MODELING SEQUENTIAL EVENT TIMES USING FAMILY DATA

(Spine title: Modeling sequential event times using family data)

(Thesis format: Monograph)

by

Balakumar Swaminathan, M.Sc.
Graduate Program in Epidemiology & Biostatistics

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

THE UNIVERSITY OF WESTERN ONTARIO

SCHOOL OF GRADUATE AND POSTDOCTORAL STUDIES

CERTIFICATE OF EXAMINATION

Supervisor

Dr. Yun-Hee Choi

Supervisory Committee

Dr. Neil Klar

Dr. GuangYong Zou

Examiners

Dr. Wenqing He

Dr. John Koval

Dr. Serge Provost

The thesis by

Balakumar Swaminathan

entitled

Modeling sequential event times using family data

is accepted in partial fulfillment of the requirements for the degree of Master of Science

Date: _____

_____

Chair of the Thesis Examination Board

Abstract

In genetic epidemiology, families harboring certain genetic mutations are predisposed to successive cancers in their lifetime. This thesis aims to provide reliable estimates of relative risk and age-dependent cumulative risks (penetrance) associated with the mutated gene for successive cancers. We develop a statistical framework for modeling sequential event times arising from family data. A shared frailty model is employed to incorporate the dependence between the two event times. Because families are ascertained through non-random sampling, an ascertainment-corrected retrospective likelihood approach is proposed to account for the non-ignorable sampling design. Simulation studies demonstrate that our proposed method provides unbiased and reliable estimates of disease risks associated with a mutated gene. The frailty approach is also compared to an independent model that ignores the dependence between the events. Finally, we illustrate our approach using 12 Lynch syndrome families and provide penetrance estimates for developing first and second colorectal cancer.

Keywords: Sequential event times, shared frailty model, penetrance, ascertainment, retrospective likelihood, Lynch syndrome, three-state progressive model.

Acknowledgments

This thesis would not have been possible without the guidance and dedication of my supervisor, Dr. Yun-Hee Choi. I am highly grateful for her patience and steadfast motivation which helped me to complete the thesis on time.

I am indebted to my thesis supervisory committee members, Dr. Neil Klar and Dr. GuangYong Zou, for their timely and thoughtful suggestions. I would like to thank the Department of Epidemiology and Biostatistics at The University of Western Ontario for providing ample facilities and financial assistance throughout my degree. Special thanks to my friend, Michael Lebenbaum, for proofreading my thesis.

Sincere thanks to Veena Vincent for her unfailing support and constant encouragement. I would also like to thank my parents for their prayers, which helped me hurdle through difficult times during my research and academic work. Last but not least, I am always grateful to the Almighty for answering all of my prayers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Lynch syndrome, popularly known as hereditary non-polyposis colorectal cancer (HNPCC), refers to a genetic disorder due to a mutation in the DNA mismatch repair gene, which is inherited from parents. Individuals with this syndrome are at high risk of early-onset colorectal cancer (CRC) compared to the general population (Lin et al., 1998). Mutation carriers are also prone to develop other cancers, such as endometrial, ovarian, and stomach cancers, as a consequence of the primary cancer (Lynch et al., 1977). As it is a heritable disease, the families are expected to contain multiple mutation carriers and require constant surveillance to prevent adverse outcomes. For instance, of the estimated 21,500 new colorectal cancer cases in Canada in 2008, cancers due to mutation accounted for 1% - 5% (Lynch et al., 2009; Lynch and Smyrk, 1996). In most genetic disorders, as in the case of Lynch syndrome, the age-of-onset of disease is highly variable and successive events are often encountered; therefore, reliable estimation of age-specific risk (penetrance) for successive cancers and relative risk associated with a mutated gene is essential to decide the appropriate prevention strategy.

The main goals of this thesis are to develop a general statistical framework for modeling successive event times based on family-based studies and to provide reliable estimates of relative and absolute risks of developing a first and second cancer associated with a mutated gene. We also aim to model the dependency between the sequential events and quantify their association. The families selected in our analysis

have been sampled in a non-random manner, so that an appropriate ascertainment correction is necessary to make population-based inference.

The rest of this chapter is structured as follows: Section 1.1 provides an introduction to the family-based design and its types and Section 1.2 describes some of the commonly used ascertainment correction approaches for non-random sampling of families. Section 1.3 outlines the three-state progressive approach to model two successive events and Section 1.4 proposes a shared frailty approach for modeling the dependence between the two event times. The scope and objectives of the thesis are provided in Sections 1.5 and 1.6, respectively. The chapter concludes with the thesis outline.

1.1 Family studies

In the study of rare genetic disorders, a simple random sample of unrelated individuals from a general population may not yield sufficient numbers of mutation carriers or disease cases and it may result in limited power to study the genetic association. A statistically powerful alternative to overcome this limitation is the use of a family-based study (Thomas, 1999; Laird and Lange, 2006). In genetic epidemiology, the role of certain genetic mutations predisposing individuals to complex diseases has been well documented with the help of family studies.

In the family-based study design, families are sampled through individuals who are affected with a particular disease. An affected individual who brings their family into the study is called a "proband". The rationale for including the proband's family is that, if the proband is affected by a genetic disorder, then the disease causing gene is more likely to be segregated within the family, whereby such a family would be more susceptible to the genetic disorder. Family studies are largely used in genetic epidemiology as family members are considered to have a common genetic background and are exposed to similar environmental factors, so that they tend to form a more

homogeneous group. It avoids population stratification issues and obviates the difficulties that may arise in finding a matched control. In what follows, we discuss some of the commonly used family-based study designs.

