no)
Heart attack	Experienced heart attack (1=yes, 0=no)
Obese	Obesity (1=yes, 0=no)
Province	Province (1=non-western, 0=western)
Duration	Follow-up time in years
Status_	Status (0=censored, 1=CVD death, 2=non-CVD death)
cardio	Status (0=censored, 1= CVD death)
ncardio	Status (0=censored, 1= non-CVD death)

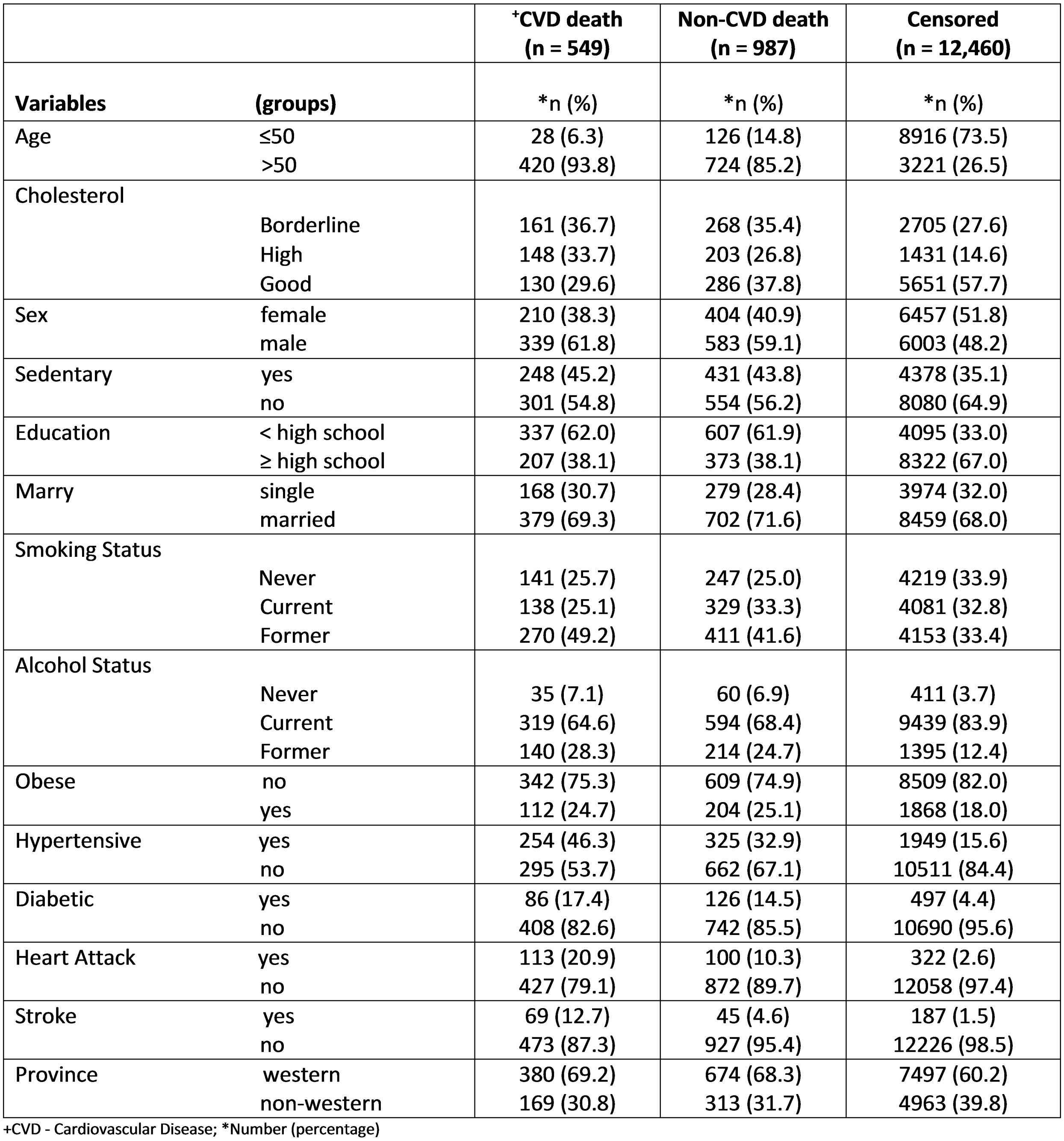
4.3 Results

4.3.1 Descriptive Statistics

In Table 4.2, I present the characteristics of the study subjects based on their status at the end of the Follow-up study. Of the 13,996 eligible subjects, 549 (4%) of them experienced

CVD deaths, 987 (7%) experienced non-CVD related deaths, and 12,460 (89%) were alive at the end of the follow-up period. Follow-up time is defined in my study as the number of years a subject accumulated since the date of the survey to the date of any death or to December 31, 2004. The median follow-up time is 14.99 years (interquartile range = 5.52 years, minimum follow-up time= 0.08 years and maximum follow-up time= 18.84 years). In my study, neither the probability weights nor the bootstrap weights were considered in the analysis. This is because the bootstrap weights were available for only 3 provinces. Moreover, my study findings are meant to be generalized to the sample used and not the entire population.

Table 4.2: Characteristics of eligible participants at the end of follow-up (N=13,996)



		$^+$CVD death (n = 549)	Non-CVD death (n = 987)	Censored (n = 12,460)
Variables	**(groups)**	*n (%)	*n (%)	*n (%)
Age	≤50	28 (6.3)	126 (14.8)	8916 (73.5)
	>50	420 (93.8)	724 (85.2)	3221 (26.5)
Cholesterol				
	Borderline	161 (36.7)	268 (35.4)	2705 (27.6)
	High	148 (33.7)	203 (26.8)	1431 (14.6)
	Good	130 (29.6)	286 (37.8)	5651 (57.7)
Sex	female	210 (38.3)	404 (40.9)	6457 (51.8)
	male	339 (61.8)	583 (59.1)	6003 (48.2)
Sedentary	yes	248 (45.2)	431 (43.8)	4378 (35.1)
	no	301 (54.8)	554 (56.2)	8080 (64.9)
Education	< high school	337 (62.0)	607 (61.9)	4095 (33.0)
	≥ high school	207 (38.1)	373 (38.1)	8322 (67.0)
Marry	single	168 (30.7)	279 (28.4)	3974 (32.0)
	married	379 (69.3)	702 (71.6)	8459 (68.0)
Smoking Status				
	Never	141 (25.7)	247 (25.0)	4219 (33.9)
	Current	138 (25.1)	329 (33.3)	4081 (32.8)
	Former	270 (49.2)	411 (41.6)	4153 (33.4)
Alcohol Status				
	Never	35 (7.1)	60 (6.9)	411 (3.7)
	Current	319 (64.6)	594 (68.4)	9439 (83.9)
	Former	140 (28.3)	214 (24.7)	1395 (12.4)
Obese	no	342 (75.3)	609 (74.9)	8509 (82.0)
	yes	112 (24.7)	204 (25.1)	1868 (18.0)
Hypertensive	yes	254 (46.3)	325 (32.9)	1949 (15.6)
	no	295 (53.7)	662 (67.1)	10511 (84.4)
Diabetic	yes	86 (17.4)	126 (14.5)	497 (4.4)
	no	408 (82.6)	742 (85.5)	10690 (95.6)
Heart Attack	yes	113 (20.9)	100 (10.3)	322 (2.6)
	no	427 (79.1)	872 (89.7)	12058 (97.4)
Stroke	yes	69 (12.7)	45 (4.6)	187 (1.5)
	no	473 (87.3)	927 (95.4)	12226 (98.5)
Province	western	380 (69.2)	674 (68.3)	7497 (60.2)
	non-western	169 (30.8)	313 (31.7)	4963 (39.8)

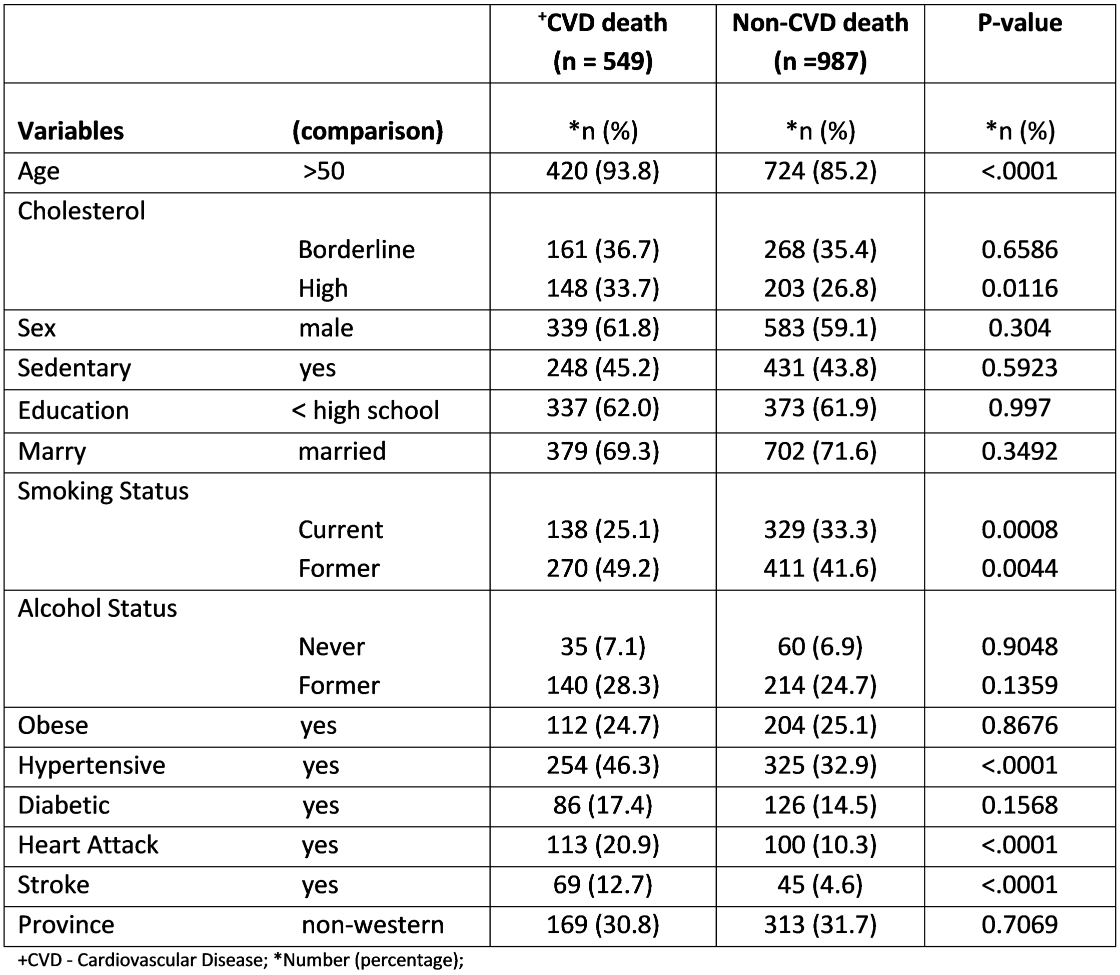
+CVD - Cardiovascular Disease; *Number (percentage)

4.3.2 Cause of Death

Among the dead subjects, I examined the associations between the causes of death and each of the potential risk factor using the chi-square test. I present the chi-square test results in

Table 4.3. The test results showed a significant difference in the distribution of death across the age groups. Subjects who are above 50 years of age are significantly more likely to die from CVD versus non-CVD events (93.8% vs 85.2%; P-value < 0.0001). Current smokers are significantly less likely to die from CVD versus non-CVD (25.1% vs 33.3%; P-value $= 0.001$). However, former smokers are more liable to die from CVD compared to non-CVD causes (49.2% vs 41.6%; P-value $= 0.004$). The proportion of CVD versus non-CVD death is not significantly different across sexes (P-value= 0.304), physical activity level (P-value $= 0.592$), education level (P-value= 0.997), marital status (P-value= 0.349), obesity status (P-value $=$ 0.868), diabetes status (P-value= 0.157), and province of residence (P-value $= 0.707$). Also, CVD versus non-CVD deaths is not significantly different among former alcohol users (P-value= 0.136) and subjects who never used alcohol (P-value= 0.905; Table 4.3). Those with high cholesterol are significantly more likely to die from CVD versus non-CVD events (33.7% vs 26.8%; P-value $= 0.012$). The proportion of subjects with borderline cholesterol who experience CVD deaths is higher than those who experience non-CVD deaths. However, the association is not significantly different in the two groups (36.7% vs 35.4%; P-value =0.659). CVD compared to non-CVD death is more likely to occur among hypertensive subjects, stroke survivors, and subjects who had witnessed a heart attack. These associations are very strong with P-values < 0.0001 (Table 4.3).

Table 4.3: Characteristics of study subjects based on cause of death



		$^{+}$CVD death (n = 549)	Non-CVD death (n =987)	P-value
Variables	**(comparison)**	*n (%)	*n (%)	*n (%)
Age	>50	420 (93.8)	724 (85.2)	<.0001
Cholesterol				
	Borderline	161 (36.7)	268 (35.4)	0.6586
	High	148 (33.7)	203 (26.8)	0.0116
Sex	male	339 (61.8)	583 (59.1)	0.304
Sedentary	yes	248 (45.2)	431 (43.8)	0.5923
Education	< high school	337 (62.0)	373 (61.9)	0.997
Marry	married	379 (69.3)	702 (71.6)	0.3492
Smoking Status				
	Current	138 (25.1)	329 (33.3)	0.0008
	Former	270 (49.2)	411 (41.6)	0.0044
Alcohol Status				
	Never	35 (7.1)	60 (6.9)	0.9048
	Former	140 (28.3)	214 (24.7)	0.1359
Obese	yes	112 (24.7)	204 (25.1)	0.8676
Hypertensive	yes	254 (46.3)	325 (32.9)	<.0001
Diabetic	yes	86 (17.4)	126 (14.5)	0.1568
Heart Attack	yes	113 (20.9)	100 (10.3)	<.0001
Stroke	yes	69 (12.7)	45 (4.6)	<.0001
Province	non-western	169 (30.8)	313 (31.7)	0.7069

+CVD - Cardiovascular Disease; *Number (percentage);

4.3.3 Kaplan-Meier Survival Estimate

The overall survival was estimated using Kaplan-Meier approach. Figure 4.2 shows that the survival probabilities at 5 years and 10 years are 0.98 and 0.94, respectively.

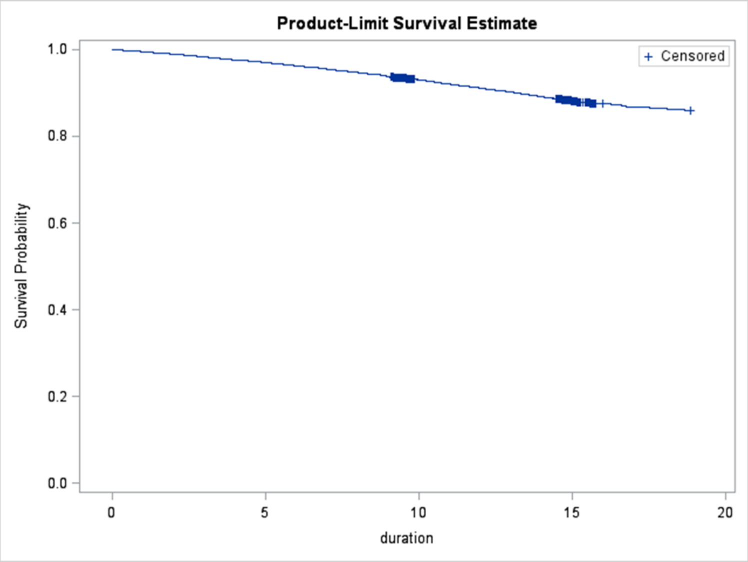


Figure 4.2: Kaplan-Meier estimate of the survival probability

4.3.4 Competing Risks Models

To answer Study Objective 2, which is to determine the risk factors for CVD using the competing risks approach, I applied the Cox multiplicative cause-specific hazards model, the F-G multiplicative subdistribution hazards model, the Aalen additive hazards model, and the L-Y additive hazards model for the competing events (CVD and non-CVD mortalities) to the study data.

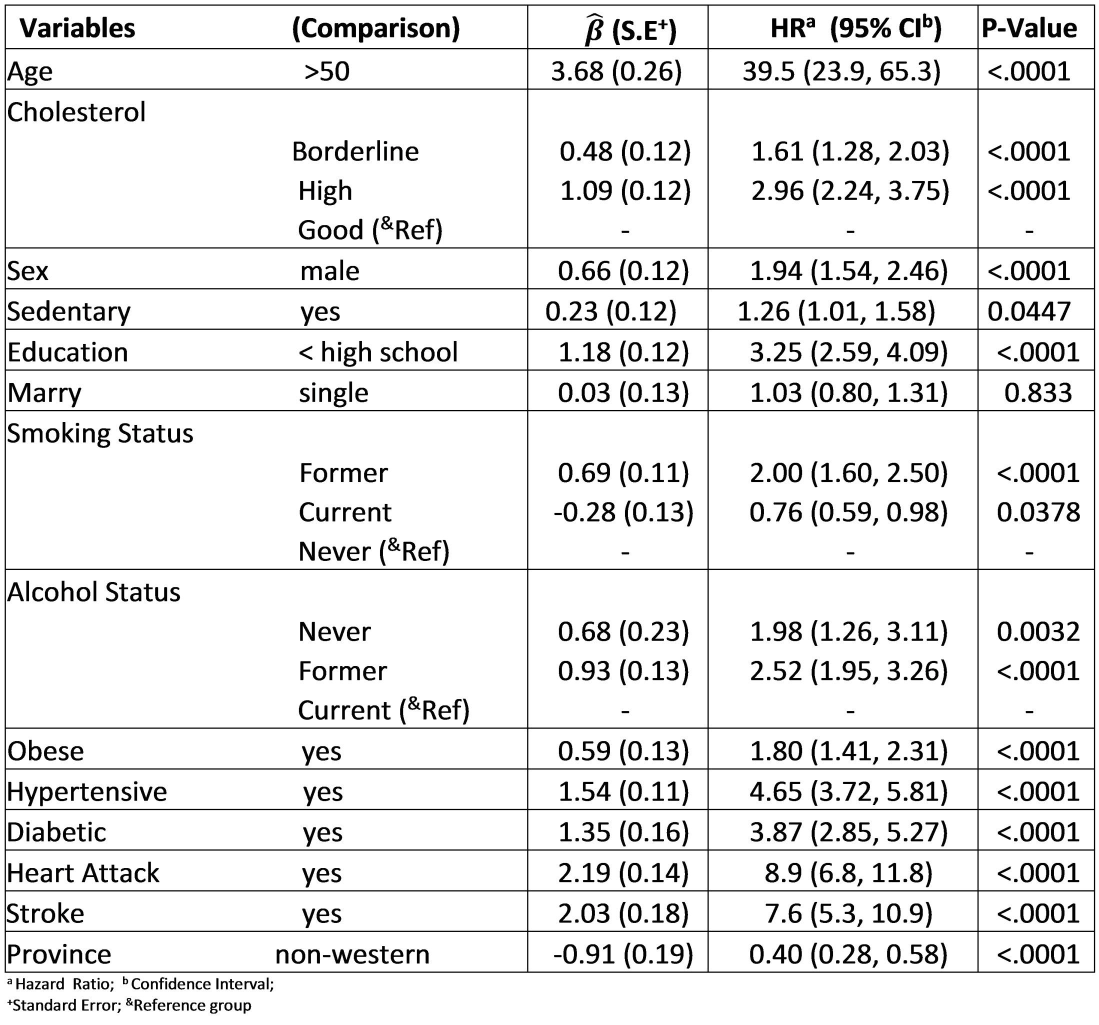
4.3.4.1 Cox Cause-Specific Hazards Model

I performed a univariate analysis for CVD and then used potential predictors to fit a multivariate model. Similarly, I performed a univariate analysis for non-CVD event, and with the relevant predictors, fit a multivariate model. The multivariate model was built based on the manual model selection recommended by Collett [1]. I started with a model consisting of all the significant (using a 5% significant level) covariates from the univariate analysis. Then I used the Wald statistic P-value to delete covariates in the potential multivariate model that were no more significant. To the model that has only significant covariates, I added covariates that were not significant in the univariate analysis one at a time, to see if they are now significant. Those that were significant were included in the multivariate model. I rechecked all covariates that were removed to be sure no significant covariate is left out of the multivariate model. I checked all two-way interactions of the covariates in the multivariate model. Significant interactions were included in the final model.

4.3.4.1.1 Cardiovascular Disease Outcome

The results of the univariate Cox CSH model for CVD death in Table 4.4 shows that Age, Cholesterol, Sex, Sedentary, Education, Smoking Status, Alcohol Status, Obese, Hypertensive, Diabetic, Heart Attack, Stroke, and Province are independently associated with CVD death. However, Marital status (Marry) is not a significant independent predictor of CVD death (P-value= 0.833; Table 4.4).

Table 4.4: Univariate Cox cause-specific hazards model for CVD



Variables	(Comparison)	$\widehat{\beta}$ (S.E^{+})	HR^{a} (95% CI^{b})	P-Value
Age	>50	3.68 (0.26)	39.5 (23.9, 65.3)	<.0001
Cholesterol				
	Borderline	0.48 (0.12)	1.61 (1.28, 2.03)	<.0001
	High	1.09 (0.12)	2.96 (2.24, 3.75)	<.0001
	Good ($^{\&}$Ref)	-	-	-
Sex	male	0.66 (0.12)	1.94 (1.54, 2.46)	<.0001
Sedentary	yes	0.23 (0.12)	1.26 (1.01, 1.58)	0.0447
Education	< high school	1.18 (0.12)	3.25 (2.59, 4.09)	<.0001
Marry	single	0.03 (0.13)	1.03 (0.80, 1.31)	0.833
Smoking Status				
	Former	0.69 (0.11)	2.00 (1.60, 2.50)	<.0001
	Current	-0.28 (0.13)	0.76 (0.59, 0.98)	0.0378
	Never ($^{\&}$Ref)	-	-	-
Alcohol Status				
	Never	0.68 (0.23)	1.98 (1.26, 3.11)	0.0032
	Former	0.93 (0.13)	2.52 (1.95, 3.26)	<.0001
	Current ($^{\&}$Ref)	-	-	-
Obese	yes	0.59 (0.13)	1.80 (1.41, 2.31)	<.0001
Hypertensive	yes	1.54 (0.11)	4.65 (3.72, 5.81)	<.0001
Diabetic	yes	1.35 (0.16)	3.87 (2.85, 5.27)	<.0001
Heart Attack	yes	2.19 (0.14)	8.9 (6.8, 11.8)	<.0001
Stroke	yes	2.03 (0.18)	7.6 (5.3, 10.9)	<.0001
Province	non-western	-0.91 (0.19)	0.40 (0.28, 0.58)	<.0001

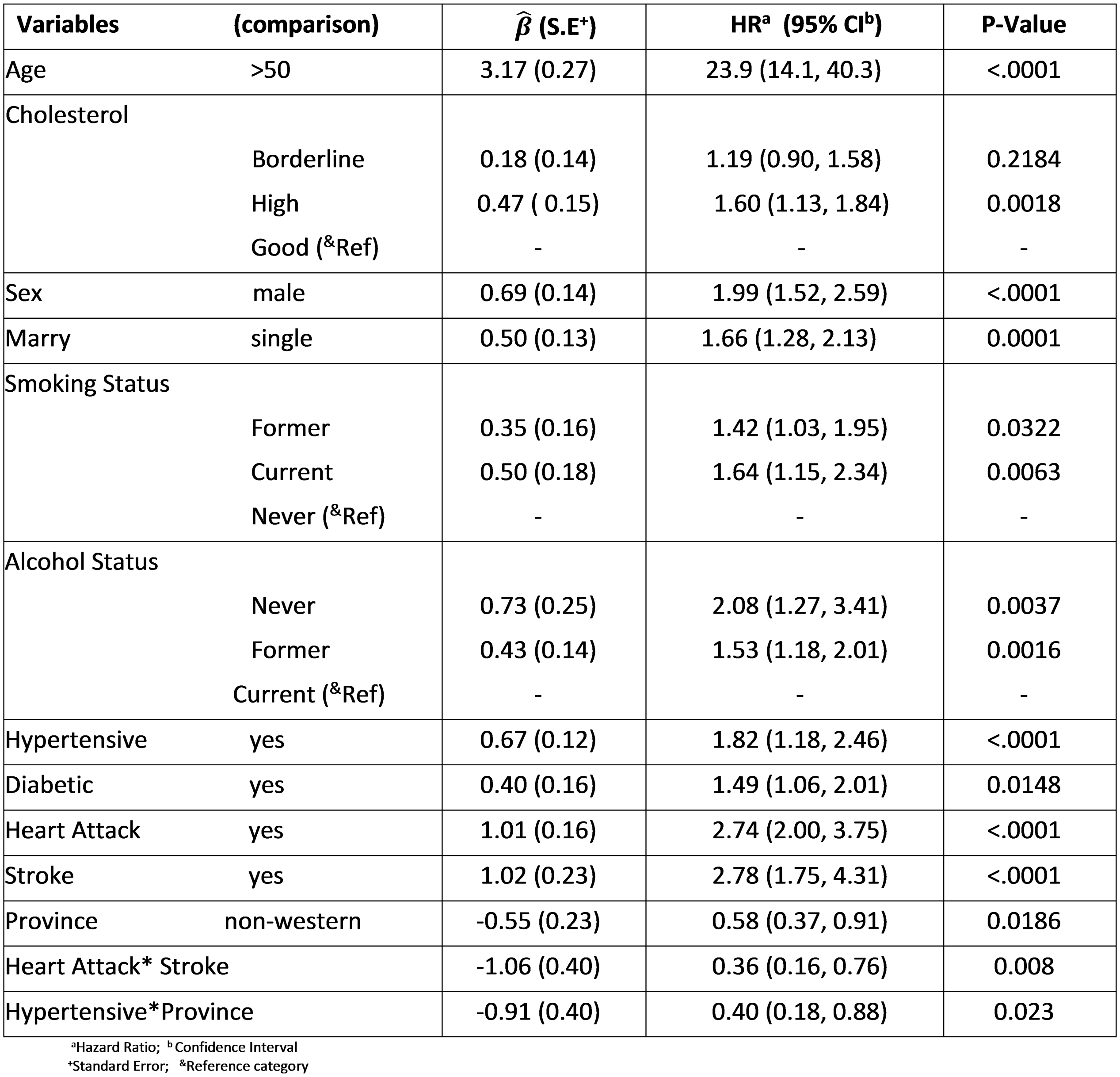
[a] Hazard Ratio; [b] Confidence Interval;
[+]Standard Error; [&]Reference group

The multivariate Cox cause-specific model for CVD presented in Table 4.5 shows that Age, Cholesterol, Sex, Marry, Smoking status, Alcohol status, Hypertensive, Diabetic, Heart Attack, Stroke, and Province are significant predictors of CVD. Interactions between Heart Attack & Stroke (P-value= 0.008), and that between Hypertensive & Province (P-value= 0.023) are significant and thus included in the final model (Table 4.5). Upon adjusting for the other covariates in the model, the hazard of CVD death increased significantly among subjects who are over 50 years of age compared to those less than 50 years of age (HR= 23.9; 95% CI: (14.1, 40.3); P-value=< 0.0001). Subjects who never used alcohol were twice at risk of CVD death compared to current alcohol users (HR= 2.08; 95% CI: (1.27, 3.41); P-value= 0.004), but those who consumed alcohol in the past have a 53% (HR= 1.53; 95% CI: (1.18, 2.01); P-value= 0.002) higher risk of CVD compared to current alcohol users. The risk of CVD mortality for subjects who had experienced a stroke and a heart attack was 2.64 times higher compared to subjects without a stroke and a heart attack

$$HR = exp(1.01 + 1.02 - 1.06) = 2.64$$

The 95% confidence interval of the hazard ratio (HR) estimate is between 1.51 and 4.90. Thus, a stroke together with a heart attack experience increased the hazard of CVD death significantly compared to neither a stroke nor a heart attack.

