10-4-2017 11:00 AM

On the estimation of penetrance in the presence of competing risks with family data

Daniel Prawira
The University of Western Ontario

Supervisor
Dr. Yun-hee Choi, Ph.D
The University of Western Ontario

Graduate Program in Epidemiology and Biostatistics
A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science

Abstract

In family studies, we are interested in estimating the penetrance function of the event of interest in the presence of competing risks. Failure to account for competing risks may lead to bias in the estimation of the penetrance function. In this thesis, three statistical challenges are addressed: clustering, missing data, and competing risks. We proposed the cause-specific model with shared frailty and ascertainment correction to account for clustering and competing risks along with ascertainment of families into study. Multiple imputation is used to account for missing data. The simulation study showed good performance of our proposed model in estimating the penetrance function under high familial correlation. However the competing risks model without frailty provided a good alternative under low familial correlation. We illustrate the proposed model using Colon Cancer Family Registry data.

Keywords: Penetrance Function, Relative Risks, Competing Risks, Frailty Model, Clustered Data, Missing Data, Family Study, Time-to event data.

Acknowledgements

I would like to express my deepest gratitude to all the people who helped me during my study at Western University.

I am deeply grateful for my supervisor Dr. Yunhee Choi for her patience and guidance towards the development of this thesis. I am also thankful for Dr. Neil Klar as the supervisory committee for his support towards this thesis.

I would like to thank all the professors who taught me very interesting course especially for Dr. Guangyong Zou, Dr. Igor Karp, Dr. Neil Klar, Dr. Wenqing He who improved my knowledge about Statistics.

I would like to thank my parents and friends in the department of Biostatistics who helped me a lot for the completion of the thesis.

I am also thankful for Dr. Ken Butler who helped me a lot during my undergraduate study and who helped me formatting the whole thesis in LaTEX.

Lastly, I would like to thank Dr. Laurent Briollais for sparing his time to read this thesis and be part of the examination committee and traveled to London to attend the thesis defense.

Contents

List of Figures

List of Tables

List of Acronyms

CLI *Clinic Based Study Design*

CLI+ *Clinic Based Study Design where affected proband is a mutation carriers and at least one parent and one sib are affected by the disease*

Colon CFR *Colon Cancer Family Registry*

CRC *Colorectal Cancer*

DNA *Deoxyribonucleic acid*

EM *Expectation-Maximization*

HR *Hazard Ratio*

KM *Kaplan-Meier*

LS *Lynch Syndrome*

MAR *Missing at Random*

ML *Maximum Likelihood*

MCAR *Missing Completely at Random*

MNAR *Missing not at Random*

MMR *Missmatch Repair*

Model1 *Competing risks model with frailty using one stage estimation*

Model2 *Competing risks model with frailty using two-stage estimation*

Model3 *Competing risks model without frailty*

Model4 *Shared frailty model without competing events taken into account*

OLS *Other Lynch Syndrome Cancer*

POP *Population Based Study Design*

POP+ *Population Based Study Design where probands are affected by disease and also a mutation carrier*

RR *Relative Risks*

Chapter 1

Introduction

Competing risks analysis is a natural extension of survival analysis, where individuals can fail from one of several competing causes (Koller et al., 2012). Investigators are often interested in the distribution of time to failure for a main event in the presence of other competing risks.

One of the key challenges of competing risk survival analysis lies in the estimation of the survival function $S(t)$ or equivalently the cumulative incidence function $F(t)$. The method of estimating cumulative incidence calculated as $1 - S(t)$ from the Kaplan-Meier (KM) curve is known to be inappropriate and may lead to the overestimation of cumulative incidence in the presence of competing events due to the fact that the Kaplan-Meier curve assumed the individuals who experienced the competing events as censored. In the Framingham Osteoporosis study (Berry et al., 2010) a standard survival analysis overestimated the five-year risk of second hip fracture by 37% and 10-year risk of second hip fracture by 75% by treating competing risks as censored. Overestimation of cumulative incidence is the result of violating the non-informative censoring assumption requiring when constructing a KM curve (Kim, 2007). Non-informative censoring occurs when the reason why participants are censored is

unrelated to the study outcome (Ranganathan and Pramesh, 2012). Non-informative censoring is required to obtain a consistent estimate of cumulative incidence from the KM survival curve. However, under competing risk survival analysis, individuals who are censored because they experienced competing events violate the non-informative censoring assumption.

The example of the competing risks can also be found in the breast cancer family study with BRCA1 mutation as a covariates (Gorfine and Hsu, 2011). In this study, Gorfine and Hsu (2011) are interested in estimating the risk of breast cancer in the presence of ovarian cancer, testis cancer, and other BRCA1 related cancers. This study introduced the cause-specific hazard model with flexible correlation structure to account for clustered competing risks data. The detail of this method is explained further in Chapter 2.

In this thesis, our interest lies in estimating the risk of colorectal cancer (CRC) in the presence of other Lynch Syndrome cancers (OLS) which arise from the Lynch Syndrome (LS) family study. Lynch Syndrome is a hereditary non-polyposis colorectal cancer (CRC) which carries a very high risk of colorectal cancer and other LS-related cancers (OLS) such as endometrial cancer, gastric cancer, etc (Lynch et al., 2009). Lynch Syndrome is indicated by germline mutation in the DNA missmatch repair (MMR) genes that causes the build up of error during DNA replication. Lynch Syndrome is estimated to account for 3% of colorectal cancer incidence. Choi et al. (2009), Kopciuk et al. (2009) and Jasperson et al. (2010) estimated the risk of developing colorectal cancer by age 70 associated with a MMR gene mutation without considering competing risks. In this thesis, we propose a statistical method to estimate the cumulative incidence function, also called penetrance function or the probability of developing CRC given the genotype, and relative risks of CRC in the presence of competing risks accounting for family study design. Penetrance estimation is useful to develop intervention

and prevention strategies for genetically susceptible individuals (Choi, 2012). In other words, the genetic counselors may use the information about the penetrance estimation to make a decision for early screening (colonoscopy) or early intervention of colonostomy to prevent colorectal cancer (CRC) for high risks individuals. We choose to estimate the penetrance at age 70 because it is approximately the midpoint of the global life expectancy between male and female according to the World Health Organization (WHO, 2017).

Three statistical challenges arising from this study are selection bias, clustering, and missing data. The first statistical challenge, which needs to be addressed, is selection bias. Families are selected based on the proband instead of a simple random sample. Probands are the first individuals in the families who experienced genetic disease. Thus, the families are unlikely to represent the general population and an ascertainment correction is required to ensure the validity of statistical inference. The second statistical challenge arises from clustered data collected from families. Familial correlation within the families will affect the cumulative incidence, the parameter estimates in the model, and their variances, so that it has to be taken into account for accurate and precise estimation. The third challenge is due to missing data. This issue arises because we do not observe genotype information from all the family members but we still want to make inference about those missing data using observed genotype and phenotype data from the families. These challenges are addressed extensively in Chapters 2 and 3.

1.1 Rationale

This thesis focuses on the cause-specific hazard model to estimate the relative and absolute risks of developing cancer associated with mutated genes in the presence of competing risks. Particularly, we provide the cumulative incidence function of CRC for LS families.

The aim of this study is to extend the competing risk models with frailty/random effect to account for familial correlation in a family study design. In particular, we extend the frailty model introduced by Choi (2012) to account for competing risks.

1.2 Scope of the thesis

In this thesis, we propose the cause-specific hazard model with frailty to account for competing risks and familial correlation. We estimate the relative and absolute risks of developing CRC in the presence of OLS adjusted for two covariates (gender and mutation status). In addition, we compare the proposed model with various models to see the effect of ignoring familial correlation and/or the competing risks.

1.3 Objectives of the study

The objectives of this thesis are:

1. To incorporate familial correlation and ascertainment correction in the cause-specific hazard model.

2. To estimate the penetrance function at age 70 based on the parameter estimates from the proposed model.

3. By simulation study

   (a) To assess the performance of the proposed model in terms of bias and precision in the model parameter and penetrance estimations.

   (b) To compare the bias and efficiency of the estimates from proposed model with those from non-competing risks model and non-frailty model to see the effect of ignoring competing risks or ignoring familial correlation.

4. To apply our proposed model to estimate the risks of developing CRC from the Lynch Syndrome families in the Colon Cancer Family Registry http://www.coloncfr.org/.

1.4 Organization of the thesis

The remainder of the thesis is structured as follows: competing risk models, random effect methods to account for familial correlation, and multiple imputation to account for missing genotypes is presented in Chapter 2. The competing risks models which incorporate random effects are presented in Chapter 3. The design of the simulation study to evaluate the performance of our statistical framework is provided in Chapter 4. The results of the simulation study are then presented in Chapter 5. The application of our proposed model to Lynch Syndrome families from the Colon Cancer Family Registry (Colon CFR) is presented in Chapter 6. Finally, some discussion, concluding remarks, and possible future topics are provided in Chapter 7.

Chapter 2

Literature Review

This chapter introduces different statistical methods to account for competing risks and will address three statistical challenges arising from family studies: selection bias, familial correlation, and missing data. This chapter is divided into five sections. We introduce the cause-specific hazard model in Section 2.1. The family sampling designs along with the ascertainment correction for selection bias are described in Sections 2.2 and 2.3. Frailty models to account for familial correlation are then introduced in Section 2.4. Finally, missing data mechanisms along with methods to account for them are described in Section 2.5

2.1 Competing risks models

There are many methods available to model competing risks data. We describe the three most commonly used statistical models for analyzing competing risks in epidemiologic data: the cause-specific hazard model, the subdistribution hazard model, and the mixture model (Lau et al., 2009).

2.1.1 Cause-specific hazard model

The cause-specific hazard model is a competing risk model introduced by Prentice et al. 1978. In competing risks, the observed outcome for individual i consists of T_i, the time to event, and δ_i, the event type, or cause, j, which takes a value $j = 1, \cdots, J$. The cause-specific hazard for cause $j = 1, \ldots, J$ for individual i, $i = 1, \ldots, n_j$ is defined as

$$\lambda_{ij}(t; \mathbf{X}_i) = lim_{\Delta t \to 0} \frac{Pr(t \leq T_i < t + \Delta t, \delta_i = j | T_i \geq t; \mathbf{X}_i)}{\Delta t}.$$

The function $\lambda_{ij}(t; \mathbf{X}_i)$ represents the instantaneous risk from cause j at time t, given the vector of covariates $\mathbf{X}_i$ in the presence of other failure types. Assuming proportional hazards, the cause-specific hazard model is written as

$$\lambda_{ij}(t; \mathbf{X}_i) = \lambda_{j0}(t) \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_i),$$

where $\lambda_{j0}(t)$ represents the baseline hazard function for cause j and $\boldsymbol{\beta}_j$ represents the vector of cause-specific effects for covariates $\mathbf{X}_i$. The cumulative cause-specific hazard function for cause $j = 1, \ldots, J$ is defined as

$$\Lambda_{ij}(t; \mathbf{X}_i) = \int_0^t \lambda_{ij}(u; \mathbf{X}_i) du.$$

The overall survival function is obtained as

$$S(t; \mathbf{X}_i) = \exp\left\{ - \sum_{j=1}^{J} \Lambda_{ij}(t; \mathbf{X}_i) \right\}.$$

Thus, the cause-specific cumulative incidence for cause j can be defined as

$$F_j(t; \mathbf{X}_i) = \int_0^t \lambda_j(u; \mathbf{X}_i) S(u; \mathbf{X}_i) du = \int_0^t \lambda_{ij}(u; \mathbf{X}_i) \exp\left\{ - \sum_{j=1}^{J} \Lambda_{ij}(u; \mathbf{X}_i) \right\} du. \tag{2.1}$$

Equation (2.1) shows that all the cause-specific hazard functions must be identified to obtain the cause-specific cumulative incidence function. If not all competing risks are identified for the event of interest, it is impossible to obtain all the cause-specific hazards from the competing events.

2.1.2 Sub-distribution hazard model

To account for competing risks Fine and Gray (1999) introduced the sub-distribution hazard model. The sub-distribution hazard function for cause j is defined as

$$h_{ij}(t;\mathbf{X}_i) = lim_{\Delta t \to 0} \frac{Pr(t \leq T_i < t + \Delta t, \delta_i = j | T_i > t \cup (T_i < t \cap \delta_i \neq j); \mathbf{X}_i)}{\Delta t},$$

where δ_i is the event type indicator for cause j, T_i is the random variable for the minimum observed time. $\mathbf{X}_i$ is the vector of covariates for individual i. In the definition of subdistribution hazard, the individuals who experienced other competing events are still at risk for the event of interest, while in the cause-specific hazard, such individuals are considered to be censored from the event of interest.

The proportional sub-distribution hazard model is then:

$$h_{ij}(t;\mathbf{X}_i) = h_{j0}(t)\exp(\boldsymbol{\beta}_j^\top \mathbf{X}_i),$$

where $h_{j0}(t)$ is the baseline sub-distribution hazard for cause j and $\boldsymbol{\beta}_j$ is the vector of regression coefficients for cause j.

The cumulative incidence from the sub-distribution hazard assuming proportional hazards is then obtained as

$$F_j(t;\mathbf{X}_i) = 1 - \exp\{-H_{ij}(t;\mathbf{X}_i)\},$$

where $H_{ij}(t;\mathbf{X}_i) = \int_0^t h_{ij}(u;\mathbf{X}_i)du$ is the cumulative sub-distribution hazard for cause j.

2.1.3 Mixture model

Larson and Dinse (1985) established a mixture model for competing risks, assuming that the number of risk-specific failures follows a multinomial distribution, with the risk of failing

from cause j from the logistic model:

$$Pr(\delta_i = j; \mathbf{X}_i) = \frac{\exp(\beta_{j0} + \boldsymbol{\beta}_j^\top \mathbf{X}_i)}{\sum_{j=1}^{J} \exp(\beta_{j0} + \boldsymbol{\beta}_j^\top \mathbf{X}_i)},$$

where β_{j0} is a scalar constant and $\boldsymbol{\beta}_j$ is the vector of regression coefficients related to cause j.

Given the event j, the conditional survival function for cause j follows

$$S_j(t; \mathbf{X}_i) = Pr(T_i > t; \mathbf{X}_i, \delta_i = j) = \exp\left\{ - \int_0^t h_{j0}(x) dx \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_i) \right\},$$

where $h_{j0}(x)$ is the baseline sub distribution hazard function for event j and the $\boldsymbol{\beta}_j$ is the vector of the regression coefficients related to cause j.

The cause-specific cumulative incidence for cause j is obtained as

$$\begin{aligned} Pr(T_i \leq t, \delta_i = j; \mathbf{X}_i) &= Pr(T_i \leq t; \delta_i = j, \mathbf{X}_i) Pr(\delta_i = j; \mathbf{X}_i) \\ &= \left[1 - \exp\left\{ - \int_0^t h_{j0}(x) dx \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_i) \right\} \right] \times \frac{\exp(\beta_{j0} + \boldsymbol{\beta}_j^\top \mathbf{X}_i)}{\sum_{j=1}^{J} \exp(\beta_{j0} + \boldsymbol{\beta}_j^\top \mathbf{X}_i)}. \end{aligned}$$

This model is called mixture model because the cause-specific cumulative incidence for cause j is estimated using both continuous and discrete probability distribution. The main limitation of the mixture model approach is the distribution of the $Pr(T \leq t, \delta = j; \mathbf{X}_i)$ has to be correctly specified for appropriate inference (Andersen and Keiding, 2012).

2.1.4 Comparison between the competing risk models

This section summarize the advantages and the disadvantages of the cause-specific hazard model, the sub-distribution hazard model, and the mixture model.

The main advantage of the cause-specific hazard model is easier interpretation of the relative risks obtained through the cause-specific hazard model. The relative risks for cause j is interpreted as the relative change in the cause-specific hazard of event j for 1-unit increase in

the covariate (Lau et al., 2009). However, independent competing events assumption needs to be satisfied for correct inference of the cause-specific hazard and the cumulative incidence function.

The main advantage of the sub-distribution hazard model is direct modeling of the covariates towards the cumulative incidence function. However, the idea of placing the individuals who experienced the competing events in the risks set maybe counterintuitive (Lau et al., 2009).