1.1.1 Case-control design

In a traditional case-control study, one randomly samples disease cases and healthy controls from a well defined population and compares the distribution of an exposure between these two groups. To yield a valid inference, cases and controls must be drawn from the same population as it is expected that they share similar characteristics. However, in genetics studies, it may be difficult or even impossible to obtain controls having similar genetic patterns as the cases. Instead of sampling unrelated controls, the family-based case-control design (Gauderman et al., 1999) samples the relatives of the cases as the controls. The controls can be the siblings, cousins, or parents of the affected case. This design offers the advantage of being robust to population stratification by sampling relatives who share similar genetic and environmental factors. Kraft and Thomas (2000) used a sibship-based case-control design for a binary outcome to estimate the disease risk. Hopper et al. (2005) extended this sibship-based design to the case-control-family design where the case families were compared to the families of random controls sampled from the same population. This study design is simple to implement and the relative risks of an exposure on an outcome can be easily obtained. However, direct estimation of absolute risk of a disease may not be possible as the number of cases is fixed in the sample and may differ from that of the general population.

1.1.2 Cohort design

When it is of interest to study multiple end-points, the case-control study design may not be appropriate. To overcome this limitation, a cohort study may be used where a group of individuals who possess a particular risk factor, such as a gene mutation,

and another group free of the risk factor are followed for a well-defined time period to collect information on disease outcomes. To avoid long follow-up times, a retrospective cohort study can be used. In the study of rare genetic mutations, a special type of retrospective cohort design, called the kin-cohort design, has been employed by Struewing et al. (1997) and Wacholder et al. (1998) to estimate the penetrance of breast cancer; penetrance is the age-specific cumulative risk of developing a disease given a person's carrier status of the mutated gene of interest (Thomas, 2004). In this design, only the probands are genotyped and their relatives are not. Struewing et al. (1997) recruited volunteers among Ashkenazi Jews in Washington to undergo genotyping for BRCA1 and BRCA2 genes and collected the history of breast cancer in their first-degree relatives. The target population was chosen because the prevalence of certain genetic mutations is elevated in Ashkenazi Jews. The cumulative risk of breast cancer by the age of 70 years was calculated by comparing the proportions of affected relatives between mutation carriers and non-carriers. Gail et al. (1999) extended the kin-cohort design by genotyping the relatives of the probands. Although it is relatively easy to implement with smaller sample sizes, the kin-cohort design suffers from bias if the proband's intention to participate is influenced by familial history of disease and if there is an under-reporting of disease history among relatives.

1.1.3 Multistage design

Multistage sampling design (White, 1982b; Whittemore and Halpern, 1997), as the name suggests, requires successive subsampling through several stages. It proceeds in a sequential manner until a feasible sample size is reached. For instance, consider a two stage sampling design where, in the first stage, individuals are randomly sampled from a population and are stratified according to some variable; in the second stage, a subsample of individuals are selected from each strata using a pre-determined proportion. The sampling proportion at each stage would depend on the information available in the previous stage. In the presence of limited resources, such a design

can be applied to a case-control or a cohort study to make them more efficient by sampling informative units that are at high risk. Multistage designs are popular in genetics studies as they reduce costs by genotyping fewer yet informative individuals. Siegmund et al. (1999) discussed a four-stage sampling approach to sample cases and controls using prostate cancer data and estimated the parameters using the Horvitz-Thompson estimator by adjusting for the sampling weights at each stage. By using a composite likelihood approach, Choi and Briollais (2011) modeled family data obtained from a two-stage sampling design and estimated the genetic relative risk of a mutated gene in the presence of missing genotypic information of family members.

1.1.4 Population-based and clinic-based study designs

Gong and Whittemore (2003) discussed two types of family-based study designs − population-based and clinic-based. In the population-based design, families are ascertained by randomly sampling probands from a population-based disease registry. The probands do not need to be a carrier of the mutated gene; however, the efficiency of selecting mutation segregating families increases if carrier probands are selected. Then, the relatives of the proband are screened for the mutated gene and interviewed for disease history. This design derived its name from the fact that the probands are randomly sampled from the diseased population. By genotyping only the case probands and examining their relatives, the population-based design can apparently be viewed as an extension of the kin-cohort design. This design with carrier probands has the merit of providing the most efficient estimates of the penetrance function (Choi et al., 2008).

In the clinic-based design, families are eligible for study if they satisfy an eligibility criterion of having multiple diseased individuals, in addition to an affected proband, within the family. This design is called 'clinic-based' as the high-risk families are mostly identified through counseling clinics due to an unusual number of affected individuals in the family. Sampling such high-risk families is largely used in

gene-characterizing studies to identify the influence of a mutated gene on a particular disease (Easton et al., 1995). The criteria required to sample such families are called ascertainment schemes. Well known ascertainment schemes to identify mutation carrying families of colorectal cancer are: Amsterdam criteria (Vasen et al., 1999) and Bethesda criteria (Umar et al., 2004). It is clear that clinic-based design tends to sample more diseased individuals than other population-based family designs and requires an ascertainment correction to make population-based estimates of disease risks. When the ascertainment correction is properly implemented, this design has the advantage of yielding unbiased and efficient estimates of both relative and absolute risk of the disease (Choi et al., 2008). However, a limitation is that complex ascertainment schemes are difficult to model and may require a method that implicitly corrects for the ascertainment, such as the retrospective likelihood (Carayol and Bonaïti-Pellié, 2004). The following section discusses several ascertainment correction approaches.

1.2 Ascertainment correction

The families for genetics studies are predominantly ascertained (sampled) in a non-random manner and this necessitates a correction during analysis to facilitate a valid inference about the parameters of interest (Le Bihan et al., 1995). The correction for ascertainment depends on the nature of the ascertainment scheme employed. There are two most commonly used ascertainment schemes − single ascertainment and complete ascertainment (Ewens and Elston, 2012). Under the single ascertainment, the probability that a family is ascertained is proportional to the number of affected individuals in the family. Under the complete ascertainment scheme, the ascertainment probability of a family is independent of the total number of affected individuals. Apart from these two types, researchers have used other attractive sampling procedures that are effective in identifying mutation segregating families. For instance,

when the age-of-onset is highly variable (as in the case of genetic mutations), it is more appropriate to include an age criterion to ascertain probands as well as their relatives (Le Bihan et al., 1995; Carayol and Bonaïti-Pellié, 2004). Similarly, when the gene mutation is rare, it is desirable to recruit multiple affected family members in an ad hoc fashion to identify mutation carriers (Choi et al., 2008).