Table 4.5: **Multivariate Cox cause-specific hazards model for CVD**



Variables	(comparison)	$\widehat{\beta}$ (S.E$^+$)	HR^a (95% CI^b)	P-Value
Age	>50	3.17 (0.27)	23.9 (14.1, 40.3)	<.0001
Cholesterol				
	Borderline	0.18 (0.14)	1.19 (0.90, 1.58)	0.2184
	High	0.47 (0.15)	1.60 (1.13, 1.84)	0.0018
	Good ($^{\&}$Ref)	-	-	-
Sex	male	0.69 (0.14)	1.99 (1.52, 2.59)	<.0001
Marry	single	0.50 (0.13)	1.66 (1.28, 2.13)	0.0001
Smoking Status				
	Former	0.35 (0.16)	1.42 (1.03, 1.95)	0.0322
	Current	0.50 (0.18)	1.64 (1.15, 2.34)	0.0063
	Never ($^{\&}$Ref)	-	-	-
Alcohol Status				
	Never	0.73 (0.25)	2.08 (1.27, 3.41)	0.0037
	Former	0.43 (0.14)	1.53 (1.18, 2.01)	0.0016
	Current ($^{\&}$Ref)	-	-	-
Hypertensive	yes	0.67 (0.12)	1.82 (1.18, 2.46)	<.0001
Diabetic	yes	0.40 (0.16)	1.49 (1.06, 2.01)	0.0148
Heart Attack	yes	1.01 (0.16)	2.74 (2.00, 3.75)	<.0001
Stroke	yes	1.02 (0.23)	2.78 (1.75, 4.31)	<.0001
Province	non-western	-0.55 (0.23)	0.58 (0.37, 0.91)	0.0186
Heart Attack* Stroke		-1.06 (0.40)	0.36 (0.16, 0.76)	0.008
Hypertensive*Province		-0.91 (0.40)	0.40 (0.18, 0.88)	0.023

[a]Hazard Ratio; [b]Confidence Interval
[+]Standard Error; [&]Reference category

4.3.4.1.2 Model Diagnostics

I examined the adequacy of the Cox CVD-specific hazards model. The proportional hazard assumption and the strength of the survival prediction of the model were also checked. The proportional CSH assumption was assessed using the cumulative hazard plot, and it was valid for all the covariates in the final model except Marital status (Marry) as shown in Figure 4.3. The overlapping of the hazard curves for the levels of marital status suggest that the hazard is not proportional.

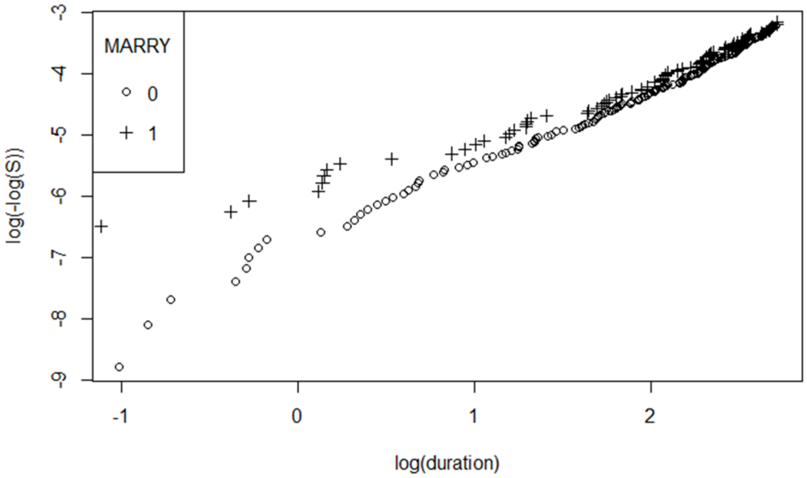


Figure 4.3: Log-cumulative hazards plot for marital status

The Cox-Snell residual plot in Figure 4.4 follows a 45^0 line with a zero intercept and an approximately unit slope. The shape of the plot indicates that the final multivariate model for the event of interest can be considered as appropriate for the dataset. The estimates in the tail of the plot lie below the 45^0 line, and this can be considered a result of long follow-up with a small number of observations remaining in the dataset.

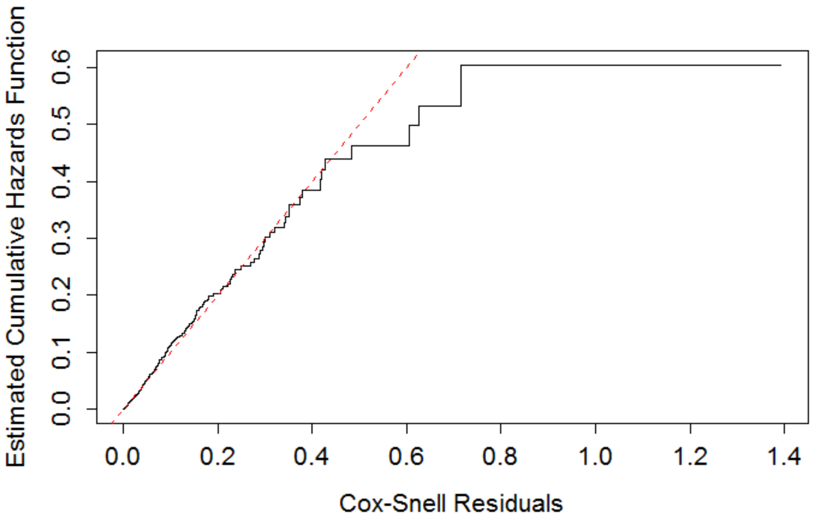


Figure 4.4: Cumulative hazards plot of the Cox-Snell residuals

The profound skewness of the martingale residual makes it less informative for possible outliers in the data. In contrast to martingale residuals, the deviance residual is more appropriate for detecting outliers. The plot of the deviance residuals in Figure 4.6 shows that a few observations have relatively large residuals. However, for most of the observations, the residuals were small as expected. Most of the participants with large positive risk scores (i.e. they have a higher than average risk of CVD death) have residuals that are close to zero. Also, those that have a lower risk of CVD death also have residuals that are near zero. However, the heavy censoring (89%) in the dataset resulted in the large clustering of points near zero, thus distorting the expected normally distributed residuals.

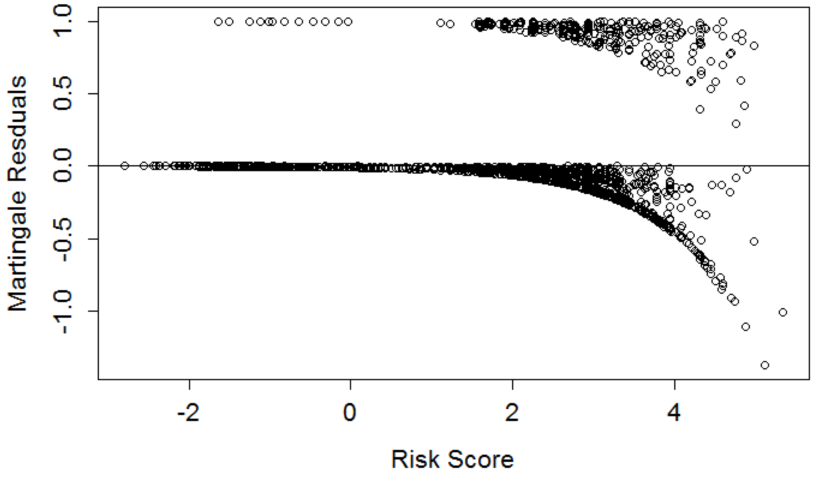


Figure 4.5: Martingale residuals plot

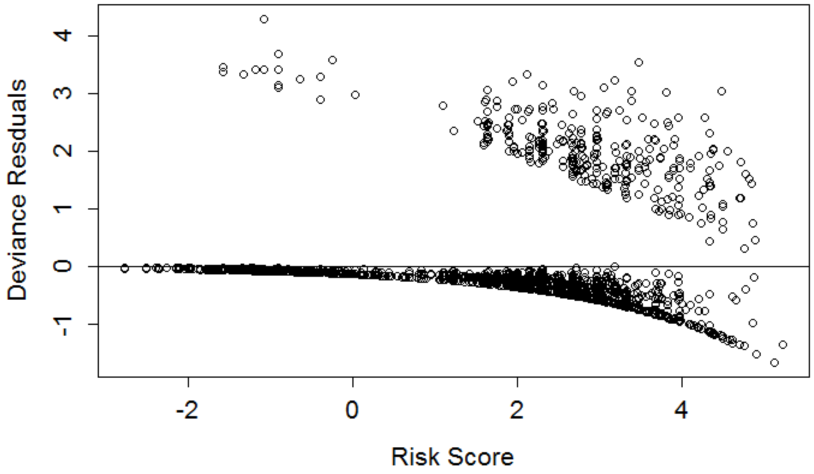
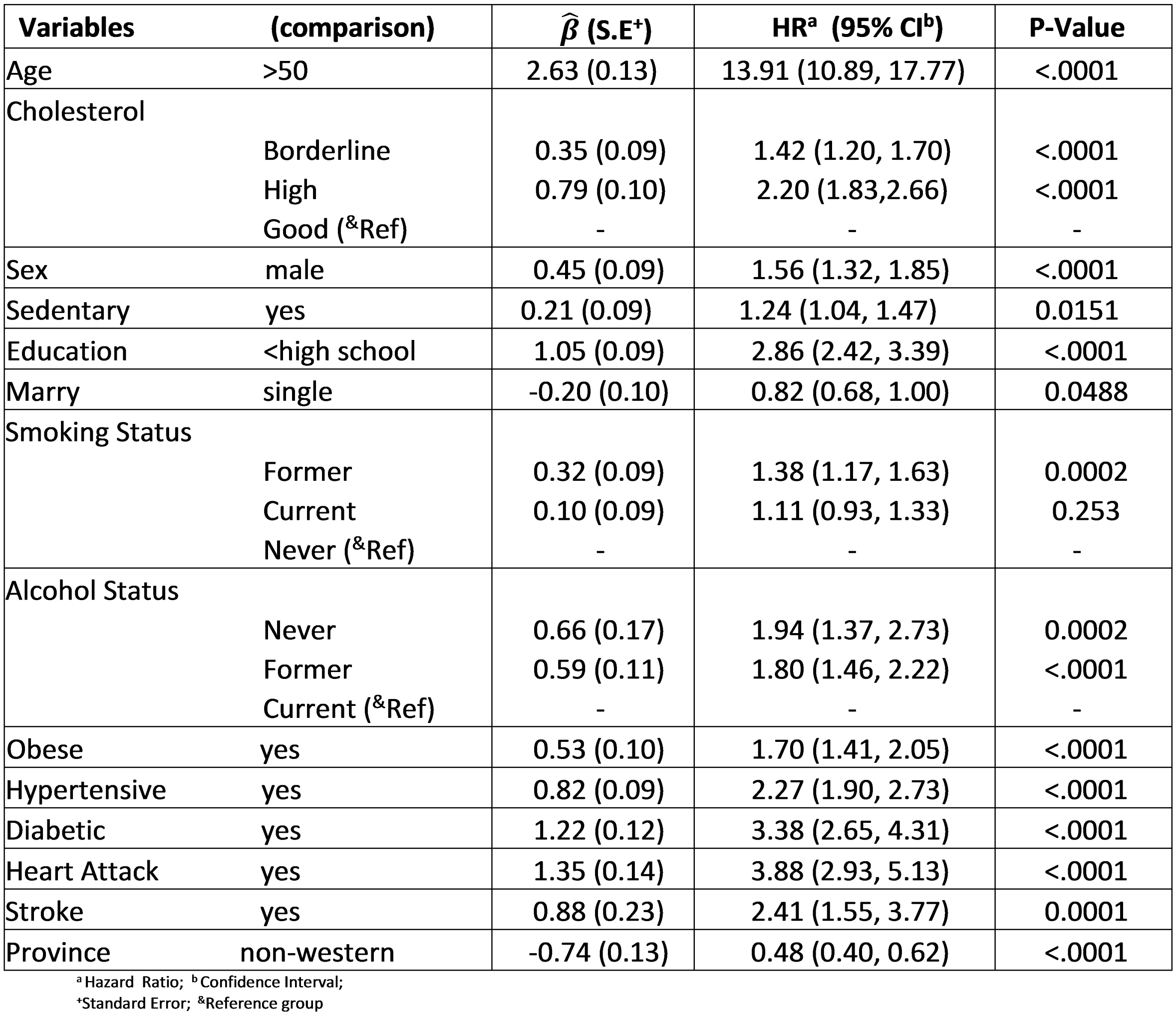


Figure 4.6: Deviance residuals plot

4.3.4.1.3 Non-Cardiovascular Disease Outcome

The univariate Cox cause-specific model for non-CVD presented in Table 4.6 identified the following as significant predictors: Age, Cholesterol, Sex, Sedentary, Education, Marry, Smoking status, Alcohol status, Obese, Hypertensive, Diabetic, Heart Attack, Stroke, and Province.

Table 4.6: Univariate Cox cause-specific hazards model for non-CVD

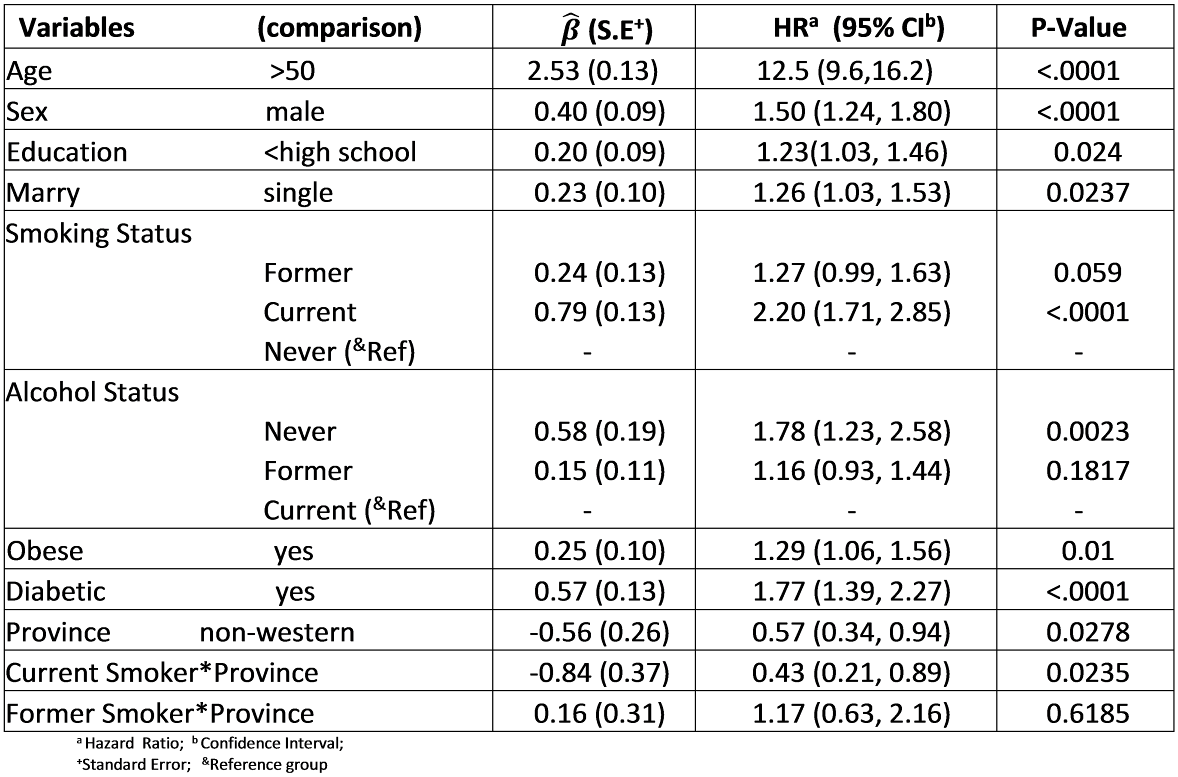


Variables	(comparison)	$\widehat{\beta}$ (S.E^{+})	HR^{a} (95% CI^{b})	P-Value
Age	>50	2.63 (0.13)	13.91 (10.89, 17.77)	<.0001
Cholesterol				
	Borderline	0.35 (0.09)	1.42 (1.20, 1.70)	<.0001
	High	0.79 (0.10)	2.20 (1.83,2.66)	<.0001
	Good ($^{\&}$Ref)	-	-	-
Sex	male	0.45 (0.09)	1.56 (1.32, 1.85)	<.0001
Sedentary	yes	0.21 (0.09)	1.24 (1.04, 1.47)	0.0151
Education	<high school	1.05 (0.09)	2.86 (2.42, 3.39)	<.0001
Marry	single	-0.20 (0.10)	0.82 (0.68, 1.00)	0.0488
Smoking Status				
	Former	0.32 (0.09)	1.38 (1.17, 1.63)	0.0002
	Current	0.10 (0.09)	1.11 (0.93, 1.33)	0.253
	Never ($^{\&}$Ref)	-	-	-
Alcohol Status				
	Never	0.66 (0.17)	1.94 (1.37, 2.73)	0.0002
	Former	0.59 (0.11)	1.80 (1.46, 2.22)	<.0001
	Current ($^{\&}$Ref)	-	-	-
Obese	yes	0.53 (0.10)	1.70 (1.41, 2.05)	<.0001
Hypertensive	yes	0.82 (0.09)	2.27 (1.90, 2.73)	<.0001
Diabetic	yes	1.22 (0.12)	3.38 (2.65, 4.31)	<.0001
Heart Attack	yes	1.35 (0.14)	3.88 (2.93, 5.13)	<.0001
Stroke	yes	0.88 (0.23)	2.41 (1.55, 3.77)	0.0001
Province	non-western	-0.74 (0.13)	0.48 (0.40, 0.62)	<.0001

^{a}Hazard Ratio; ^{b}Confidence Interval;
$^{+}$Standard Error; $^{\&}$Reference group

Table 4.7 shows that the final multivariate Cox cause-specific model for non-CVD includes: Age, Sex, Education, Marry, Smoking Status, Alcohol Status, Obese, Diabetic, and Province as significant predictors of non-CVD mortality. Also, the model include the interaction of Smoking Status & Province. Controlling for the other covariates in the final model, the risk of dying from non-CVD for subjects above 50 years is 12.5 (95% CI: (9.6, 16.2); P-value= < 0.0001) relative to those less than 50 years of age. Also, subjects who never used alcohol have a 78% (HR= 1.78; 95% CI:(1.23, 2.58); P-value= 0.002) increased risk of non-CVD mortality compared to current alcohol users. However, the risk of non-CVD mortality is not significantly different for former and current alcohol users (P-value= 0.182, Table 4.7).

Table 4.7: Multivariate Cox cause-specific hazards model for non-CVD



Variables	(comparison)	$\widehat{\beta}$ (S.E$^+$)	HR^a (95% CI^b)	P-Value
Age	>50	2.53 (0.13)	12.5 (9.6,16.2)	<.0001
Sex	male	0.40 (0.09)	1.50 (1.24, 1.80)	<.0001
Education	<high school	0.20 (0.09)	1.23(1.03, 1.46)	0.024
Marry	single	0.23 (0.10)	1.26 (1.03, 1.53)	0.0237
Smoking Status				
	Former	0.24 (0.13)	1.27 (0.99, 1.63)	0.059
	Current	0.79 (0.13)	2.20 (1.71, 2.85)	<.0001
	Never (&Ref)	-	-	-
Alcohol Status				
	Never	0.58 (0.19)	1.78 (1.23, 2.58)	0.0023
	Former	0.15 (0.11)	1.16 (0.93, 1.44)	0.1817
	Current (&Ref)	-	-	-
Obese	yes	0.25 (0.10)	1.29 (1.06, 1.56)	0.01
Diabetic	yes	0.57 (0.13)	1.77 (1.39, 2.27)	<.0001
Province	non-western	-0.56 (0.26)	0.57 (0.34, 0.94)	0.0278
Current Smoker*Province		-0.84 (0.37)	0.43 (0.21, 0.89)	0.0235
Former Smoker*Province		0.16 (0.31)	1.17 (0.63, 2.16)	0.6185

[a] Hazard Ratio; [b] Confidence Interval;
[+] Standard Error; [&] Reference group

4.3.4.1.4 Model Diagnostics

I examined the fitness of the multivariate model for the competing non-CVD event. The proportional cause-specific hazard assumption was checked and the strength of the survival prediction was investigated. The cumulative hazards plot was employed to see if the proportional CSH assumption holds for all the covariates. Figure 4.7 might raise some doubt about the multivariate Cox cause-specific model for non-CVD based on the overlapping of the hazard curves for covariates Marital status (Marry) and Smoking Status.

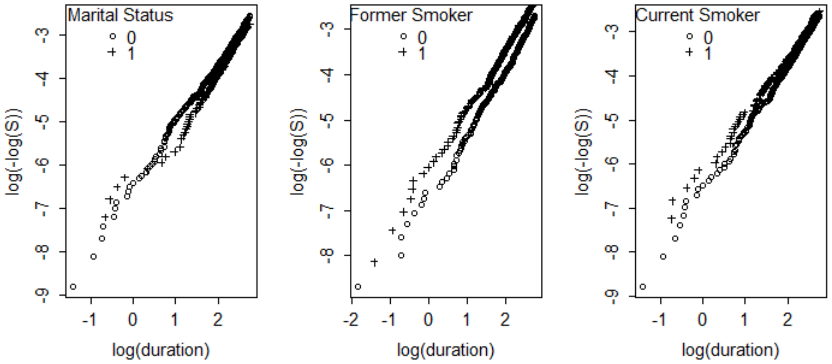


Figure 4.7: Log-cumulative hazards plot for marital status and smoking status

However, the Cox-Snell residual plot in Figure 4.8, with a zero intercept and a unit slope following a 45^0 line suggests that the multivariate Cox model for non-CVD event is a good fit to the data.

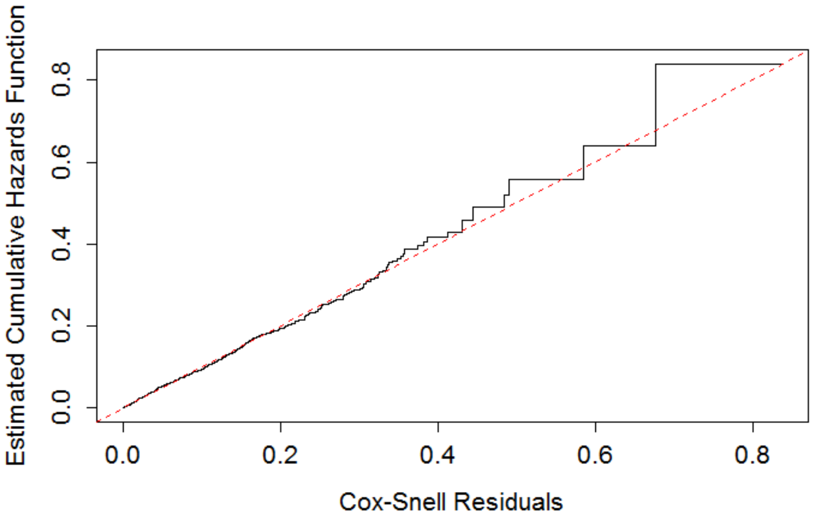


Figure 4.8: Cumulative hazards plot of the Cox-Snell residuals

Figure 4.9 indicates that there are no outliers and all subjects have residuals that are not far away from zero. Outliers are not easily detected in a martingale residuals plot. Contrary to Figure 4.9, the plot of the deviance residuals in Figure 4.10 shows some relatively large residuals, suggesting outliers might exist in the data. However, for most of the observations, the residuals are small as expected in a good model. Most of the participants with large positive risk scores have residuals that are close to zero. Those that have a lower risk of CVD death also have residuals that are approximately zero. Thus, indicating that the observations do not deviate too much from a well-fitted model. The impact of the heavy censoring in the dataset is still visible as the residuals are concentrated near zero.

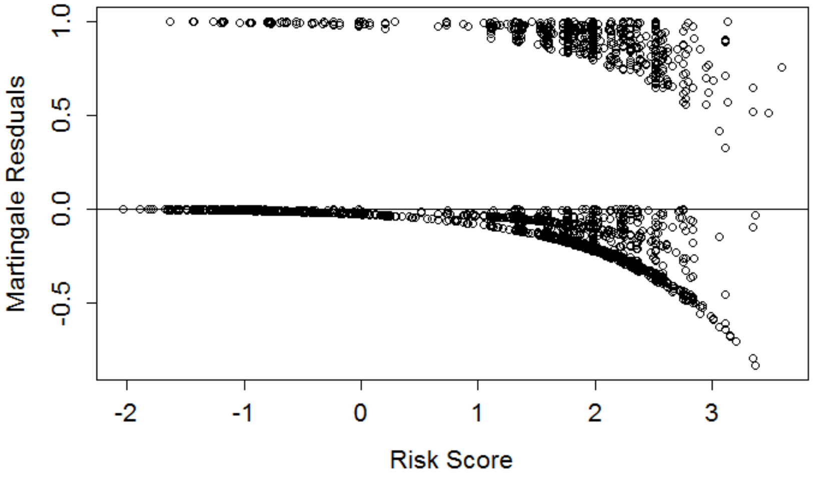


Figure 4.9: Martingale residuals plot

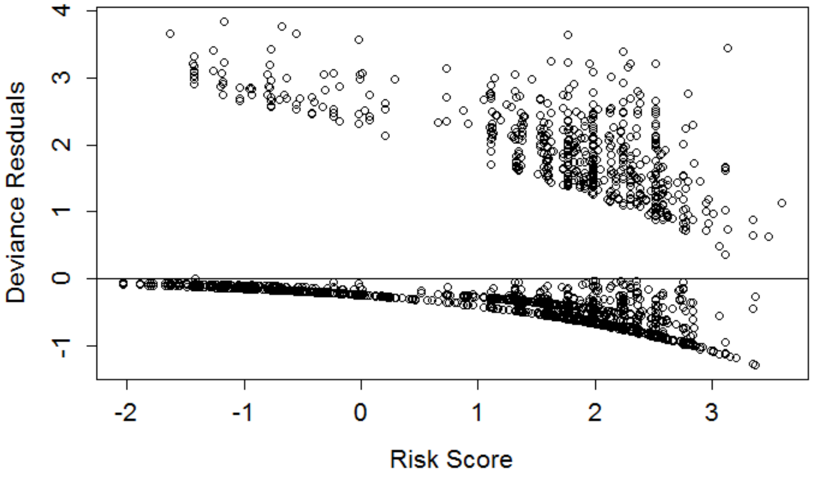


Figure 4.10: Deviance residuals plot

4.3.4.1.5 Summary of Cox Cause-Specific Model for CVD and non-CVD

Figure 4.11 compares the significant predictors included in the final Cox model for CVD and non-CVD outcomes. Age, Sex, Marry, Smoking status, Alcohol status, Diabetic, and Province show significant common association with both CVD and non-CVD death outcomes. Education and Obese are related with non-CVD death outcome. Whereas, Cholesterol, Hypertensive, Heart Attack, and Stroke are predictors of CVD mortality (Figure 4.11).