The main advantage of the mixture model is the ability to derive the cause-specific hazard and the subdistribution hazard and are not constrained to be constant over time (Lau et al., 2009). However, both the distributions of the event of interest and the competing events have to be correctly specified for correct inference of obtaining the cause-specific hazard and the subdistribution hazard model.

In this thesis, we choose the cause-specific hazard model to account for the competing risks because the cause-specific hazards and cumulative incidence provide better understanding of the effect of the risks factor towards the population as a whole (Hinchliffe and Lambert, 2013).

2.2 Sampling design

The main objective of this thesis is to assess the performance of the cause-specific hazard model in estimating the penetrance function of colorectal cancer for individuals in Lynch Syndrome (LS) families. For this purpose, it is important to understand the distinguishing features of collecting families into the study. Gong and Whittemore (2003) described two types of family based design for collecting families into the study: clinic based design and population

based design. In the clinic based design, a family is eligible for the study if it meets criteria concerning multiple disease occurences among its members. These families typically are identified in clinics for counselling and high risk of the disease. The main limitation of clinic based design is the penetrance estimate from these families may not reflect the risk level in general population. Therefore, it motivated the designs in which families are sampled by identifying single affected individual or probands. Probands are sampled from general population in a given period of time and include their family members and this is called the population-based design.

Gong and Whittemore (2003) pointed out that the bias in risk estimate in both population-based design and clinic based-design is relatively small compared to the standard error provided that the disease requirements to ascertain families are not stringent specifically, three individuals or more diagnosed with the disease before 100 years old (non stringent). In addition, the parameter estimates from clinic-based families are more precise compared to the population-based families and reflected the larger identified number of mutation carriers in the family. Also, the upward bias is relatively large if there is a large variance in the risks among carriers.

Extending the work of Gong and Whittemore (2003), Choi et al. (2008) evaluated four-family based designs in terms of efficiency and accuracy of estimating relative risks and penetrances under several genetic models based on different ascertainment-corrected likelihood approaches. In this study, population-based study designs, POP and POP+, and two clinic-based designs, CLI and CLI+, were considered. POP study design indicates that proband is affected by the disease, while POP+ indicates both proband is affected by the disease and also a mutation carrier. CLI study design indicates that the proband and at least one parent and one sib are affected by the disease, while CLI+ indicates that the affected proband is a mutation carrier and at least one parent and one sib are affected by the disease. Choi et al. (2008) concluded

that design efficiency depends on the research objectives. For the research mainly focused on the estimation of genetic relative risks, CLI design yields the most efficient estimate. However, for the research focused on penetrance estimation, POP+ provides the most efficient estimates. In addition, the presence of second gene effect can lead into some bias in the risk estimation.

Choi (2012), accounting for study design, incorporated familial correlation by using frailty model, and proposed frailty-based ascertainment corrected likelihood approach for estimating absolute (penetrance) and relative risks of disease associated with mutated genes and handled missing genotypes using a modified segregation-based method. Choi (2012) concluded that family-specific frailty model performed well in estimating penetrance and relative risks under high to moderate correlation but under low familial correlation, independent model provided a reasonably good result and the frailty-based likelihood approach was shown effective implementation under population-based family registry.

In this thesis, we further extend Choi (2012)'s method to estimate the penetrance of colorectal cancer in the presence of the competing risks based on the prospective ascertainment-corrected likelihood.

2.3 Ascertainment corrected likelihood

In the case of population based sampling designs, the likelihood is modified by the probability of being ascertained through the proband at the age of examination. The probability for family f being ascertained into study, is denoted by $A_f(\boldsymbol{\theta})$,

$$P(T_{fp} < a_{fp}; \mathbf{X}_{fp}) = 1 - S(a_{fp}; \mathbf{X}_{fp}) = A_f(\boldsymbol{\theta}), \tag{2.2}$$

where p indexes the proband and a_{fp} represents the proband's age of examination and $\boldsymbol{\theta}$ is the vector of parameters needed to construct the model.

Then, the ascertainment corrected likelihood can be obtained from dividing the likelihood contribution with the probability of being ascertained

$$L^c(\boldsymbol{\theta}) = \prod_{f=1}^{n} \frac{L_f(\boldsymbol{\theta})}{A_f(\boldsymbol{\theta})}. \tag{2.3}$$

It is also important to note that depending on the sampling design, we will have different expression of $A_f(\boldsymbol{\theta})$. We may obtain the maximum likelihood estimates of the parameters involved in the model by maximizing the ascertainment-corrected likelihood. In Chapter 3, we will discuss the construction of ascertainment-corrected likelihood and the step to obtain the maximum likelihood estimates of the parameters involved in the model.

2.4 Frailty model

A frailty model is a random effect model, where the random effect (frailty) has multiplicative effect on the hazard (Hougaard, 1995). Frailty model can be used for univariate survival analysis, but it is also useful for multivariate failure times as the conditional of independent times given the random effect.

2.4.1 Shared frailty model for clustered data

The shared frailty model is a type of frailty model, where subjects from specific cluster share the same frailty factor (Duchateau and Janssen, 2008). The shared frailty model is usually relevant for event times for related individuals such as sibling or repeated measurement for

the same individual (Wienke, 2011). This model was first introduced by Clayton (1978) and extended to univariate gamma frailty model by Hougaard (2000). Under proportional hazard assumption, the conditional hazard for individual i in the cluster f is written as

$$h_{fi}(t) = h_0(t)\omega_f \exp(\boldsymbol{\beta}^\top \mathbf{X}_{fi}), \tag{2.4}$$

where $h_0(t)$ is the baseline hazard function, $\mathbf{X}_{fi}$ is the vector of covariates for individual i in family f, $\boldsymbol{\beta}$ is the vector of corresponding regression coefficients, and ω_f is the random effect for cluster f.

The common choice for the distribution of shared frailty ω_f is a one-parameter gamma distribution with shape parameter k and scale parameter $\frac{1}{k}$, Gamma$(k,\frac{1}{k})$, whose density function has the form

$$f_{\omega_f}(\omega_f) = \frac{\omega_f^{k-1}\exp(-k\omega_f)}{k^{-k}\Gamma(k)}. \tag{2.5}$$

The interesting property of this one-parameter gamma distribution is the simplicity of mean and variance. The mean of this variable ω_f is 1, and the variance is $\frac{1}{k}$. Thus ω_f is the measure on whether or not a particular cluster f is more frail relative to other clusters, and $\frac{1}{k}$ is the variability of the cluster f in the population of clusters (Duchateau and Janssen, 2008).

The shared frailty model assumes that the same random effect is shared by all family members within families so that the correlations between the event times of any two individuals within the same family are the same. This is the main limitation of the shared-frailty model because we assume the correlation of the event times between grandparents and grandchildren to be the same with the correlation of the event times between parents and children within the same family.

2.4.2 Shared frailty competing risks model for clustered data

Competing risks analysis for cluster data is a special type of competing risk analysis, where there is a correlation between the time-to-event outcomes within cluster (Zhou et al., 2012). Cluster data often arise in the observational study and clinical trial as the part of study design. Gorfine and Hsu (2011) introduced the frailty-based competing risks model for clustered data. The estimation of model parameter can be done in either one-stage or two-stage estimation. Hsu et al. (1999) introduced one stage estimation of model parameter in the semi-parametric setting by maximizing the partial likelihood with respect to the baseline parameters and $\boldsymbol{\beta}$ in correlated survival data. Hsu et al. (2004) proposed two-stage estimation method: first, estimate the dependent parameter k and then, estimate the rest of the parameters in the model assuming the dependence parameter is known.

Gorfine and Hsu (2011) introduced correlated frailties to account for three types of dependence: dependence of failure times for the same event between individuals in the same cluster, dependence of failure times for the different event between individuals in the same cluster, and dependence of failure times for different failure types within the subject in the same cluster.

2.4.3 Intracluster correlation

Another interesting property of the shared frailty model is that there is a relationship between Kendall's τ (non parametric measure of correlation) and the variance of the frailty under Gamma $(k,\frac{1}{k})$ (Munda et al., 2012). Under Gamma $(k,\frac{1}{k})$ and fixed covariates, Kendall's τ can be written as

$$\tau = \frac{1}{1+2k}. \tag{2.6}$$

The large value of k represents the small familial correlation, and the small value of k represents the high familial correlation. For example, for k=1, 2, 5, and 10 the corresponding Kendall's τ are 0.33, 0.20, 0.09, 0.04.

2.5 Missing data

Missing data is one of the statistical questions that we address in this thesis. To infer missing data, the missing data mechanism has to be assumed (Ibrahim et al., 2012). Reasons for missingness is an important factor to be considered to obtain valid statistical inference. There are three classifications for missing data mechanisms such as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

Data is considered to be MCAR if missingness does not depend on observed and unobserved data. Under MCAR, using a complete-case analysis will not create bias in the parameter estimation, however there is a loss in efficiency.

Data is considered to be MAR if the failure to observe data conditioning on the observed data, the missingness is independent of unobserved data (Little and Rubin, 1987). In most cases of MAR, complete-case analysis will create both bias and loss in efficiency when estimating parameters.

Data is considered to be MNAR if the failure to observe the data does depend on both observed data or unobserved data. Under MNAR, complete case analysis will create both bias and loss of efficiency in estimating parameters. In addition, valid inference on MNAR requires determination of the correct model for missing data mechanisms. Ibrahim et al. (2012) also mentioned that the assumptions of missing data mechanism cannot be determined solely by the

data at hand but sensitivity analysis can be conducted between MNAR and MAR.

Little (1992) introduced some methods to account for missing data in regression analysis; but in this thesis, the attention is limited to 3 methods:

1. Complete-case analysis
2. Imputations
3. Maximum-likelihood

For complete-case analysis, MCAR assumption must be satisfied to obtain the unbiased parameter estimate. Imputations and maximum-likelihood approach shared two assumptions: The joint distribution of the data is multivariate normal and missing data mechanism is ignorable or missing at random (Pigott, 2001)

Complete-case analysis

Complete-case analysis is a method in which any cases with any missing values are simply discarded. This method is easy for implementation, but considering we have about 70 % of missing data in the genetic mutation information of Lynch Syndrome patients, complete-case analysis seems to be inappropriate for our study because we will lose all the information that the patients with missing genotype mutation may have.

Imputations

Another method to account for missing data is through imputations based on least squares on imputed data and multiple imputation. The least square method imputes missing data in

three ways, by (1) unconditional mean, (2) conditional mean based on observed covariates X, (3) conditional mean based on observed covariates X and outcome value.

First, the unconditional mean imputation substitutes the missing value in X's with its unconditional sample means. This method yields an inconsistent estimate of μ, and under MCAR the sample variance of X_{mis} is biased by a factor of $\frac{(n^{mis}-1)}{n-1}$. so that this method is not generally recommended.

Second, the missing values in X's can be imputed by the conditional mean of missing values given the observed value of X's, often estimated by linear regression on the observed X from the complete case. This method inflates the residual variance and introduces a correlation between the incomplete observations.

Third, the missing value of X can be imputed by the conditional mean imputation estimated from linear regression on the observed Y's and X's. There is a bias in the regression estimates results from this methods, however there are some researchers Afifi and Elashoff (1969) who proposed the bias corrected version of this method in the case of univariate X. There is an issue when estimating standard error of the regression estimates from this method as the standard error tends to be too small and the formula of the standard error are hard to derive. To solve this issue, a bootstrap methodology is used to estimate the standard error of the regression estimates.

As an extension of the least square method, which impute each missing value by a single mean, the multiple imputation repeats multiple times the imputation by drawing missing values from an appropriate model and the complete data analysis with each imputation substituted, then the parameter estimates are obtained from aggregating the values of the parameter estimates across multiple imputations (Rubin, 1987). The main advantage of this method over

the least square method, it takes into account the errors in the imputation. Thus, the standard errors of the regression estimates is accounted for using multiple imputation.

Maximum likelihood

Instead of direct imputation of missing values, the maximum likelihood approach can account for missing data by using the expectation and maximization (EM) algorithm (Little, 1992). There are two steps, the E-step and M-step, required for the EM algorithm. In the E-step, construct the expected log-likelihood function using the joint probability summed over all the possible values of the variable with missing data (in discrete case) or using the integrals in the place of summation in the continous case and in the M-step obtain maximum likelihood estimates for the parameter, then repeat the two-steps until convergence. Under the same assumption about missing data mechanism, both the multiple imputations and the EM algorithm approaches produce consistent, asymptotically efficient, and normal estimates (Allison, 2012).

There are two advantages of the multiple imputation approach compared to maximum likelihood approach discussed by Dong and Peng (2013). First, when dealing with categorical variables, the multiple imputation outperformed EM in efficiency for both small and large sample size (Peng and Zhu, 2008). Second, once missing data have been imputed, fitting multiple model to the single data set does not require application of multiple imputation (Sinharay et al., 2001). Since the missing data on genotype can be classified as binary variable, therefore we limit our attention to multiple imputation in accounting for missing data.

Chapter 3

Statistical Models

The purpose of this chapter is to develop the cause-specific hazard model to account for familial correlation with family design. This chapter is divided into five sections. Section 3.1 describes the competing risks model for clustered data. Likelihood constructions from the cause-specific model to estimate the model parameters is provided in Section 3.2. The estimation procedure to estimate model parameters and the penetrance function is provided in Section 3.3. The description on the imputation method to account for missing data for data analysis is provided in Section 3.4. Finally, the bootstrap method to obtain the variance of the penetrance function and the relative risks for family data is described in Section 3.5.

3.1 Shared frailty competing risks model for clustered data

The cause-specific hazard model introduced in Chapter 2 only accounts for the competing risks for the independent subjects. In this thesis, we introduce a shared frailty model for the competing risks data from family based study design to account for familial correlation.

Let T_{fi} be the minimum observed time to event for subject i in the family f where $i = 1,\ldots,n_f$, $f = 1,\ldots,n$ and δ_{fi} be the event indicator for a specific cause of failure, which takes value $j, j = 1,\ldots,J$. We let ω_f denote the shared frailty for family f and $\mathbf{X}_{fi}$ denote the vector of covariates for individual i in family f.

The cause-specific hazard for cause j in family f given covariates $\mathbf{X}_{fi}$ and shared frailty ω_f can be written as

$$\lambda_{fij}(t;\mathbf{X}_{fi},\omega_f) = lim_{\Delta t\to 0}\frac{Pr(t \le T_{fi} < t+\Delta t, \delta_{fi} = j | T_{fi} \ge t;\mathbf{X}_{fi},\omega_f)}{\Delta t}.$$

Assuming proportional hazards, the cause-specific hazard model with frailty can be written as

$$\lambda_{fij}(t;\mathbf{X}_{fi},\omega_f) = \lambda_{j0}(t)\omega_f \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi}),$$

where $\lambda_{j0}(t)$ is the baseline cause-specific hazard for cause j, where $\boldsymbol{\beta}_j$ is a vector of regression coefficients which correspond to log of the cause-specific hazard ratio.

The conditional cumulative cause-specific hazard of cause j for subject i in family f given shared frailty ω_f can be calculated as

$$\Lambda_{fij}(t;\mathbf{X}_{fi},\omega_f) = \int_0^t \lambda_{fij}(u;\mathbf{X}_{fi},\omega_f)du.$$

The conditional overall survival function is obtained as

$$S_{fi}(t;\mathbf{X}_{fi},\omega_f) = \exp\left\{-\sum_{j=1}^{J}\Lambda_{fij}(t;\mathbf{X}_{fi},\omega_f)\right\}.$$

Thus, the conditional cause-specific cumulative incidence for cause j for subject i in family f given shared frailty ω_f can be obtained as

$$\begin{aligned} F_{fij}(t;\mathbf{X}_{fi},\omega_f) &= \int_0^t \lambda_{fij}(u;\mathbf{X}_{fi},\omega_f)S_{fi}(u;\mathbf{X}_{fi},\omega_f)du \\ &= \int_0^t \lambda_{fij}(u;\mathbf{X}_{fi},\omega_f)\exp\left\{-\sum_{j=1}^{J}\Lambda_{fij}(u;\mathbf{X}_{fi},\omega_f)\right\}du. \end{aligned}$$

The marginal cause-specific cumulative incidence function can be expressed as:

$$F_{fij}(u,\mathbf{X}_{fi}) = \int_0^\infty F_{fij}(t;\mathbf{X}_{fi},\omega_f)g(\omega_f)d\omega_f$$
$$= \int_0^\infty \int_0^t \lambda_{fij}(u;\mathbf{X}_{fi},\omega_f)\exp\left\{-\sum_{j=1}^J \Lambda_{fij}(u;\mathbf{X}_{fi},\omega_f)\right\} du g(\omega_f)d\omega_f.$$

We could reverse the order of integration by changing the region of integration as follows:

$$F_{fij}(u,\mathbf{X}_{fi}) = \int_0^t \int_0^\infty \lambda_{fij}(u;\mathbf{X}_{fi},\omega_f)\exp\left\{-\sum_{j=1}^J \Lambda_{fij}(u;\mathbf{X}_{fi},\omega_f)\right\} g(\omega_f)d\omega_f du$$
$$= \int_0^t \int_0^\infty \lambda_{j0}(u)\exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi})\omega_f \exp\{-\sum_{j=1}^J \Lambda_{fij}(u;\mathbf{X}_{fi};\omega_f)\}g(\omega_f)d\omega_f du$$
$$= \int_0^t \lambda_{j0}(u)\exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi})$$
$$\int_0^\infty \omega_f \exp\{-\sum_{j=1}^J \Lambda_{j0}(u)\omega_f \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi})\}g(\omega_f)d\omega_f du$$
$$= \int_0^t \lambda_{j0}(u)\exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fij})(-1)^d \phi^{(d)}\left(\sum_{j=1}^J \Lambda_{j0}(u)\exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi})\right) du.$$

There is no close form for the marginal cause-specific cumulative incidence, and therefore numerical integration is used to estimate the cause-specific penetrance function.