The concept of ascertainment correction is almost a century old and there is extensive literature discussing this issue. Weinberg (1912) proposed a simple method to correct for ascertainment by excluding the proband information from the analysis with the rationale that the proband is responsible for the inclusion of his/her family. Recently, Alarcon et al. (2008) extended this idea to time-to-event data arising from families with at least one affected individual. Using the proband's phenotype exclusion likelihood, the age-specific penetrance function was estimated from the Weibull model. They also proposed a non-parametric approach called the "Index Discarding EuclideAn Likelihood (IDEAL)" as a validating tool to check departures from the assumed parametric baseline distribution and provided unbiased estimates of penetrance along with their confidence bands (Alarcon et al., 2009).

Most of the ascertainment correction procedures discussed in the literature are likelihood-based approaches pioneered by Fisher (1934). He corrected for single ascertainment by providing weights to the sampled families based on the number of affected individuals in the family. Following Fisher, several authors (for example, Elston and Sobel, 1979; Bonney, 1998; Clayton, 2003) extended this approach to various sampling schemes. The likelihood-based approach is an attractive choice to correct for ascertainment as it is capable of incorporating the baseline hazard distribution and also adjusting for known risk factors.

Kraft and Thomas (2000) considered three likelihood approaches − prospective, retrospective, and joint likelihoods - and compared their efficiencies in the estimation of genetic relative risks for a binary outcome. They used a case-control design with controls as siblings and assumed conditional independence given their genotypes. The

prospective likelihood models the conditional probability of phenotypes of all sampled members on their genotypes and ascertainment process, i.e. $P(D|G, A)$, where D and G are the vectors of phenotypes and genotypes of sampled individuals, respectively, and A is the ascertainment process. Difficulty in modeling arises in the presence of a complex or unknown ascertainment scheme. On the other hand, the retrospective likelihood corrects for ascertainment (when the ascertainment is based on disease status only) by modeling the genotypes of individuals given their phenotypes, i.e. $P(G|D)$. By conditioning on all observed phenotypes, the retrospective likelihood implicitly corrects for ascertainment and hence is robust to ascertainment bias. Nevertheless, there is a loss in statistical efficiency by conditioning on all sampled individuals instead of conditioning only on those responsible for the family's ascertainment (Kraft and Thomas, 2000). In order to overcome this limitation, Carayol and Bonaïti-Pellié (2004) proposed a modified retrospective likelihood approach in which the conditioning was based only on those individuals who were involved in the ascertainment process. Lastly, the joint likelihood models the joint probability of phenotype and genotype of the individuals given the ascertainment scheme, $P(D, G|A)$. It has the weakest condition compared to the other two likelihood approaches and hence is capable of providing the most efficient risk estimates among these three likelihood-based approaches. Nevertheless, similar to the prospective likelihood, the joint likelihood also requires the modeling of the ascertainment process. Choi et al. (2008) broadened the scope of the above mentioned likelihood approaches to time-to-event data and evaluated their efficiencies for population- and clinic-based study designs.

For large pedigrees, Chatterjee and Wacholder (2001) proposed a composite likelihood approach to circumvent the problem of summing over all possible genotypes of all family members. In this approach, a pedigree consisting of a proband and his/her M relatives was broken down into M pairs of the proband and one of the sampled family members. Assuming independence between the pairs within a pedigree, the likelihood was constructed as the product of all possible pairs of the family.

We employ the retrospective likelihood, which is also known as 'ascertainment-free-assumption', to model the ascertainment correction in this thesis. This approach, although statistically less efficient than others, can model complex or unknown ascertainment schemes and also produce reliable estimates of relative risk of the disease gene (Choi et al., 2008).

1.3 Modeling sequential events

In the follow-up of complex diseases, several recurrent events can occur as a consequence of a primary event; for example, patients who had a heart attack are often at risk of subsequent attacks. Analyzing such successive events in their order of occurrence can provide valuable insight into the overall disease progression and severity. The multi-state model (Putter et al., 2007) is a familiar way to model a stochastic process that progresses through a set of distinct states over time. These states are determined by the conditions experienced by an individual over the progression of the disease. A transition occurs when there is a movement between states. There are two types of states − absorbing and transient states. Absorbing states are states beyond which there exists no transition and transient states are the possible intermediate events, experienced prior to attaining an absorbing state. Usually, in a multi-state model, the states are graphically represented by boxes and the possible transition(s) are indicated by arrows pointing at the next possible event.

The simplest form of a multi-state model is the two-state model, popularly known as the mortality model (see Figure 1.1). In the mortality model, there exist two states − 'Alive' and 'Dead' - with one possible transition from the former to the latter. For example, in the context of human immunodeficiency virus (HIV) infection, the 'Alive' state corresponds to the event of contracting the HIV virus and the 'Dead' state corresponds to the event of death with no chance of recovery. The time taken to transit from the 'Alive' state to the 'Dead' state can be modeled using survival

analysis techniques.