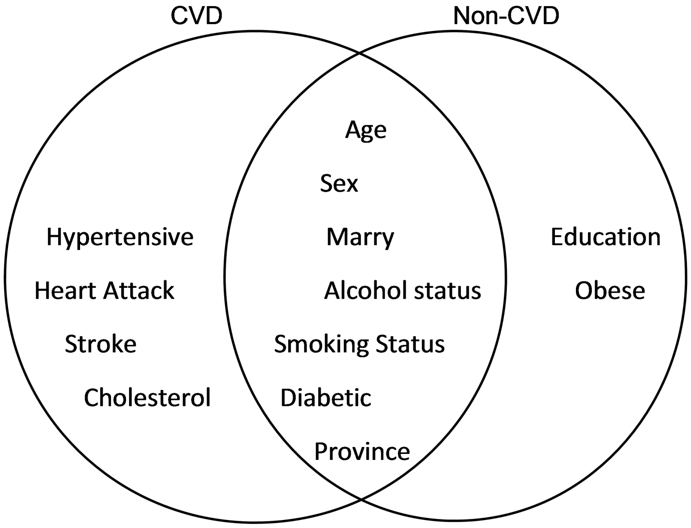


Figure 4.11: Venn diagram comparing risk factors for CVD and Non-CVD risk factors under the Cox cause-specific hazards model

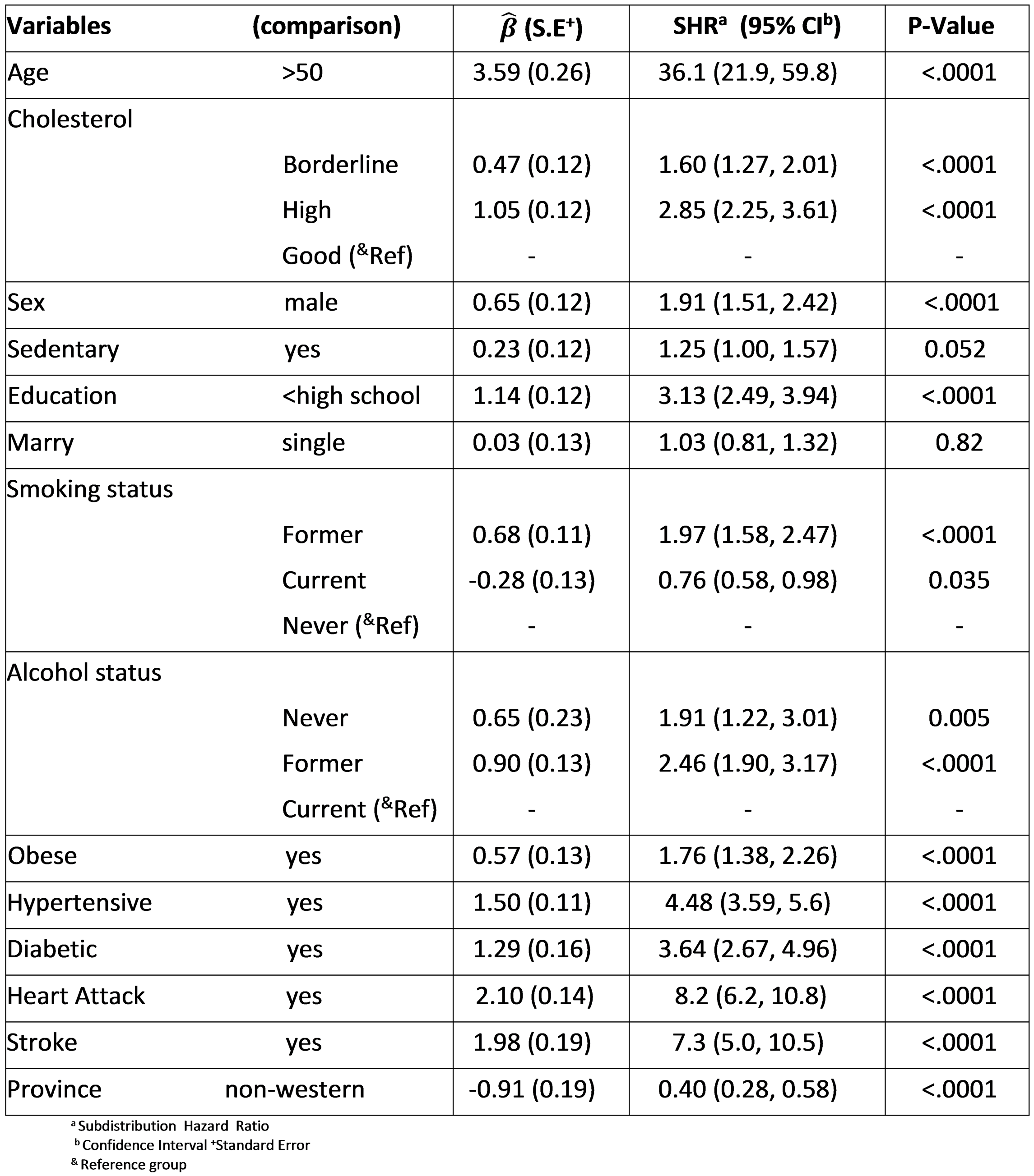
4.3.4.2 Fine & Gray Subdistribution Hazards Model

The F-G model was applied to the CHHS follow-up data. Each variable was assessed for significant association with the study events. For CVD and non-CVD mortalities, I present the univariate and multivariate analysis below.

4.3.4.2.1 Cardiovascular Disease Outcome

Table 4.8 shows that significant predictors in the univariate F-G model for CVD includes: Age, Cholesterol, Sex, Sedentary, Education, Smoking status, Alcohol status, Obese, Hypertensive, Diabetic, Heart Attack, Stroke, and Province. All but Marry (P-value = 0.82) show significant relationship with CVD death (Table 4.8).

Table 4.8: Univariate Fine & Gray subdistribution hazards model for CVD



Variables	(comparison)	$\widehat{\beta}$ (S.E[+])	SHR[a] (95% CI[b])	P-Value
Age	>50	3.59 (0.26)	36.1 (21.9, 59.8)	<.0001
Cholesterol				
	Borderline	0.47 (0.12)	1.60 (1.27, 2.01)	<.0001
	High	1.05 (0.12)	2.85 (2.25, 3.61)	<.0001
	Good ([&]Ref)	-	-	-
Sex	male	0.65 (0.12)	1.91 (1.51, 2.42)	<.0001
Sedentary	yes	0.23 (0.12)	1.25 (1.00, 1.57)	0.052
Education	<high school	1.14 (0.12)	3.13 (2.49, 3.94)	<.0001
Marry	single	0.03 (0.13)	1.03 (0.81, 1.32)	0.82
Smoking status				
	Former	0.68 (0.11)	1.97 (1.58, 2.47)	<.0001
	Current	-0.28 (0.13)	0.76 (0.58, 0.98)	0.035
	Never ([&]Ref)	-	-	-
Alcohol status				
	Never	0.65 (0.23)	1.91 (1.22, 3.01)	0.005
	Former	0.90 (0.13)	2.46 (1.90, 3.17)	<.0001
	Current ([&]Ref)	-	-	-
Obese	yes	0.57 (0.13)	1.76 (1.38, 2.26)	<.0001
Hypertensive	yes	1.50 (0.11)	4.48 (3.59, 5.6)	<.0001
Diabetic	yes	1.29 (0.16)	3.64 (2.67, 4.96)	<.0001
Heart Attack	yes	2.10 (0.14)	8.2 (6.2, 10.8)	<.0001
Stroke	yes	1.98 (0.19)	7.3 (5.0, 10.5)	<.0001
Province	non-western	-0.91 (0.19)	0.40 (0.28, 0.58)	<.0001

[a] Subdistribution Hazard Ratio
[b] Confidence Interval [+]Standard Error
[&] Reference group

Table 4.9 shows that the multivariate F-G subdistribution model for CVD includes: Age, Cholesterol, Sex, Marry, Smoking status, Alcohol status, Hypertensive, Heart Attack, Stroke, and Province. Interactions between Heart Attack & Stroke (P-value=0.023) and that between Hypertensive & Province (P-value=0.023) are also included in the model. Adjusting for the other variables in the model, subjects who are over 50 years were 22 times more at risk of CVD death compared to those below 50 years of age(SHR= 22.2; 95% CI: (13.0, 37.9);

P-value < 0.0001). The subdistribution hazard of CVD death increased by 89% (SHR= 1.89; 95% CI: (1.18, 3.04); P-value= 0.009) and 55% (SHR= 1.55; 95% CI: (1.19, 2.04); P-value= 0.001) for never and former alcohol consumers, respectively compared to current alcohol users. The subdistribution hazard of CVD mortality for subjects who had experienced both a stroke and a heart attack was 2.97 times higher compared to subjects with neither a stroke nor a heart attack

$$SHR = exp(1.04 + 0.98 - 0.93) = 2.97.$$

The 95% confidence interval of the SHR estimate is between 1.72 and 5.26, which implies that those who have had a stroke and a heart attack before the survey time have a hazard of death that is significantly distinct from those who neither experience a stroke nor a heart attack.

Table 4.9: **Multivariate Fine & Gray subdistribution hazards model for CVD**

Variables	(comparison)	$\widehat{\beta}$ (SE[+])	SHR[a] (95% CI[b])	P-Value
Age	>50	3.10 (0.27)	22.2 (13.0, 37.9)	<.0001
Cholesterol				
	Borderline	0.16 (0.15)	1.17 (0.88, 1.58)	0.275
	High	0.43 (0.16)	1.53 (1.13, 2.08)	0.0058
	Good ([&]Ref)	-	-	-
Sex	male	0.63 (0.13)	1.86 (1.44, 2.43)	<.0001
Marry	single	0.48 (0.13)	1.62 (1.25, 2.10)	0.0002
Smoking status				
	Former	0.35 (0.16)	1.42 (1.04, 1.92)	0.0286
	Current	0.43 (0.18)	1.53 (1.08, 2.18)	0.0183
	Never ([&]Ref)	-	-	-
Alcohol status				
	Never	0.64 (0.24)	1.89 (1.18, 3.04)	0.0087
	Former	0.44 (0.14)	1.55 (1.19, 2.04)	0.0013
	Current ([&]Ref)	-	-	-
Hypertensive	yes	0.68 (0.13)	1.99 (1.55, 2.54)	<.0001
Heart Attack	yes	0.98 (0.17)	2.65 (1.92, 3.67)	<.0001
Stroke	yes	1.04 (0.24)	2.82 (1.76, 4.52)	<.0001
Province	non-western	-0.56 (0.23)	0.57 (0.36, 0.90)	0.016
Heart Attack* Stroke		-0.93 (0.41)	0.40 (0.18, 0.88)	0.023
Hypertensive*Province		-0.92 (0.40)	0.40 (0.18, 0.88)	0.023

[a] Subdistribution Hazard Ratio; [b] Confidence Interval
[+] Standard Error; [&] Reference group

4.3.4.2.2 Model Diagnostics

The plots of log (- log (1 - CIF)) against log (time) for the covariates in the F-G models were examined to check the assumption of proportional subdistribution hazard. The levels of the covariates were represented by curves and those were parallel for all but covariate ”Marital status”, suggesting that marital status violates the proportional subdistribution hazards assumption (Figure 4.12).

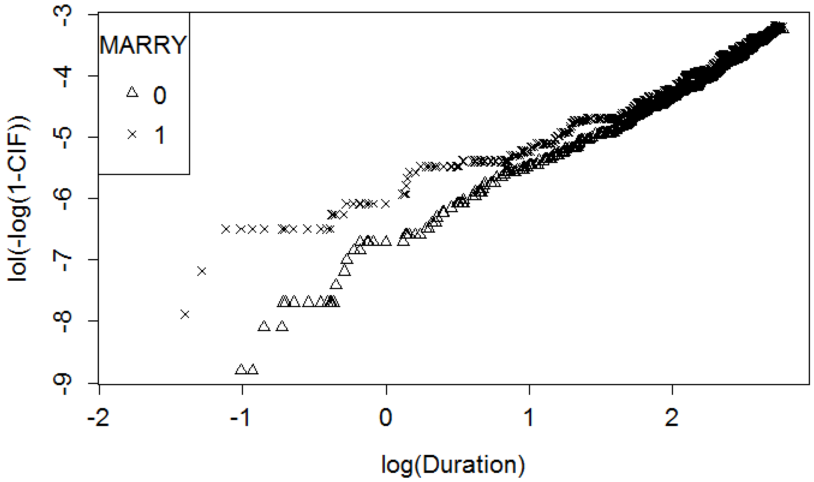
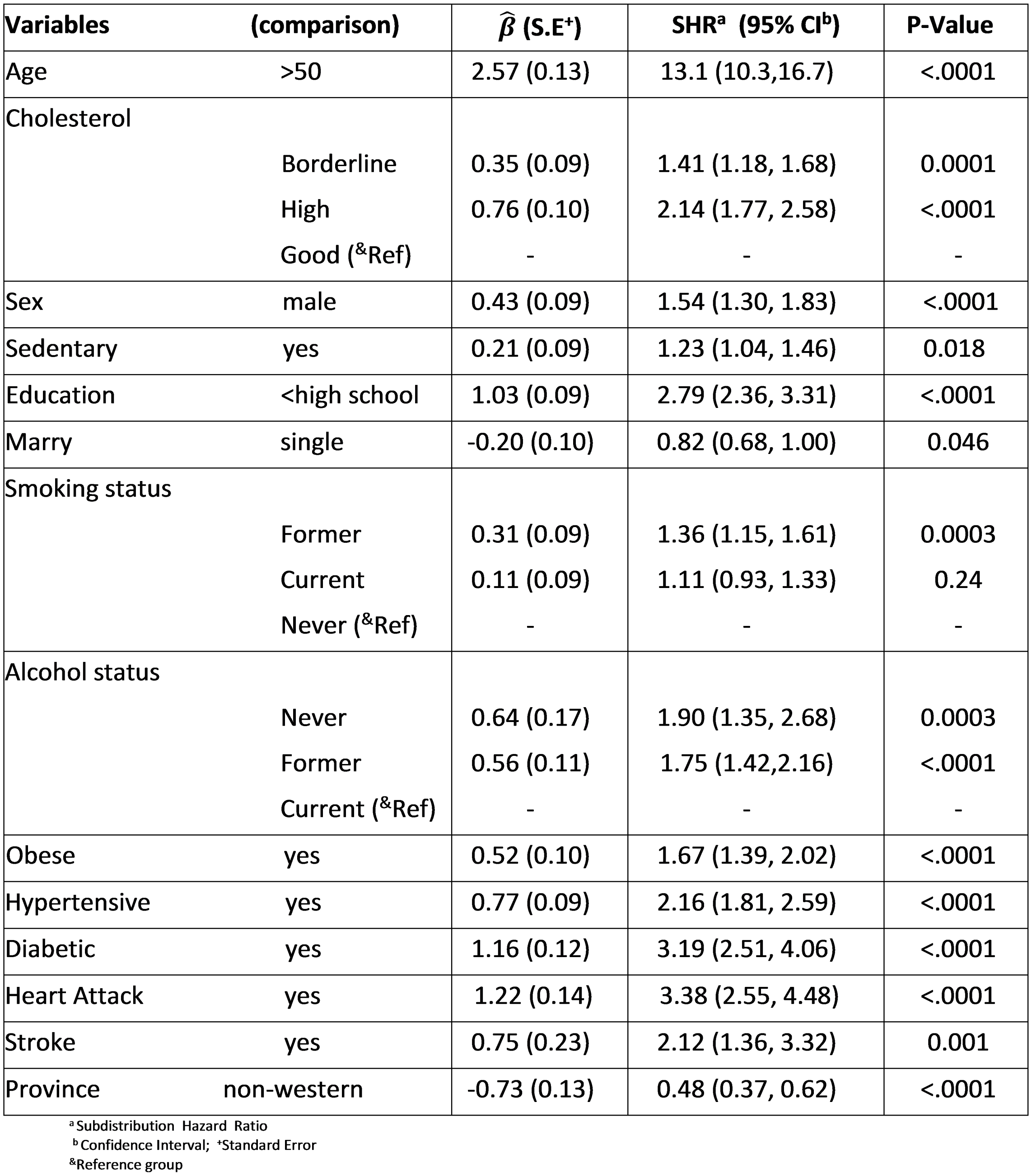


Figure 4.12: Proportionality of the subdistribution hazards for marital status

4.3.4.2.3 Non-Cardiovascular Disease Outcome

Table 4.10 shows that significant predictors in the univariate F-G model for non-CVD death are: Age, Cholesterol, Sex, Sedentary, Education, Marry, Smoking status, Alcohol status, Obese, Hypertensive, Diabetic, Heart Attack, Stroke, and Province.

Table 4.10: Univariate Fine & Gray subdistribution hazards model for Non-CVD

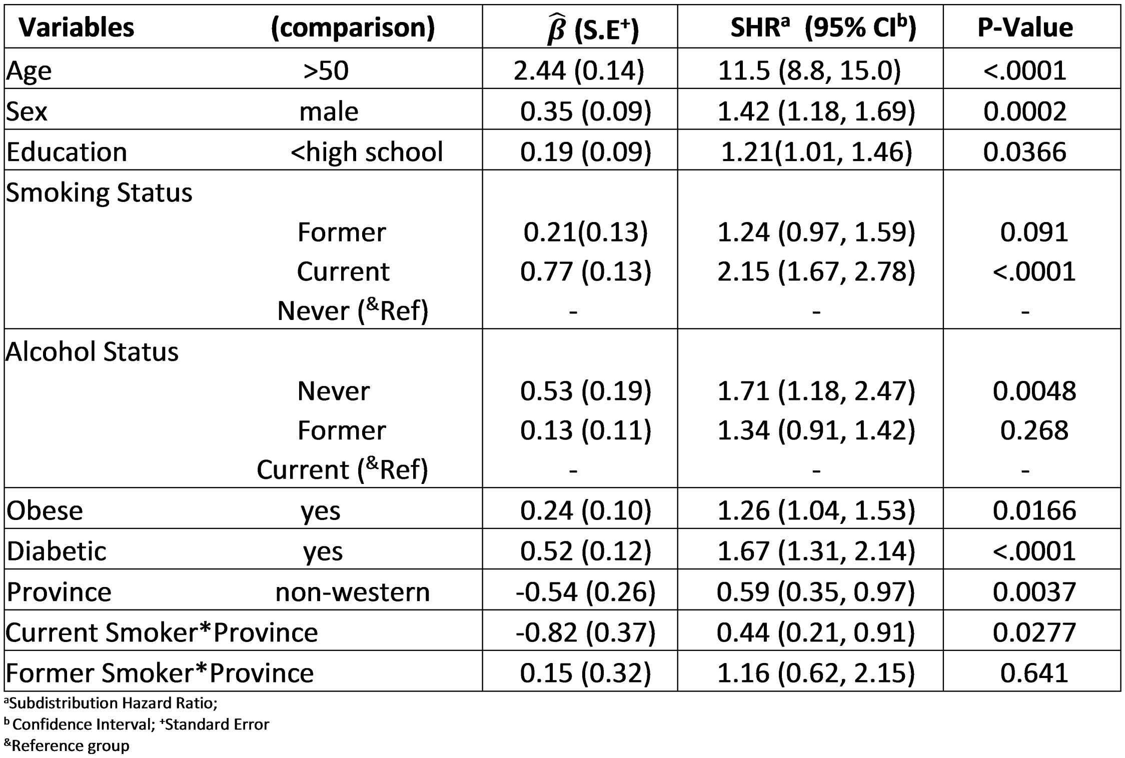


Variables	(comparison)	$\widehat{\beta}$ (S.E$^+$)	SHR[a] (95% CI[b])	P-Value
Age	>50	2.57 (0.13)	13.1 (10.3,16.7)	<.0001
Cholesterol				
	Borderline	0.35 (0.09)	1.41 (1.18, 1.68)	0.0001
	High	0.76 (0.10)	2.14 (1.77, 2.58)	<.0001
	Good (&Ref)	-	-	-
Sex	male	0.43 (0.09)	1.54 (1.30, 1.83)	<.0001
Sedentary	yes	0.21 (0.09)	1.23 (1.04, 1.46)	0.018
Education	<high school	1.03 (0.09)	2.79 (2.36, 3.31)	<.0001
Marry	single	-0.20 (0.10)	0.82 (0.68, 1.00)	0.046
Smoking status				
	Former	0.31 (0.09)	1.36 (1.15, 1.61)	0.0003
	Current	0.11 (0.09)	1.11 (0.93, 1.33)	0.24
	Never (&Ref)	-	-	-
Alcohol status				
	Never	0.64 (0.17)	1.90 (1.35, 2.68)	0.0003
	Former	0.56 (0.11)	1.75 (1.42,2.16)	<.0001
	Current (&Ref)	-	-	-
Obese	yes	0.52 (0.10)	1.67 (1.39, 2.02)	<.0001
Hypertensive	yes	0.77 (0.09)	2.16 (1.81, 2.59)	<.0001
Diabetic	yes	1.16 (0.12)	3.19 (2.51, 4.06)	<.0001
Heart Attack	yes	1.22 (0.14)	3.38 (2.55, 4.48)	<.0001
Stroke	yes	0.75 (0.23)	2.12 (1.36, 3.32)	0.001
Province	non-western	-0.73 (0.13)	0.48 (0.37, 0.62)	<.0001

[a] Subdistribution Hazard Ratio
[b] Confidence Interval; +Standard Error
&Reference group

Table 4.11 shows that predictors in the multivariate F-G subdistribution model for non-CVD are: Age, Sex, Education, Smoking status, Alcohol status, Obese, Diabetic, Province, and the interaction of Smoking status & Province. Controlling for the other covariates in the model, the risk of non-CVD mortality for subjects who are above 50 years of age is significantly higher than that of subjects below 50 years of age (SHR= 11.5; 95% CI: (8.8, 15.0); P-value< 0.0001). Subjects who never used alcohol compared to current alcohol users have a 71% higher risk of non-CVD mortality (SHR= 1.71; 95% CI: (1.18, 2.47); P-value= 0.005). However, there is no significant difference in the hazard of non-CVD mortality for former and current alcohol consumers (P-value = 0.268; Table 4.11).

Table 4.11: Multivariate Fine & Gray subdistribution hazards model for Non-CVD



Variables	(comparison)	$\hat{\beta}$ (S.E$^+$)	SHR^a (95% CI^b)	P-Value
Age	>50	2.44 (0.14)	11.5 (8.8, 15.0)	<.0001
Sex	male	0.35 (0.09)	1.42 (1.18, 1.69)	0.0002
Education	<high school	0.19 (0.09)	1.21(1.01, 1.46)	0.0366
Smoking Status				
	Former	0.21(0.13)	1.24 (0.97, 1.59)	0.091
	Current	0.77 (0.13)	2.15 (1.67, 2.78)	<.0001
	Never ($^\&$Ref)	-	-	-
Alcohol Status				
	Never	0.53 (0.19)	1.71 (1.18, 2.47)	0.0048
	Former	0.13 (0.11)	1.34 (0.91, 1.42)	0.268
	Current ($^\&$Ref)	-	-	-
Obese	yes	0.24 (0.10)	1.26 (1.04, 1.53)	0.0166
Diabetic	yes	0.52 (0.12)	1.67 (1.31, 2.14)	<.0001
Province	non-western	-0.54 (0.26)	0.59 (0.35, 0.97)	0.0037
Current Smoker*Province		-0.82 (0.37)	0.44 (0.21, 0.91)	0.0277
Former Smoker*Province		0.15 (0.32)	1.16 (0.62, 2.15)	0.641

[a]Subdistribution Hazard Ratio;
[b] Confidence Interval; [+]Standard Error
[&]Reference group

4.3.4.2.4 Summary of F-G Subdistribution Hazards Model for CVD and non-CVD

Figure 4.13 compares the significant predictors included in the multivariate F-G subdistribution model for CVD and non-CVD outcomes. Age, Sex, Smoking status, Alcohol status, and Province show common significant association with both CVD and non-CVD deaths. Education, Diabetic, and Obese are related with non-CVD death outcome. Whereas, Cholesterol, Marry, Hypertensive, Heart Attack, and Stroke are predictors for CVD mortality (Figure 4.13).

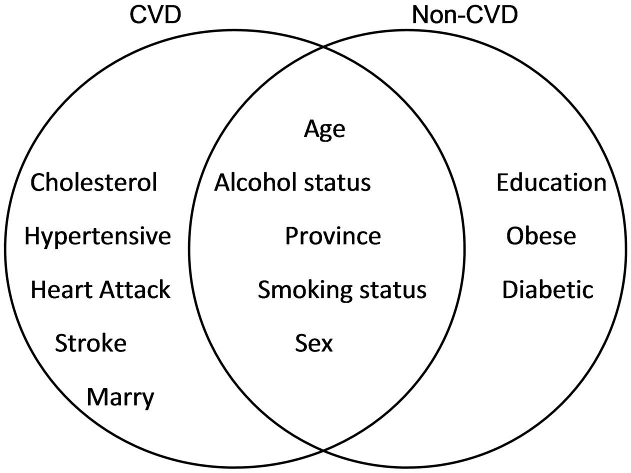


Figure 4.13: Venn diagram comparing risk factors for CVD and Non-CVD events under F-G subdistribution model

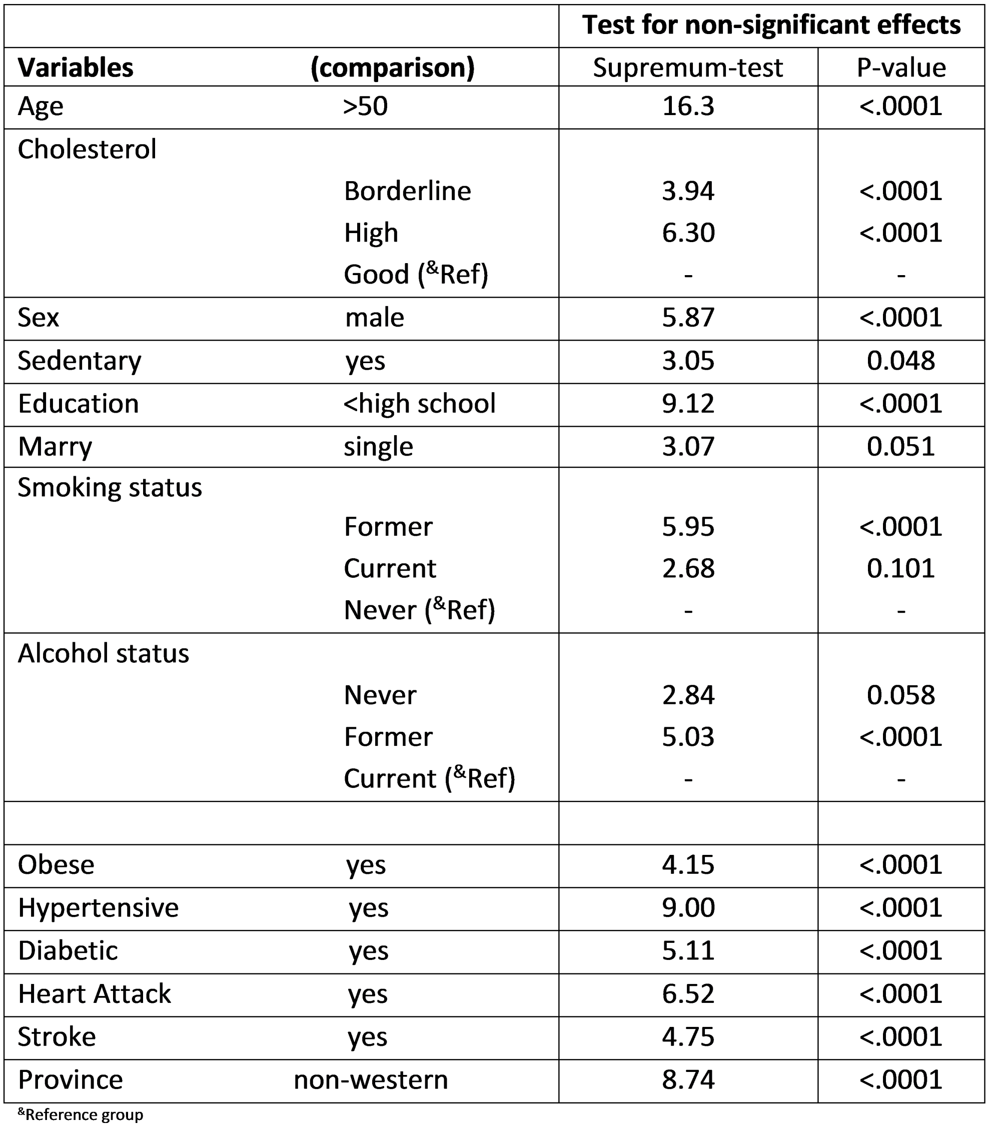
4.3.4.3 Aalen Additive Hazards Model

Additive hazards models were applied to the CHHS dataset. First, I fit the Aalen model and then the Lin & Ying model. The result of testing the covariate effects in the unadjusted and adjusted Aalen models for the competing events are presented below. For the adjusted model, the result of testing the association of time with each variable is also presented.