3.2 Likelihood construction with ascertainment correction

According to the cause-specific hazard model given the frailty and observed covariates, we obtain the likelihood function for family f by integrating over the frailty distribution (Choi,

2012) as follows:

$$
\begin{aligned}
L_{Cf}(\boldsymbol{\theta}) &= \int \prod_{i=1}^{n_f} \left\{ \prod_{j=1}^{J} \lambda_{fij}(t_{fi};\mathbf{X}_{fi},\omega_f)^{I\{\delta_{fi}=j\}} \right\} S_{fi}(t_{fi};\mathbf{X}_{fi},\omega_f) g(\omega_f) d\omega_f \\
&= \int \prod_{i=1}^{n_f} \left[\prod_{j=1}^{J} \{\lambda_{j0}(t_{fi}) \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi}) \omega_f\}^{I\{\delta_{fi}=j\}} \right] \exp\{-\sum_{j=1}^{J} \Lambda_{fij}(t_{fij};\mathbf{X}_{fi};\omega_f)\} g(\omega_f) d\omega_f \\
&= \prod_{i=1}^{n_f} \prod_{j=1}^{J} \{\lambda_{j0}(t_{fi}) \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi})\}^{I\{\delta_{fi}=j\}} \\
&\int \prod_{i=1}^{n_f} \left[\prod_{j=1}^{J} (\omega_f)^{I\{\delta_{fi}=j\}} \right] \exp\{-\sum_{j=1}^{J} \Lambda_{j0}(t_{fi}) \omega_f \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi})\} g(\omega_f) d\omega_f \\
&= \prod_{i=1}^{n_f} \prod_{j=1}^{J} \{\lambda_{j0}(t_{fi}) \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi})\}^{I\{\delta_{fi}=j\}} \\
&\int (\omega_f)^{\sum_{i=1}^{n_f} \sum_{j=1}^{J} I\{\delta_{fi}=j\}} \exp\{-\omega_f \sum_{i=1}^{n_f} \sum_{j=1}^{J} \Lambda_{j0}(t_{fi}) \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi})\} g(\omega_f) d\omega_f \\
&= \prod_{i=1}^{n_f} \prod_{j=1}^{J} \{\lambda_{j0}(t_{fi}) \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi})\}^{I\{\delta_{fi}=j\}} (-1)^{d_f} \phi^{(d_f)} \left(\sum_{i=1}^{n_f} \sum_{j=1}^{J} \Lambda_{j0}(t_{fi}) \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi}) \right),
\end{aligned}
$$

where $\phi(z)$ is the Laplace transform of the frailty distribution $g(\omega_f)$, $\phi^{(d)}(z)$ is the dth derivative of $\phi(z)$ with respect to z, and $d_f = \sum_{i=1}^{n_f} \sum_{j=1}^{J} I\{\delta_{fi} = j\}$ and $\boldsymbol{\theta} = \{\lambda_{j0}(t_{fi}), \boldsymbol{\beta}_j, j = 1, \cdots, J, \log(k)\}$. The Laplace transform of the frailty distribution and dth derivative have the following forms

$$
\begin{aligned}
\phi(z) &= \int_0^\infty \exp(-\omega_f z) g(\omega_f) d\omega_f \\
\phi^{(d)}(z) &= (-1)^d \int \omega_f^d \exp(-\omega_f z) g(\omega_f) d\omega_f.
\end{aligned}
$$

For the data sampled through the affected probands, a correction for sampling bias is required to obtain the unbiased parameter estimates (Choi, 2012). In the case of prospective ascertainment-corrected likelihood in the population based design (POP, POP+), the ascertainment correction is done by the cumulative distribution function for the proband ($i = p$) who is affected by any of the event j at her or his age at examination a_{fp}. The ascertainment proba-

bility for family f can be expressed as

$$\begin{aligned}
P(T_{fp} < a_{fp}; \mathbf{X_{fp}}) &= \int \{1 - S(a_{fp}; \mathbf{X_{fp}}, \omega_f)\} g(\omega_f) d\omega_f \\
&= 1 - \int \exp\{-\sum_{j=1}^{J} \Lambda_{fpj}(a_{fp}; \mathbf{X_{fp}}, \omega_f)\} g(\omega_f) d\omega_f \\
&= 1 - \phi\left(\sum_{j=1}^{J} \Lambda_{j0}(a_{fp}) \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fp})\right) \\
&= A_{Cf}(\boldsymbol{\theta}).
\end{aligned}$$

The ascertainment-corrected likelihood from n families can be obtained by dividing each family's likelihood contribution by its probability of being ascertained, which we express as:

$$L_C^c(\boldsymbol{\theta}) = \prod_{f=1}^{n} \frac{L_{Cf}(\boldsymbol{\theta})}{A_{Cf}(\boldsymbol{\theta})}. \tag{3.1}$$

Refering to equation (3.1), we can calculate the ascertainment-corrected likelihood for all the families as

$$\begin{aligned}
L_C^c(\boldsymbol{\theta}) &= \prod_{f=1}^{n} \frac{\prod_{i=1}^{n_f} \prod_{j=1}^{J} \{\lambda_{j0}(t_{fi}) \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi})\}^{I\{\delta_{fi}=j\}} (-1)^{d_f} \phi^{(d_f)} \left(\sum_{i=1}^{n_f} \sum_{j=1}^{J} \Lambda_{j0}(t_{fi}) \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi})\right)}{1 - \int \exp\{-\sum_{j=1}^{J} \Lambda_{fij}(a_{fp}; \mathbf{X}_{fp}, \omega_f)\} g(\omega_f) d\omega_f} \\
&= \prod_{f=1}^{n} \frac{\prod_{i=1}^{n_f} \prod_{j=1}^{J} \{\lambda_{j0}(t_{fi}) \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi})\}^{I\{\delta_{fi}=j\}} (-1)^{d_f} \phi^{(d_f)} \left(\sum_{i=1}^{n_f} \sum_{j=1}^{J} \Lambda_{j0}(t_{fi}) \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi})\right)}{1 - \phi\{\sum_{j=1}^{J} \Lambda_{j0}(a_{fp}) \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fp})\}}.
\end{aligned}$$

Then, the corresponding ascertainment-corrected log-likelihood for cause-specific hazard model can be obtained as

$$\begin{aligned}
l_C^c(\boldsymbol{\theta}) &= \sum_{f=1}^{n} \{\log L_{Cf}(\boldsymbol{\theta}) - \log A_{Cf}(\boldsymbol{\theta})\} \\
&= \sum_{f=1}^{n} \sum_{i=1}^{n_f} \sum_{j=1}^{J} I\{\delta_{fi} = j\} \{\log(\lambda_{j0}(t_{fi})) + (\boldsymbol{\beta}_j^\top \mathbf{X}_{fi})\} \\
&+ \sum_{f=1}^{n} \log\left\{(-1)^{d_f} \phi^{(d_f)} \left(\sum_{i=1}^{n_f} \sum_{j=1}^{J} \Lambda_{j0}(t_{fi}) \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fi})\right)\right\} \\
&- \sum_{f=1}^{n} \log\left\{1 - \phi\left(\sum_{j=1}^{J} \Lambda_{j0}(a_{fp}) \exp(\boldsymbol{\beta}_j^\top \mathbf{X}_{fp})\right)\right\}.
\end{aligned}$$

For the special case of J=2, assuming $\lambda_{j0}(t_{fi})$ follows Weibull(λ_j, ρ_j), j=1,2, and the shared

frailty ω_f follows Gamma$(k,\frac{1}{k})$ with $\phi(z)=(1+\frac{z}{k})^{-k}$ and $\phi^{(d)}(z)=(-1)^d\frac{(k+d-1)!}{k!k^{d-1}}(1+\frac{z}{k})^{-k-d}$.

The ascertainment-corrected log-likelihood for all the families can be written as

$$\begin{aligned}
l_C^c(\boldsymbol{\theta}) &= \sum_{f=1}^{n}\sum_{i=1}^{n_f} I\{\delta_{fi}=1\}\left(\log(\lambda_1\rho_1)+(\rho_1-1)\log(\lambda_1 t_{fi})+\boldsymbol{\beta}_1^\top \mathbf{X}_{fi}\right)\\
&+\sum_{f=1}^{n}\sum_{i=1}^{n_f} I\{\delta_{fi}=2\}\left(\log(\lambda_2\rho_2)+(\rho_2-1)\log(\lambda_2 t_{fi})+\boldsymbol{\beta}_2^\top \mathbf{X}_{fi}\right)\\
&+\sum_{f=1}^{n}\log\{(k+d_f-1)!\}-\log(k!)-(d_f-1)\log(k)\\
&+\sum_{f=1}^{n}(-k-d_f)\log\left(1+\frac{\sum_{i=1}^{n_f}(\lambda_1 t_{fi})^{\rho_1}\exp(\boldsymbol{\beta}_1^\top \mathbf{X}_{fi})+\sum_{i=1}^{n_f}(\lambda_2 t_{fi})^{\rho_2}\exp(\boldsymbol{\beta}_2^\top \mathbf{X}_{fi})}{k}\right)\\
&-\sum_{f=1}^{n}\log\left(1-\left\{1+\frac{(\lambda_1 a_{fp})^{\rho_1}\exp(\boldsymbol{\beta}_1^\top \mathbf{X}_{fp})+(\lambda_2 a_{fp})^{\rho_2}\exp(\boldsymbol{\beta}_2^\top \mathbf{X}_{fp})}{k}\right\}^{-k}\right).
\end{aligned}$$

As shown in Section 3.1, there is no close form in estimating the penetrance function, therefore numerical integration is used to estimate the cause-specific penetrance function.

3.3 Estimation procedures

In this thesis, we estimate the model parameters in two ways: one-stage and two-stage estimations. In one-stage estimation, we estimate all the model parameters simultaneously by maximizing the log likelihood constructed in Section 3.2. For the two stage estimation, we estimate the baseline parameters and regression coefficients at the first stage from the competing risk model without frailty and then at the second stage, fixing those parameters estimate in the first stage, we estimate the frailty parameter k. Then, we combine the estimated frailty parameter k to the rest of parameters to estimate the penetrance function in second stage. The main advantage of two-stage estimation compared to the one-stage estimation is simple formulation of the model and fast computation (Hsu et al., 2004).

3.4 Missing data

The method to account for missing data is based on the multiple imputation method. In this study, we implemented the carrierprob function from **FamEvent** package (Choi et al., 2016) to compute mutation carrier probabilities for the individuals with missing genotypes based on observed data specific to disease status, gender, and relationship to probands.

After we obtain the mutation carrier probabilities for individuals with missing genotype, we sample carrier status with the carrier probability to impute the missing values. For the purpose of obtaining reliable parameter estimates, we replicate the imputation 1000 times and obtain average parameter estimates and penetrance estimates over 1000 imputed datasets. The results of the parameter estimation from imputed data are presented in Chapter 6.

3.5 Variance estimation

One of the challenging aspect of clustering is the variance estimation of parameters. Usually the parameter estimate of $\hat{\boldsymbol{\theta}}$ obtained from maximizing the ascertainment corrected log-likelihood is a consistent estimator of $\boldsymbol{\theta}$ for correlated cluster failure time, however the hessian matrix does not provide the correct estimate of the variance of the parameter $\hat{\boldsymbol{\theta}}$ because of the violation of independence assumption (Peng et al., 2007). Therefore we introduce bootstrap-based variance estimator to correctly estimate the variance of the parameters

3.5.1 Bootstrap-based variances estimation

We employ cluster bootstrap to estimate the variance of the penetrance function and the relative risks. The cluster boostrap refers to the technique of the sampling which resamples the cluster with replacement and include all the members in the cluster for analysis.

For B number of bootstraps resamples, the cause-specific hazard model for event j with P covariates yields a $B \times P$ matrix of regression coefficients. For each coefficient β_{jp}, $j = 1, \cdots J, p = 1, \cdots P$, the bootstrap-based SE (SE^B) is estimated as standard deviation of the log hazard ratio obtained through bootstraps $\hat{\beta_{jp}}^{*1}, ..., \hat{\beta_{jp}}^{*B}$

$$SE^B = \left\{ \frac{\sum_{b=1}^{B} (\hat{\beta_{jp}}^{*b} - \hat{\beta_{jp}}^{**})^2}{B-1} \right\}^{1/2}, \tag{3.2}$$

where $\hat{\beta}_{jp}^{**}$ is the mean of the log hazard ratio obtained through bootstraps. Assuming Wald Confidence Interval justified through normality assumption under Central Limit Theorem, the 95 % CI is constructed as $\hat{\beta_{jp}} \pm 1.96 SE^B$ where $\hat{\beta_{jp}}$ is the point estimate of the relative risks from the original data (Xiao and Abrahamowicz, 2010).

Similarly for the penetrance estimate for event j, the bootstrap-based SE (SE^B) is estimated as the standard deviation of the penetrance estimates obtained through bootstraps $\hat{P}_j(70;\mathbf{X})^{*1}, \cdots, \hat{P}_j(70;\mathbf{X})^{*B}$ with fixed covariates $\mathbf{X}$

$$SE^B = \left[\frac{\sum_{b=1}^{B} \{\hat{P}_j(70;\mathbf{X})^{*b} - \hat{P}_j(70;\mathbf{X})^{**}\}^2}{B-1} \right]^{1/2}, \tag{3.3}$$

where $\hat{P}_j(70;\mathbf{X})^{**}$ is the mean of the penetrance estimates obtained through bootstrap. Assuming Wald Confidence Interval justified through normality assumption under Central Limit Theorem, the 95 % CI is constructed as $\hat{P}_j(70;\mathbf{X}) \pm 1.96 SE^B$ where $\hat{P}_j(70;\mathbf{X})$ is the point estimate of the penetrance estimated from the original data.

The main advantages of bootstraps methods in estimating variance are avoiding distri-

butional assumptions of the model and assuming only exchangeability of clustered being re-sampled and it can approximate variance in a complex analysis of clustered data (Xiao and Abrahamowicz, 2010).

Chapter 4

Design of the Simulation Study

The purpose of this chapter is to describe the design of the simulation study for estimation of the penetrance function and relative risks accounting for the clustered competing risks. The simulation study is designed with the guidelines provided by Burton et al. (2006). There are four sections in this chapter. The objectives for this simulation study are provided in Section 4.1. The study design and data generation methods used in the simulation study are provided in Sections 4.2 and 4.3. Finally the evaluation criteria for the methods proposed in the simulation study are summarized in Section 4.4.

4.1 Objectives

The two objectives of the simulation study are:

1. To assess the performance of the cause-specific hazard model with frailty in estimating the penetrance by age 70 and relative risks associated with mutated genes and gender

when strong to low correlation is present within families.

2. To evaluate the performance of the following modelling approaches in the estimation of penetrance and relative risks for family data in the presence of competing risks.

Four models are considered:

Model1: Competing risks model with frailty using one stage estimation approach,

Model2: Competing risks model with frailty using two-stage estimation approach,

Model3: Competing risks model without frailty,

Model4: Shared frailty model without competing events taken into account.