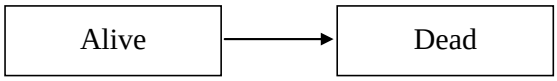

Figure 1.1: Mortality model

We can extend this simple two-state model to accommodate three states with a possible transition from one state to another in a progressive manner with no possibility of returning to the previous state(s). Such a model is called a three-state progressive model (Joly and Commenges, 1999). Figure 1.2 depicts a three-state progressive model with the states – 'Healthy', 'Event 1', and 'Event 2'. These models are popular in the context of cancer and HIV studies due to their ability to model the effect of an intermediate state on an absorbing state. Meira-Machado et al. (2009) studied the impact of tumor recurrence on death due to breast cancer using a three-state progressive model with the states 'Alive and Disease Free', 'Alive With Recurrence', and 'Death' and adjusted for known prognostic factors in their analysis. Frydman (1992, 1995) used this method to study disease progression in the treatment of AIDS using information from hemophilia patients. She constructed a progressive three state model where an individual initially starts as HIV negative from which he/she may transit into the HIV positive state and potentially progress to clinically proven AIDS.

For modeling the times taken to transit from one state to another, Putter et al. (2007) described two distinct ways: 'Clock Forward' and 'Clock Reset' approaches. In the 'Clock Forward' approach, the event times are measured continually from the start of the initial state until the occurrence of the absorbing state. On the other hand, in the 'Clock Reset' approach, the event times are reset to zero after the occurrence of

Figure 1.2: Three-state progressive model

an intermediate event. The second approach is highly useful in the study of sequential events to better understand the role of intermediate state(s) on the absorbing state. Therefore, we consider the 'Clock Reset' approach where we define T_1 as the time spent in the 'healthy' state prior to attaining 'event 1' and define T_2 as the time spent in the state 'event 1' prior to attaining 'event 2'. The event times T_1 and T_2 can also be viewed as the 'gap times' (Cook and Lawless, 2010). Thus, we define our sequential event times T_1 and T_2 as the time to a first event and the time to a second event since the first event, respectively.

An important assumption in the multi-state model is the Markov property, which states that the occurrence of a future event is only dependent on the current event and independent of the past event(s). This assumption could fail when the event times are dependent. In such a scenario, we can relax the Markov assumption and use a Semi-Markov Model, allowing more flexibility compared to the Markov Model (Janssen and Limnios, 1999).

1.4 Frailty models

In medical research, a primary study outcome is the time to an event of interest, where the event can be death, occurrence of disease, such as cancer, or a transition from one state to another like remission to relapse. Survival analysis models time-to-event data accounting for censored observations (Fleming and Lin, 2000). Censoring arises if the event time of an individual is unknown or missing and it is only known

that he/she has survived until a specific time point. This is called right censoring. It is predominantly assumed in survival analysis that the censoring time is independent of the failure time and is called non-informative (random) censoring. It is worth mentioning that for sequential events, the independent censoring assumption no longer holds in the presence of a dependence between the event times. For example, consider two sequential event times, T_1 and T_2, where T_1 is the time to a first event and T_2 is the time to a second event since the first event. If the second event is censored by the time a, then the censoring time for T_2 is given by $a - T_1$. Therefore, when T_1 and T_2 are dependent, T_2 is also dependent on the censoring time of T_2, so that, we cannot assume the independent censoring for T_2. In this section, we discuss the shared frailty model that can effectively model bivariate event times in the presence of dependence. We begin by introducing the univariate frailty model followed by the shared frailty model.

Vaupel et al. (1979) first introduced the term 'frailty' in demographics to model heterogeneity among individuals in population mortality data and illustrated that in the presence of individual heterogeneity, the population mortality rates were largely underestimated. Successively, several authors studied the impact of population heterogeneity using frailty models (Vaupel and Yashin, 1985; Aalen, 1994; Aalen and Tretli, 1999). In the univariate frailty model, a non-negative random effect (frailty) variable, Z, is introduced into the proportional hazards (PH) model such that it acts multiplicatively on the baseline hazard function. The frailty is used in the model to account for the unobserved heterogeneity in the population. It is assumed to be constant over time and vary across individuals in the population. The univariate frailty model can be viewed as the survival data analogue of the random effects model.

In addition to their ability to model unobserved heterogeneity in univariate time-to-event data, the frailty models can also be extended to model the association between event times in multivariate survival data. Such data are commonly encountered in the study of recurrent events, such as cancer, or in the study of related individuals

like twin studies, where the correlation among the event times cannot be ignored in the analysis. The shared frailty model provides an efficient way to model this correlation by introducing a non-negative frailty variable, Z, in the PH model. The introduced frailty is considered to be shared among the members within a cluster to induce the dependence among them. Here, the cluster can be a group of related individuals or multiple observations from an individual at different time points. Conditional on the frailty, the event times within a cluster are assumed to be independent. Thus, the conditional joint survival distribution of the event times within a cluster given the frailty can be simply written as the product of their individual conditional survival functions. Then, the marginal (unconditional) joint survival function can be derived by integrating the conditional joint distribution over the frailty distribution.

Some popular choices for the frailty distribution include: gamma (Vaupel et al., 1979), compound poisson (Aalen and Tretli, 1999), inverse gaussian (Hougaard, 1984), log normal (Vaupel and Yashin, 1983), and positive stable (Hougaard, 1986) distributions. The gamma distribution is generally employed to model frailties due to its relatively simple expression of the Laplace transform, required to construct the unconditional joint distribution. Further, the variance of the frailty distribution serves as a measure of dependence between the event times in the shared frailty model, while it serves as a measure of heterogeneity in the population for the univariate frailty model. The smaller the variance, the smaller the dependence or heterogeneity and vice versa. Shared frailty modeling approaches were applied to model familial correlation in penetrance estimation using case-control family design (Chen et al., 2009) and population-based family design (Choi, 2012). To overcome the distributional assumptions, Horowitz (1999) proposed a semi-parametric approach to model the frailty distribution and the baseline hazard function in an univariate PH model.