4.3.4.3.1 Cardiovascular Disease Outcome

Table 4.12 shows that independent predictors for CVD under the Aalen additive model are: Age, Cholesterol, Sex, Education, Smoking status, Alcohol status, Obese, Hypertensive, Diabetic, Heart Attack, Stroke, and Province. Sedentary and Marital status (Marry) are also marginally significant in the Aalen univariate model for CVD (Table 4.12).

Table 4.12: Univariate Aalen additive model for CVD



		Test for non-significant effects	
Variables	**(comparison)**	Supremum-test	P-value
Age	>50	16.3	<.0001
Cholesterol			
	Borderline	3.94	<.0001
	High	6.30	<.0001
	Good ($^{\&}$Ref)	-	-
Sex	male	5.87	<.0001
Sedentary	yes	3.05	0.048
Education	<high school	9.12	<.0001
Marry	single	3.07	0.051
Smoking status			
	Former	5.95	<.0001
	Current	2.68	0.101
	Never ($^{\&}$Ref)	-	-
Alcohol status			
	Never	2.84	0.058
	Former	5.03	<.0001
	Current ($^{\&}$Ref)	-	-
Obese	yes	4.15	<.0001
Hypertensive	yes	9.00	<.0001
Diabetic	yes	5.11	<.0001
Heart Attack	yes	6.52	<.0001
Stroke	yes	4.75	<.0001
Province	non-western	8.74	<.0001

$^{\&}$Reference group

Table 4.13 shows the test for non-significant effects (columns 2 & 3) and the test for time-invariant effects (columns 4 & 5) for the multivariate Aalen model for CVD. The final multivariate Aalen model for CVD includes: Age, Sex, Marry, Smoking status, Hypertensive, Heart Attack, Stroke, Province, the interaction of Hypertensive & Province, and interaction of Age & Marry. Based on the supremum test (T_{sup} discussed in Section 3.2.2.1), which tests the null hypothesis of insignificant effects of the covariates, only covariate Marry has P-value that favor the null of insignificant effect. Although Marry is not significant, the interaction of Age and Marry is significant (P-value=0.019, column 3 in Table 4.13). This suggests that the risk of CVD mortality for married and single subjects may vary by age. The test also suggest that the risk of CVD mortality for hypertensive compared to non-hypertensive subjects may vary by province (P-value $<$ 0.0001; column 3 in Table 4.13).

The Kolmogorov-Smirnov test (T_{KS} discussed in Section 3.2.2.1), which tests the null hypothesis of time-invariant covariate effects showed that only the effect of Age (P-value <0.0001; column 5 in Table 4.13) and Province (P-value <0.0001; column 5 in Table 4.13) varied strongly with time. There was no evidence to reject the null hypothesis of constant effect for the other covariates.

Table 4.13: Multivariate Aalen additive model for CVD

		Test for non-significant effects		Test for time invariant effects	
Variables	**(comparison)**	**Supremum-test**	**P-value**	**Kolmogorov-Smirnov test**	**P-value**
(Intercept)		5.91	<.0001	0.004	0.309
Age	>50	10.21	<.0001	0.025	<.0001
Sex	male	4.84	<.0001	0.003	0.827
Marry	single	1.15	0.914	0.004	0.723
Smoking Status					
	Former	3.65	0.01	0.006	0.247
	Current	3.66	0.004	0.006	0.116
	Never ([&]Ref)	-	-	-	-
Hypertensive	yes	6.1	<.0001	0.01	0.613
Heart Attack	yes	4.34	0.001	0.022	0.758
Stroke	yes	3.04	0.033	0.032	0.599
Province	non-western	3.71	0.005	0.008	<.0001
Hypertensive*Province		6.46	<.0001	0.008	0.895
Age*Marry		3.35	0.019	0.02	0.575

[&]Reference group

4.3.4.3.2 Aalen Cumulative Plots for CVD

To assess the influence of the significant covariates on cardiovascular disease survival, the graph of the cumulative regression coefficients against the survival time was plotted showing the 95% pointwise confidence intervals (Figures 4.14 - 4.23). The slope of the plot informs about the behavior of the covariate on the hazard of CVD mortality over time. The plot in

Figure 4.14 is the estimated baseline cumulative hazards function (intercept). With its negative slope, the plot suggests that when there are no covariates, there is an overall significant and decreasing hazard of CVD mortality.

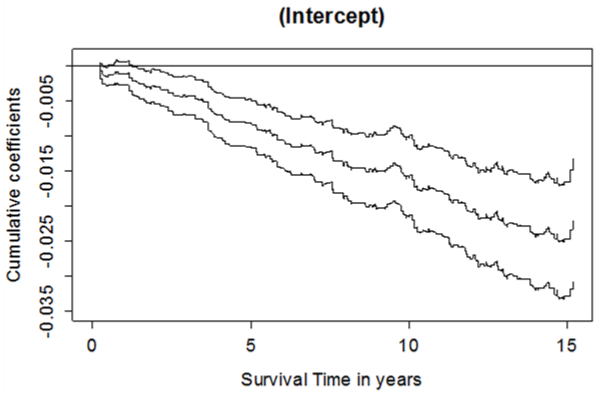


Figure 4.14: Baseline cumulative regression function for CVD with 95% confidence interval

Figure 4.15 shows the plot for Age, and it increased steadily with a positive slope throughout the follow-up time. The slope of the plot suggests that older age increases the risk of CVD mortality.

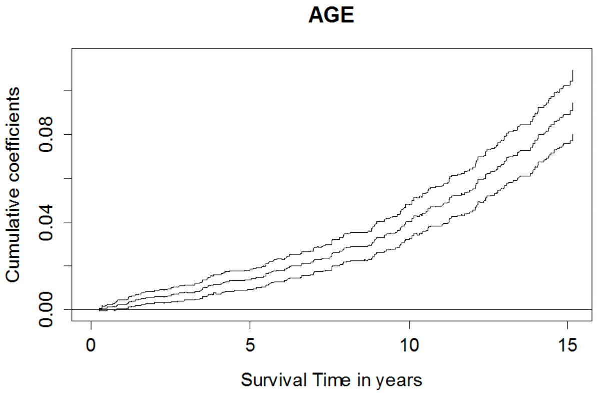


Figure 4.15: CVD cumulative regression function for AGE with 95% confidence interval

The regression function for Sex shown in Figure 4.16. During the first 5 years of follow-up, Sex did not show any significant effect on cardiovascular death because the lower confidence interval includes the zero line. However, after 5 years, males have significant increase in the risk of CVD mortality because of the positive slope.

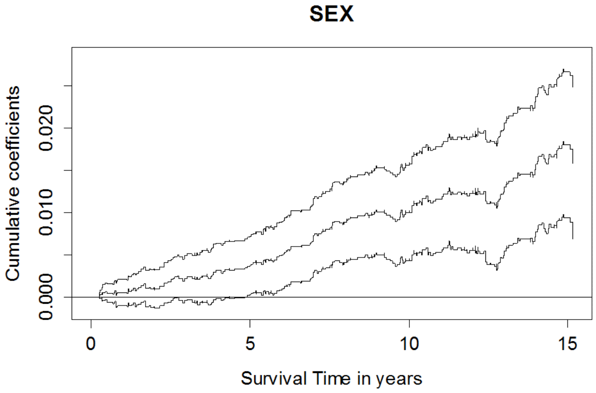


Figure 4.16: CVD cumulative regression function for SEX with 95% confidence interval

The regression function for Marital status (Marry) in Figure 4.17 suggests that the main effect of marry is not a significant predictor of CVD mortality. Overall, the plot has a zero slope and the confidence interval also includes the zero line.

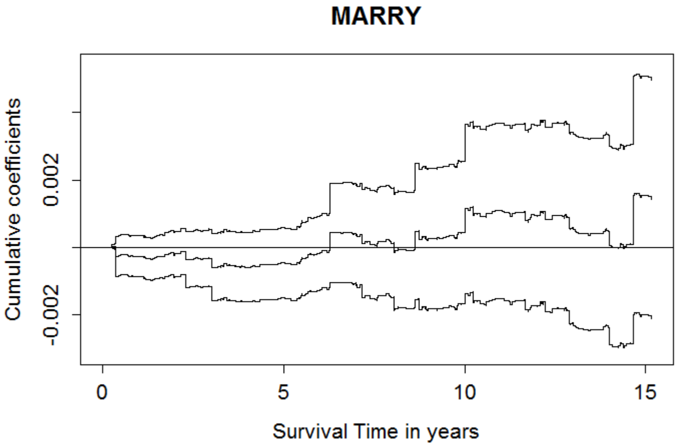


Figure 4.17: CVD cumulative regression function for Marry with 95% confidence interval

Figure 4.18 represents the regression plot for former smoker. In the first 5 years of follow-up, subjects who smoked in the past have no risk for CVD mortality as the zero line is contained in lower confidence interval. However, a former smoker who survives past year 10 has a significant increased risk of CVD death.

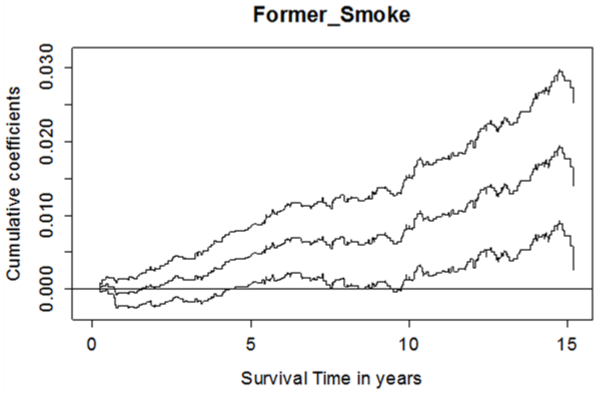


Figure 4.18: CVD cumulative regression function for Former Smoking with 95% confidence interval

Figure 4.19 represents the regression plot for current smoker, and with an overall positive slope, it suggests that current smokers have an increased additive risk of CVD mortality.

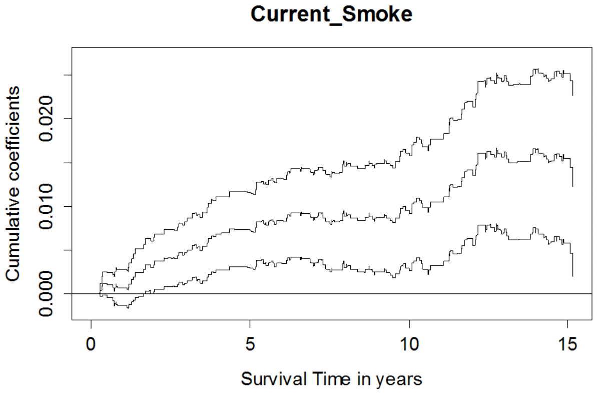


Figure 4.19: CVD cumulative regression function for Current Smoking with 95% confidence interval

Figure 4.20 represents the plot for Hypertensive. For the first 3 years, hypertensive has a marginal effect on CVD death risk but by the fourth year, hypertensive linearly increased the hazard of CVD mortality.

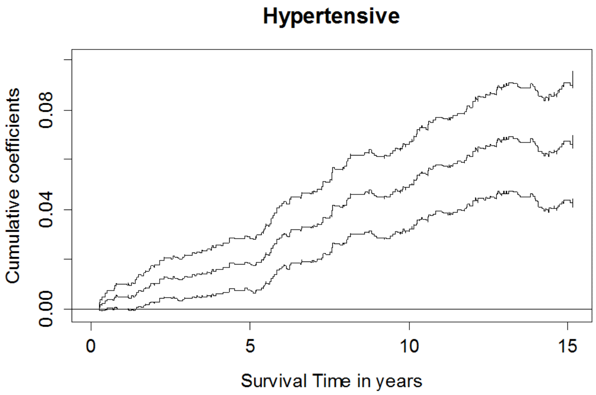


Figure 4.20: CVD cumulative regression function for Hypertensive with 95% confidence interval

Figure 4.21 presents the plot for Heart Attack. The plot is approximately linear with a positive slope by the start of year 5. Although initially, a heart attack does not influence the risk of CVD mortality, but as survival progresses, it increases the risk of CVD mortality significantly.

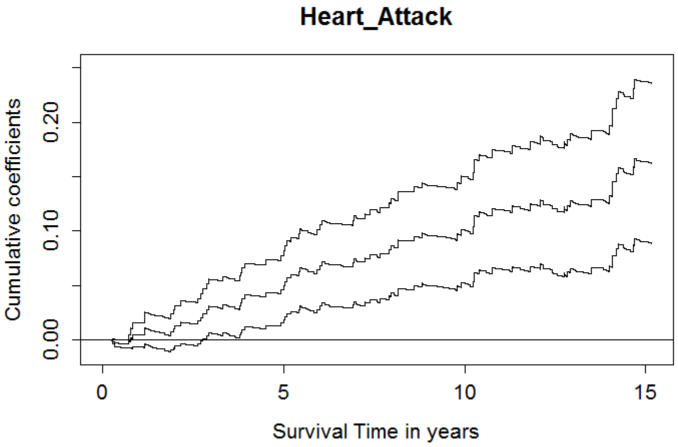


Figure 4.21: CVD cumulative regression function for Heart Attack with 95% confidence interval

Figure 4.22 represents the plot for Stroke. For most of the follow-up, the lower confidence interval nears the zero line, but with a P-value of 0.032 (Table 4.13), one can conclude that the harmful effect of a stroke on CVD mortality is significantly different from a no-stroke.

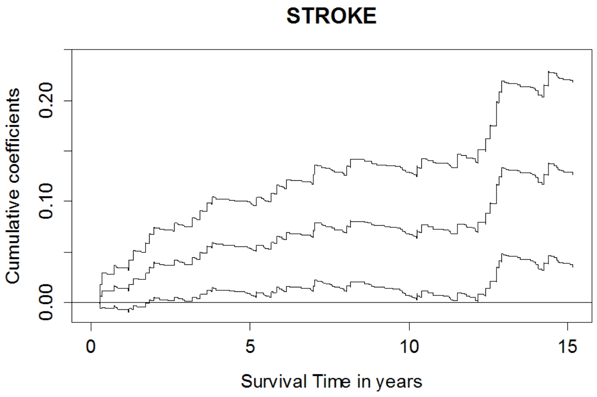


Figure 4.22: CVD cumulative regression function for Stroke with 95% confidence interval

Figure 4.23 shows that province of residence initially did not have any relevant effect on CVD death. However, non-western province participants who survived beyond year 10 have a decreased hazard of CVD death. The conclusion is based on the negative slope of the plot after 10 years.

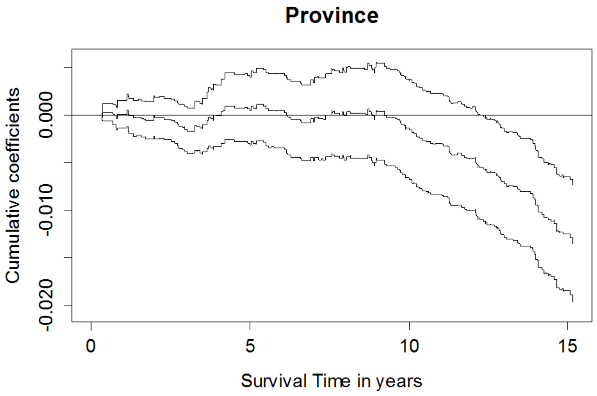


Figure 4.23: CVD cumulative regression function for Province with 95% confidence interval

Summarily, Figures 4.14 - 4.23 show that all the covariates included in the multivariate Aalen additive hazards model except Marry have significant main effects on the CVD death; despite most of the lower confidence intervals including the zero-function at some time-points. Moreover, the alternative hypothesis of significant covariate effects in the Aalen model need only to be valid for some survival time-points and not necessarily all the time-points [102].

The performance of the plots remain consistent with the P-values for testing for insignificant effect obtained using the supremum test. However, only the effects of Age and Province have strong time-varying effects on the hazard of CVD mortality (Table 4.13).

4.3.4.3.3 Model Diagnostics

I examined the fitness of the Aalen additive model for the CVD event of interest using a series of cumulative martingale residuals with 500 random simulations under the null (Figures 4.24 - 4.26). The scale of the y-axis turned out extremely small that if it is rescaled, the residuals would fall on the zero line. However, Martinussen and Scheike recommended using P-values to ascertain the extent of the departure from the null that is observed when a large size of residuals are calculated [9]. With P-values of 0.404 and 0.202 for the observed cumulative residuals for Heart attack and Stroke, respectively, that confirm that Heart Attack and Stroke fit the data well (Figure 4.26). The observed cumulative residuals for the other predictors have P-values that favor the alternative hypothesis that the residuals are significantly different from zero. Thus, leading to unacceptable fits as shown in the figures below.

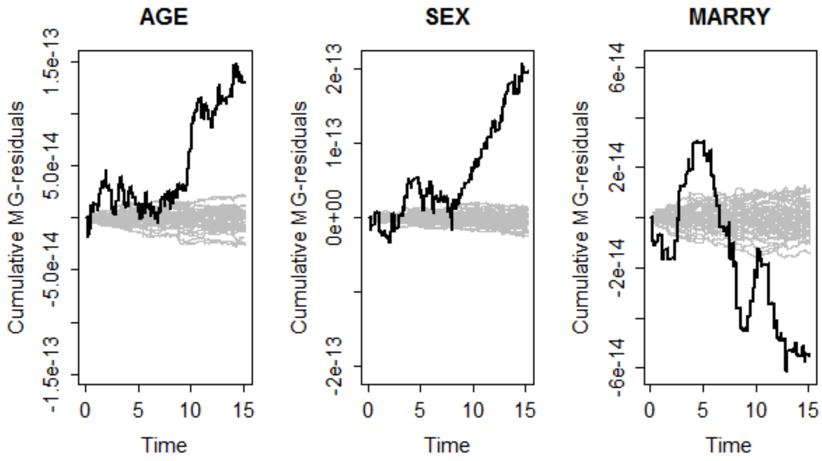


Figure 4.24: Observed cumulative residuals with 500 random simulations for Age, Sex, and Marry

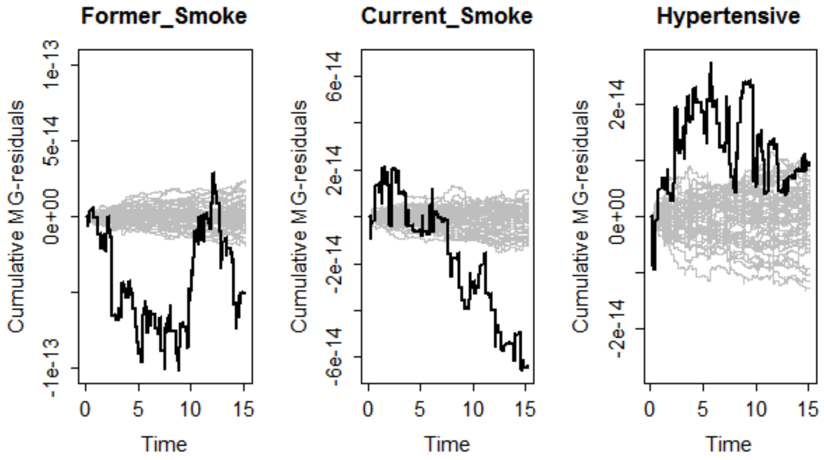


Figure 4.25: Observed cumulative residuals with 500 random simulations for Smoking Status and Hypertensive

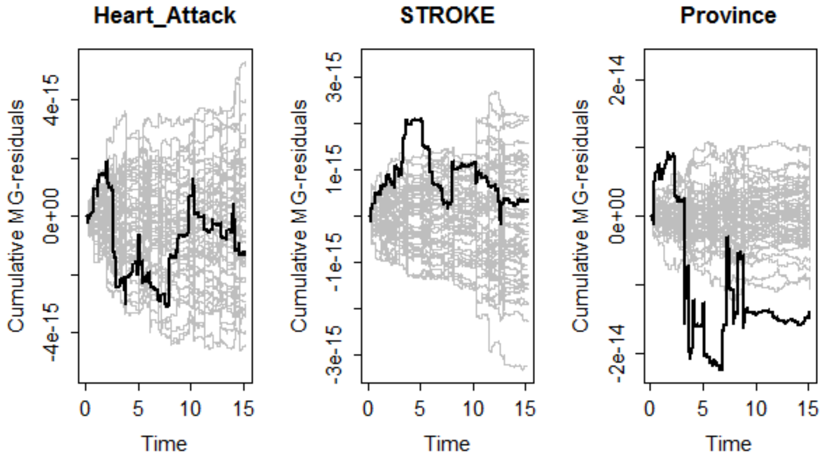
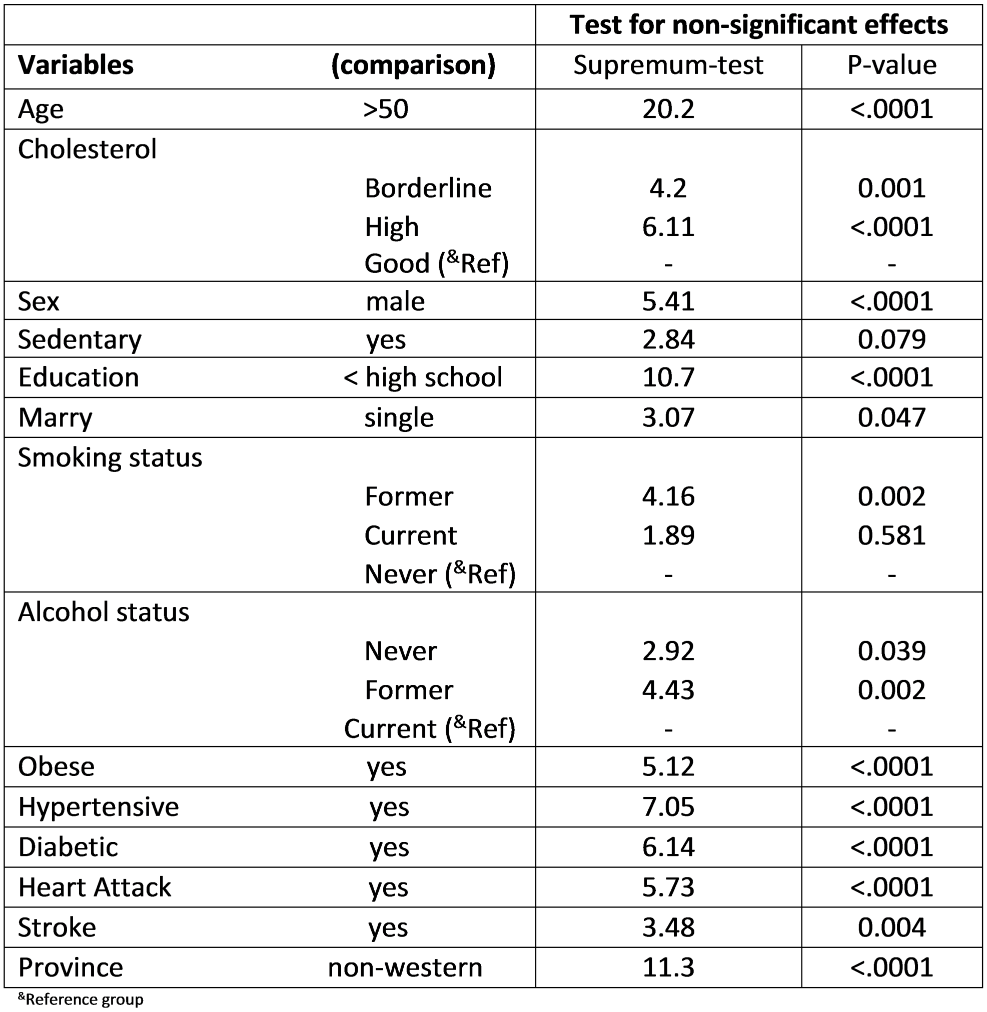


Figure 4.26: Observed cumulative residuals with 500 random simulations for Heart Attack, Stroke, and Province

4.3.4.3.4 Non-Cardiovascular Disease Outcome

Table 4.14 shows that significant covariates in the univariate Aalen model for non-CVD are: Age, Cholesterol, Sex, Education, Smoking status, Alcohol status, Obese, Hypertensive, Diabetic, Heart Attack, Stroke, and Province are independent predictors of non-CVD. However, Sedentary (P-value= 0.079) is not a significant predictor of non-CVD and Marry (P-value=0.047) is only marginally significant (Table 4.14).

Table 4.14: Univariate Aalen additive model for non-CVD



		Test for non-significant effects	
Variables	**(comparison)**	Supremum-test	P-value
Age	>50	20.2	<.0001
Cholesterol			
	Borderline	4.2	0.001
	High	6.11	<.0001
	Good ($^{\&}$Ref)	-	-
Sex	male	5.41	<.0001
Sedentary	yes	2.84	0.079
Education	< high school	10.7	<.0001
Marry	single	3.07	0.047
Smoking status			
	Former	4.16	0.002
	Current	1.89	0.581
	Never ($^{\&}$Ref)	-	-
Alcohol status			
	Never	2.92	0.039
	Former	4.43	0.002
	Current ($^{\&}$Ref)	-	-
Obese	yes	5.12	<.0001
Hypertensive	yes	7.05	<.0001
Diabetic	yes	6.14	<.0001
Heart Attack	yes	5.73	<.0001
Stroke	yes	3.48	0.004
Province	non-western	11.3	<.0001

$^{\&}$Reference group

Table 4.15 shows the test for non-significant effects (columns 2 & 3) and the test for time-invariant effect (columns 4 & 5) for the multivariate Aalen model for non-CVD. Using the supremum test statistic, significant predictors in the final multivariate Aalen model for non-CVD include: Age, Sex, Education, Smoking status, Diabetic, Province, and the interaction of Smoking, and Province. The supremum test suggest that the risk of non-CVD mortality for current smokers compared to subjects who never smoked may vary across provinces (P-value < 0.0001; column 3 in Table 4.15). However, there is no significant difference in non-CVD mortality risk for former smokers compared to those who never smoked across the provinces (P-value =0.761; column 3 in Table 4.15).