In Model 1, all the parameters in the model are estimated simultaneously, and the penetrances at given ages are estimated by plugging the model parameter estimates into the penetrance function. In Model 2, the two-stage estimation approach is employed for the competing risks model with frailty: in the first stage, estimate the the baseline parameters and regression coefficients from Model 3, then in second stage, estimate the frailty parameter by fixing the parameters estimated from the first stage. In Models 3 and 4, all the parameters in the model are estimated simultaneously. For Model 4 the competing cause of failure (OLS) outome is considered as the censored outcomes.

Table 4.1: The choices of k parameter with its corresponding Kendall's τ

k	Kendall's τ
1	0.33
2	0.20
5	0.09
10	0.04

4.2 Selection of parameter values

We generate event times for two competing events based on the cause-specific hazard models with frailty and two binary covariates, x_{sex} and x_{gen},

$$\begin{aligned}\lambda_1(t|\mathbf{X},\omega_f) &= \lambda_{10}\omega_f \exp(\beta_{1sex}x_{sex}+\beta_{1gen}x_{gen})\\ \lambda_2(t|\mathbf{X},\omega_f) &= \lambda_{20}\omega_f \exp(\beta_{2sex}x_{sex}+\beta_{2gen}x_{gen}),\end{aligned}$$

where $\lambda_{10}=\lambda_1\rho_1(\lambda_1 t)^{\rho_1-1}$ and $\lambda_{20}=\lambda_2\rho_2(\lambda_2 t)^{\rho_2-1}$ are the baseline hazard functions for event 1 and event 2, respectively and ω_f is the family-specific frailty that follows the gamma distribution with mean 1 and variance $1/k$. The nine parameters involved in the model are $\boldsymbol{\theta}=\{\lambda_1,\rho_1,\lambda_2,\rho_2,\beta_{1sex},\beta_{1gen},\beta_{2sex},\beta_{2gen},k\}$.

The parameter values used in the simulation study are based on the parameter estimates from LS family data set introduced in Chapter 1. We vary the choice of k parameter into 1, 2, 5, and 10, which represents high to low familial residual correlations; the corresponding value of Kendall's τ is summarized in Table 4.1. The sizes of the families are considered 500, 779, and 1000 to investigate how all the statistical models performed under different family sizes. Thus, there are 12 scenarios considered for four distinct k values and three different family sizes. For each scenario, we generated 500 datasets and analyzed with the four models due to model complexity.

Table 4.2: The parameters values used to generate Time to Event data in the Cause-Specific Hazard Model

Parameters	Values
λ_1	0.0042
ρ_1	2.40
λ_2	0.0092
ρ_2	2.92
β_{1sex}	0.41
β_{1gen}	2.86
β_{2sex}	-0.72
β_{2gen}	1.27

4.3 Data generation

4.3.1 Cause-specific times to event data

The generation of competing risks data based on the cause-specific hazard model follows the algorithm proposed by Beyersmann et al. (2009):

1. Specify the cause-specific hazard functions $\lambda_{fi1}(t;\mathbf{X}_{fi},\omega_f)$ and $\lambda_{fi2}(t;\mathbf{X}_{fi},\omega_f)$ as a function of frailty and covariate values.

2. Simulate survival time T with all-cause specific hazard $\lambda_{fi1}(t;\mathbf{X}_{fi},\omega_f)+\lambda_{fi2}(t;\mathbf{X}_{fi},\omega_f)$ given the frailty and covariates values.

3. For a given simulated survival time T run a binomial or multinomial (if we have more

than 1 competing event) experiment which is decided by probability of

$$\frac{\lambda_{fi1}(t;\mathbf{X}_{fi},\omega_f)}{\lambda_{fi1}(t;\mathbf{X}_{fi},\omega_f)+\lambda_{fi2}(t;\mathbf{X}_{fi},\omega_f)}. \tag{4.1}$$

4. Generate censoring time C.

In terms of generating event times with hazard rate $\lambda(t|\mathbf{X}_{fi}) = \lambda_{fi1}(t;\mathbf{X}_{fi},\omega_f)+\lambda_{fi2}(t;\mathbf{X}_{fi},\omega_f))$, we use the inversion method (Bender et al., 2005).

4.3.2 Family data

In simulating family data, we generated the family structure consists of three generations of family members with two parents and two to five offsprings, with one of the offsprings to be proband. For each offspring has a spouse and they have two to five children. The gender of each members were generated with equal probability of being a male and female. The age of examination of the members from the first generation was generated using normal distribution with the mean of 65 and variance of 2.5. The age of examination of the members from the second generation was generated using the normal distribution with the mean of 45 and variance of 2.5. For each family the shared frailty was generated from Gamma distribution with a given value of the frailty parameter k. For the genotype variable, we generated the genotype of the probands first based on gender, age at examination, disease status, and shared frailty then based on proband's genotype, we generated the genotype for the rest of the family members. We generated time-to-onset for probands based on the competing risks model but adjusting the proband is affected before the age at examination. Then, we generated the age of onset for the rest of the family members unconditionally with the minimum age onset of 14 years and maximum age for follow up of 90 years. Finally, we determined the disease status by comparing the age at onset with the age of examination and if the age at onset is smaller than the age of

examination then equation (4.1) was used to determine disease status of 1 or 2. In this thesis, we modified the "simfam" commands from the **FamEvent** package (Choi et al., 2016) in R to generate the familial time-to-event data described above.

4.4 Evaluation criteria

We evaluate and compare the accuracy and precision for the penetrance and relative risks estimators from the four statistical models: the competing risks model with one-stage estimation (Model 1), the competing risks with two stage estimation (Model 2), the competing risks model without shared-frailty (Model 3), and the shared-frailty model without competing risks (Model 4).

4.4.1 Mean bias

The bias of an estimate is computed as the difference between the estimate and the true value. Then, we summarize the mean bias over simulations.

4.4.2 Empirical standard error

The empirical standard errors are obtained by the sample standard deviation of the estimates from 500 simulations.

Chapter 5

Results of the simulation study

In this chapter we focus our attention on the simulation results based on 779 families, as our Colon CFR data includes 779 families; they are summarized in Tables 5.1-5.3 and graphically in Figures 5.1–5.5. The simulation results based on 500 and 1000 families are presented in Tables A.1-A.7 in the Appendix. Section 5.1 describes the performance of different statistical models in estimating the relative risks and $\log(k)$ parameter. Section 5.2 describes the performance of different statistical models in estimating the colorectal cancer (CRC) penetrance at age 70.

5.1 Relative risks

The results of the simulation studies are summarized in Table 5.1 to evaluate the performance of various models under different familial correlations for estimating the log relative risks $\beta_{1gen}, \beta_{1sex}, \beta_{2gen}, \beta_{2sex}$ and the log-transformed frailty parameter $\log(k)$. The accuracy and precision of the parameter estimators from different models are also graphically displayed

in Figures 5.1 – 5.3. The four statistical models are the competing risks model with frailty using one stage estimation (Model 1), the competing risks model with frailty using two-stage estimation between k and the rest of the parameters (Model 2), the competing risks model without frailty (Model 3), and the shared frailty model without competing risks (Model 4).

Under strong familial correlation ($k = 1$ or $k = 2$), Model 2 performed the best in estimating the log of relative risks of major gene towards CRC (β_{1gen}) with bias of -0.09 or -0.01 with the standard error of 0.16. Model 2 also performed the best in estimating the log of relative risks of gender towards CRC (β_{1sex}) with bias of 0.05 and standard error of 0.09, compared to Model 1 (bias of 0.07 and standard error of 0.14). If we ignored the shared-frailty (Model 3), it provided similar estimation compared to Model 2. However, if we ignored the competing risks, it provided less precise estimate compared to Model 2. When estimating the frailty parameter k, Model 1 outperformed all the other statistical models with bias of -0.21 and standard error of 1.27. When estimating the baseline parameter shown in Table 5.3, Model 1 failed to capture the accurate estimate of both λ_1 and λ_2 parameter, with the bias of -0.002 and -0.0035 which is almost three times of Model 2. This might be due to Model 1 had to estimate 9 parameters simultaneously while Model 2 had to estimate those 9 parameters in 2 stages. Therefore Model 1 may have some issue in estimating the penetrance function because of relatively high bias in baseline parameter under strong familial correlation.

Under weak familial correlation ($k = 5$), Model 2 outperformed all the other statistical models in estimating both the log of the relative risks of covariates towards CRC (β_{1gen}, β_{1sex}) with bias of 0.03 and -0.02 and the standard error of 0.09 and 0.16. Model 2 also performed well in estimating the log of the relative risks of covariates towards OLS (β_{2gen}, β_{2sex}) with the bias of 0.05 and 0.03. However Model 1 still outperformed all the other statistical models in estimating k parameter with bias of 1.41 and the standard error of 26. Interestingly, Model 1

can capture the accurate and precise estimate of both λ parameter with bias of -0.0005 and -0.0009. If we ignored shared-frailty (Model 3), it performed relatively similar to competing risks with two-stage estimation in estimating β_{1gen}, β_{1sex}, β_{2sex}, β_{2gen} with similar bias and slightly higher standard error. If we ignored the competing risks (Model 4), it failed to accurately and precisely estimate β_{1gen}, β_{1sex}, β_{2sex}, β_{2gen}.

In general, as k increases, the standard error of all the parameters tend to decrease.

These results hold true in 3 different family sizes: 500, 779, and 1000, but generally the empirical standard error is lower as the family sizes increases.

5.2 Penetrance estimation

The penetrance functions for the colorectal cancer were estimated using four different statistical models: The competing risks model with frailty using one stage estimation (Model 1), the competing risks model with frailty using two-stage estimation between k and the rest of the parameters (Model 2), the competing risks model without frailty (Model 3), and the shared frailty model without competing risks (Model 4).

The simulation results in terms of accuracy and precision for estimating the penetrance function are summarized in Table 5.2 and graphically displayed in Figure 5.4–5.5. Under strong familial correlation ($k = 1$ or $k = 2$), the competing risks model with frailty using one stage estimation (Model 1) failed to capture the accurate estimate of the colorectal cancer penetrance by age 70 for both male and female carriers; biases for male and female carriers under $k = 1$ are -0.228 (SE = 0.099) and -0.164 (SE = 0.062) and under $k = 2$ are -0.145 (SE = 0.06) and -0.110 (SE = 0.037). The competing risks model with frailty using two-

stage estimation between k and the rest of the parameters (Model 2) performed the best in estimating the colorectal cancer by age 70 with bias for male and female carriers CRC: -0.016 (SE = 0.018) and -0.027 (SE = 0.014), respectively. If we ignored the frailty (Model 3), the statistical model can capture the accurate penetrance estimate for male and female carriers with a slightly larger empirical standard error. If we ignored the competing risks, the empirical standard error for all the penetrance estimate increases.

Under weak familial correlation ($k = 5$ or $k = 10$), the competing risks model with frailty using two-stage estimation between k and the rest of the parameters (Model 2) still performed the best in estimating the colorectal cancer for the male and female carriers; biases for male and female carriers under $k = 5$ are -0.020 (SE = 0.020) and -0.031 (SE = 0.016) and under $k = 10$ are -0.020 (SE = 0.021) and -0.031 (SE = 0.018), respectively. However, the competing risks model with frailty using one stage estimation (Model 1) performed reasonably well in estimating the CRC for male and female carriers; biases for male and female carriers under $k = 5$ are -0.069 (SE = 0.031) and -0.063 (SE = 0.022) and under $k = 10$ are -0.050 (SE = 0.021) and -0.051 (SE = 0.020), respectively.

In general, as k increases, all the penetrance estimates show some decrease in bias and standard error. These results hold true in 3 different family sizes: 500, 779, and 1000, but generally the empirical standard error is lower as the family sizes increases.

Figure 5.1: The accuracy and precision in the estimation of the log-relative risks towards CRC based on $\mathbf{n = 779}$ families using 4 different statistical models; the point and the interval estimates of the bias were based on the simulation study.

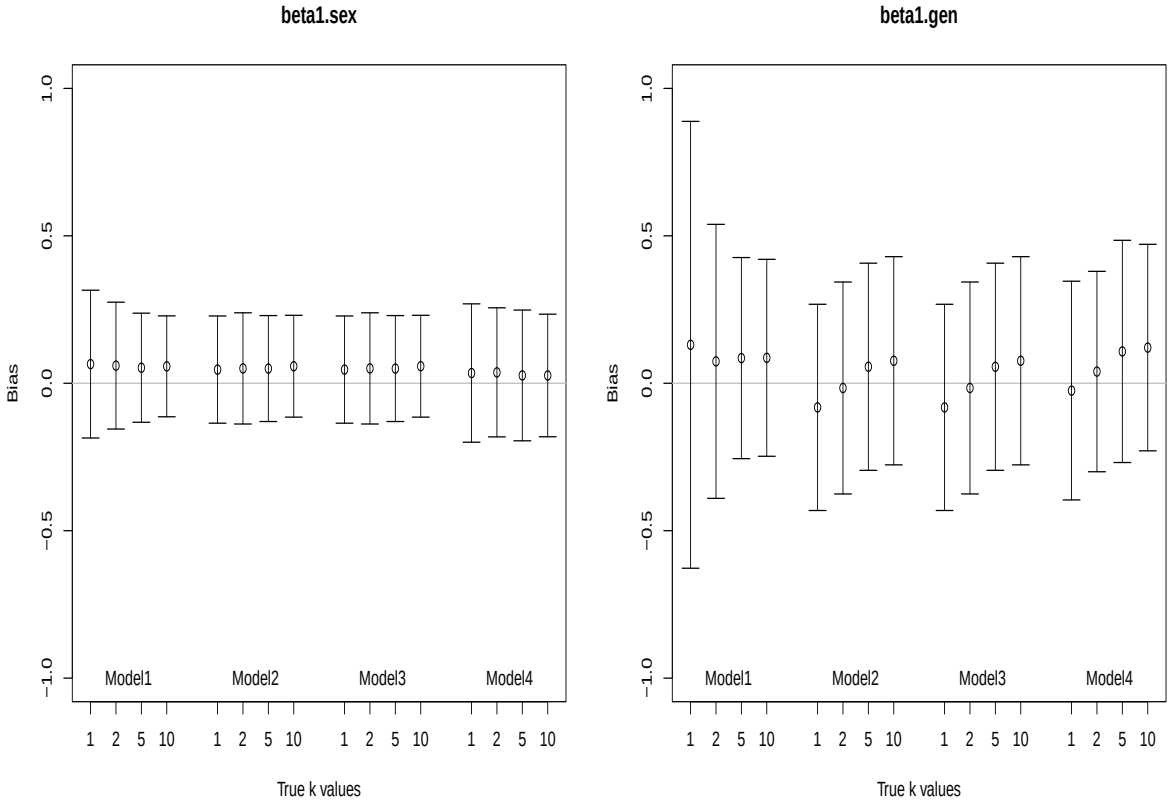


Figure 5.2: The accuracy and precision in the estimation of the log-relative risks towards OLS based on $\mathbf{n} = \mathbf{779}$ families using 3 different statistical models; the point and the interval estimates of the bias were based on the simulation study.