Alternatively, one can model the dependency between the multivariate event times using copula models (Nelsen, 2010). In the copula model, the joint distribution function of the event times is generated as a function of the marginal distributions. The

function used to join the marginals is called a copula, and it determines the dependence structure between the event times. The marginals can be modeled using the PH model either by assuming a parametric or a nonparametric form. For multivariate event time data, He and Lawless (2003) applied both frailty-based and copula-based approaches to model the dependency among the event times along with piecewise-constant or spline-approximated baseline hazard functions. Their approach handled interval censoring and sequential event times. To overcome the limitations of parametric baseline hazard assumptions, Lawless and Yilmaz (2011) provided a semi-parametric approach for modeling sequential event times using copula models. Although the resulting joint survival distributions from the copula model and the shared frailty model are equivalent for certain choices of baseline and frailty distributions, the derived marginal survival functions using these two methods are unique and this causes a difference in the interpretation of dependence measures between these two methods (Goethals et al., 2008).

1.5 Scope of the thesis

In this thesis, we propose a bivariate modeling approach for two sequential event times based on a shared frailty model. We estimate the disease risks (absolute and relative) associated with a mutated gene using family data that arise from two types of family designs − population-based and clinic-based. The families are selected through affected probands following which their relatives' disease history and mutation status are collected retrospectively. We assume the genotype and phenotype information of all recruited families to be fully observed.

We restrict our attention to right censoring for the two sequential event times, where the second event time is subject to informative censoring. We adopt parametric models for the baseline hazard function and frailty. The frailties, that are used to model the dependence between two successive events experienced by an individual

are assumed to be time invariant and vary across individuals. Lastly, we assume conditional independence between family members given their genotypes and covariates (Choi et al., 2008; Kraft and Thomas, 2000; Le Bihan et al., 1995). Our work would serve as an extension of the existing approach for univariate time-to-event data arising from family-based designs (Choi et al., 2008).

1.6 Objectives of the thesis

The objectives of the thesis are:

1. Develop a statistical framework to model two sequential survival times arising from family-based study designs (population- and clinic-based) by
   (a) accounting for the dependence between the event times using a shared frailty model and
   (b) incorporating the necessary ascertainment correction using the retrospective likelihood approach;
2. Establish the robust variance estimators of the age-specific penetrance function and relative risk of a mutated gene for the sequential events using our proposed method;
3. Assess the performance of our frailty method in terms of accuracy and precision in a large sample setting using simulation studies;
4. Illustrate our approach using real data from Newfoundland consisting of 12 large high-risk Lynch syndrome families.

Objectives 1 and 2 are covered in Chapter 2 and objectives 3 and 4 are addressed in Chapters 3 and 4, respectively.

1.7 Outline of the thesis

The remainder of the thesis is structured as follows: Chapter 2 establishes the statistical framework for modeling sequential survival times using a shared frailty model to estimate the age-specific penetrance function and genetic relative risks using family data. Chapter 3 presents the simulation studies used to evaluate our proposed model and Chapter 4 provides an application of our approach to Lynch syndrome families from Newfoundland. Finally, Chapter 5 concludes the thesis by summarizing the results and discussing possible future research work.

Chapter 2

ASCERTAINMENT CORRECTED LIKELIHOOD FOR SEQUENTIAL EVENTS

This chapter provides a general framework for modeling sequential event times associated with a mutated disease-causing gene based on the data arising from a family-based design. In genetic epidemiology, when a disease causing gene has been identified, one is interested in estimating the relative and absolute risks of developing diseases associated with the disease gene (Choi et al., 2008; Le Bihan et al., 1995). Certain complex diseases have the hallmark of several recurrent events following a primary event. For instance, patients treated for breast cancer are predisposed to local and/or distant recurrence of cancer.

In this thesis, we consider family data obtained using two types of study designs − population-based design and clinic-based (Gong and Whittemore, 2003). The population-based design samples families through a single affected proband and the clinic-based design samples high risk families with multiple affected individuals. Since the families are sampled in a non-random manner, an ascertainment correction is required to obtain population-based inference.

We first develop a bivariate distribution based on a shared frailty model (Wienke, 2009) for the dependent sequential event times, then we employ the retrospective likelihood approach (Carayol and Bonaïti-Pellié, 2004; Kraft and Thomas, 2000) to correct for the complex ascertainment procedure involved in obtaining the family data.

The chapter begins by defining two types of disease risks associated with a mutated gene that we aim to estimate via our modeling of sequential event times. Section 2.2 provides an introduction to the univariate frailty model and it is later extended to the bivariate modeling of sequential events using a shared frailty model in Section 2.3. Section 2.4 discusses the ascertainment corrected retrospective likelihood in detail. In Section 2.5, we present the robust variance approach to estimate the parameter variances.

2.1 Measures of disease risks

In the genetic analysis of time-to-event data, the disease risk associated with a mutated gene can be expressed relatively or absolutely. The following are the two measures of disease risk of interest.

- The relative risk of mutation carriers compared to non-mutation carriers can be expressed as the hazard ratio of the mutated gene obtained in the proportional hazards (PH) model, i.e. the ratio of hazard for a mutated gene carrier to the hazard for a non-carrier, such that

$$Hazard\ ratio = \exp(\beta_g),$$

  where β_g is the regression coefficient of the mutated gene from the PH model. The hazard ratio obtained from a shared frailty model should be expressed by conditioning on all risk factors and frailty.
- The absolute risk of the disease can be derived by the penetrance function, defined as the cumulative risk of developing a disease by age t given observed covariates, X. This cumulative function can be expressed as the complement of the survival function ($S(t|X)$), such as

$$Penetrance = P(T < t|X) = 1 - S(t|X),$$

where X is the vector of measured risk factors including the mutated gene. We are interested to calculate the penetrance for the first event by 70 years and a 10 year penetrance for the second event, i.e. the cumulative risk of developing a second event in 10 years after the first event.