The Kolmogorov-Smirnov test, which tests the hypothesis of time-invariant covariate effect shows that only the effect of Age (P-value < 0.0001), Diabetic (P-value=0.007), and Province (P-value < 0.0001) varied significantly with time (column 5 in Table 4.15).

Table 4.15: Multivariate Aalen additive model for non-CVD

	Test for non-significant effects		Test for time invariant effects	
Variables (comparison)	**Supremum-test**	**P-value**	**Kolmogorov-Smirnov test**	**P-value**
(Intercept)	4.59	<.0001	0.009	0.281
Age >50	17.8	<.0001	0.027	<.0001
Sex male	4.9	<.0001	0.005	0.711
Education <high school	3.36	0.022	0.005	0.85
Smoking status				
Former	3.18	0.035	0.011	0.294
Current	5.7	<.0001	0.008	0.763
Never $^{(\&}$Ref)	-	-	-	-
Diabetic yes	4.24	<.0001	0.058	0.007
Province non-western	3.7	0.004	0.024	<.0001
Current Smoke*Province	3.68	<.0001	0.006	0.975
Former Smoke*Province	1.6	0.761	0.009	0.75

$^{\&}$Reference group

4.3.4.3.5 Aalen Cumulative Plots for non-CVD

The cumulative regression plots for all the main covariates in the final model with the 95% pointwise confidence intervals are depicted in Figures 4.27 - 4.34. Figure 4.27 contains the plot of the estimated baseline cumulative regression function (intercept). The slope of the plot is negative for the entire follow-up time, which indicates that when there are no covariates in the model, the risk of non-CVD death decreased linearly.

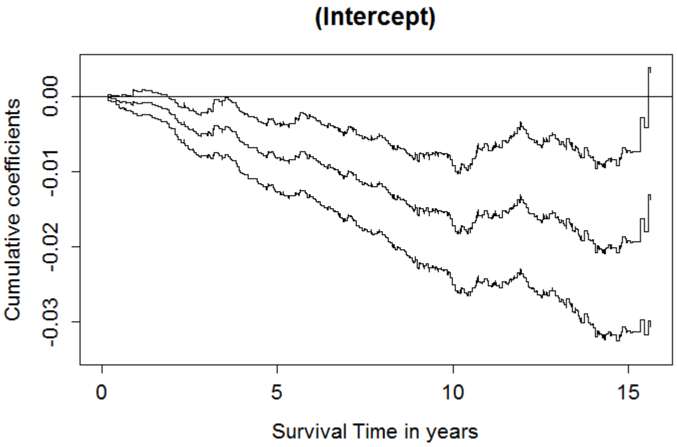


Figure 4.27: Non-CVD Baseline cumulative regression function with 95% confidence interval

Figure 4.28 shows the cumulative plot for Age. The plot is approximately a straight line with a positive slope. It suggests a constant increase in the risk of non-CVD mortality for subjects older than 50 years for the entire follow-up period.

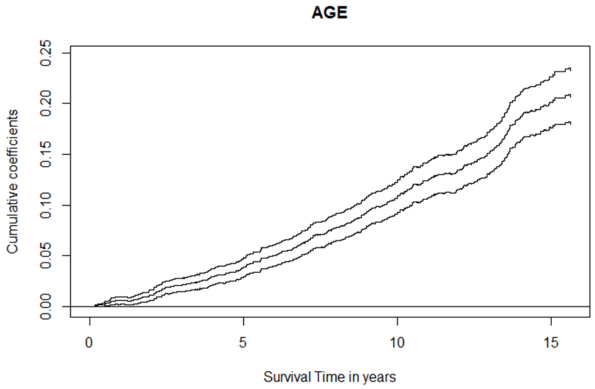


Figure 4.28: Non-CVD cumulative regression function for Age with 95% confidence interval

Figure 4.29 represents the plot for Sex. Overall the plot has a positive slope which corresponds to increased non-CVD death hazard for males throughout follow-up.

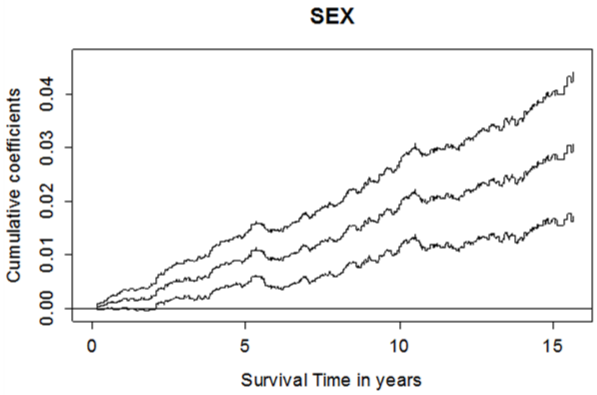


Figure 4.29: Non-CVD cumulative regression function for Sex with 95% confidence interval

Figure 4.30 shows the plot for education level. During the first 10 years of follow-up, education level does not affect non-CVD death hazard as the lower confidence interval contains the zero function. After 10 years, low education level increased non-CVD mortality.

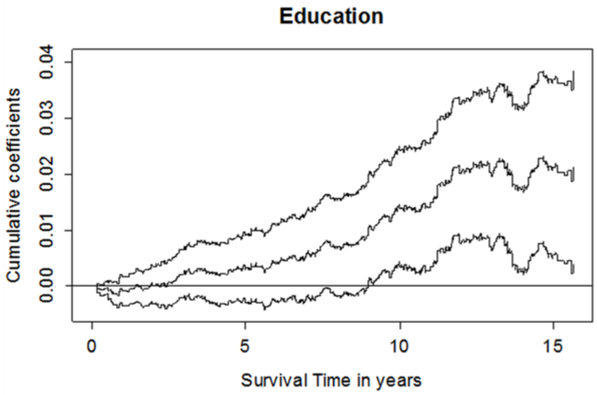


Figure 4.30: Non-CVD cumulative regression function for Education with 95% confidence interval

Figure 4.31 shows the plot for former smoking. For the first 6 years, former smoking did not affect the risk of non-CVD mortality since the lower confidence interval included the zero line. After year 6, the slope shows upward and downward bumps, which suggests that at some time, former smoking increased the risk of non-CVD mortality. The slope suggests that the effect of former smoking on non-CVD mortality is marginally significant. This is consistent with the test of non-significant effects for former smoking, which has a P-value of

0.035 (column 3 of Table 4.15).

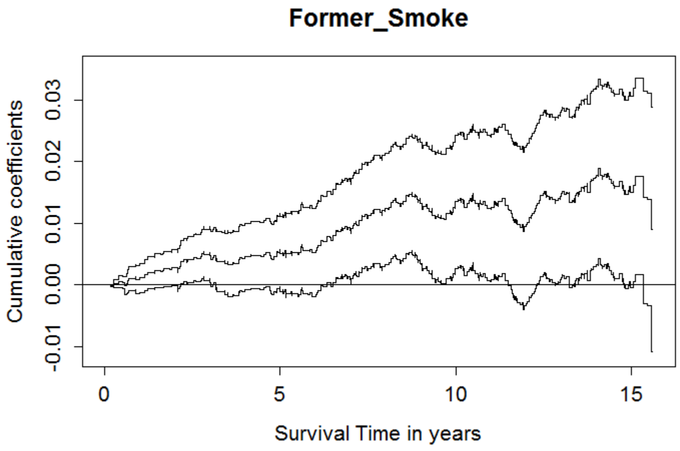


Figure 4.31: Non-CVD cumulative regression function for Former smoking with 95% confidence interval

Figure 4.32 shows the plot for current smoking. In the first 2 years, current smokers do not have any significant risk for non-CVD death since the lower confidence interval contains the zero line. Afterwards, current smoking linearly increases the hazard of non-CVD mortality based on the positive slope of the plot.

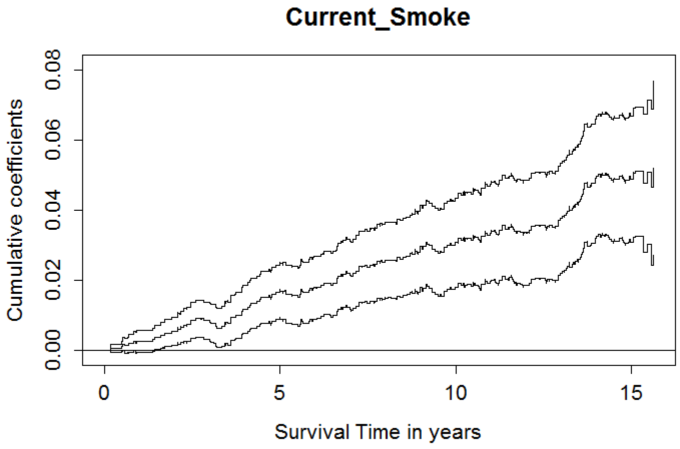


Figure 4.32: Non-CVD cumulative regression function for Current Smoke with 95% confidence interval

Figure 4.33 represents the plot for diabetic. During the first 10 years of follow-up, diabetic shows no effect on the hazard of non-CVD death as the lower confidence interval of the plot includes the zero line. After 10 years, diabetic exhibits a significant increased additive risk for non-CVD death (P-value < 0.0001; column 3 in Table 4.15).

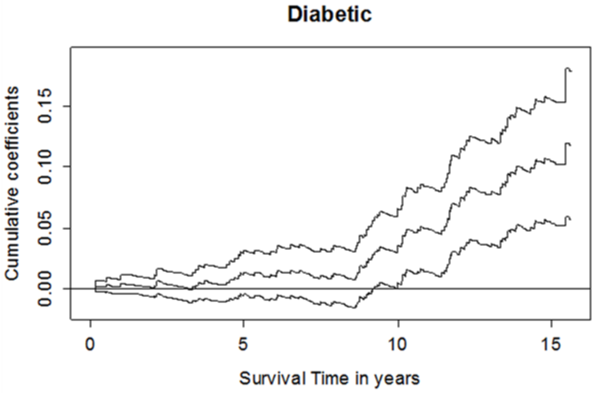


Figure 4.33: Non-CVD cumulative regression function for Diabetic with 95% confidence interval

Figure 4.34 represents the plot for province. For the first ten 10 years, province of residence did not pose any risk for non-CVD mortality since the slope of the plot is zero. However, in the last years of follow-up, the plot has a negative slope, which implies that residing in a non-western province decreases the risk of non-CVD mortality significantly (P-value= 0.004; column 3 in Table 4.15).

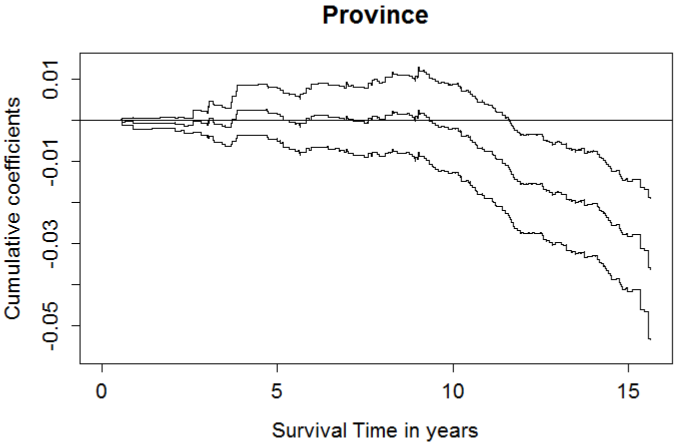


Figure 4.34: Non-CVD cumulative regression function for Province with 95% confidence interval

In summary, Figures 4.27 - 4.34 suggest that the covariates included in the multivariate

Aalen additive hazards model have significant effects on the hazard of non-CVD death. These performances are consistent with the result for testing for non-significant effects obtained using the supremum test (columns 2 & 3 in Table 4.15). However, only Age, Diabetic, and Province have time-varying effects in the multivariate Aalen model for non-CVD (columns 4 & 5 in Table 4.15).

4.3.4.3.6 Model Diagnostics

I examined the fitness of the Aalen model for non-CVD event using series of martingale residuals along with 500 random simulations under the null (Figures 4.35 and 4.36). Again, the scale of the y-axis was very small; if it is rescaled, the residuals would fall on the zero line. Following Martinussen and Scheike [9], we used P-values to judge the consistency of the residuals with the true model. The cumulative residuals for all the covariates but Diabetic (P-value=0.06) have P-values that indicate lack of fit. The residuals as shown below are significantly inconsistent with the true model (Figures 4.35 and 4.36).

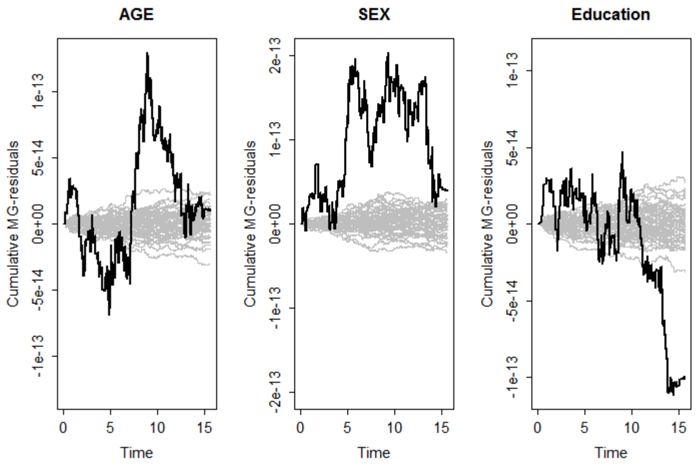


Figure 4.35: Observed cumulative residuals with 500 random simulations for age, sex, and education

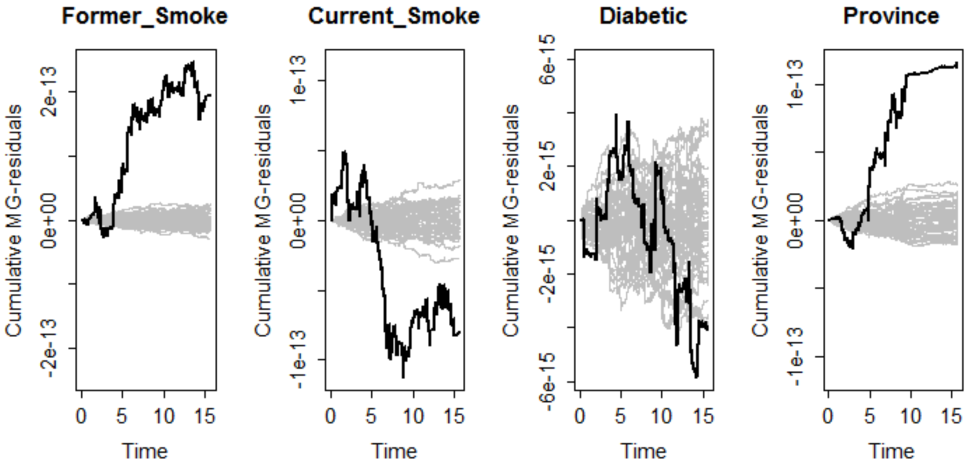


Figure 4.36: Observed cumulative residuals with 500 random simulations for Smoking status, Diabetic, and Province

4.3.4.3.7 Summary of Aalen Model for CVD and non-CVD

Figure 4.37 shows the summary of significant predictors in the Aalen additive models. Age, Sex, Smoking status, and Province have significant relationship with both the event of interest and the competing risk event. Heart Attack, Hypertensive, and Stroke are predictors for CVD death, while Diabetic and Education are relevant to non-CVD outcome.

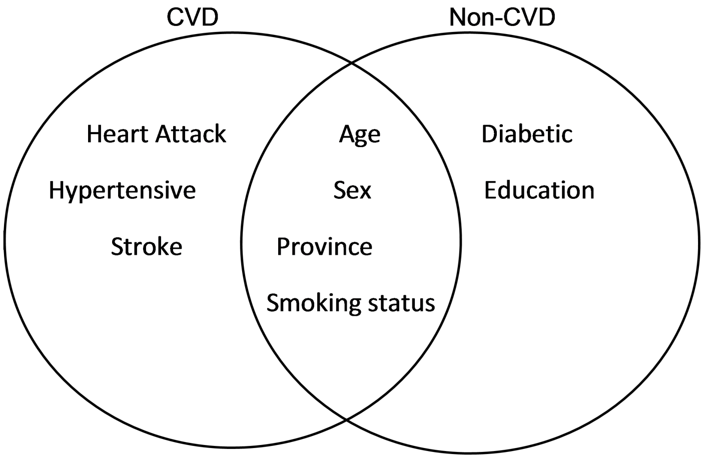


Figure 4.37: Venn diagram comparing risk factors for CVD and Non-CVD risk factors under Aalen additive model

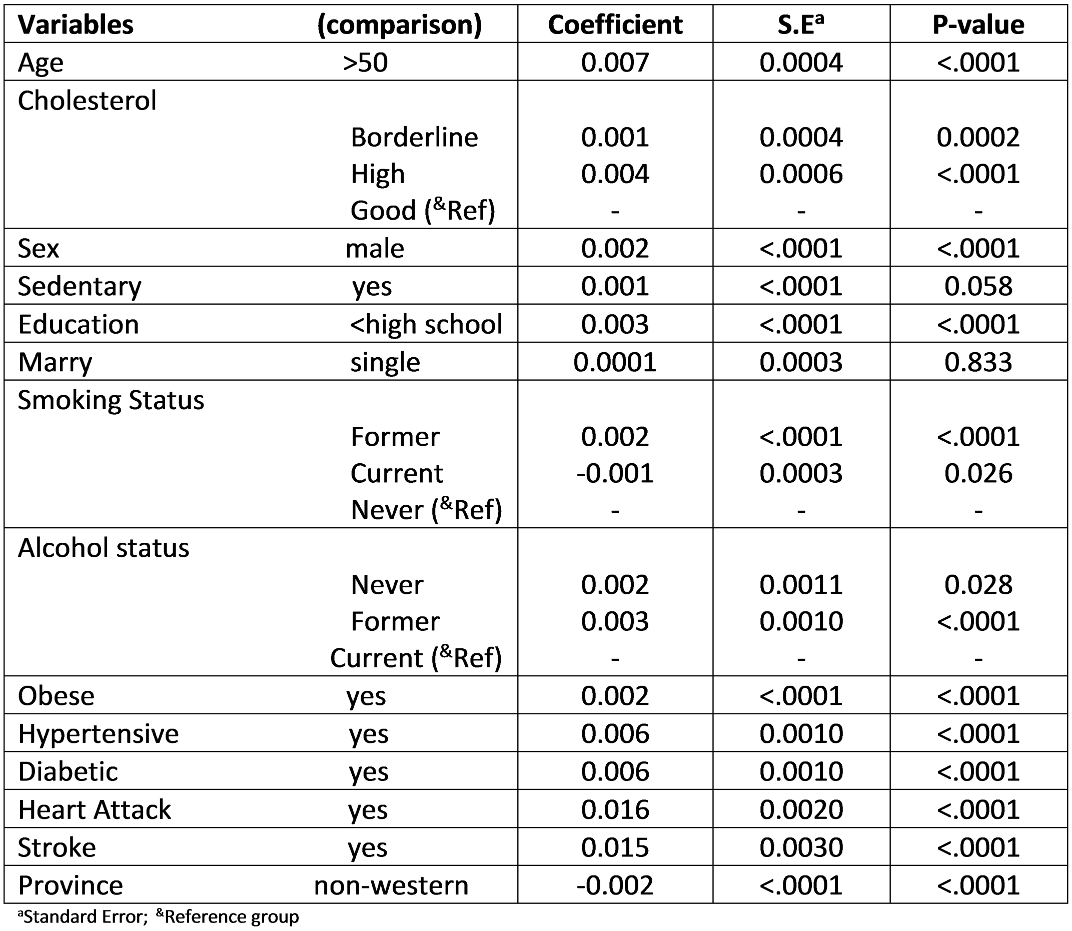
4.3.4.4 Lin & Ying Additive Hazards Model

Finally, I applied the L-Y model to the CHHS follow-up dataset. The results of testing the covariate effects in the unadjusted and adjusted L-Y models for the CVD and the non-CVD outcomes are discussed below.

4.3.4.4.1 Cardiovascular Disease Outcome

Table 4.16 shows the results of the univariate analysis for CVD using the L-Y additive model. Significant predictors in the univariate model are: Age, Cholesterol, Sex, Education, Smoking status, Alcohol status, Obese, Hypertensive, Diabetic, Heart Attack, Stroke, and Province. Marry is not significant in the univariate analysis (P-value=0.833) and Sedentary (P-value=0.058) is marginally associated with the risk of CVD mortality (Table 4.16).

Table 4.16: **Univariate Lin and Ying additive hazards model for CVD**

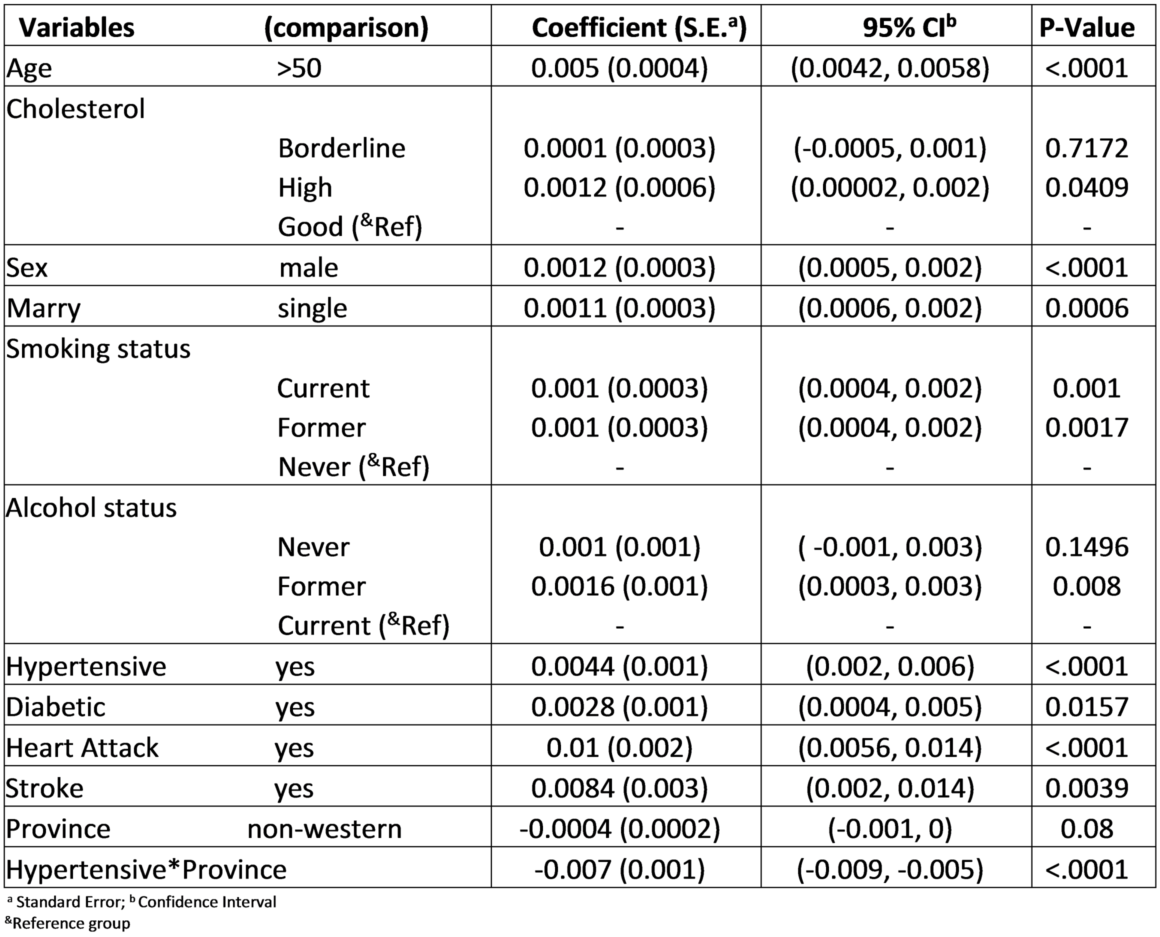


Variables	(comparison)	Coefficient	S.E^{a}	P-value
Age	>50	0.007	0.0004	<.0001
Cholesterol				
	Borderline	0.001	0.0004	0.0002
	High	0.004	0.0006	<.0001
	Good (&Ref)	-	-	-
Sex	male	0.002	<.0001	<.0001
Sedentary	yes	0.001	<.0001	0.058
Education	<high school	0.003	<.0001	<.0001
Marry	single	0.0001	0.0003	0.833
Smoking Status				
	Former	0.002	<.0001	<.0001
	Current	-0.001	0.0003	0.026
	Never (&Ref)	-	-	-
Alcohol status				
	Never	0.002	0.0011	0.028
	Former	0.003	0.0010	<.0001
	Current (&Ref)	-	-	-
Obese	yes	0.002	<.0001	<.0001
Hypertensive	yes	0.006	0.0010	<.0001
Diabetic	yes	0.006	0.0010	<.0001
Heart Attack	yes	0.016	0.0020	<.0001
Stroke	yes	0.015	0.0030	<.0001
Province	non-western	-0.002	<.0001	<.0001

[a]Standard Error; [&]Reference group

In Table 4.17, I present the adjusted L-Y model for CVD. The following significant predictors are included in the model: Age, Sex, Marry, Smoking status, Alcohol status, Diabetic, Heart Attack, and Stroke. Interaction of Hypertensive and Province is significant and therefore included in the final model. The interpretation of the regression coefficients under the L-Y model is illustrated for some covariates as follows: controlling for the other covariates in the final model, having a heart attack increased the hazard of CVD death by 0.01. Thus, a heart attack experience compared to no heart attack increased CVD mortality by 100 additional cases per 10,000 person years (PY) (P-value < 0.0001; 95% CI: (56, 140)). Also, having a stroke increased the hazard of CVD mortality by 0.0084. This effect implies that stroke added 84 new cases of CVD mortality compared to no stroke per 10,000 PY (P-value=0.004; 95% CI: (20, 140)).

Table 4.17: Multivariate Lin & Ying additive hazards model for CVD



Variables	(comparison)	Coefficient (S.E.[a])	95% CI[b]	P-Value
Age	>50	0.005 (0.0004)	(0.0042, 0.0058)	<.0001
Cholesterol				
	Borderline	0.0001 (0.0003)	(-0.0005, 0.001)	0.7172
	High	0.0012 (0.0006)	(0.00002, 0.002)	0.0409
	Good ([&]Ref)	-	-	-
Sex	male	0.0012 (0.0003)	(0.0005, 0.002)	<.0001
Marry	single	0.0011 (0.0003)	(0.0006, 0.002)	0.0006
Smoking status				
	Current	0.001 (0.0003)	(0.0004, 0.002)	0.001
	Former	0.001 (0.0003)	(0.0004, 0.002)	0.0017
	Never ([&]Ref)	-	-	-
Alcohol status				
	Never	0.001 (0.001)	(-0.001, 0.003)	0.1496
	Former	0.0016 (0.001)	(0.0003, 0.003)	0.008
	Current ([&]Ref)	-	-	-
Hypertensive	yes	0.0044 (0.001)	(0.002, 0.006)	<.0001
Diabetic	yes	0.0028 (0.001)	(0.0004, 0.005)	0.0157
Heart Attack	yes	0.01 (0.002)	(0.0056, 0.014)	<.0001
Stroke	yes	0.0084 (0.003)	(0.002, 0.014)	0.0039
Province	non-western	-0.0004 (0.0002)	(-0.001, 0)	0.08
Hypertensive*Province		-0.007 (0.001)	(-0.009, -0.005)	<.0001

[a] Standard Error; [b] Confidence Interval
[&]Reference group

4.3.4.4.2 Model Diagnostics

The fit of the L-Y model was checked using series of martingale residuals, the same approach used for the Aalen model. This is so because the L-Y model is a sub-model of the nonparametric Aalen additive model [9]. The large size of the residual is reflected in the plots (Figures 4.38 - 4.41), but for a clearer picture, I reproduced the residual plots using a sample of 500 randomly chosen observations (Figures 4.42 - 4.45). Based on these new plots in Figures 4.42 - 4.45, the model might be adjudged as having a good fit.