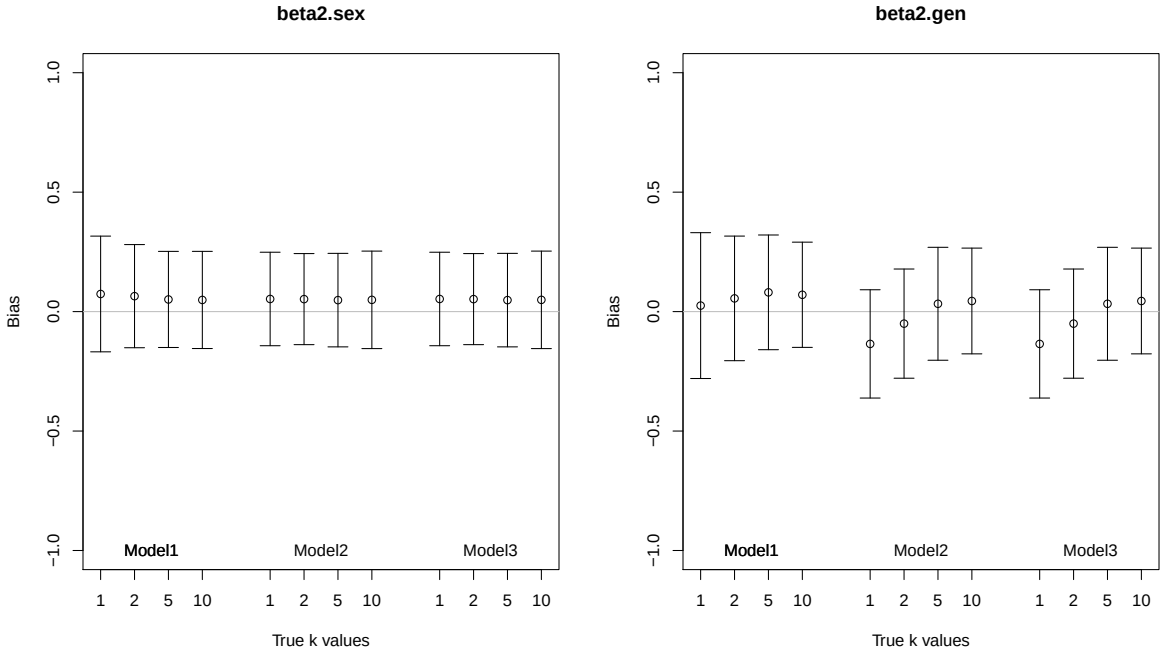


Figure 5.3: The accuracy and precision in the estimation of the log-transformed frailty parameter, $\log(k)$ based on $\mathbf{n} = \mathbf{779}$ families using Model 1 (left), Model 2 (mid), and Model 4 (right). The blue diamond inside the boxplot represents the mean bias of $\log(k)$ and the black line inside the box represents the median bias of $\log(k)$.

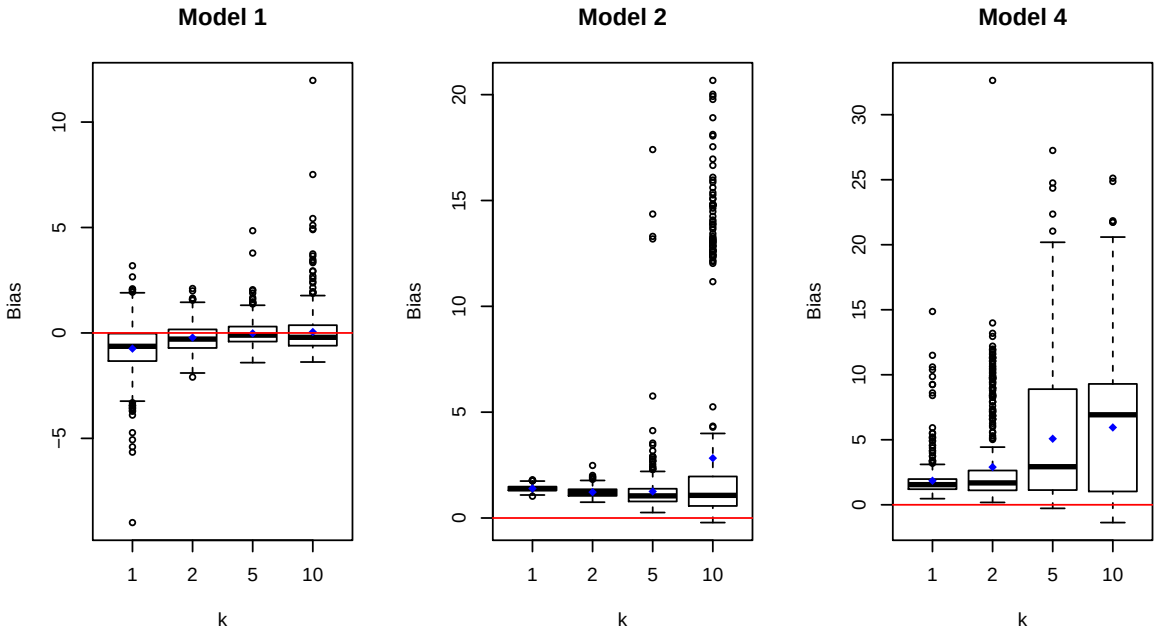


Figure 5.4: The accuracy and precision in the CRC penetrance estimation at age 70 based on $\mathbf{n = 779}$ families using 4 different statistical models; the point and the interval estimates of the bias were based on the simulation study.

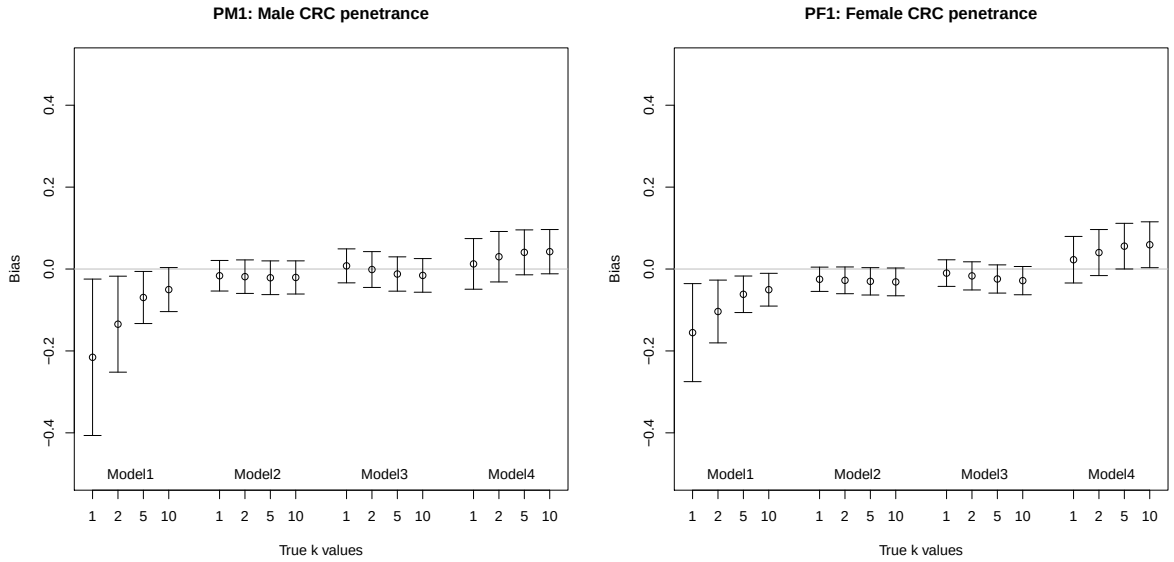


Figure 5.5: The accuracy and precision in the OLS penetrance estimation at age 70 based on $\mathbf{n = 779}$ families using 3 different statistical models; 0oint and the interval estimates of the bias were based on the simulation study.

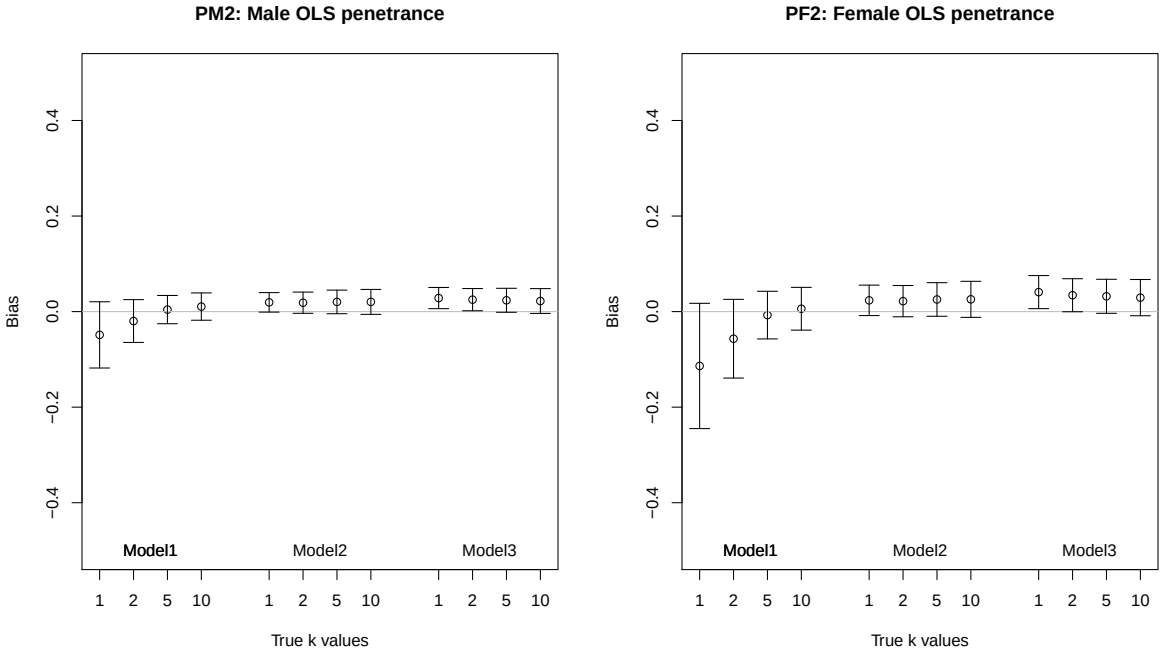


Table 5.1: Mean bias and empirical standard error (SE) for log relative risk $\beta's$ and frailty parameter $\log(k)$ estimates from various assumed models (Model1–Model4) for family data simulated in the presence of competing risks under different familial correlations ($k = 1, 2, 5, 10$); for each assumed k, $\mathbf{n} = \mathbf{779}$ families were simulated.

		True	Model 1		Model 2		Model 3		Model 4	
		Value	Bias	SE	Bias	SE	Bias	SE	Bias	SE
$k=1$	β_{1sex}	0.41	0.07	0.14	0.05	0.09	0.05	0.09	0.03	0.12
	β_{1gen}	2.86	0.14	0.35	−0.09	0.16	−0.09	0.16	−0.01	0.18
	β_{2sex}	−0.72	0.07	0.13	0.05	0.10	0.04	0.10	–	–
	β_{2gen}	1.28	0.05	0.16	−0.12	0.11	−0.12	0.12	–	–
	$\log(k)$	0	−1.11	2.22	1.25	0.12	–	–	1.31	0.67
$k=2$	β_{1sex}	0.41	0.06	0.10	0.05	0.09	0.05	0.09	0.03	0.11
	β_{1gen}	2.86	0.10	0.22	−0.01	0.16	−0.01	0.16	0.04	0.18
	β_{2sex}	−0.71	0.06	0.11	0.05	0.10	0.05	0.10	–	–
	β_{2gen}	1.28	0.07	0.12	−0.03	0.11	−0.03	0.10	–	–
	$\log(k)$	0.69	−0.35	0.57	1.07	0.18	–	–	1.80	2.47
$k=5$	β_{1sex}	0.41	0.06	0.09	0.06	0.09	0.06	0.09	0.02	0.11
	β_{1gen}	2.86	0.08	0.16	0.05	0.16	0.05	0.16	0.11	0.18
	β_{2sex}	−0.72	0.05	0.10	0.05	0.10	0.05	0.10	–	–
	β_{2gen}	1.28	0.08	0.11	0.03	0.11	0.03	0.11	–	–
	$\log(k)$	1.61	-0.07	0.53	1.05	1.11	–	–	3.48	3.88
$k=10$	β_{1sex}	0.41	0.06	0.09	0.06	0.09	0.06	0.09	0.03	0.11
	β_{1gen}	2.86	0.09	0.16	0.08	0.16	0.08	0.17	0.13	0.18
	β_{2sex}	−0.72	0.05	0.10	0.05	0.10	0.05	0.10	–	–
	β_{2gen}	1.28	0.08	0.11	0.05	0.11	0.05	0.11	–	–
	$\log(k)$	2.30	< 0.01	1.62	1.87	3.37	–	–	4.66	4.19

Table 5.2: Mean Bias and empirical standard error (SE) for penetrance estimates by age 70 for mutation carriers specific to gender and competing event from various assumed models (Model1–Model4) for family data simulated in the presence of competing risks under different familial correlations ($k = 1, 2, 5, 10$); for each assumed k, $\mathbf{n} = \mathbf{779}$ families were simulated; $P_1(X)$ represents the gender-specific penetrance estimate by age 70 for the first colorectal cancer with X taking male (M) and female (F) and $P_2(X)$ is the corresponding penetrance estimate for the other LS related cancer.

		True	Model 1		Model 2		Model 3		Model 4	
		Value	Bias	SE	Bias	SE	Bias	SE	Bias	SE
$k=1$	$P_1(M)$	0.402	−0.228	0.099	−0.016	0.018	0.012	0.027	0.010	0.032
	$P_1(F)$	0.273	−0.164	0.062	−0.027	0.014	−0.009	0.020	0.023	0.031
	$P_2(M)$	0.115	−0.053	0.037	0.018	0.010	0.028	0.011	–	–
	$P_2(F)$	0.242	−0.122	0.069	0.023	0.016	0.043	0.022	–	–
$k=2$	$P_1(M)$	0.446	−0.145	0.06	−0.021	0.019	−0.001	0.021	0.031	0.033
	$P_1(F)$	0.302	−0.110	0.037	−0.030	0.015	−0.017	0.016	0.043	0.020
	$P_2(M)$	0.129	−0.023	0.021	0.018	0.011	0.026	0.012	–	–
	$P_2(F)$	0.271	−0.063	0.039	0.022	0.017	0.036	0.018	–	–
$k=5$	$P_1(M)$	0.480	−0.069	0.031	−0.020	0.020	−0.010	0.020	0.039	0.029
	$P_1(F)$	0.325	−0.063	0.022	−0.031	0.016	−0.024	0.017	0.058	0.029
	$P_2(M)$	0.140	0.004	0.013	0.020	0.012	0.024	0.012	–	–
	$P_2(F)$	0.293	−0.009	0.23	0.024	0.017	0.032	0.018	–	–
$k=10$	$P_1(M)$	0.493	−0.050	0.021	−0.020	0.021	−0.014	0.021	0.043	0.028
	$P_1(F)$	0.333	−0.051	0.020	−0.031	0.018	−0.028	0.018	0.059	0.029
	$P_2(M)$	0.144	0.011	0.014	0.021	0.013	0.023	0.013	–	–
	$P_2(F)$	0.302	0.006	0.022	0.026	0.019	0.030	0.019	–	–

Table 5.3: Mean bias and empirical standard error (SE) for baseline parameter estimates from various assumed models (Model1–Model4) for family data simulated in the presence of competing risks under different familial correlations ($k = 1, 2, 5, 10$); for each assumed k, $\mathbf{n} = \mathbf{779}$ families were simulated.

		True	Model 1		Model 2		Model 3		Model 4	
		Value	Bias	SE	Bias	SE	Bias	SE	Bias	SE
$k=1$	λ_1	0.0042	−0.002	0.0009	−0.0007	0.0003	−0.0007	0.0004	−0.0008	0.0003
	ρ_1	2.40	0.05	0.10	−0.09	0.06	− 0.09	0.08	−0.09	0.07
	λ_2	0.0092	−0.0035	0.0017	−0.001	0.0005	0.001	0.0005	–	–
	ρ_2	2.92	−0.15	0.13	−0.31	0.08	−0.31	0.10	–	–
$k=2$	λ_1	0.0043	−0.0009	0.00048	−0.00047	0.00032	−0.00047	0.00032	−0.0005	0.0003
	ρ_1	2.40	0.06	0.08	−0.04	0.06	−0.04	0.06	−0.05	0.07
	λ_2	0.0092	−0.0019	0.0005	−0.0008	0.0004	0.0008	0.0004	–	–
	ρ_2	2.92	−0.15	0.11	−0.25	0.09	−0.24	0.09	–	–
$k=5$	λ_1	0.0042	−0.0005	0.0003	−0.0003	0.0003	−0.0003	0.0003	−0.0004	0.0004
	ρ_1	2.40	0.05	0.07	0.01	0.06	0.01	0.06	−0.01	0.07
	λ_2	0.0092	−0.0009	0.0005	−0.0005	0.0005	−0.0006	0.0005	–	–
	ρ_2	2.92	−0.15	0.09	−0.20	0.09	−0.20	0.09	–	–
$k=10$	λ_1	0.0042	−0.0004	0.0003	−0.0003	0.0003	−0.0003	0.0003	−0.0003	0.0004
	ρ_1	2.40	0.05	0.06	0.03	0.06	0.03	0.06	0.04	0.07
	λ_2	0.0092	−0.0007	0.0004	−0.0004	0.0004	−0.0004	0.0004	–	–
	ρ_2	2.92	−0.14	0.09	−0.17	0.09	−0.17	0.09	–	–

Chapter 6

Application to Lynch Syndrome Families

This chapter describes the analysis of LS family data identified from Colon Cancer Family Registry (Colon CFR). Section 6.1 provides some background information about the Colon Cancer Family. Section 6.2 provides the descriptive analysis from the data obtained from Colon Cancer Family. Section 6.3 provides model assumption for the baseline of the cumulative hazard. Section 6.4 provides the parameter estimation using four different statistical models. Section 6.5 provides the estimation of relative risks towards the development of CRC or OLS based on four statistical models. Section 6.6 provides the estimation of the penetrance by age 70 for Lynch Syndrome family members. Section 6.7 provides the summary of the results.