2.2 Univariate frailty models

In the analysis of univariate survival data, an individual's unobserved heterogeneity can be modeled by introducing a random variable Z, also known as frailty, into the Cox PH model (Cox, 1972). The frailty Z is a non-negative random variable that accounts for the heterogeneity among individuals. The conditional hazard function of the survival time T given the frailty Z is provided by

$$\lambda(t|X,Z) = Z\lambda_0(t)e^{\beta^\top X},$$

where $X = (x_1, x_2, ..., x_p)$ is a vector of p risk factors and $\beta^\top = (\beta_1, \beta_2, ..., \beta_p)$ is the corresponding regression coefficients (log relative risks). The term $\lambda_0(t)$ is the baseline hazard function, which can be interpreted as the individual's hazard when all the Xs equal to zero. The conditional survival function is given by

$$S(t|X,Z) = e^{-\int_0^t \lambda(u|X,Z)du} = e^{-Z\Lambda_0(t)e^{\beta^\top X}},$$

where $\Lambda_0(t)$ represents the cumulative baseline hazard function.

Since the frailty is an unobserved variable, we can obtain the marginal (unconditional) distribution by integrating out the frailty. The marginal survival function can be expressed in terms of the Laplace transform of the frailty distribution as follows

$$S(t|X) = E_Z\left[S(t|X,Z)\right] = E_Z\left[e^{-Z\Lambda_0(t)e^{\beta^\top X}}\right] = \mathscr{L}\left(\Lambda_0(t)e^{\beta^\top X}\right), \tag{2.1}$$

where $\mathscr{L}(.)$ is the Laplace transform of the frailty distribution. The corresponding marginal density and hazard functions can also be derived in terms of the Laplace

transform of the frailty distribution using the relation, $\lambda(t) = f(t)/S(t)$, such that

$$\begin{aligned} f(t|X) &= -\lambda_0(t)e^{\beta^\top X}\mathscr{L}'(\Lambda_0(t)e^{\beta^\top X}) \\ \lambda(t|X) &= -\lambda_0(t)e^{\beta^\top X}\frac{\mathscr{L}'(\Lambda_0(t)e^{\beta^\top X})}{\mathscr{L}(\Lambda_0(t)e^{\beta^\top X})}, \end{aligned}$$

where $\mathscr{L}'(u)$ is the first order derivative of the Laplace transform with respect to u.

2.3 Shared frailty models

As an extension of the univariate frailty model, the shared frailty model is employed for modeling the two sequential event times T_1 and T_2 from a three-state progressive model with the states – 'Healthy', 'Event 1', and 'Event 2'. Let T_1 be a non-negative continuous random variable that measures the time spent in the 'Healthy' state prior to experiencing 'Event 1' and let T_2 be another non-negative continuous random variable that measures the time spent in the state 'Event 1' prior to experiencing 'Event 2', such that T_2 represents the gap time between the two events (see Figure 1.2). We also let Z be a random frailty variable that measures the amount of dependence between two events experienced by each individual, such that conditional on the frailty, his/her event times are independent. Therefore, the conditional bivariate hazard function of (T_1, T_2) given Z is written as

$$\lambda(t_1, t_2|X_1, X_2, Z) = Z\lambda_{01}(t_1)\lambda_{02}(t_2)e^{\beta_1^\top X_1}e^{\beta_2^\top X_2},$$

where $\lambda_{01}(t_1)$ and $\lambda_{02}(t_2)$ are the baseline hazard functions for T_1 and T_2, respectively, $X_1 = (x_{11}, x_{12}, ..., x_{1p})$ and $X_2 = (x_{21}, x_{22}, ..., x_{2p})$ are the risk factors associated with events 1 and 2, respectively, and their corresponding regression coefficients are $\beta_1^\top = (\beta_{11}, \beta_{12}, ..., \beta_{1p})$ and $\beta_2^\top = (\beta_{21}, \beta_{22}, ..., \beta_{2p})$.

We assume the baseline hazard functions for the event times T_1 and T_2 follow Weibull distributions, respectively, as

$$\begin{aligned} \lambda_{01}(t_1) &= \nu_1\varphi_1{t_1}^{\varphi_1-1}; (\nu_1 > 0, \varphi_1 > 0) \text{ and} \\ \lambda_{02}(t_2) &= \nu_2\varphi_2{t_2}^{\varphi_2-1}; (\nu_2 > 0, \varphi_2 > 0), \end{aligned}$$

where (ν_1, φ_1) and (ν_2, φ_2) are the scale and shape parameters for T_1 and T_2, respectively. We also assume a gamma distribution for the frailty with expectation 1 and variance $1/k$. The probability density function of this one parameter gamma distribution is

$$f(z;k) = \frac{1}{\Gamma(k)} k^k z^{k-1} e^{-kz}.$$

The conditional bivariate survival function for (T_1, T_2) given frailty Z can be expressed as

$$S(t_1, t_2|X_1, X_2, Z) = \exp\left\{-Z\left(\Lambda_{01}(t_1)e^{\beta_1^\top X_1} + \Lambda_{02}(t_2)e^{\beta_2^\top X_2}\right)\right\},$$

where $\Lambda_{01}(t_1)$ and $\Lambda_{02}(t_2)$ are the cumulative baseline hazard functions for T_1 and T_2, respectively. As mentioned in equation (2.1), the bivariate survival function can be derived by integrating out the unobserved frailty as follows

$$\begin{aligned} S(t_1, t_2|X_1, X_2) &= E_Z\left[S(t_1, t_2|X_1, X_2, Z)\right] \\ &= E_Z\left[\exp\left\{-Z\left(\Lambda_{01}(t_1)e^{\beta_1^\top X_1} + \Lambda_{02}(t_2)e^{\beta_2^\top X_2}\right)\right\}\right] \\ &= \mathscr{L}\left(\Lambda_{01}(t_1)e^{\beta_1^\top X_1} + \Lambda_{02}(t_2)e^{\beta_2^\top X_2}\right), \end{aligned}$$