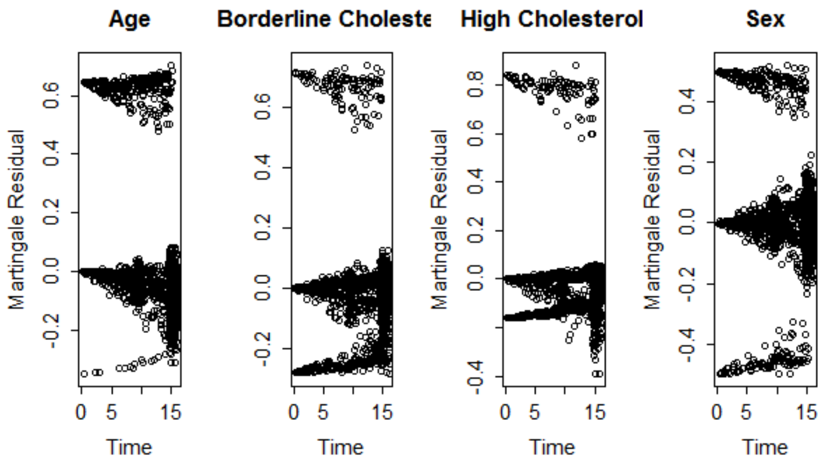


Figure 4.38: Observed cumulative residuals for Age, Cholesterol, and Sex

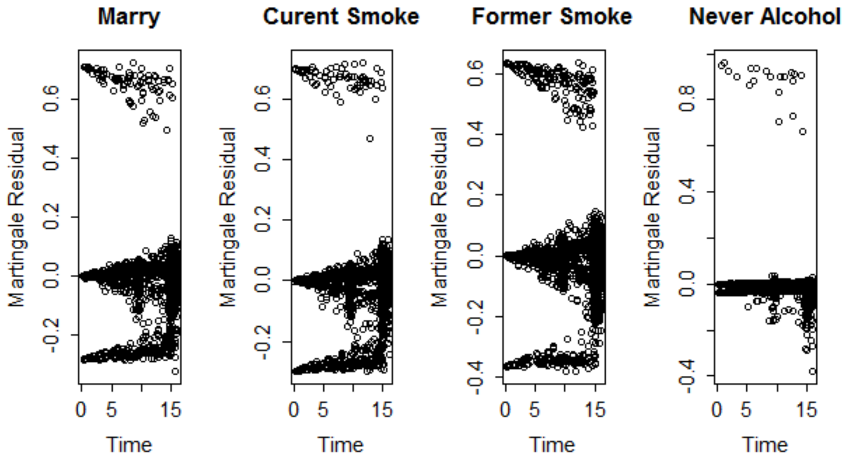


Figure 4.39: Observed cumulative residuals for Marital status, Smoking status, and Alcohol status

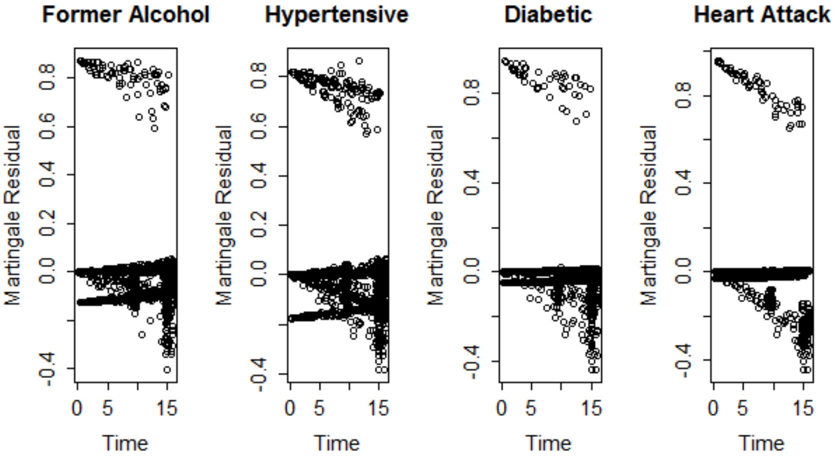


Figure 4.40: Observed cumulative residuals for Hypertensive, Diabetic, and Heart Attack

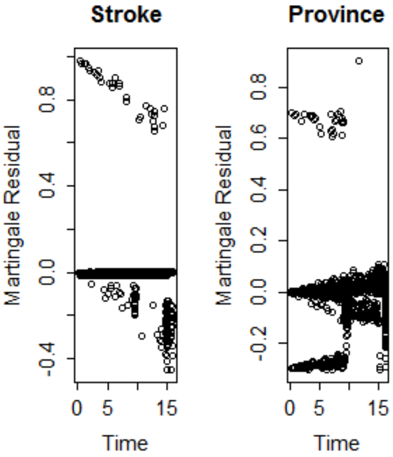


Figure 4.41: Observed cumulative residuals for Stroke and Province

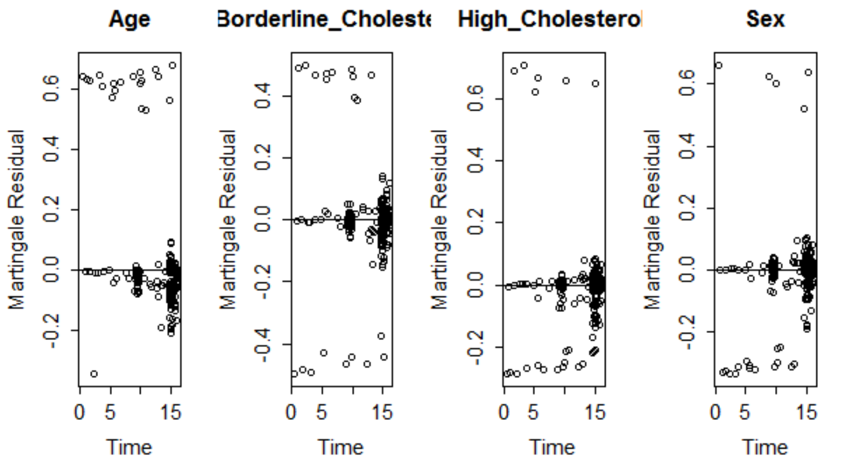


Figure 4.42: Observed cumulative residuals for Age, Sex, and Marry using 500 random observations

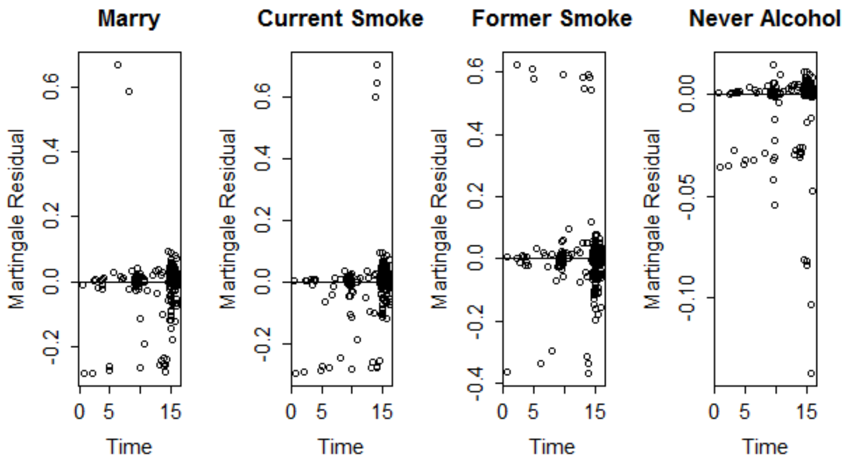


Figure 4.43: Observed cumulative residuals for Current Smoke, Former Smoke, and Former Alcohol using 500 random observations

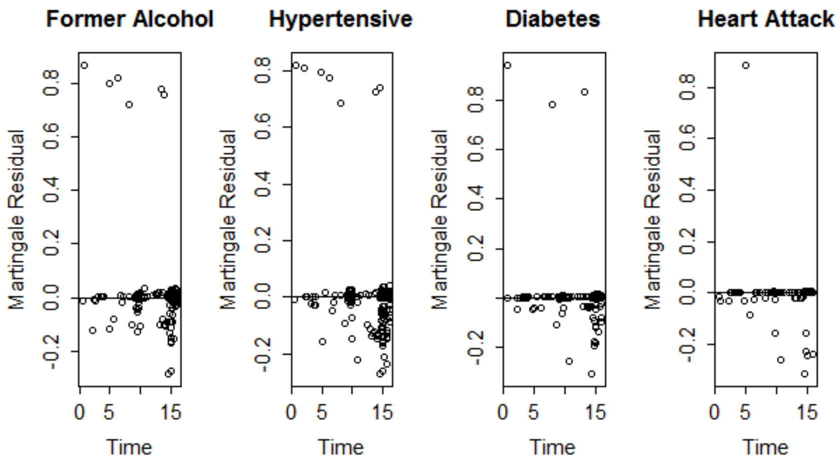


Figure 4.44: Observed cumulative residuals for Hypertensive, Diabetes, and Heart Attack using 500 random observations

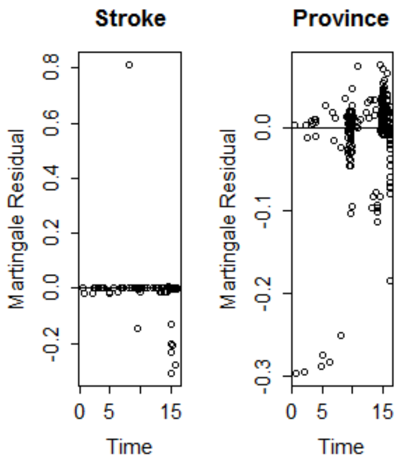
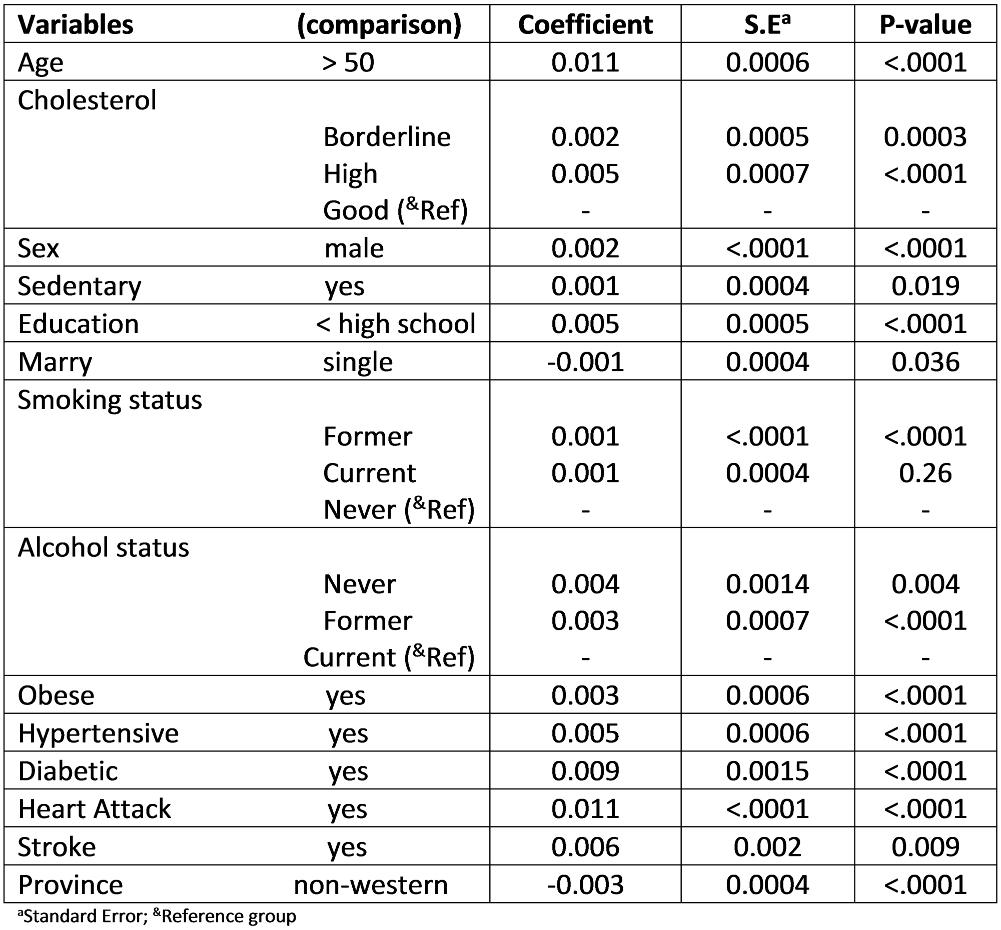


Figure 4.45: Observed cumulative residuals for Stroke and Province using 500 random observations

4.3.4.4.3 Non-Cardiovascular Disease Outcome

Table 4.18 shows the univariate analysis of L-Y model for non-CVD, and all the variables under consideration are significant predictors of non-CVD mortality.

Table 4.18: Univariate Lin and Ying additive hazards model for non-CVD

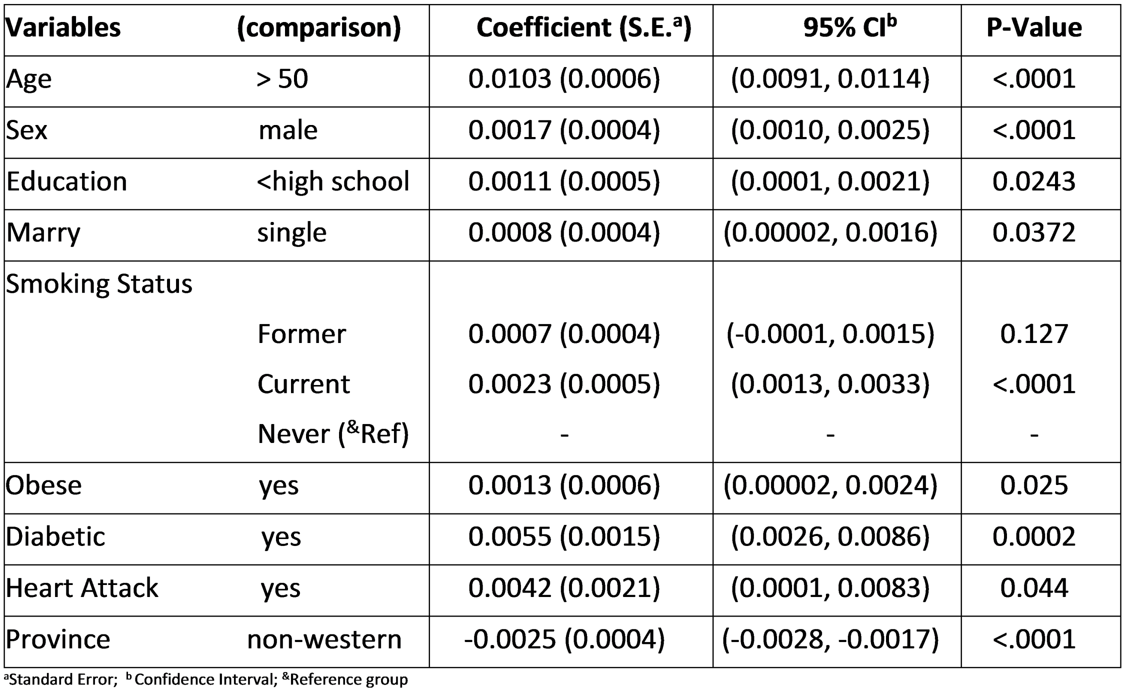


Variables	(comparison)	Coefficient	S.E[a]	P-value
Age	> 50	0.011	0.0006	<.0001
Cholesterol				
	Borderline	0.002	0.0005	0.0003
	High	0.005	0.0007	<.0001
	Good ([&]Ref)	-	-	-
Sex	male	0.002	<.0001	<.0001
Sedentary	yes	0.001	0.0004	0.019
Education	< high school	0.005	0.0005	<.0001
Marry	single	-0.001	0.0004	0.036
Smoking status				
	Former	0.001	<.0001	<.0001
	Current	0.001	0.0004	0.26
	Never ([&]Ref)	-	-	-
Alcohol status				
	Never	0.004	0.0014	0.004
	Former	0.003	0.0007	<.0001
	Current ([&]Ref)	-	-	-
Obese	yes	0.003	0.0006	<.0001
Hypertensive	yes	0.005	0.0006	<.0001
Diabetic	yes	0.009	0.0015	<.0001
Heart Attack	yes	0.011	<.0001	<.0001
Stroke	yes	0.006	0.002	0.009
Province	non-western	-0.003	0.0004	<.0001

[a]Standard Error; [&]Reference group

Table 4.19 contains the adjusted L-Y model for non-CVD. The model includes Age, Sex, Education, Marry, Smoking Status, Obese, Diabetic, Heart Attack, and Province (Table 4.19). No interaction effect is significant for inclusion in the final model. As mentioned earlier, the interpretation of covariates effect under the L-Y model is directly through the regression coefficients. Adjusting the other variables in the final model, a heart attack experience increased the hazard of CVD death by 0.0042 compared to no heart attack. Thus, a heart attack significantly increased non-CVD death events by adding 42 new cases per 10,000 PY (P-value= 0.032; 95% CI: (1, 83)). Contrarily, residing in a non-western province reduced CVD mortality by 0.0025. There were 25 fewer cases of non-CVD mortality per 10,000 PY in a non-western province compared to a western province (P-value $<$ 0.0001; 95% CI: (17, 28), Table 4.19).

Table 4.19: Multivariate Lin & Ying additive hazards model for non-CVD



Variables	(comparison)	Coefficient (S.E.[a])	95% CI[b]	P-Value
Age	> 50	0.0103 (0.0006)	(0.0091, 0.0114)	<.0001
Sex	male	0.0017 (0.0004)	(0.0010, 0.0025)	<.0001
Education	<high school	0.0011 (0.0005)	(0.0001, 0.0021)	0.0243
Marry	single	0.0008 (0.0004)	(0.00002, 0.0016)	0.0372
Smoking Status				
	Former	0.0007 (0.0004)	(-0.0001, 0.0015)	0.127
	Current	0.0023 (0.0005)	(0.0013, 0.0033)	<.0001
	Never ([&]Ref)	-	-	-
Obese	yes	0.0013 (0.0006)	(0.00002, 0.0024)	0.025
Diabetic	yes	0.0055 (0.0015)	(0.0026, 0.0086)	0.0002
Heart Attack	yes	0.0042 (0.0021)	(0.0001, 0.0083)	0.044
Province	non-western	-0.0025 (0.0004)	(-0.0028, -0.0017)	<.0001

[a]Standard Error; [b] Confidence Interval; [&]Reference group

4.3.4.4.4 Model Diagnostics

The adequacy of the L-Y model for non-CVD event is also examined with martingale residuals as shown in Figures 4.46 - 4.48. These figures are similar to those obtained for the CVD main event. As a result, I did not reproduce the residuals with 500 random observations. However, one may say that the model has a good fit as high percentage of the residuals are concentrated around the zero line in all the figures.

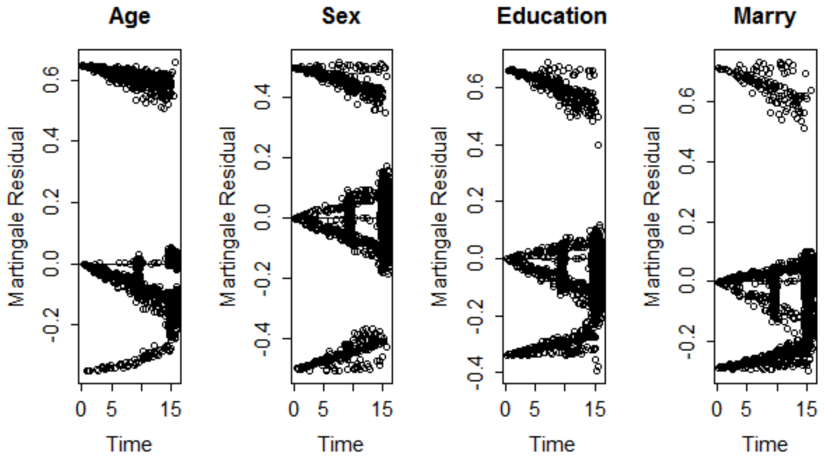


Figure 4.46: Observed cumulative residuals for Age, Sex, Education, and Marry

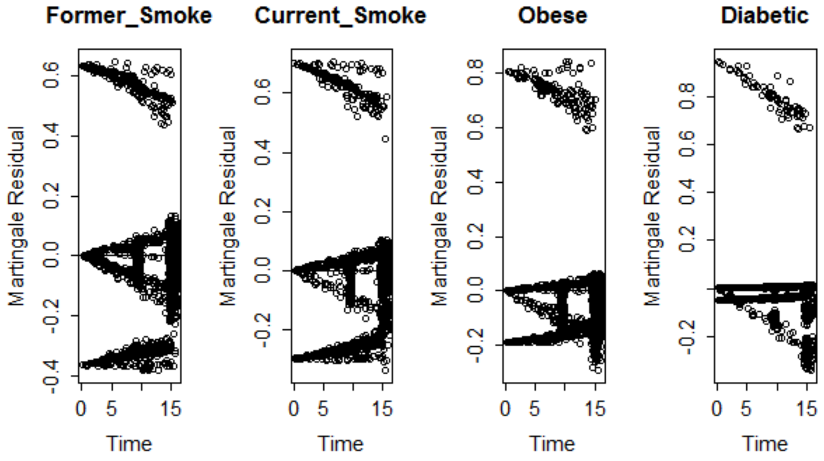


Figure 4.47: Observed cumulative residuals for Smoking Status, Obese, and Diabetic

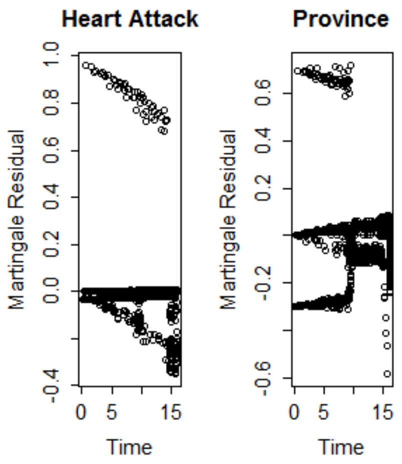


Figure 4.48: Observed cumulative residuals for Heart Attack and Province

4.3.4.4.5 Summary of Lin & Ying Model for CVD and non-CVD

The Venn diagram in Figure 4.49 compares the significant predictors included in the final L-Y model for CVD and non-CVD. Age, Sex, Smoking status, Marry, Diabetic, and Heart Attack affect both CVD and non-CVD survival probability. Stroke, Hypertensive, Cholesterol, and Alcohol status are associated with CVD mortality. Whereas, Education, Province and Obese are risk factors for non-CVD death.

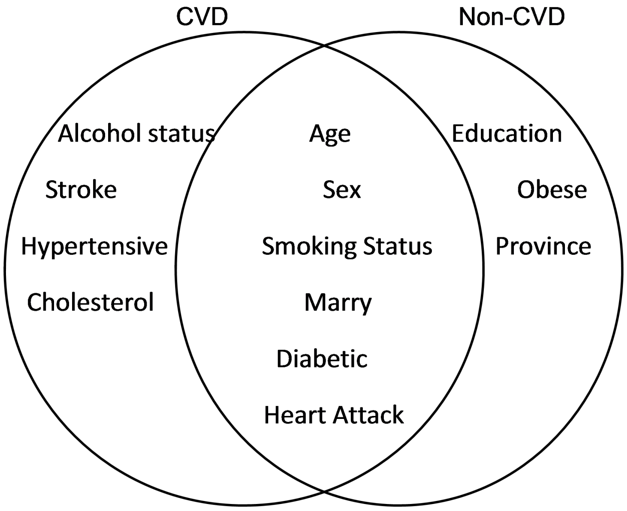


Figure 4.49: Venn diagram comparing risk factors for CVD and Non-CVD risk factors for Lin & Ying model

4.3.5 Comments

Considering the second and third objectives of my study (determining risk factors for cardiac disease mortality and comparing risks factors selected in the multiplicative and the additive hazards models), the outcome of the competing risks analyses of the CHHS Follow-up study dataset is summarized in Table 4.20. The covariates considered for the analysis and the four models are listed in the rows and columns of Table 4.20 respectively. A cell is marked (✓) if the covariate is significant in the multivariate model of the corresponding cause of death. A blank cell corresponds to covariate which is not in the model.

The following covariates are consistently significant through the models for CVD and non-CVD outcomes: Age, Sex, and Smoking Status. Also, the four models selected Hypertensive, Heart Attack, and Stroke as predictors of CVD mortality. The interaction of Hypertensive and Province is included in all the four models for CVD. The Cox CSH, F-G, and L-Y models selected Cholesterol and Alcohol status as predictors of CVD mortality, these did not apply to the Aalen model. Unlike the F-G and the Aalen models, the Cox CSH and the L-Y models selected Diabetic as a risk factor for CVD mortality. The interaction of Heart Attack and Stroke is included in the Cox CSH and the F-G models for CVD while the interaction of Age and Marital status is included in the Aalen model for CVD. None of the Cox CSH, F-G, Aalen, or L-Y models included Education and Obese as predictors of CVD mortality.

All the four models included Education and Diabetic as risk factors for non-CVD mortality. All but the Aalen model included Obese as a risk factor for non-CVD mortality. The Cox CSH and the F-G but not the additive models selected Alcohol status as a predictor of non-CVD mortality. In addition, the Cox CSH and L-Y models but not F-G and Aalen models included Marital status as a risk factor for non-CVD mortality. Only the L-Y model selected Heart Attack as a risk factor for non-CVD mortality. All the models but L-Y included the interaction of Smoking and Province as risk factor for non-CVD mortality. None of the models selected Stroke, Hypertensive, and Cholesterol as predictors of non-CVD mortality. Of all the covariates under consideration, only Sedentary is not included in any of the multivariate models for CVD and non-CVD outcomes.

In general, risk factors for CVD mortality, which are included in at least one of either the multiplicative or the additive hazards models are Age, Cholesterol, Sex, Marital status, Smoking status, Alcohol status, Hypertension, Diabetes, Heart Attack, Stroke, Province, interaction of Heart Attack and Stroke, Hypertensive and Province, and that between Age and Marital status. Similarly, risk factors for non-CVD mortality included in at least one of the four models are Age, Sex, Education level, Marital status, Smoking status, Alcohol status, Obesity, Diabetes, Heart Attack, Province, and the interaction of Smoking and Province.