6.1 Motivations based on LS family data

Family studies have established the genetic research on colorectal cancer. These family studies especially the multiple-case families discovered the colorectal cancer susceptibility syndromes formally known as Lynch Syndrome (LS). The association of genetic towards the

Lynch Syndrome is evident with the high-penetrance identified in kindreds as mutation in DNA missmatch repair (MMR). Despite the mutations in MMR gene is the established cause of colorectal cancer, there are many questions remain to be answered in particular the effect of age and sex towards colorectal cancer. The Colon CFR build and maintain high-risks colorectal cancer patients and families record including epidemiologic risk factors, biological specimens and follow up the participants for colorectal cancer outcome (Newcomb et al., 2007).

Lynch Syndrome data from Colon CFR data contains information about the presence of colorectal cancer, and OLS related cancers. Thus it is a perfect scenario for competing risks analysis. In this thesis, we are interested in estimating the penetrance of colorectal cancer (CRC) at age 70 in the presence of other related Lynch Syndrome cancer (OLS). We considered two important covariates: gender and mutation carrier status. The aim of the analysis is to apply our proposed model to estimate the penetrance function of CRC in the presence of OLS as the competing risks along with the relative risks of developing CRC and OLS based on gender and mutation carrier status.

6.2 Data description

This section provides some information about data description of the Lynch Syndrome family data and some basic descriptive analysis for the data.

The Lynch Syndrome family data from Colon CFR consists of 7657 Individuals from 779 LS families. This data contains missing values, especially the indicator variable for the genetic mutation. Table 6.2 summarizes 7657 Lynch Syndrome family members based on the events of interests, 1305 developed CRC (738 males and 567 females), and 962 developed OLS (268

Table 6.1: The contingency table summarizing the incidence of colorectal cancer (CRC) and other Lynch Syndrome Cancer (OLS) and no event.

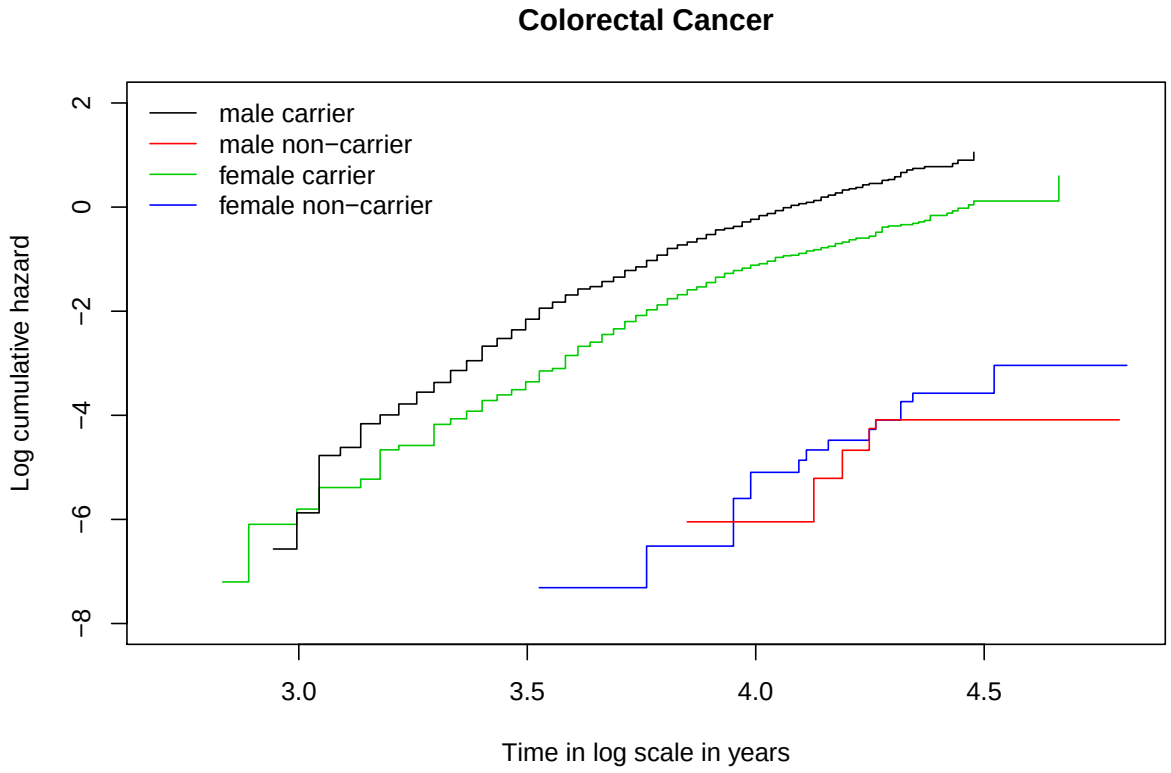
	CRC		OLS		No event	
	Male	Female	Male	Female	Male	Female
Mut=1	430	351	53	216	223	316
Mut=0	9	13	15	33	330	418
Mut=N/A	299	203	200	345	2257	1928
Total	738	567	268	594	2810	2662

males and 594 females). As shown in Table 6.1, we have 502 CRC cases where the genetic mutation information is missing, and 545 OLS cases where the genetic mutation information is missing. The mean age for CRC is about 45.18 years with standard deviation of 13 years, and the mean age for the OLS is about 51.52 years with standard deviation of 14 years. The missing data of genetic mutation does not depend on gender because the missing rate between male and female is similar. The missing data of genetic mutation may be depend on the event because the missing rate of CRC is roughly 38% and the missing rate of OLS is roughly 63% and the missing rate of no event is roughly 76%. However, we are not sure if the missing data in genetic mutation depends on the status of genetic mutation itself. Thus, we assume missing at random (MAR) as the missing data mechanism in this thesis.

6.3 Baseline model assumptions

In this thesis, we employ the cause-specific hazard model as a method to account for competing risks. In order to obtain a reliable result from the cause-specific hazard model, we have to check the distributional assumption with plots based on the observed data.

Figure 6.1: The log-cumulative hazard(Y-axis) and the time in log-scale(X-axis) for Colorectal Cancer and Other Lynch Syndrome cancers.



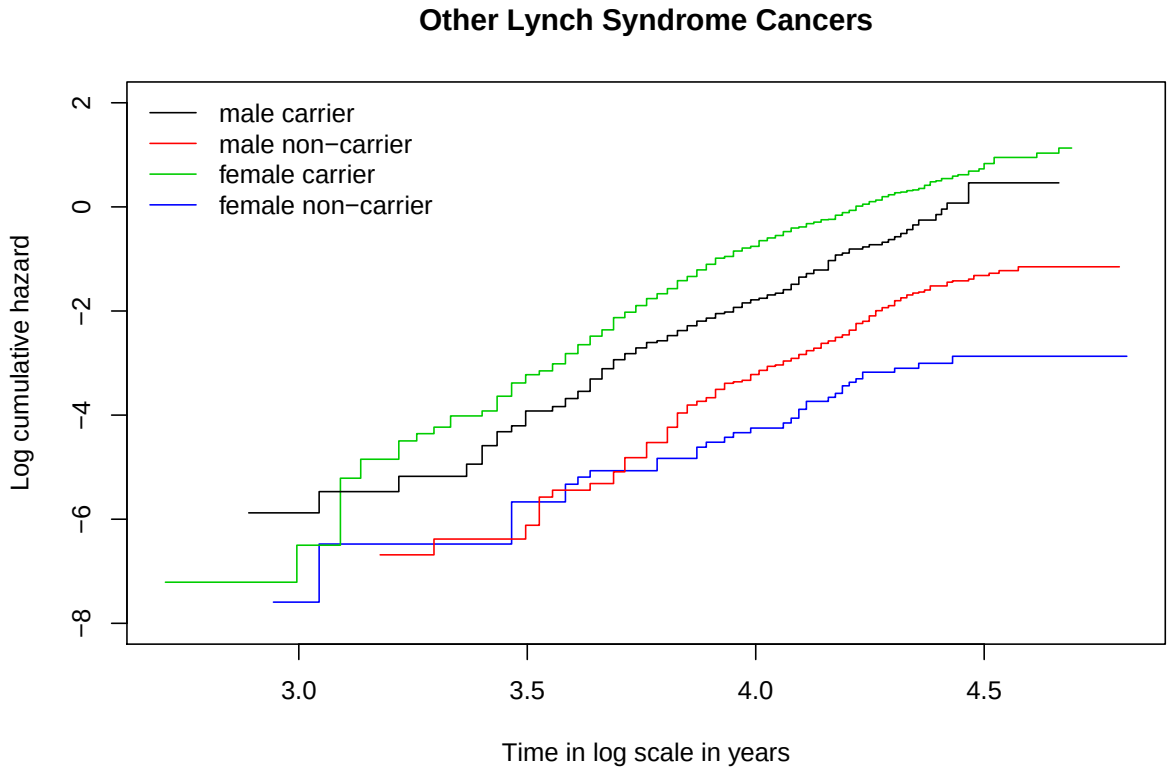


Figure 6.1 shows an approximate linear relationship between the log of cumulative hazard with respect to $\log(T)$, indicating that it is reasonable to assume that the baseline hazards follow Weibull distribution. Also, there is an approximate parallel pattern between the male carriers and female carriers beyond $\log(T) = 3$ making it reasonable to satisfy the proportional hazard assumption.

6.4 Model specifics

We fitted the data and estimated the penetrance function by age 70 and the relative risks using 4 different statistical models: The competing risks with shared-frailty estimated in one stage (Model 1), the two-stage estimation between k and the rest of the parameters (Model 2), the shared-frailty without competing risks (Model 3), and the competing risks model without frailty (Model 4). To account for the missing genotype, the multiple imputation methods based on the observed data is used. The empirical standard errors from all the parameters were obtained through 1000 bootstraps runs. Results of the relative risks and the penetrance estimation are presented in Table 6.2 and in Figure 6.2.

In this thesis, we are interested in estimating β_{1sex}, β_{1gen}, β_{2sex}, and β_{2gen}, where β_{1sex} corresponds to log of relative risks between male and female in developing CRC, β_{1gen} corresponds to log of relative risks between the mutation carriers and noncarriers in developing CRC, and β_{2sex} and β_{2gen} are the corresponding log relative risks for developing OLS cancer.

We are also interested in estimating the penetrance function $P_1(X)$ which correspond to gender specific penetrance estimate by age 70 for first CRC with X taking male (M) and female (F) and $P_2(X)$ which correspond to gender specific penetrance estimate by age 70 for other LS

(OLS) related cancer with X taking male(M) and female(F). We obtain the penetrance estimates by plugging the parameter estimates into the penetrance function derived in Section 3.2.

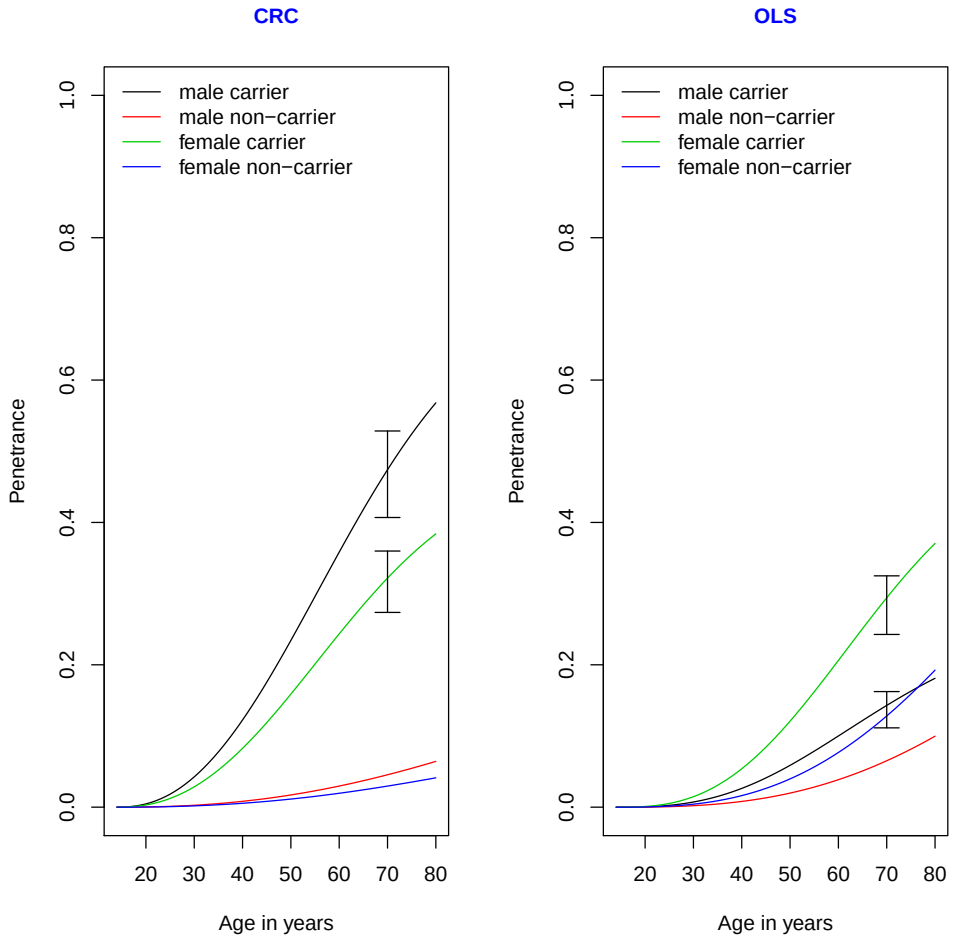
Table 6.2: Parameter estimation from the fitted data and the bootstrap-based standard error (SE^B) obtained through 1000 bootstrap runs. $P_1(X)$ and $P_2(X)$ represent the penetrance estimates of CRC and OLS, respectively, by age 70 for a given gender X. $-$loglik represents the negative log-likelihood value at maximum.