Thus, the unconditional bivariate survival function can be expressed by the Laplace transform of the frailty distribution as a function of the sum of the cumulative baseline hazards of the event times, T_1 and T_2. The gamma distribution that possesses a mathematically simple form of the Laplace transform, given by

$$\mathscr{L}(u) = E\left[e^{-Tu}\right] = \left(1 + \frac{u}{k}\right)^{-k},$$

provides the following joint survival function,

$$S(t_1, t_2|X_1, X_2) = \left[\frac{k + \nu_1 {t_1}^{\varphi_1} e^{\beta_1^\top X_1} + \nu_2 {t_2}^{\varphi_2} e^{\beta_2^\top X_2}}{k}\right]^{-k}. \tag{2.2}$$

2.3.1 Likelihood construction for sequential event times

Consider two sequential event times T_1 and T_2 from an individual. If the age at examination is a, then the event times can be defined as $(t_1, t_2) = (\min(T_1, a), \min(T_2, a - t_1))$ and their censoring indicators are $(\delta_1, \delta_2) = (I(T_1 = t_1), I(T_2 = t_2))$, where I is an indicator function. We construct the likelihood function for this bivariate event times based on the unconditional survival function derived in equation (2.2). The likelihood function derived for the two sequential event times accounts for the following three event occurrence possibilities:

Case 1 An individual survived both the events, i.e. $\delta_1 = 0$ and $\delta_2 = 0$;

Case 2 An individual experienced both the events, i.e. $\delta_1 = 1$ and $\delta_2 = 1$;

Case 3 An individual experienced the first event but has not experienced the second event, i.e. $\delta_1 = 1$ and $\delta_2 = 0$.

We do not consider the case where an individual survived the first event but experienced the second event.

For Case 1, an individual has not experienced either events by his/her age at examination a, so we observe $T_1 = a$ and $T_2 = 0$ with the corresponding censoring indicators $\delta_1 = 0$ and $\delta_2 = 0$. We model Case 1 using the bivariate survival function provided in equation (2.2), i.e. $S(t_1, t_2|X_1, X_2) = P(T_1 > t_1, T_2 > t_2|X_1, X_2)$.

For Case 2, an individual has experienced both the events by his/her age at examination, so we observe $T_1 = t_1$ and $T_2 = t_2$ with the corresponding censoring indicators $\delta_1 = 1$ and $\delta_2 = 1$. We model Case 2 using the bivariate density function $f(t_1, t_2|X_1, X_2) = P(T_1 = t_1, T_2 = t_2|X_1, X_2)$, which can be derived as follows

$$f(t_1,t_2|X_1,X_2) = \frac{\partial^2}{\partial t_1 \partial t_2}\mathscr{L}\left[\Lambda_{01}(t_1)e^{\beta_1^\top X_1}+\Lambda_{02}(t_2)e^{\beta_2^\top X_2}\right]$$
$$= \mathscr{L}''\left[\Lambda_{01}(t_1)e^{\beta_1^\top X_1}+\Lambda_{02}(t_2)e^{\beta_2^\top X_2}\right]\lambda_{01}(t_1)e^{\beta_1^\top X_1}\lambda_{02}(t_2)e^{\beta_2^\top X_2}$$
$$= \frac{k+1}{k}\left(\nu_1\varphi_1 {t_1}^{\varphi_1-1}e^{\beta_1^\top X_1}\right)\left(\nu_2\varphi_2 {t_2}^{\varphi_2-1}e^{\beta_2^\top X_2}\right)\times$$
$$\left[\frac{k+\nu_1 {t_1}^{\varphi_1}e^{\beta_1^\top X_1}+\nu_2 {t_2}^{\varphi_2}e^{\beta_2^\top X_2}}{k}\right]^{-(k+2)},$$

where $\mathscr{L}''(u)$ is the second order derivative of the Laplace function.

Lastly, for Case 3, an individual has experienced only the first event at time t_1 but not the second event, then $T_1 = t_1$ with $\delta_1 = 1$ and $T_2 = a - t_1$ with $\delta_2 = 0$, which can be modeled as

$$\begin{aligned} P(T_1 = t_1, T_2 > t_2|X_1,X_2) &= \frac{-\partial}{\partial t_1}\mathscr{L}\left[\Lambda_{01}(t_1)e^{\beta_1^\top X_1}+\Lambda_{02}(t_2)e^{\beta_2^\top X_2}\right] \\ &= -\mathscr{L}'\left[\Lambda_{01}(t_1)e^{\beta_1^\top X_1}+\Lambda_{02}(t_2)e^{\beta_2^\top X_2}\right]\lambda_{01}(t_1)e^{\beta_1^\top X_1} \\ &= \left[\frac{k+\nu_1 {t_1}^{\varphi_1}e^{\beta_1^\top X_1}+\nu_2 {t_2}^{\varphi_2}e^{\beta_2^\top X_2}}{k}\right]^{-(k+1)}\nu_1\varphi_1 {t_1}^{\varphi_1-1}e^{\beta_1^\top X_1}, \end{aligned}$$

where $\mathscr{L}'(u)$ is the first order derivative of the Laplace function of the frailty distribution with respect to u.