Table 4.20: Comparing significant risk factors in models for CVD and non-CVD

	Multiplicative Models				Additive Models			
Variables	**Cox CSH**		**F-G SDH**		**Aalen**		**L-Y**	
	CVD	**Non-CVD**	**CVD**	**Non-CVD**	**CVD**	**Non-CVD**	**CDV**	**Non-CVD**
Age	✓	✓	✓	✓	✓	✓	✓	✓
Cholesterol	✓		✓				✓	
Sex	✓	✓	✓	✓	✓	✓	✓	✓
Sedentary								
Education		✓		✓		✓		✓
Marital status	✓	✓	✓				✓	✓
Smoking Status	✓	✓	✓	✓	✓	✓	✓	✓
Alcohol status	✓	✓	✓	✓			✓	
Obese		✓		✓				✓
Hypertensive	✓		✓		✓		✓	
Diabetic	✓	✓		✓		✓	✓	✓
Heart Attack	✓		✓		✓		✓	✓
Stroke	✓		✓		✓		✓	
Province	✓	✓	✓	✓	✓	✓		✓
Heart Attack*Stroke	✓		✓					
Hypertensive*Province	✓		✓		✓		✓	
Smoking Status*Province		✓		✓		✓		
Age*Marital status					✓			

Cox CSH - Cox Cause-Specific Hazards model; F-G SDH - Fine and Gray Subdistribution Hazards model; L-Y - Lin and Ying model; CVD - Cardiovascular Disease; non-CVD - Non-Cardiovascular Disease

Chapter 5
Discussion

In this study, I have three objectives: (i) To apply the additive and multiplicative hazards models to the Canadian Heart Health Survey Follow-up data in the competing risks setting; (ii) To determine risk factors for cardiovascular disease using the competing risks approach; (iii) To compare risk factors named under the additive and multiplicative models in the context of competing risks. In the competing risks analysis of real life cardiovascular data, I performed two multiplicative hazards models (the Cox cause-specific hazards and the Fine & Gray subdistribution hazards models) and two additive hazards models (the Aalen hazards model and the Lin & Ying hazards model) The competing events in my data are CVD and non-CVD deaths. In estimating the cause-specific hazard of CVD mortality, non-CVD is considered the competing risk event, and vice versa. The Cox CSH model, the Aalen additive model, and the L-Y additive model were used to estimate cause-specific hazard rates, while the F-G SH model was used to model the hazard of subdistribution.

The main distinction between the F-G subdistribution model and the Cox cause-specific model is that each time an event of non-interest occurs, it partakes in the former's risk set, highlighting the uniqueness of the subdistribution approach for competing risks analysis. In contrast, the event of non-interest reduces the size of the risk set in the cause-specific analysis [3]. The Cox CSH model estimates the hazard of a specific event in those who are still alive [110]. Thus, it is appropriate for epidemiological studies, where disease etiology is paramount. Conversely, the F-G subdistribution model, which estimates the relative incidence of a specific event among those who are still alive and those who have died from a competing risk event is best for risk prognosis [6, 14, 15]. Whether or not to retain those who have died from a competing risk event in the risk set has formed the basis of debate among researchers. Andersen *et al.* admitted that the subdistribution approach is mathematically sound but objected to keeping the observed competing risk event in the risk set because it jeopardizes the interpretation of the subdistribution hazard as an instantaneous failure risk [111]. Pintilie, however, argued that removing the competing risk event from the risk set distorts the impact of the covariates on the outcomes [11]. Generally, researchers agree that the F-G SH and the Cox CSH models should be used together for the event of interest and the competing risk event, for a comprehensive understanding of the association between the predictors and the events [110, 112]. Moreover, the solutions of the two models are usually distinct [15, 113] because the F-G model depends largely on the distribution of the competing risk event since the occurrence of such event prevents the occurrence of the event of interest [51].

My study result shows that covariates effect within the F-G SH and the Cox CSH models are similar in magnitude and direction, agreeing with some past studies [18, 34, 47, 50, 52]. A study may obtain a similar covariate effect in both models when such covariate does not impact the cause specific hazard of the competing rsk event or when there is huge censoring [52, 114, 115]. For the CHHS Follow-up data, the multiplicative models identified identical sets of covariates as risk factors for CVD and non-CVD, and the covariates' effects are similar. In addition, the residual plots suggest that both models fit the CHHS Follow-up data in my study reasonably well. The F-G and the Cox CSH models in my study suggest that smoking cigars, pipes, or cigarettes regularly or occasionally significantly increases the risk of cardiac mortality. These results conform with findings from a cohort study among Germans, which found that current smokers have a relative hazard of 2.45 for CVD mortality compared to those who have never smoked [92]. Similarly, the INTERHEART study found that smokers are three times more likely to die from acute myocardial infarction (AMI) than those who never smoked [91]. Also, the multiplicative models in my study show that subjects who smoked regularly or occasionally compared to those who never smoked have an increased risk of death from non-CVD. In addition, the Cox CSH and the F-G models in my study indicate that both lifetime alcohol abstainers and former alcohol consumers are at a higher risk of CVD mortality. These results are consistent with those from a cohort study among Canadians, which showed that consuming alcohol regularly but moderately protected against CVD outcome [94]. In another study in Spain, the odds of coronary heart disease (CHD) mortality among abstainers compared to moderate consumers was found to increase by 86%

[95]. Similarly, alcohol consumers in the United States had a 42% reduced risk of CHD compared to long-term abstainers [96]. The multiplicative models my study also found that long-time alcohol abstainers have a greater risk of dying from non-CVD relative to alcohol users.

The Cox CSH and F-G models in my study show that being above 50 years of age increased the risk of both CVD and non-CVD mortalities. Also, the Cox CSH and F-G models show that men have a higher risk of CVD and non-CVD mortalities. My results agree with findings in a population-based study in the Netherlands in which women had a 4% lower risk for CVD and a 24% lower risk for non-CVD deaths compared to men [41]. Similarly, a study among Finnish residents showed a five-fold increase in CHD mortality risk among men compared with women [85]. In my study, single subjects relative to married subjects have an increased CVD mortality in the multiplicative models. These conform with a study in Scotland where being unmarried increased the risk of CVD death by two-folds in men and women [72]. In a related population study the U.S.A, the odds of dying from CVD increased by 38% in single versus married adults [73]. Another study in Japan found a three-fold risk of CVD death among never-married versus married subjects [74]. In my study, marital status also affected non-CVD mortality in the Cox CSH model. Education status is another socio-demographic predictor of CVD event. In a cohort of Melbourne residents, individuals who attained primary education have an 18% higher risk of CVD mortality compared to postsecondary degree holders [90]. In a similar study, low education level was associated with a significant 24% increase in CVD death among men but not women [75]. My study results deviate from these findings as educational status did not influence the risk of CVD mortality in the multiplicative hazards models. However, low education substantially increases the risk of non-CVD mortality in the Cox CSH and F-G models.

Many studies agree with my study result in which the multiplicative hazards models show that diseases such as hypertension, stroke, and diabetes are risk factors for cardiac death. A Chinese study concurs, finding that compared to non-hypertensive individuals, the risk of CVD mortality doubled among subjects that are hypertensive [76]. Similarly, the Framingham Heart Study revealed that high blood pressure doubled the risk of CVD mortality among the study participants [77]. In a related research, a case-control study in

America found that the risk of CVD mortality increased by 180% among residents who have had a stroke [81]. The Cox CSH model in my study identifies diabetes as a predictor of CVD and non-CVD events. Similarly, the APAC cohort study showed that diabetes was associated with a 97% increased risk of CVD mortality and a 56% increased risk of non-CVD mortality [80]. A study among Britons found that diabetes raised the risk of CVD death in men and women by four-fold and seven-fold respectively [78]. Another study in Finland showed that individuals suffering from diabetes are five times more likely to die from cardiac causes and two times at risk for non-CVD than non-diabetic subjects [79]. Even though series of published cardiovascular disease studies show similar risk factors with mine, few of the studies adjusted for competing risk events [41], and none of the studies employed the additive hazards models.

Additive hazards models are alternative regression models to the multiplicative hazards models [19]. Unlike in the Cox model, the total of the cause-specific hazards modeled through additive hazards models, is the all-cause hazard [18]. Consequently, additive hazards models have been recommended for competing risks analysis [13]. Also, additive hazards models give an absolute covariate effect, which is interpreted as risk difference. Such interpretation is alluring to public health practitioners because it directly infers the excess risk due to the covariate [23]. The semiparametric L-Y additive model is synonymous to the Cox hazards model [19]. CVD predictors selected in both the Cox CSH and the L-Y models are similar in my study. Regression coefficients estimated from the two models are difficult to compare because the Cox model relates multiplicatively to the unknown baseline hazard function, while additive models relate in an additive way to the baseline hazard. Covariate effects estimated from the L-Y additive hazards model are usually smaller than effects from the Cox model since the latter gives hazard ratios while additive models reflect the risk due to the covariates in absolute terms [116]. Contrary to the semiparametric L-Y additive model, the nonparametric Aalen additive model allows the effect of the covariate to change with time. However, one of the setbacks of the nonparametric additive model is that its results cannot be easily interpreted numerically, although the covariate effects over time can be visualized [13]. Covariate effects in the Aalen nonparametric additive model are expected to be time-dependent. However, in this study, most of the covariate have constant effects, and, in conformity, the martingale residual plots suggest that the Aalen model does not provide a good fit to the CHHS Follow-up dataset.

By and large, my study has significant strengths and weaknesses. One of its strengths is that it uses data from a cohort of men and women unlike several cardiovascular disease risk factor studies, which are restricted to a single sex. Furthermore, participation rates were similar for men and women in my study. Another strength of the CHHS study population is that many of the environmental risk factors for CVD including physical activity level, education status, alcohol use, and smoking were assessed during the survey. The large sample size, together with the long follow-up time of subjects in the CHHS Follow-up data are other strengths of my study. One limitation of my study is that the questionnaire did not assess family history of cardiovascular disease in most of the provinces surveyed. Another limitation of any study of this kind is recall bias where survey subjects fail to remember events in the past, and this can lead to misclassification of the potential risk factors. My study is also limited by the self-reported diabetes status and the lack of information to distinguish between type 1 and type 2 diabetes among study subjects.

Chapter 6
Conclusion

Cardiovascular disease is a non-communicable illness that burdens countries whatever their level of advancement. Identifying the predictors of the disease would assist in understanding the pathology of the disease. Possessing the right knowledge of the role of competing risk event in CVD risk prediction has become imperative for the effective management of the disease. My study explores additive and multiplicative hazards regression models in the competing risks analysis of CVD mortality precluded by non-CVD mortality. Multiplicative hazards models estimate the covariate effect in relative terms, while additive hazards models give the risk due to the covariate in absolute additive terms. Thus, the two models provide distinct facts about a risk factor, which make it desirable to use them together, not as alternatives. Moreover, it may be tough to decide beforehand whether the additive or the multiplicative hazards model will provide a good fit to any dataset. However in public health, the additive hazards model is easier to interpret for the general population since there is no need for comparison group.

The competing risk analysis of the CHHS Follow-up data found that the Cox cause-specific, the F-G subdistribution, and the L-Y additive models provide good fits to the data based on the residual plots, and the CVD-specific predictors identified in these models are identical. In addition, my study suggests that CVD death is common among the survey participants who are above 50 years old, men, hypertensive, those with a previous heart attack, singles, those who have previously experienced a stroke, and participants who reside in a western province, regardless of the handling of the competing non-CVD event. Other factors that foretell CVD mortality in this sample include cholesterol level, smoking status, alcohol status, and diabetes.

Researchers will benefit from user-friendly software for additive hazards models, which will handle all aspects of the data analysis including estimation, inference, and goodness of fit procedures. In the future, research can explore the competing risks analysis of the CHHS Follow-up dataset using the Cox-Aalen model proposed by Scheike and Zhang [65]. Their model allows additive nonparametric time-dependent covariate effects and multiplicative semiparametric time-independent covariate effects concurrently in a single hazards model. The Cox-Aalen model is capable of providing useful public health information for the prevention and control of CVD mortalities precluded by non-CVD mortalities. A weighted competing risk analysis resulting in estimates that can be generalized to the Canadian populace is also worthy of future research.

References

[1] Collett D. Modelling Survival Data in Medical Research. 3rd ed. CRC Press; 2015.

[2] Putter H, Fiocco M, Geskus R, et al. Tutorial in biostatistics: competing risks and multi-state models. Statistics in medicine. 2007;26(11):2389.

[3] Pintilie M. Analysing and Interpreting Competing Risk Data. Statistics in Medicine. 2007;26(6):1360–1367.

[4] Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. 2nd ed. John Wiley & Sons; 2002.

[5] Gooley TA, Leisenring W, Crowley J, Storer BE, et al. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. Statistics in Medicine. 1999;18(6):695–706.

[6] Pintilie M. Competing Risks: A Practical Perspective. vol. 58. 1st ed. John Wiley & Sons; 2006.

[7] Tai BC, Machin D, White I, Gebski V. Competing risks analysis of patients with osteosarcoma: a comparison of four different approaches. Statistics in Medicine. 2001;20(5):661–684.

[8] Southern DA, Faris PD, Brant R, Galbraith PD, Norris CM, Knudtson ML, et al. Kaplan–Meier methods yielded misleading results in competing risk scenarios. Journal of clinical epidemiology. 2006;59(10):1110–1114.

[9] Martinussen T, Scheike TH. Dynamic Regression Models for Survival Data. Springer Science & Business Media; 2007.

[10] Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association. 1999;94(446):496–509.

[11] Pintilie M. An Introduction to Competing Risks Analysis. Revista Española de Cardiología (English Edition). 2011;64(7):599–605.

[12] Klein JP, Van Houwelingen HC, Ibrahim JG, Scheike TH. Handbook of survival analysis. CRC Press; 2013.

[13] Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. 2nd ed. Springer; 2003.

[14] Noordzij M, Leffondré K, van Stralen KJ, Zoccali C, Dekker FW, Jager KJ. When do we need competing risks methods for survival analysis in nephrology? Nephrology Dialysis Transplantation. 2013;28(11):2670–2677.

[15] Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. American Journal of Epidemiology. 2009;170(2):244–256.

[16] de Glas NA, Kiderlen M, Vandenbroucke JP, de Craen AJ, Portielje JE, van de Velde CJ, et al. Performing Survival Analyses in the Presence of Competing Risks: A Clinical Example in Older Breast Cancer Patients. Journal of the National Cancer Institute. 2016;108(5):djv366.

[17] Klein JP. Modelling competing risks in cancer studies. Statistics in Medicine. 2006;25(6):1015–1034.

[18] Zhang X, Akcin H, Lim HJ. Regression analysis of competing risks data via semi-parametric additive hazard model. Statistical Methods & Applications. 2011;20(3):357–381.

[19] Lin D, Ying Z. Semiparametric analysis of the additive risk model. Biometrika. 1994;81(1):61–71.

[20] Shen Y, Cheng S. Confidence bands for cumulative incidence curves under the additive risk model. Biometrics. 1999;55(4):1093–1100.

[21] Sun J, Sun L, Flournoy N. Additive hazards model for competing risks analysis of the case-cohort design. Communications in Statistics-Theory and Methods. 2004;33(2):351–366.

[22] Scheike TH, ZHANG MJ. An additive–multiplicative Cox–Aalen regression model. Scandinavian Journal of Statistics. 2002;29(1):75–88.

[23] Lin D, Ying Z. Additive hazards regression models for survival data. In: Proceedings of the First Seattle Symposium in Biostatistics. Springer; 1997. p. 185–198.

[24] Haller B, Schmidt G, Ulm K. Applying competing risks regression models: an overview. Lifetime data analysis;p. 1–26.

[25] Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society. 1972;34(2):187–220.

[26] Kim E, Kim JS, Choi M, Thomas CRJ. Conditional Survival in Anal Carcinoma Using the National Population-Based Survey of Epidemiology and End Results Database (1988-2012). Diseases of the Colon & Rectum. 2016;59(4):291–298.

[27] Lee JJ, Lin MY, Chang JS, Hung CC, Chang JM, Chen HC, et al. Hepatitis C virus infection increases risk of developing end-stage renal disease using competing risk analysis. PloS one. 2014;9(6):e100790.

[28] de Mutsert R, Sun Q, Willett WC, Hu FB, van Dam RM. Overweight in early adulthood, adult weight change, and risk of type 2 diabetes, cardiovascular diseases, and certain cancers in men: a cohort study. American journal of epidemiology. 2014;179(11):1353–1365.

[29] Wada N, Jacobson LP, Cohen M, French A, Phair J, Muñoz A. Cause-specific life expectancies after 35 years of age for human immunodeficiency syndrome-infected and human immunodeficiency syndrome-negative individuals followed simultaneously in long-term cohort studies, 1984–2008. American journal of epidemiology. 2013;177(2):116–125.

[30] Vejakama P, Ingsathit A, Attia J, Thakkinstian A. Epidemiological study of chronic kidney disease progression: a large-scale population-based cohort study. Medicine. 2015;94(4):e475–483.

[31] Hess CN, Roe MT, Clare RM, Chiswell K, Kelly J, Tcheng JE, et al. Relationship Between Cancer and Cardiovascular Outcomes Following Percutaneous Coronary Intervention. Journal of the American Heart Association. 2015;4(7):e001779.

[32] Wickramasinghe CD, Ayers CR, Das S, de Lemos JA, Willis BL, Berry JD. Prediction of 30-Year Risk for Cardiovascular Mortality by Fitness and Risk Factor Levels The Cooper Center Longitudinal Study. Circulation: Cardiovascular Quality and Outcomes. 2014;7(4):597–602.

[33] Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. The Annals of Statistics. 1988;p. 1141–1154.

[34] Schöttker B, Herder C, Rothenbacher D, Perna L, Müller H, Brenner H. Serum 25-hydroxyvitamin D levels and incident diabetes mellitus type 2: a competing risk analysis in a large population-based cohort of older adults. European journal of epidemiology. 2013;28(3):267–275.

[35] Fawcett VJ, Flynn-O'Brien KT, Shorter Z, Davidson GH, Bulger E, Rivara FP, et al. Risk factors for unplanned readmissions in older adult trauma patients in Washington state: a competing risk analysis. Journal of the American College of Surgeons. 2015;220(3):330–338.

[36] Marashi-Pour S, Morrell S, Cooke-Yarborough C, Arcorace M, Baker D. Competing risk analysis of mortality from invasive cutaneous melanoma in New South Wales: a population-based study, 1988–2007. Australian and New Zealand Journal of Public Health. 2012;36(5):441–445.

[37] Dasgupta P, Youlden DR, Baade PD. An analysis of competing mortality risks among colorectal cancer survivors in Queensland, 1996–2009. Cancer Causes & Control. 2013;24(5):897–909.

[38] Shen W, Sakamoto N, Yang L. Cancer-specific mortality and competing mortality in patients with head and neck squamous cell carcinoma: a competing risk analysis. Annals of surgical oncology. 2015;22(1):264–271.

[39] Tang Z, Zhou T, Luo Y, Xie C, Huo D, Tao L, et al. Risk factors for cerebrovascular disease mortality among the elderly in Beijing: a competing risk analysis. PloS one. 2014;9(2):e87884.

[40] Abdollah F, Sun M, Thuret R, Jeldres C, Tian Z, Briganti A, et al. A competing-risks analysis of survival after alternative treatment modalities for prostate cancer patients: 1988–2006. European urology. 2011;59(1):88–95.

[41] Leening MJ, Ferket BS, Steyerberg EW, Kavousi M, Deckers JW, Nieboer D, et al. Sex differences in lifetime risk and first manifestation of cardiovascular disease: prospective population based cohort study. BMJ. 2014;349:g5992.

[42] Melberg T, Nygård OK, Kuiper KKJ, Nordrehaug JE. Competing risk analysis of events 10 years after revascularization. Scandinavian Cardiovascular Journal. 2010;44(5):279–288.

[43] Zhou T, Li X, Tang Z, Xie C, Tao L, Pan L, et al. Risk factors of CVD mortality among the elderly in Beijing, 1992–2009: an 18-year cohort study. International Journal of Environmental Research and Public Health. 2014;11(2):2193–2208.

[44] Bellera CA, MacGrogan G, Debled M, de Lara CT, Brouste V, Mathoulin-Pélissier S. Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. BMC medical research methodology. 2010;10(1):20.

[45] Lim HJ, Zhang X, Dyck R, Osgood N. Methods of competing risks analysis of end-stage renal disease and mortality among people with diabetes. BMC medical research methodology. 2010;10(1):97.

[46] Tabrizi R, Moosazadeh M, Sekhavati E, Jalali M, Afshari M, Akbari M, et al. Competing Risk Analyses of Patients with End-Stage Renal Disease. Electronic physician. 2015;7(7):1458.

[47] Taghipour S, Banjevic D, Fernandes J, Miller AB, Montgomery N, Jardine AK, et al. Predictors of competing mortality to invasive breast cancer incidence in the Canadian National Breast Screening study. BMC cancer. 2012;12(1):299.

[48] Patel RM, Knezevic A, Shenvi N, Hinkes M, Keene S, Roback JD, et al. Association of Red Blood Cell Transfusion, Anemia, and Necrotizing Enterocolitis in Very Low-Birth-Weight Infants. Journal of the American Medical Association. 2016;315(9):889–897.

[49] Ong DS, Spitoni C, Klouwenberg PMK, Lunel FMV, Frencken JF, Schultz MJ, et al. Cytomegalovirus reactivation and mortality in patients with acute respiratory distress syndrome. Intensive care medicine;.

[50] Li G, Cook DJ, Levine MA, Guyatt G, Crowther M, Heels-Ansdell D, et al. Competing Risk Analysis for Evaluation of Dalteparin Versus Unfractionated Heparin for Venous Thromboembolism in Medical-Surgical Critically Ill Patients. Medicine. 2015;94(36).

[51] Teixeira L, Rodrigues A, Carvalho MJ, Cabrita A, Mendonça D. Modelling competing risks in nephrology research: an example in peritoneal dialysis. BMC nephrology. 2013;14(1):110.

[52] Wolbers M, Koller MT, Witteman JC, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology. 2009;20(4):555–561.

[53] Shastri S, Tangri N, Tighiouart H, Beck GJ, Vlagopoulos P, Ornt D, et al. Predictors of sudden cardiac death: a competing risk approach in the hemodialysis study. Clinical Journal of the American Society of Nephrology. 2012;7(1):123–130.

[54] Dignam JJ, Zhang Q, Kocherginsky M. The use and interpretation of competing risks regression models. Clinical Cancer Research. 2012;18(8):2301–2308.

[55] Kutikov A, Egleston BL, Wong YN, Uzzo RG. Evaluating overall survival and competing risks of death in patients with localized renal cell carcinoma using a comprehensive nomogram. Journal of Clinical Oncology. 2010;28(2):311–317.

[56] Aalen O. A model for nonparametric regression analysis of counting processes. In: Mathematical Statistics and Probability Theory. Springer; 1980. p. 1–25.

[57] Aalen OO. A linear regression model for the analysis of life times. Statistics in Medicine. 1989;8(8):907–925.

[58] Aalen OO. Further results on the non-parametric linear regression model in survival analysis. Statistics in Medicine. 1993;12(17):1569–1588.

[59] McKeague IW, Sasieni PD. A partly parametric additive risk model. Biometrika. 1994;81(3):501–514.

[60] Huffer FW, McKeague IW. Weighted least squares estimation for Aalen's additive risk model. Journal of the American Statistical Association. 1991;86(413):114–129.

[61] Lim HJ, Zhang X. Additive and multiplicative hazards modeling for recurrent event data analysis. BMC medical research methodology. 2011;11(1):101.

[62] Abadi A, Saadat S, Yavari P, Bajdik C, Jalili P. Comparison of Aalen's additive and Cox proportional hazards models for breast cancer survival: analysis of population-based data from British Columbia, Canada. Asian Pacific Journal of Cancer Prevention. 2011;12:3113–3116.

[63] Sarker S. Applicability of multiplicative and additive hazards regression models in survival analysis. University of Saskatchewan; 2011.

[64] Lin D, Ying Z. Semiparametric analysis of general additive-multiplicative hazard models for counting processes. The annals of Statistics. 1995;p. 1712–1734.

[65] Scheike TH, Zhang MJ. Extensions and applications of the Cox-Aalen survival model. Biometrics. 2003;59(4):1036–1045.

[66] Who Health Organization. Cardiovascular Diseases (CVDs); 2017. Retrieved from: `http://www.who.int/mediacentre/factsheets/fs317/en`.

[67] Roth GA, Huffman MD, Moran AE, Feigin V, Mensah GA, Naghavi M, et al. Global and Regional Patterns in Cardiovascular Mortality From 1990 to 2013. Circulation. 2015;132(17):1667–1678.

[68] Public Health Agency of Canada. Tracking heart disease and stroke in Canada, 2009; 2009. Retrieved from: `http://www.phac-aspc.gc.ca/publicat/2009/cvd-avc/pdf/cvd-avs-2009-eng.pdf`.

[69] Manuel DG, Leung M, Nguyen K, Tanuseputro P, Johansen H. Burden of cardiovascular disease in Canada. Canadian Journal of Cardiology. 2003;19(9):997–1004.

[70] In: Kirch W, editor. Multifactorial Disease. Dordrecht: Springer Netherlands; 2008. p. 970–971. Available from: `https://doi.org/10.1007/978-1-4020-5614-7_2248`.

[71] Azeez M, Taylor PC. In: El Miedany Y, editor. Impact of Comorbidity. Cham: Springer International Publishing; 2017. p. 33–52. Available from: `https://doi.org/10.1007/978-3-319-59963-2_2`.

[72] Molloy GJ, Stamatakis E, Randall G, Hamer M. Marital status, gender and cardiovascular mortality: behavioural, psychological distress and metabolic explanations. Social science & medicine. 2009;69(2):223–228.

[73] Kaplan RM, Kronick RG. Marital status and longevity in the United States population. Journal of Epidemiology and Community Health. 2006;60(9):760–765.

[74] Ikeda A, Iso H, Toyoshima H, Fujino Y, Mizoue T, Yoshimura T, et al. Marital status and mortality among Japanese men and women: the Japan Collaborative Cohort Study. BMC Public Health. 2007;7(1):73.

[75] Harald K, Pajunen P, Jousilahti P, Koskinen S, Vartiainen E, Salomaa V. Modifiable risk factors have an impact on socio-economic differences in coronary heart disease events. Scandinavian Cardiovascular Journal. 2006;40(2):87–95.

[76] Kelly TN, Gu D, Chen J, Huang Jf, Chen Jc, Duan X, et al. Hypertension subtype and risk of cardiovascular disease in Chinese adults. Circulation. 2008;118(15):1558–1566.

[77] Vasan RS, Larson MG, Leip EP, Evans JC, O'Donnell CJ, Kannel WB, et al. Impact of high-normal blood pressure on the risk of cardiovascular disease. New England journal of medicine. 2001;345(18):1291–1297.

[78] Soedamah-Muthu SS, Fuller JH, Mulnier HE, Raleigh VS, Lawrenson RA, Colhoun HM. High risk of cardiovascular disease in patients with type 1 diabetes in the UK A cohort study using the general practice research database. Diabetes Care. 2006;29(4):798–804.

[79] Juutilainen A, Lehto S, Rönnemaa T, Pyörälä K, Laakso M. Similarity of the impact of type 1 and type 2 diabetes on cardiovascular mortality in middle-aged subjects. Diabetes care. 2008;31(4):714–719.

[80] Collaboration APCS, et al. The effects of diabetes on the risks of major cardiovascular diseases and death in the Asia-Pacific region. Diabetes care. 2003;26(2):360–366.