Parameter	Model 1		Model 2		Model 3		Model 4	
	Est	SE^B	Est	SE^B	Est	SE^B	Est	SE^B
λ_1	0.0042	0.00023	0.0040	0.00027	0.0040	0.00020	0.0033	0.00021
ρ_1	2.40	0.047	2.28	0.041	2.28	0.040	2.25	0.047
λ_2	0.0092	0.00025	0.0091	0.00025	0.0091	0.00025	–	–
ρ_2	2.92	0.081	2.78	0.073	2.78	0.073	–	–
β_{1sex}	0.41	0.066	0.42	0.063	0.41	0.061	0.65	0.081
β_{1gen}	2.86	0.012	2.72	0.011	2.72	0.011	2.59	0.013
β_{2sex}	-0.72	0.087	-0.71	0.082	-0.71	0.081	–	–
β_{2gen}	1.27	0.095	1.15	0.088	1.15	0.085	–	–
$\log(k)$	1.67	0.21	1.83	0.32	–	–	0.85	0.20
$-loglik$	11409	–	11415	–	11444	–	6195	–
Penetrance Estimates								
$P_1(M)$	0.4799	0.018	0.4677	0.031	0.4978	0.018	0.4051	0.021
$P_1(F)$	0.3248	0.012	0.3167	0.022	0.3343	0.012	0.2475	0.015
$P_2(M)$	0.1414	0.010	0.1369	0.013	0.1445	0.010	–	–
$P_2(F)$	0.2959	0.014	0.2838	0.021	0.2934	0.014	–	–

6.5 Relative risks

The relative risks of gene and sex for CRC and OLS were estimated using 4 different statistical models; the results are summarized in Table 6.2. Hazard ratio (HR) is the relative risk obtained through exponential transformation of the regression coefficient in the model. The competing risks with shared-frailty estimated in one-stage (Model 1) estimated the log-transformation of relative risks of sex and gene towards CRC as $\beta_{1sex} = 0.41$ (SE = 0.066) and $\beta_{1gen} = 2.86$ (SE = 0.012) and OLS as $\beta_{2sex} = -0.72$ (SE = 0.087) and $\beta_{2gen} = 1.27$ (SE = 0.095), respectively. The result indicates that being a mutation carrier increases the cause-specific hazard of developing colorectal cancer by approximately 17.5 times compared to the non-mutation carriers adjusting for gender and accounting for familial correlation (HR=17.46, 95% CI between 17.05 and 17.87). Also, being a male, compared to a female, increases the cause-specific hazard of developing colorectal cancer by 1.5 times adjusting for mutation status and incorporating familial correlation (HR=1.51, 95% CI between 1.32 and 1.71). The result shows that being a male, compared to a female, has a protective effect towards the OLS cancer by reducing the risks of developing OLS adjusting for mutation status and incorporating familial correlation by 51% (HR=0.49, 95% CI between 0.41 and 0.58). Also being a mutation carrier, compared to the non-mutation carrier, increases the risks of developing OLS cancer by 3.5 times adjusting for gender and incorporating familial correlation (HR = 3.56, 95% CI between 2.96 and 4.29). The estimate of frailty parameter k is around 5, which correspond to Kendall's tau around 0.09 indicating low familial correlation. The competing risks model without shared-frailty (Model 3) slightly underestimated relative risks of gene towards CRC and OLS compared to Model 1 with $\beta_{1gen} = 2.72$ (SE = 0.011) and $\beta_{2gen} = 1.15$ (SE = 0.085), respectively. However, the relative risks of sex towards CRC and OLS remains the same, however model 3 slightly overestimated the frailty parameter k to be around 6, which

Figure 6.2: The penetrance functions estimated for CRC (left) and OLS (right) based on the competing risks model with frailty using two-stage estimation; 95% CIs at age 70 for male and female carriers are displayed.



in turns underestimated the kendall's tau to be around 0.08. If we ignored the competing risks (Model 4), the relative risk of gene towards CRC was underestimated with $\beta_{1gen} = 2.59$ (SE = 0.013). Also, the frailty parameter k was underestimated to be around 2, which overestimated the Kendall's tau to be around 0.20. However, the relative risk of sex towards CRC was overestimated with $\beta_{1sex} = 0.65$ (SE = 0.081).

6.6 Penetrance estimation

As shown in Table 6.2, the penetrance estimates at age 70 are similar for Model 1 and Model 2. Refering to Figure 6.2, and Table 6.2 the CRC penetrance estimate for male carriers from two-stage estimation was estimated at 0.468 with the standard error of 0.031 (95 % CI of 0.407 to 0.529). The CRC penetrance estimate for female carriers from two-stage estimation was estimated at 0.317 with the standard error of 0.022 (95% CI of 0.274 to 0.360). The OLS penetrance estimate for male carriers from two-stage estimation was estimated at 0.137 with the standard error of 0.013 (95% CI of 0.111 to 0.162). The OLS penetrance estimate for female carriers from two-stage estimation was estimated at 0.284 with the standard error of 0.021 (95% CI of 0.243 to 0.325). If we ignored the familial correlation or frailty (Model 3), the penetrance estimates were similar to Model 1 with similar precision. This is because the estimate of the frailty parameter k is around 5 which corresponds to low familial correlation with Kendall's τ of 0.09. If we ignored the competing risks model (Model 4), the penetrance estimates were underestimated with worse precision compared to Model 1.

6.7 Summary

In summary, both gender and genetic mutation are important covariates in estimating the cause-specific hazard and cumulative incidence of colorectal cancer and other OLS cancer. If we ignored shared-frailty (Model 3) the model parameter estimates and the penetrance estimates were similar to the competing risks model because of low familial correlation. If we ignored the competing risks (Model 4), the model parameter estimates and the penetrance estimates were generally underestimated and had higher standard error.

Chapter 7

Discussion

7.1 Summary

This thesis presented the cause-specific hazard model to account for competing risks, introduced frailty concept to account for familial correlation, used multiple imputation as a method to account for missing data, and use ascertainment correction to account for study design. Then, we compared four different statistical models in estimating the penetrance functions by age 70 and the log of relative risks towards CRC in the presence of competing risks in the simulation studies. The simulation results show that under strong familial correlation ($k = 1$ or $k = 2$), the competing risks model with two-stage estimation (Model 2) outperformed all the other models in estimating the log of relative risks towards CRC (β_{1sex}, β_{1gen}). The advantage of this method is reflected further in the estimation of the penetrance function by age 70 by providing almost the unbiased estimates. Under moderate to weak familial correlation ($k = 5$ or $k = 10$) however, the competing risks model with frailty using one-stage estimation (Model 1) performed relatively well in estimating the penetrance function by age

70. Therefore we recommend Model 2 if the research objective is to accurately and precisely estimating the penetrance function of the event interest along with the competing events under strong familial correlation. However, under weak familial correlation, the competing risks model with frailty using one-stage estimation (Model 1) and the competing risks model without shared frailty (Model 3) can be used as an alternative method to two-stage estimation when estimating the penetrance functions of the event of interest and the competing events. If we ignore shared-frailty (Model 3), the empirical standard errors for all the penetrance estimates are higher compared to the two-stage estimation under strong familial correlation ($k = 1$ or $k = 2$). As a result, Model 3 failed to precisely estimate the penetrance functions under strong familial correlation. If we ignore the competing risks (Model 4), the bias and empirical standard errors for all the penetrance estimates are higher compared to the two-stage estimation for all k. As a result, Model 4 failed to accurately and precisely estimate the penetrance function under all assumptions of familial correlation.

The analysis from Colon CFR data presented the case under weak familial correlation, where competing risks without the shared frailty can be used as an alternative to the competing risks model with shared frailty (Model 1 and Model 2). By ignoring the frailty in Model 3, there is a minor/slight increase in all the penetrance estimates with relatively consistent bootstrap standard error compared to Model 1. There is a slight decrease in some of the log of the relative risks estimates from Model 1 and Model 3. However, the results from Model 1 and Model 3 arerelatively consistent within the 95% Confidence Interval. By ignoring the competing risks in Model 4, there is a significant decrease in the penetrance estimates with significant increase in β_{1sex}.

There is a limitation in our study as 12% non convergence rates were identified when $k = 1$ for Model 3 and 5% non convergence issues were identified for Model 1. However as

k increases, the non-convergence issue diminish to less than 2%. In other words, we have to generate around 560 datasets to obtain 500 convergence dataset for $k = 1$, but we only need around 510 datasets to obtain 500 convergence dataset for $k = 2, 5$, and 10. Further research is required to investigate why Model 3 have a higher rates of non-convergence when k is small.

The main advantage of our approach compared Gorfine and Hsu (2011) is the application of ascertainment correction towards the likelihood to account for ascertainment bias. However, the main advantage of the model proposed by Gorfine and Hsu (2011) is the assumption of correlated frailty to account for three types of dependence: dependence of failure times for the same event between individuals in the same cluster, dependence of failure times for the different events between individuals in the same cluster, and dependence of failure times for different events within the subject in the same cluster.

In general, this study reached similar conclusion with Choi (2012) that the shared-frailty model (Model 1 and Model 2) performed well in estimating the penetrance function and relative risks under strong to moderate familial correlation, but independent model (Model 3) performed well in estimating the penetrance function and relative risks under weak familial correlation.

7.2 Further work

We have only considered the case where the method to account for competing risks is limited to cause-specific hazard model. However, there are other methods such as subdistribution hazard model, the mixture model, and the pseudovalue regression model that may be used instead of the cause-specific hazard model. Also, for the missing data, we have only con-

sidered the multiple imputation as a method to account for missing data. Future work should also consider the EM-algorithm to account for missing genotype. In addition to that, Garibotti et al.(2006) also have extended the shared-frailty model into the correlated frailty model with the kinship coefficient matrix. Further work on the correlated frailty as a method to account for familial correlation could be considered. In addition, copula model to account for familial correlation instead of frailty may be considered as an alternative to account for familial correlation. In addition, further research in estimating the 95% confidence interval of the penetrance function, coverage with % missing on the left, and % missing on the right may be studied.

Bibliography

[1] Afifi, A. and Elashoff, R. (1969). Missing observation in multivariate statistics: Iii large sample analysis of simple linear regression. *Journal of the American Statistical Association*, 64(325):337–358.

[2] Allison, P. (2012). Handling missing data by maximum likelihood. Presented in SAS Global Forum.

[3] Andersen, P. and Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine*, 31(11-12):1074–1088.

[4] Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723.

[5] Berry, S. D., Ngo, L., Samelson, E. J., and Kiel, D. (2010). Competing risk of death: An important consideration in studies of older adults. *Journal of American Geriatrics Society*, 58(4):783–787.

[6] Beyersmann, J., Latouche, A., Buchholz, A., and Schumacher, M. (2009). Simulating competing risks data in survival analysis. *Statistics in Medicine*, 28(6):956–971.

[7] Burton, A., Altman, D., Royston, P., and Holder, R. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292.

[8] Choi, Y.-h. (2012). A frailty-model based method for estimating age dependent penetrance from family data. *Journal of Biometrics and Biostatistics*, S4:001. doi:10.4172/2155-6180.S4-001.

[9] Choi, Y.-h., Cotterchio, M., Mckeown-Eyssen, G., Neerav, M., Bapat, B., Boyd, K., Gallinger, S., Mclaughlin, Aronson, M., and Briollais, L. (2009). Penetrance of colorectal cancer among MLH1/MSH2 carriers participating in the colorectal cancer familial registry in Ontario. *Hereditary Cancer in Clinical Practice*, 7(1):14.

[10] Choi, Y.-h., Kopciuk, K., and Briollais, L. (2008). Estimating disease risks associated with mutated genes in family-based designs. *Human Hereditary*, 798(66):238–251.

[11] Choi, Y.-H., Kopciuk, K., He, W., and Briollais, L. (2016). *FamEvent: Simulation of Time-to-Event Family Data and Penetrance Estimation*. R package version 1.1.

[12] Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151.

[13] Dong, Y. and Peng, C.-y. J. (2013). *Principled missing data methods for researchers*. SpringerPlus, 2, 222.

[14] Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. Springer, New York.

[15] Fine, J. and Gray, R. (1999). A proportional hazard model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509.

[16] Garibotti, G., Smith, K., Kerber, R., and KM, B. (2006). Longevity and correlated frailty in multigenerational families. *Journal of Gerontology Biological Science Medical Science*, 61(12):1253–1261.

[17] Gong, G. and Whittemore, A. (2003). Optimal designs for estimating penetrance of rare mutations of a disease-susceptibility gene. *Genetic Epidemiology*, 24(3):173–180.

[18] Gorfine, M. and Hsu, L. (2011). Frailty-based competing risks model for multivariate survival data. *Biometrics*, 67(2):415–426.

[19] Hinchliffe, S. and Lambert, P. (2013). Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. *BMC Medical Research Methodology*, 13:13.

[20] Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis*, 1(3):255–273.

[21] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer, New York.

[22] Hsu, L., Chen, L., Gorfine, M., and Malone, K. (2004). Semiparametric estimation of marginal hazard function from case–control family studies. *Biometrics*, 60(4):936–944.

[23] Hsu, L., Prentice, R., Zhao, L., and Fan, J. (1999). On dependence estimation using correlated failure time data from case-control family studies. *Biometrika*, 86(4):743–753.

[24] Ibrahim, J., Chu, H., and Chen, M. (2012). Missing data in clinical studies: Issues and methods. *Journal of Clinical Oncology*, 30(26):3297–3303.

[25] Jasperson, K., Tuohy, T., Neklason, D., and Burt, R. (2010). Hereditary and familial colon cancer. *Gastroenterology*, 138(6):2044–2058.

[26] Kim, H. (2007). Cumulative incidence in competing risks data and competing risks regression analysis. *Clinical Cancer Research*, 13(2):559–565.

[27] Koller, M. T., Raatz, H., Steyerberg, W., and Wolbers, M. (2012). Competing risks and the clinical community:irrelevance or ignorance. *Statistics in Medicine*, 31(11-12):1089–1097.

[28] Kopciuk, K. A., Choi, Y.-h., Parkhomenko, E., Parfey, P., Mclaughlin, J., Green, J., and Briollais, L. (2009). Penetrance of HNPCC-related cancers in a retrolective cohort of 12 large Newfoundland families carrying a MSH2 founder mutation: an evaluation using modified segregation models. *Hereditary Cancer in Clinical Practice*, 7(1):16.

[29] Larson, M. and Dinse, G. (1985). A mixture model for the regression analysis of competing risks data. *Journal of the Royal Statistical Society*, 34(3):201–211.

[30] Lau, B., Cole, S., and Gange, S. (2009). Competing risk regression models for epidemiology data. *American Journal of Epidemiology*, 170(2):244–256.

[31] Little, R. (1992). Regression with missing X's: A review. *Journal of the American Statistical Assocation*, 87(420):1227–1237.

[32] Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data*. John Wiley and Sons, New York.

[33] Lynch, H., Lynch, P., Lanspa, S., Synder, C., Lynch, J., and Boland, C. (2009). Review of the lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clinical Genetics*, 76(1):1–18.

[34] Munda, M., Rotolo, F., and Legrand, C. (2012). parfm: Parametric frailty models in R. *Journal of Statistical Software*, 51(11):1–20.

[35] Newcomb, P., Baron, J., Cotterchio, M., Gallinger, S., Grove, J., Haile, R., Hall, D., Hopper, J., Jass, J., Le Marchand, L., Limburg, P., Lindor, N., Potter, J., Templeton, A., Thibodeau, S., and Seminara, D. (2007). Colon cancer family registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiology Biomarker*, 16(11):2331–2343.

[36] Peng, C. and Zhu, J. (2008). Comparison of two approaches for handling missing covariates in logistic regression. *Educational and Psychological Measurement*, 68(1):58–77.

[37] Peng, Y., Taylor, J., and Yu, B. (2007). A marginal regression model for multivariate failure time data with a surviving fraction. *Lifetime Data Analysis*, 13:351–369.

[38] Pigott, T. (2001). A review of methods for missing data. *Educational Research and Evaluation*, 7(4):353–383.

[39] Prentice, R., Kalbfleisch, J., Peterson, A., Fluornoy, N., Farewell, V., and Breslow, N. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–554.

[40] Ranganathan, P. and Pramesh, C. S. (2012). Censoring in survival analysis: potential for bias. *Perspective in Clinical Research*, 3(1):40.

[41] Rubin, D. (1987). *Multiple Imputation for Non Response In Surveys*. John Wiley, New York.

[42] Sinharay, S., Stern, H., and Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, 6(4):317–329.

[43] WHO (2017). Global health observatory data-life expectancy. Technical report.

[44] Wienke, A. (2011). *Frailty Models in Survival Analysis*. CRC Press, Boca Raton, FL.

[45] Xiao, Y. and Abrahamowicz, M. (2010). Bootstrap-based methods for estimating standard errors in cox's regression analyses of clustered event times. *Statistics in Medicine*, 29(7-8):915–923.

[46] Zhou, B., Fine, J., Latouche, A., and Labopin, M. (2012). Competing risks regression for clustered data. *Biostatistics*, 13(3):371–383.

Appendix A

Appendix: Simulation Results based on 500 and 1000 families.

Table A.1: Mean bias and median bias of the frailty parameter, $\log(k)$ from various assume model (Model1, Model 2, and model 4) for family data simulated in the presence of competing risks under different familial correlations (k=1,2,5,10)

	True	Model 1		Model 2		Model 4	
	Value	Mean bias	Median bias	Mean bias	Median bias	Mean bias	Median bias
n=500							
$\log(k)$	log(1)	-0.97	-0.72	1.26	1.25	1.35	1.18
	log(2)	-0.34	-0.37	1.08	1.05	1.94	1.21
	log(5)	-0.05	-0.13	1.12	0.89	3.53	1.34
	log(10)	-0.06	-0.17	2.57	0.92	4.97	3.45
n=779							
$\log(k)$	0	-1.11	-0.86	1.24	1.24	1.31	1.21
	0.69	-0.35	-0.42	1.07	1.05	1.80	1.18
	1.61	-0.07	-0.12	1.05	0.89	3.49	1.53
	2.30	$-1.96 \cdot 10^{-4}$	-0.23	1.87	0.87	4.66	3.67
n=1000							
$\log(k)$	0	-1.00	-0.81	1.25	1.25	1.23	1.19
	0.69	-0.28	-0.37	1.07	1.06	1.47	1.24
	1.61	-0.03	-0.07	1.06	0.95	3.57	1.62
	2.30	-0.03	-0.22	1.59	0.89	4.81	3.81

Table A.2: Mean Bias and empirical standard error (SE) for log relative risk $\beta's$ and frailty parameter $\log(k)$ estimates from various assumed models (Model1–Model4) for family data simulated in the presence of competing risks under different familial correlations ($k = 1, 2, 5, 10$); for each assumed k, $\mathbf{n} = \mathbf{500}$ families were simulated.

		True	Model 1		Model 2		Model 3		Model 4	
		Value	Bias	SE	Bias	SE	Bias	SE	Bias	SE
$k=1$	β_{1sex}	0.41	0.07	0.15	0.05	0.11	0.05	0.11	0.03	0.14
	β_{1gen}	2.86	0.12	0.40	−0.09	0.21	−0.09	0.22	−0.01	0.22
	β_{2sex}	−0.72	0.07	0.15	0.05	0.13	0.05	0.13	–	–
	β_{2gen}	1.28	0.04	0.17	−0.12	0.13	−0.12	0.13	–	–
	$\log(k)$	0	−0.97	1.55	1.26	0.16	–	–	1.35	0.97
$k=2$	β_{1sex}	0.41	0.07	0.13	0.06	0.12	0.06	0.12	0.04	0.14
	β_{1gen}	2.86	0.12	0.27	0.03	0.22	0.03	0.22	0.09	0.22
	β_{2sex}	−0.72	0.06	0.13	0.05	0.12	0.05	0.12	–	–
	β_{2gen}	1.28	0.08	0.15	−0.03	0.14	−0.03	0.14	–	–
	$\log(k)$	0.69	−0.34	0.68	1.08	0.24	–	–	1.94	2.50
$k=5$	β_{1sex}	0.41	0.05	0.11	0.05	0.11	0.05	0.11	0.03	0.13
	β_{1gen}	2.86	0.09	0.21	0.06	0.21	0.05	0.21	0.11	0.21
	β_{2sex}	−0.72	0.05	0.13	0.05	0.13	0.05	0.13	–	–
	β_{2gen}	1.28	0.08	0.11	0.03	0.11	0.03	0.11	–	–
	$\log(k)$	1.61	−0.05	0.66	1.12	1.23	–	–	3.53	4.20
$k=10$	β_{1sex}	0.41	0.05	0.10	0.05	0.11	0.06	0.11	0.03	0.13
	β_{1gen}	2.86	0.09	0.20	0.08	0.20	0.08	0.20	0.13	0.20
	β_{2sex}	−0.72	0.05	0.12	0.05	0.12	0.05	0.12	–	–
	β_{2gen}	1.28	0.04	0.17	0.12	0.13	0.12	0.13	–	–
	$\log(k)$	2.30	0.06	1.25	2.57	4.41	–	–	4.97	4.57

Table A.3: Mean Bias and empirical standard error (SE) for penetrance estimates by age 70 for mutation carriers specific to gender and competing event from various assumed models (Model1–Model4) for family data simulated in the presence of competing risks under different familial correlations ($k = 1, 2, 5, 10$); for each assumed k, $\mathbf{n} = \mathbf{500}$ families were simulated; $P_1(X)$ represents the gender-specific penetrance estimate by age 70 for the first colorectal cancer with X taking male (M) and female(F) and $P_2(X)$ is the corresponding penetrance estimate for the other LS related cancer.

		True	Model 1		Model 2		Model 3		Model 4	
		Value	Bias	SE	Bias	SE	Bias	SE	Bias	SE
$k=1$	$P_1(M)$	0.402	−0.225	0.096	−0.017	0.024	−0.011	0.027	−0.001	0.037
	$P_1(F)$	0.273	−0.162	0.060	−0.025	0.018	−0.008	0.020	0.014	0.031
	$P_2(M)$	0.115	−0.053	0.035	0.018	0.014	−0.029	0.015	–	–
	$P_2(F)$	0.242	−0.121	0.068	0.022	0.020	−0.041	0.022	–	–
$k=2$	$P_1(M)$	0.446	−0.141	0.064	−0.021	0.024	−0.0003	0.026	0.021	0.036
	$P_1(F)$	0.302	−0.112	0.041	−0.031	0.020	−0.019	0.021	0.032	0.034
	$P_2(M)$	0.129	−0.022	0.024	0.019	0.013	0.027	0.014	–	–
	$P_2(F)$	0.271	−0.060	0.046	0.024	0.021	0.037	0.022	–	–
$k=5$	$P_1(M)$	0.480	−0.072	0.038	−0.021	0.026	−0.011	0.026	0.033	0.038
	$P_1(F)$	0.325	−0.063	0.026	−0.031	0.021	−0.024	0.021	0.049	0.034
	$P_2(M)$	0.140	0.003	0.017	0.020	0.015	0.024	0.016	–	–
	$P_2(F)$	0.293	−0.010	0.030	0.024	0.022	0.032	0.022	–	–
$k=10$	$P_1(M)$	0.493	−0.050	0.033	− 0.020	0.025	−0.015	0.025	0.041	0.032
	$P_1(F)$	0.333	−0.049	0.026	−0.030	0.0216	−0.027	0.022	0.057	0.034
	$P_2(M)$	0.144	0.011	0.016	0.021	0.0155	0.023	0.016	–	–
	$P_2(F)$	0.302	0.006	0.025	0.026	0.0228	0.030	0.023	–	–

Table A.4: Mean bias and empirical standard error (SE) for baseline parameter estimates from various assumed models (Model1–Model4) for family data simulated in the presence of competing risks under different familial correlations ($k = 1, 2, 5, 10$); for each assumed k, $\mathbf{n = 500}$ families were simulated.

		True	Model 1		Model 2		Model 3		Model 4	
		Value	Bias	SE	Bias	SE	Bias	SE	Bias	SE
$k=1$	λ_1	0.0042	−0.0017	0.0009	−0.0006	0.0004	−0.0006	0.0004	−0.0008	0.0004
	ρ_1	2.40	0.06	0.11	−0.09	0.08	−0.09	0.08	−0.07	0.08
	λ_2	0.0092	−0.0035	0.0016	0.0011	0.0005	−0.0011	0.0005	–	–
	ρ_2	2.92	−0.16	0.14	−0.32	0.10	−0.32	0.10	–	–
$k=2$	λ_1	0.0042	−0.0010	0.0006	−0.0006	0.0004	−0.0005	0.0004	−0.0007	0.0004
	ρ_1	2.40	0.06	0.10	−0.04	0.08	−0.04	0.08	−0.03	0.08
	λ_2	0.0092	−0.0020	0.0008	−0.0008	0.0006	−0.0008	0.0006	–	–
	ρ_2	2.92	−0.15	0.12	−0.25	0.11	−0.25	0.10	–	–
$k=5$	λ_1	0.0042	−0.0004	0.0004	−0.0003	0.0004	−0.0003	0.0004	−0.0004	0.0004
	ρ_1	2.40	0.06	0.08	0.02	0.08	0.02	0.08	0.01	0.08
	λ_2	0.0092	−0.0010	0.0006	−0.0006	0.0006	−0.0006	0.0006	–	–
	ρ_2	2.92	−0.15	0.08	−0.17	0.08	−0.17	0.08	–	–
$k=10$	λ_1	0.0042	−0.0003	0.0004	−0.0003	0.0004	−0.0003	0.0004	−0.0003	0.0004
	ρ_1	2.40	0.06	0.08	0.03	0.08	0.03	0.08	0.03	0.13
	λ_2	0.0092	−0.0006	0.0006	−0.0004	0.0005	−0.0004	0.0005	–	–
	ρ_2	2.92	−0.14	0.11	−0.17	0.12	−0.17	0.12	–	–

Table A.5: Mean bias and empirical standard error (SE) for log relative risk $\beta's$ and frailty parameter $\log(k)$ estimates from various assumed models (Model1–Model4) for family data simulated in the presence of competing risks under different familial correlations ($k = 1, 2, 5, 10$); for each assumed k, $\mathbf{n} = \mathbf{1000}$ families were simulated.

		True	Model 1		Model 2		Model 3		Model 4	
		Value	Bias	SE	Bias	SE	Bias	SE	Bias	SE
$k=1$	β_{1sex}	0.41	0.07	0.13	0.05	0.08	0.05	0.08	0.04	0.10
	β_{1gen}	2.86	0.10	0.35	0.08	0.15	0.09	0.15	−0.01	0.16
	β_{2sex}	−0.72	0.08	0.13	0.05	0.08	0.04	0.08	–	–
	β_{2gen}	1.28	0.05	0.14	−0.11	0.09	−0.11	0.09	–	–
	$\log(k)$	0	−1.00	1.75	1.25	0.11	–	–	1.23	0.38
$k=2$	β_{1sex}	0.41	0.07	0.09	0.06	0.08	0.06	0.08	0.04	0.09
	β_{1gen}	2.86	0.11	0.21	0.003	0.15	0.04	0.15	0.07	0.15
	β_{2sex}	−0.72	0.07	0.09	0.05	0.09	0.05	0.09	–	–
	β_{2gen}	1.28	0.08	0.11	−0.03	0.10	−0.03	0.10	–	–
	$\log(k)$	0.69	−0.28	0.56	1.08	0.16	–	–	2.47	1.30
$k=5$	β_{1sex}	0.41	0.06	0.08	0.06	0.08	0.06	0.08	0.04	0.09
	β_{1gen}	2.86	0.08	0.15	0.06	0.15	0.06	0.15	0.11	0.15
	β_{2sex}	−0.72	0.06	0.09	0.06	0.09	0.06	0.09	–	–
	β_{2gen}	1.28	0.08	0.10	0.03	0.10	0.03	0.10	–	–
	$\log(k)$	1.61	−0.03	0.57	1.06	0.93	–	–	3.57	3.75
$k=10$	β_{1sex}	0.41	0.06	0.08	0.06	0.08	0.06	0.08	0.03	0.09
	β_{1gen}	2.86	0.08	0.14	0.06	0.14	0.06	0.14	0.11	0.14
	β_{2sex}	−0.72	0.05	0.09	0.05	0.09	0.05	0.09	–	–
	β_{2gen}	1.28	0.08	0.10	0.05	0.10	0.05	0.10	–	–
	$\log(k)$	2.30	−0.03	0.95	1.59	2.81	–	–	4.81	4.18

Table A.6: Mean bias and empirical standard error (SE) for penetrance estimates by age 70 for mutation carriers specific to gender and competing event from various assumed models (Model1–Model4) for family data simulated in the presence of competing risks under different familial correlations ($k = 1, 2, 5, 10$); for each assumed k, $\mathbf{n} = \mathbf{1000}$ families were simulated; $P_1(X)$ represents the gender-specific penetrance estimate by age 70 for the first colorectal cancer with X taking male (M) and female(F) and $P_2(X)$ is the corresponding penetrance estimate for the other LS related cancer.

		True	Model 1		Model 2		Model 3		Model 4	
		Value	Bias	SE	Bias	SE	Bias	SE	Bias	SE
$k=1$	$P_1(M)$	0.402	−0.228	0.095	−0.016	0.017	0.012	0.019	0.0001	0.026
	$P_1(F)$	0.273	−0.162	0.062	−0.026	0.014	−0.009	0.015	−0.012	0.024
	$P_2(M)$	0.115	−0.053	0.036	0.017	0.009	0.028	0.010	–	–
	$P_2(F)$	0.242	−0.122	0.066	0.021	0.014	0.041	0.015	–	–
$k=2$	$P_1(M)$	0.446	−0.139	0.053	−0.019	0.017	0.001	0.018	0.022	0.028
	$P_1(F)$	0.302	−0.108	0.033	−0.030	0.014	−0.017	0.014	0.034	0.025
	$P_2(M)$	0.129	−0.021	0.019	0.019	0.009	0.027	0.011	–	–
	$P_2(F)$	0.271	−0.059	0.036	0.023	0.015	0.037	0.016	–	–
$k=5$	$P_1(M)$	0.480	−0.067	0.027	−0.021	0.019	−0.010	0.019	0.038	0.026
	$P_1(F)$	0.325	−0.062	0.020	−0.031	0.015	−0.025	0.015	0.051	0.024
	$P_2(M)$	0.140	0.005	0.012	0.021	0.011	0.024	0.011	–	–
	$P_2(F)$	0.293	−0.007	0.020	0.025	0.016	0.032	0.016	–	–
$k=10$	$P_1(M)$	0.493	−0.050	0.023	−0.021	0.017	−0.015	0.017	0.040	0.023
	$P_1(F)$	0.333	−0.050	0.017	−0.032	0.015	−0.028	0.015	0.055	0.024
	$P_2(M)$	0.144	0.011	0.012	0.020	0.011	0.023	0.011	–	–
	$P_2(F)$	0.302	0.005	0.019	0.025	0.016	0.029	0.017	–	–

Table A.7: Mean bias and empirical standard error (SE) for baseline parameter estimates from various assumed models (Model1–Model4) for family data simulated in the presence of competing risks under different familial correlations ($k = 1, 2, 5, 10$); for each assumed k, $\mathbf{n} = \mathbf{1000}$ families were simulated.

		True	Model 1		Model 2		Model 3		Model 4	
		Value	Bias	SE	Bias	SE	Bias	SE	Bias	SE
$k=1$	λ_1	0.0042	−0.0017	0.0009	−0.0006	0.0003	−0.0006	0.0003	−0.0008	0.0003
	ρ_1	2.40	0.06	0.10	−0.09	0.05	−0.09	0.05	−0.07	0.06
	λ_2	0.0092	−0.0035	0.0016	−0.0011	0.0004	−0.0011	0.0004	–	–
	ρ_2	2.92	−0.16	0.13	−0.32	0.08	−0.32	0.08	–	–
$k=2$	λ_1	0.0042	−0.0009	0.0004	−0.0005	0.0003	−0.0005	0.0003	−0.0006	0.0003
	ρ_1	2.40	0.06	0.07	−0.04	0.06	−0.04	0.06	−0.04	0.06
	λ_2	0.0092	−0.0019	0.0006	−0.0008	0.0004	−0.0008	0.0004	–	–
	ρ_2	2.92	−0.15	0.10	−0.25	0.07	−0.25	0.07	–	–
$k=5$	λ_1	0.0042	−0.0005	0.0003	−0.0003	0.0003	−0.0003	0.0003	−0.0004	0.0003
	ρ_1	2.40	0.05	0.06	0.005	0.06	0.005	0.06	−0.006	0.06
	λ_2	0.0092	−0.0009	0.0004	−0.0005	0.0004	−0.0005	0.0004	–	–
	ρ_2	2.92	−0.15	0.08	−0.20	0.08	−0.20	0.08	–	–
$k=10$	λ_1	0.0042	−0.0003	0.0003	−0.0003	0.0003	−0.0003	0.0003	−0.0003	0.0003
	ρ_1	2.40	0.05	0.06	0.03	0.06	0.03	0.06	0.01	0.06
	λ_2	0.0092	−0.0006	0.0004	−0.0005	0.0004	−0.0005	0.0004	–	–
	ρ_2	2.92	−0.15	0.08	−0.18	0.08	−0.18	0.08	–	–

Curriculum Vitae

Name:	Daniel Prawira
Post-Secondary Education and Degrees:	University of Toronto Scarborough Honours. Bachelors of Science 2010 - 2014 B.Sc University of Western Ontario London, ON 2014 - 2017 M.Sc.
Honours and Awards:	Schulich Graduate Scholarship 2014-2017
Related Work Experience:	Teaching Assistant for Multivariable Biostatistics The University of Western Ontario 2016-2016

Publications:

Prawira, D. (2015). Towards the Philosophical Debate of Frequentist and Bayesian Approaches in Competing Risk Survival Analysis. Western Journal of Graduate Research, 1:5-8.

Presentation:

Prawira, D. (2015). Towards the Philosophical Debate of Frequentist and Bayesian Approaches in Competing Risk Survival Analysis. Western Research Forum.