[81] Vernino S, Brown RD, Sejvar JJ, Sicks JD, Petty GW, O'Fallon WM. Cause-Specific Mortality After First Cerebral Infarction A Population-Based Study. Stroke. 2003;34(8):1828–1832.

[82] Ntaios G, Papavasileiou V, Makaritsis K, Milionis H, Michel P, Vemmos K. Association of ischaemic stroke subtype with long-term cardiovascular events. European Journal of Neurology. 2014;21(8):1108–1114.

[83] Who Heart Federation. Cardiovascular disease risk factors;. Retrieved from: `http://www.world-heart-federation.org/cardiovascular-health/cardiovascular-disease-risk-factors/`.

[84] Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the Epidemiology of Cardiovascular Disease: A Historical Perspective. The Lancet. 2014;383(9921):999–1008.

[85] Jousilahti P, Vartiainen E, Tuomilehto J, Puska P. Sex, age, cardiovascular risk factors, and coronary heart disease A prospective follow-up study of 14 786 middle-aged men and women in Finland. Circulation. 1999;99(9):1165–1172.

[86] Ho JE, Paultre F, Mosca L. The gender gap in coronary heart disease mortality: is there a difference between blacks and whites? Journal of women's health. 2005;14(2):117–127.

[87] Andresdottir M, Sigurdsson G, Sigvaldason H, Gudnason V. Fifteen percent of myocardial infarctions and coronary revascularizations explained by family history unrelated to conventional risk factors. The Reykjavik Cohort Study. European heart journal. 2002;23(21):1655–1663.

[88] Murabito JM, Pencina MJ, Nam BH, D'Agostino RB, Wang TJ, Lloyd-Jones D, et al. Sibling cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults. Jama. 2005;294(24):3117–3123.

[89] Weijmans M, van der Graaf Y, Reitsma J, Visseren F. Paternal or maternal history of cardiovascular disease and the risk of cardiovascular disease in offspring. A systematic review and meta-analysis. International journal of cardiology. 2015;179:409–416.

[90] Beauchamp A, Peeters A, Wolfe R, Turrell G, Harriss LR, Giles GG, et al. Inequalities in cardiovascular disease mortality: the role of behavioural, physiological and social risk factors. Journal of epidemiology and community health. 2010;64(6):542–548.

[91] Teo KK, Ounpuu S, Hawken S, Pandey M, Valentin V, Hunt D, et al. Tobacco use and risk of myocardial infarction in 52 countries in the INTERHEART study: a case-control study. The Lancet. 2006;368(9536):647–658.

[92] Gellert C, Schöttker B, Müller H, Holleczek B, Brenner H. Impact of smoking and quitting on cardiovascular outcomes and risk advancement periods among older adults. European journal of epidemiology. 2013;28(8):649–658.

[93] Gallo V, Neasham D, Airoldi L, Ferrari P, Jenab M, Boffetta P, et al. Second-hand smoke, cotinine levels, and risk of circulatory mortality in a large cohort study of never-smokers. Epidemiology. 2010;21(2):207–214.

[94] Snow WM, Murray R, Ekuma O, Tyas SL, Barnes GE. Alcohol use and cardiovascular health outcomes: a comparison across age and gender in the Winnipeg Health and Drinking Survey Cohort. Age and ageing. 2009;38(2):206–212.

[95] Schröder H, Masabeu A, Marti MJ, Cols M, Lisbona JM, Romagosa C, et al. Myocardial infarction and alcohol consumption: a population-based case-control study. Nutrition, Metabolism and Cardiovascular Diseases. 2007;17(8):609–615.

[96] Mukamal KJ, Chung H, Jenny NS, Kuller LH, Longstreth W, Mittleman MA, et al. Alcohol consumption and risk of coronary heart disease in older adults: the Cardiovascular Health Study. Journal of the American Geriatrics Society. 2006;54(1):30–37.

[97] Skov-Ettrup LS, Eliasen M, Ekholm O, Grønbæk M, Tolstrup JS. Binge drinking, drinking frequency, and risk of ischaemic heart disease: A population-based cohort study. Scandinavian journal of public health. 2011;39(8):880–887.

[98] Hamer M, Chida Y. Active commuting and cardiovascular risk: a meta-analytic review. Preventive medicine. 2008;46(1):9–13.

[99] LaCroix AZ, Leveille SG, Hecht JA, Grothaus LC, Wagner EH. Does walking decrease the risk of cardiovascular disease hospitalizations and death in older adults? Journal of the American Geriatrics Society. 1996;44(2):113–120.

[100] Collaboration APCS, et al. Cholesterol, coronary heart disease, and stroke in the Asia Pacific region. International journal of epidemiology. 2003;32(4):563–572.

[101] Cox DR, Snell EJ. A general definition of residuals. Journal of the Royal Statistical Society Series B (Methodological). 1968;p. 248–275.

[102] Cao HL. A comparison Between the Additive and Multiplicative Risk Models; 2005.

[103] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2015. Available from: `http://www.R-project.org/`.

[104] Promoting Heart Health in Canada : Linking Research , Policy & Action;. Retrieved from: `http://odesi2.scholarsportal.info/documentation/chhd1986-92/user_guide.pdf`.

[105] Maclean DR, Petrasovits A, Nargundkar M, et al. Canadian heart health surveys: a profile of cardiovascular risk. Survey methods and data analysis. Canadian Medical Association Journal. 1992;146(11):1969–1974.

[106] Fair M. Generalized record linkage system–Statistics Canada's record linkagesSoftware. Austrian Journal of Statistics. 2004;33(1&2):37–53.

[107] Katzmarzyk PT, Reeder BA, Elliott S, Joffres MR, Pahwa P, Raine KD, et al. Body mass index and risk of cardiovascular disease, cancer and all-cause mortality. Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique. 2012;p. 147–151.

[108] Statistics Canada. Microdata linkage program;. Retrieved from: `http://www.statcan.gc.ca/eng/health/link#cmdb`.

[109] Statistics Canada. Vital Statistics- Death Database (CVSD);. Retrieved from: `http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3233`.

[110] Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation. 2016;133(6):601–609.

[111] Andersen PK, Keiding N. Interpretability and importance of functionals in competing risks and multistate models. Statistics in medicine. 2012;31(11-12):1074–1088.

[112] Latouche A, Allignol A, Beyersmann J, Labopin M, Fine JP. A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. Journal of clinical epidemiology. 2013;66(6):648–653.

[113] Latouche A, Boisson V, Chevret S, Porcher R. Misspecified regression model for the subdistribution hazard of a competing risk. Statistics in medicine. 2007;26(5):965–974.

[114] Grambauer N, Schumacher M, Beyersmann J. Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified. Statistics in medicine. 2010;29(7-8):875–884.

[115] Beyersmann J, Dettenkofer M, Bertz H, Schumacher M. A competing risks analysis of bloodstream infection after stem-cell transplantation using subdistribution hazards and cause-specific hazards. Statistics in medicine. 2007;26(30):5360–5369.

[116] Lim HJ, Zhang X. Semi-parametric additive risk models: application to injury duration study. Accident Analysis & Prevention. 2009;41(2):211–216.

Appendix A

R Codes for Cox Cause-Specific Hazards Models

```
nomisv  <- read.csv("C:/Users/temitope/Desktop/R/nomissv.csv", header = TRUE)
```

```
library("survival")
library("KMsurv")
#############CVD########################
fit.cvd <-coxph(Surv(nomisv$duration,nomisv$cardio)˜nomisv$AGE+
nomisv$Borderline_Cholesterol+nomisv$High_Cholesterol+nomisv$SEX+
nomisv$MARRY+nomisv$Former_Smoke+nomisv$Current_Smoke+
nomisv$Never_Alcohol+nomisv$Former_Alcohol+nomisv$Hypertensive
+nomisv$Diabetic+nomisv$Heart_Attack+nomisv$STROKE+nomisv$Province
+nomisv$STROKE*nomisv$Heart_Attack+nomisv$Hypertensive*nomisv$Province,
method=’breslow’)
summary(fit.cvd,digits=4)
print.table(fit.cvd, digits=4)
#proportionality assumption test for Cox model

#marry(MARRY)
fetm <-survfit(Surv(nomisv$duration,nomisv$cardio)˜nomisv$MARRY)
sumfit <- summary(fetm)
lls0=log(-log (sumfit$surv[sumfit$strata==’nomisv$MARRY=0’]))
t0 <- sumfit$time[sumfit$strata==’nomisv$MARRY=0’]
plot(log(t0), lls0, xlab=’log(duration)’,ylab=’log(-log(S))’)
llsl=log(-log(sumfit$surv[sumfit$strata==’nomisv$MARRY=1’]))
t1= sumfit$time[sumfit$strata==’nomisv$MARRY=1’]
points(log(t1),  llsl, pch=3)
legend("topleft", legend=levels(factor(nomisv$MARRY)),bty="y",
 pch=c(1,3),
 title="MARRY")

###########Cox-Snell residuals##############
fit.cvd <- coxph(Surv(nomisv$duration,nomisv$cardio)˜nomisv$AGE+
nomisv$Borderline_Cholesterol+nomisv$High_Cholesterol+nomisv$SEX+
nomisv$MARRY+nomisv$Former_Smoke+nomisv$Current_Smoke+
nomisv$Never_Alcohol+nomisv$Former_Alcohol+nomisv$Hypertensive+
nomisv$Diabetic+nomisv$Heart_Attack+nomisv$STROKE+nomisv$Province
+nomisv$STROKE*nomisv$Heart_Attack+nomisv$Hypertensive*nomisv$Province,
method=’breslow’)
```

```
summary(fit.cvd)
coxsnellres <-nomisv$cardioresid(fit.cvd,type="martingale")
fitres <- survfit(coxph(Surv(coxsnellres,nomisv$cardio)~1,method=’breslow’),
type=’aalen’)
plot(fitres$time,-log(fitres$surv),type=’s’,xlab=’Cox-Snell Residuals’,
ylab=’Estimated Cumulative
 Hazards Function’)
abline(0,1,col=’red’,lty=2)

############Martingale residual###############
fit.cvd <- -coxph(Surv(nomisv$duration,nomisv$cardio)~nomisv$AGE+
nomisv$Borderline_Cholesterol+nomisv$High_Cholesterol+nomisv$SEX+
nomisv$MARRY+nomisv$Former_Smoke+nomisv$Current_Smoke+
nomisv$Never_Alcohol+nomisv$Former_Alcohol+nomisv$Hypertensive+
nomisv$Diabetic+nomisv$Heart_Attack+nomisv$STROKE+nomisv$Province
+nomisv$STROKE*nomisv$Heart_Attack+nomisv$Hypertensive*nomisv$Province,
method=’breslow’)
remart <- resid(fit.cvd,type="martingale")
plot(fit.cvd$linear.predictor, remart, xlab="Risk Score", ylab
 ="Martingale Resduals")
abline(0,0, lty=1, col=’black’)

###########Deviance residual#################
fit.cvd <- -coxph(Surv(nomisv$duration,nomisv$cardio)~nomisv$AGE+
nomisv$Borderline_Cholesterol+nomisv$High_Cholesterol+nomisv$SEX+
nomisv$MARRY+nomisv$Former_Smoke+nomisv$Current_Smoke+
nomisv$Never_Alcohol+nomisv$Former_Alcohol+nomisv$Hypertensive+
nomisv$Diabetic+nomisv$Heart_Attack+nomisv$STROKE+nomisv$Province+
nomisv$STROKE*nomisv$Heart_Attack+nomisv$Hypertensive*nomisv$Province,
method=’breslow’)
redev <- resid(fit.cvd,type="deviance")
plot(fit.cvd$linear.predictor, redev, xlab="Risk Score", ylab
 ="Deviance Resduals")
abline(0,0, lty=1, col=’black’)


############### non-CVD#################
fit.ncvd <- coxph(Surv(nomisv$duration,nomisv$ncardio)~nomisv$AGE
+nomisv$SEX+nomisv$Education+nomisv$MARRY+nomisv$Former_Smoke
+nomisv$Current_Smoke+nomisv$Never_Alcohol+nomisv$Former_Alcohol
+nomisv$OBESE+nomisv$Diabetic+nomisv$Province+nomisv$Current_Smoke
*nomisv$Province+nomisv$Former_Smoke*nomisv$Province,method=’breslow’)
summary(fit.ncvd,digits=4)
print.table(fit.cvd, digits=4)
```

```
#proportionality assumption test for Cox non-CVD-specific event
#MARRY (Marital status)
fetm <-survfit(Surv(nomisv$duration,nomisv$ncardio)˜nomisv$MARRY)
sumfit <- summary(fetm)
lls0=log(-log (sumfit$surv[sumfit$strata==’nomisv$MARRY=0’]))
t0 <- sumfit$time[sumfit$strata==’nomisv$MARRY=0’]
plot(log(t0), lls0, xlab=’log(duration)’,ylab=’log(-log(S))’)
llsl=log(-log(sumfit$surv[sumfit$strata==’nomisv$MARRY=1’]))
t1= sumfit$time[sumfit$strata==’nomisv$MARRY=1’]
points(log(t1),  llsl, pch=3)
legend("topleft", legend=levels(factor(nomisv$MARRY)),bty="n",
pch=c(1,3),
 title="Marital status")
#Former_Smoke (Former smoking)
fete <- survfit(Surv(nomisv$duration,nomisv$ncardio)˜nomisv$Former_Smoke)
sumfit <- summary(fete)
lls0=log(-log (sumfit$surv[sumfit$strata==’nomisv$Former_Smoke=0’]))
t0 <- sumfit$time[sumfit$strata==’nomisv$Former_Smoke=0’]
plot(log(t0), lls0, xlab=’log(duration)’,ylab=’log(-log(S))’)
llsl=log(-log(sumfit$surv[sumfit$strata==’nomisv$Former_Smoke=1’]))
t1= sumfit$time[sumfit$strata==’nomisv$Former_Smoke=1’]
points(log(t1), llsl, pch=3)
legend("topleft", legend=levels(factor(nomisv$Former_Smoke)),bty="n", pch=c(1,3),
 title="Former Smoker")
#Current_Smoke (Current Smoking)
fete2 <- survfit(Surv(nomisv$duration,nomisv$ncardio)˜nomisv$Current_Smoke)
sumfit <- summary(fete2)
lls0=log(-log (sumfit$surv[sumfit$strata==’nomisv$Current_Smoke=0’]))
t0 <- sumfit$time[sumfit$strata==’nomisv$Current_Smoke=0’]
plot(log(t0), lls0, xlab=’log(duration)’,ylab=’log(-log(S))’)
llsl=log(-log(sumfit$surv[sumfit$strata==’nomisv$Current_Smoke=1’]))
t1= sumfit$time[sumfit$strata==’nomisv$Current_Smoke=1’]
points(log(t1), llsl, pch=3)
legend("topleft", legend=levels(factor(nomisv$Current_Smoke)),bty="n", pch=c(1,3),
 title="Current Smoker")

##############Cox-Snell residuals############
fit.ncvd <- coxph(Surv(nomisv$duration,nomisv$ncardio)˜nomisv$AGE
+nomisv$SEX+nomisv$Education+nomisv$MARRY+nomisv$Former_Smoke
+nomisv$Current_Smoke+nomisv$Never_Alcohol+nomisv$Former_Alcohol
+nomisv$OBESE+nomisv$Diabetic+nomisv$Province+nomisv$Current_Smoke
*nomisv$Province+nomisv$Former_Smoke*nomisv$Province,method=’breslow’)
summary(fit.ncvd)
coxsnellres <-nomisv$ncardio-resid(fit.ncvd,type="martingale")
fitres <-survfit(coxph(Surv(coxsnellres,nomisv$ncardio)˜1,
method=’breslow’),
type=’aalen’)
plot(fitres$time,-log(fitres$surv),type=’s’,xlab=’Cox-Snell Residuals’,
ylab=’Estimated
 Cumulative Hazards Function’)
abline(0,1,col=’red’,lty=2)

###########Martingale residual###############
fit.ncvd <- coxph(Surv(nomisv$duration,nomisv$ncardio)~nomisv$AGE
+nomisv$SEX+nomisv$Education+nomisv$MARRY+nomisv$Former_Smoke
+nomisv$Current_Smoke+nomisv$Never_Alcohol+nomisv$Former_Alcohol+
nomisv$OBESE+nomisv$Diabetic+nomisv$Province+nomisv$Current_Smoke
*nomisv$Province+nomisv$Former_Smoke*nomisv$Province,method=’breslow’)
resmatncvd <- resid(fit.ncvd,type="martingale")
plot(fit.ncvd$linear.predictor, resmatncvd, xlab="Risk Score",
 ylab ="Martingale Resduals")
abline(0,0, lty=1, col=’black’)

############Deviance residual################
fit.ncvd <- coxph(Surv(nomisv$duration,nomisv$ncardio)~nomisv$AGE
+nomisv$SEX+nomisv$Education+nomisv$MARRY+nomisv$Former_Smoke
+nomisv$Current_Smoke+nomisv$Never_Alcohol+nomisv$Former_Alcohol
+nomisv$OBESE+nomisv$Diabetic+nomisv$Province+nomisv$Current_Smoke
*nomisv$Province+nomisv$Former_Smoke*nomisv$Province,method=’breslow’)
redevncvd <- resid(fit.ncvd,type="deviance")
plot(fit.ncvd$linear.predictor, redevncvd, xlab="Risk Score",
 ylab ="Deviance Resduals")
abline(0,0, lty=1, col=’black’)
```

Appendix B

R Codes for Fine and Gray Subdistribution Hazards Models

```
library("cmprsk")
##############CVD######################
y <- cbind( nomisv$AGE,nomisv$High_Cholesterol,nomisv$SEX,nomisv$MARRY,
nomisv$Former_Smoke,nomisv$Current_Smoke,nomisv$Never_Alcohol,
nomisv$Former_Alcohol,nomisv$Hypertensive,nomisv$Heart_Attack,
nomisv$STROKE,nomisv$Province,nomisv$Heart_Attack*nomisv$STROKE
,nomisv$Hypertensive*nomisv$Province)
fit_all <- crr(nomisv$duration,nomisv$status_, y,failcode=1, cencode=0,
variance=TRUE)
summary(fit_all, digits=4)
print.table(fit_all, digits=4)
#proportionality assumption test for Fine & Gray
#MARRY
fity <- cuminc(nomisv$duration, nomisv$status_, nomisv$MARRY)
a <- timepoints(fity, times=nomisv$duration)
cif=t(a$est[1:2,])
llcif=log(-log(1-cif))
matplot(log(unique(sort(nomisv$duration))), llcif, pch=c(2,4),col=1,
xlab='log(duration)',
ylab='lol(-log(1-cif))')
legend("topleft", legend=levels(factor(nomisv$MARRY)), pch=c(2,4),
 title="MARRY")

##############non-CVD#################
y <- cbind(nomisv$AGE,nomisv$SEX,nomisv$Education,nomisv$Former_Smoke
,nomisv$Current_Smoke,nomisv$Never_Alcohol,nomisv$Former_Alcohol,
nomisv$OBESE,nomisv$Diabetic,nomisv$Province,nomisv$Former_Smoke*
nomisv$Province,nomisv$Current_Smoke*nomisv$Province)fit_all <-
 crr(nomisv$duration,nomisv$status_,y,failcode=2,cencode=0,
variance=TRUE)
summary(fit_all, digits=4)
print.table(fit_all, digits=4)
```

Appendix C

R Codes for Aalen Additive Models

```
library("timereg")
################CVD#################
all.sig.var1 <- aalen(Surv(duration,cardio)~AGE+SEX+MARRY+Former_Smoke+
Current_Smoke+Hypertensive+Heart_Attack+STROKE+Province+
Province*Hypertensive+AGE*MARRY,data=nomisv)
print(all.sig.var1, digits=7)
plot(all.sig.var1,xlab="Survival Time in years")

#############Martingale residual##############
all.fit <- aalen(Surv(duration,cardio)~AGE+SEX+MARRY+Former_Smoke+
Current_Smoke+Hypertensive+Heart_Attack+STROKE+Province
+Province*Hypertensive+AGE*MARRY,data=nomisv, residuals=1, n.sim=0)
par(mfrow=c(1,3))
 X1<-model.matrix(~-1+AGE,nomisv)
re<- cum.residuals (all.fit,data=nomisv,cum.resid=0,n.sim=500,X1)
plot(re,score=1)
summary(re)
y<-model.matrix(~-1+SEX,nomisv)
re <- cum.residuals (all.fit,data=nomisv,cum.resid=0, n.sim=500,y)
plot(re, score=1)
summary(re)
Xx<-model.matrix(~-1+MARRY,nomisv)
rex <- cum.residuals (all.fit,data=nomisv,cum.resid=0,n.sim=500,Xx)
plot (rex, score=1 )
summary(rex)
X2<-model.matrix(~-1+Former_Smoke,nomisv)
re2 <- cum.residuals (all.fit,data=nomisv,cum.resid=0,n.sim=500,X2)
plot(re2,score=1)
summary(re2)
X4<-model.matrix(~-1+Current_Smoke,nomisv)
re4 <- cum.residuals (all.fit,data=nomisv,cum.resid=0,n.sim=500,X4)
plot(re4, score=1)
summary(re4)
X5<-model.matrix(~-1+Hypertensive,nomisv)
re5 <- cum.residuals (all.fit,data=nomisv,cum.resid=0,n.sim=500,X5)
plot(re5, score=1)
summary(re5)
X6<-model.matrix(~-1+Heart_Attack,nomisv)
re6 <- cum.residuals (all.fit,data=nomisv,cum.resid=0, n.sim=500,X6)
```

```
plot(re6, score=1)
summary(re6)
X7<-model.matrix(~-1+ STROKE,nomisv)
re7 <-cum.residuals (all.fit,data=nomisv,cum.resid=0,n.sim=500,X7)
plot(re7, score=1)
summary(re7)
X8<-model.matrix(~-1+ Province,nomisv)
re8 <-cum.residuals (all.fit,data=nomisv,cum.resid=0,n.sim=500,X8)
plot(re8, score=1)
summary(re8)

###########non-CVD####################
all.fit2 <- aalen(Surv(duration,ncardio)~AGE+SEX+Education+Former_
Smoke+Current_Smoke+Diabetic+Province+Former_Smoke*Province
+Current_Smoke*Province,data=nomisv)
print(all.fit2)
plot(all.fit2,xlab="Survival Time in years")
##########Martingale Residual################
all.fit2 <- aalen(Surv(duration,ncardio)~AGE+SEX+Education+Former_
Smoke+Current_Smoke+Diabetic+Province+Former_Smoke*Province+
Current_Smoke*Province,data=nomisv, residuals=1, n.sim=0)
par(mfrow=c(1,3))
 X1<-model.matrix(~-1+AGE,nomisv)
re<- cum.residuals (all.fit2,data=nomisv,cum.resid=0,n.sim=500,X1)
plot(re,score=1)
summary(re)
y<-model.matrix(~-1+SEX,nomisv)
re <- cum.residuals (all.fit2,data=nomisv,cum.resid=0, n.sim=500,y)
plot(re, score=1)
summary(re)
Xx<-model.matrix(~-1+Education,nomisv)
rex <- cum.residuals (all.fit2,data=nomisv,cum.resid=0,n.sim=500,Xx)
plot (rex, score=1 )
summary(rex)
X4a<-model.matrix(~-1+Former_Smoke,nomisv)
re4a <- cum.residuals (all.fit2,data=nomisv,cum.resid=0,n.sim=500,X4a)
plot(re4a, score=1)
summary(re4a)

re4 <- cum.residuals (all.fit2,data=nomisv,cum.resid=0,n.sim=500,X4)
plot(re4, score=1)
X5<-model.matrix(~-1+Diabetic,nomisv)
re5 <- cum.residuals (all.fit2,data=nomisv,cum.resid=0,n.sim=500,X5)
plot(re5, score=1)
summary(re5)
```

```
X6<-model.matrix(~-1+Province,nomisv)
re6 <- cum.residuals (all.fit2,data=nomisv,cum.resid=0, n.sim=500,X6)
plot(re6, score=1)
summary(re6)
```

Appendix D

R Codes for Lin and Ying Additive Models

```
library("Matrix")
library("ahaz")
#################CVD##################
a <- cbind(nomisv$AGE,nomisv$Borderline_Cholesterol,nomisv$High_Cholesterol,
nomisv$SEX,nomisv$MARRY,nomisv$Current_Smoke,nomisv$Former_Smoke,
nomisv$Never_Alcohol,nomisv$Former_Alcohol,nomisv$Hypertensive,
nomisv$Diabetic,nomisv$Heart_Attack,nomisv$STROKE,nomisv$Province,
 nomisv$Hypertensive*nomisv$Province)
y <-ahaz(Surv(time,nomisv$cardio),a,univariate=FALSE,robust=TRUE)
summary(y)

##############Martingale Residuals#########
reside <- predict(y,type = "residuals")
par(mfrow=c(1,3))
plot(time,reside[,1], xlab="Time", ylab="Martingale Residual",
 main="Age")
abline(0,0, col='black')
plot(time,reside[,4], xlab="Time", ylab="Martingale Residual",
 main="Sex")
abline(0,0, col='black')
plot(time,reside[,5], xlab="Time", ylab="Martingale Residual",
 main="Marry")
abline(0,0, col='black')
plot(time,reside[,10], xlab="Time", ylab="Martingale Residual",
 main="Hypertensive")
abline(0,0, col='black')
plot(time,reside[,11], xlab="Time", ylab="Martingale Residual",
 main="Diabetic")
abline(0,0, col='black')
plot(time,reside[,14], xlab="Time", ylab="Martingale Residual",
 main="Province")
abline(0,0, col='black')
###Martingale residuals for 500 random observations########
timres <- data.frame(time, reside)
resran <- timres[sample(1:nrow(timres), 500, replace=FALSE), ]
plot(resran[,1],resran[,2], xlab="Time", ylab="Martingale Residual",
main="Age")
abline(0,0, col='black')
plot(resran[,1],resran[,5], xlab="Time", ylab="Martingale Residual",
main="Sex")
abline(0,0, col='black')
plot(resran[,1],resran[,6], xlab="Time", ylab="Martingale Residual",
 main="Marry")
abline(0,0, col='black')
plot(resran[,1],resran[,15], xlab="Time", ylab="Martingale Residual",
main="Province")
abline(0,0, col='black')

###########non-CVD####################
b <- cbind(nomisv$AGE,nomisv$SEX,nomisv$Education,nomisv$MARRY,
nomisv$Former_Smoke,nomisv$Current_Smoke,nomisv$OBESE,nomisv$Diabetic,
nomisv$Heart_Attack,nomisv$Province)
z <-ahaz (Surv(time,nomisv$ncardio), b, univariate=FALSE,
 robust =TRUE)
summary(z)
#########Martingale Residuals###########
reside <- predict(z,type = "residuals")
plot(time,reside[,1], xlab="Time", ylab="Martingale Residual",
 main="Age")
abline(0,0, col='black')
plot(time,reside[,2], xlab="Time", ylab="Martingale Residual",
 main="Sex")
abline(0,0, col='black')
plot(time,reside[,5], xlab="Time", ylab="Martingale Residual",
 main="Former_Smoke")
abline(0,0, col='black')
plot(time,reside[,6], xlab="Time", ylab="Martingale Residual",
 main="Current_Smoke")
abline(0,0, col='black')
plot(time,reside[,10], xlab="Time", ylab="Martingale Residual",
 main="Province")
abline(0,0, col='black')
```