Thus, the likelihood function for the bivariate event times can be formulated using the functions derived as a consequence of the aforementioned three cases. Assuming the family members are independent given their genotype and other risk factors, the likelihood contribution of the f^{th} family, $f = 1,\ldots,n$, with n_f family members, $i = 1,\ldots,n_f$, can be written as

$$L_f(\theta) = \prod_{i=1}^{n_f} S(t_{fi1},t_{fi2}|X_{fi1},X_{fi2})^{(1-\delta_{fi1})(1-\delta_{fi2})} f(t_{fi1},t_{fi2}|X_{fi1},X_{fi2})^{\delta_{fi1}\delta_{fi2}}$$
$$P(T_{fi1} = t_{fi1}, T_{fi2} > t_{fi2}|X_{fi1},X_{fi2})^{\delta_{fi1}(1-\delta_{fi2})},$$

where $\theta = (\nu_1, \varphi_1, \nu_2, \varphi_2, k, \beta_1^\top, \beta_2^\top)$. For a shared gamma frailty model with Weibull baseline hazards, the likelihood can be simplified to

$$L_f(\theta) = \prod_{i=1}^{n_f} \left(\frac{k+1}{k}\right)^{\delta_{fi1}\delta_{fi2}} \left(\frac{k + \nu_1 {t_{fi1}}^{\varphi_1} e^{\beta_1^\top X_{fi1}} + \nu_2 {t_{fi2}}^{\varphi_2} e^{\beta_2^\top X_{fi2}}}{k}\right)^{-\left(k+\delta_{fi1}+\delta_{fi2}\right)} \left(\nu_1\varphi_1 {t_{fi1}}^{\varphi_1-1} e^{\beta_1^\top X_{fi1}}\right)^{\delta_{fi1}} \left(\nu_2\varphi_2 {t_{fi2}}^{\varphi_2-1} e^{\beta_2^\top X_{fi2}}\right)^{\delta_{fi2}}. \tag{2.3}$$

2.4 Ascertainment corrected likelihood

In the previous section, we derived the likelihood function for bivariate event outcomes using a shared frailty model, assuming a gamma distribution for the frailties. In what follows, we discuss an ascertainment correction approach used to account for the non-random sampling of family data. We consider the data arising from n families ($f = 1, \ldots, n$), where each family consists of n_f members, $i = 1, \ldots, n_f$, and the families are selected based on a study design. Then, a general form of the ascertainment corrected likelihood (Le Bihan et al., 1995) is as follows

$$L = \prod_{f=1}^{n} L_f^c = \prod_{f=1}^{n} \frac{N_f}{A_f},$$

where L_f^c is the ascertainment corrected likelihood function of the f^{th} family. The numerator, N_f, is the contribution of the members of family f to the likelihood and the denominator, A_f, is the probability of family f being ascertained.

We consider the retrospective likelihood approach that models the genotypes of the pedigree members given their phenotypes. For individual i in family f ($f = 1, \ldots, n_f; i = 1, \ldots, n_f$), we observe the following vector: $\{Y_{fi} = (t_{fi1}, \delta_{fi1}, t_{fi2}, \delta_{fi2}), G_{fi}\}$ where Y_{fi} is the phenotype containing the event times and censoring indicators for the first and second event and G_{fi} is the genotype which is coded as 1 if mutation carrier and 0 if non-mutation carrier. For simplicity, in this section, we adjust only for the genetic effect G for the event times T_1 and T_2. However, the following procedure can easily accommodate other risk factors. We assume the family members are

independent given their genotypes, i.e. conditional independence. Then, the ascertainment corrected retrospective likelihood (Carayol and Bonaïti-Pellié, 2004; Kraft and Thomas, 2000) for family f is writen as

$$\begin{aligned} L_f^c &= P(G_f|Y_f, Asc_f) \\ &= \frac{P(Asc_f|Y_f, G_f)P(Y_f|G_f,)P(G_f)}{P(Y_f, Asc_f)} \\ &\propto \frac{P(Y_f|G_f)P(G_f)}{\sum_{\omega\in\Omega} P(Y_f, Asc_f|G_{f\omega})P(G_{f\omega})}, \end{aligned} \tag{2.4}$$

where $P(Asc_f|Y_f, G_f) = 1$ if the family satisfies the ascertainment scheme, $P(Y_f|G_f)$ can be expressed as the likelihood function provided in equation (2.3), and the probability of genotype, $P(G_f)$, can be obtained as

$$P(G_f) = \prod_{i=1}^{n_f} \begin{cases} P(G_{fi}), & \text{if individual } i \text{ is a founder,} \\ P(G_{fi}|G_{fd}, G_{fm}), & \text{if individual } i \text{ is a nonfounder.} \end{cases}$$

The probability of genotype, $P(G_{fi})$, for a founder is calculated using the Hardy-Weinberg Equilibrium (HWE), which is based on the prevalence of the mutated gene. For a non-founder, whose parents are in the sampled pedigree, $P(G_{fi}|G_{fd}, G_{fm})$ is obtained by the Mendelian transmission probability using the genotypes of the father (G_{fd}) and the mother (G_{fm}). Details of the derivation of these genotypic probabilities are provided in Appendix A.1.

The denominator of (2.4) is the ascertainment probability of observing the phenotypes of the members through whom the family is ascertained into the study (Carayol and Bonaïti-Pellié, 2004). It can be obtained by the sum of the conditional probabilities, $P(Y_f, Asc_f|G_{\omega f})$, over all possible genotypic configurations, Ω, of all ascertained members. The calculation of this ascertainment probability is further explained in the following section.

2.4.1 Ascertainment probability for different study designs

Depending on the study design, the ascertainment probability can be derived with the knowledge of underlying ascertainment process used for sampling the families. In our thesis, we consider two types of family designs − population-based and clinic-based study designs. For the population-based design, a family is ascertained through an affected, mutation carrying proband and for the clinic-based study design, a family is ascertained if at least one of the proband's parents and at least one of proband's sibling are affected, in addition to the proband being an affected carrier.

Population-based design

The ascertainment for the population-based design is only based on the probands, who are randomly sampled from a diseased population. Therefore, the ascertainment probability for family f can be obtained simply by calculating the probability that the proband is affected by the first event prior to his/her age at examination, which can written as

$$\begin{aligned} P(Y_f, Asc_f|G_f) &= P(T_1 < a_{fp}|G_{fp}) \\ &= 1 - S_1(a_{fp}|G_{fp}) \\ &= 1 - \left[\frac{k + \nu_1 a_{fp}^{\varphi_1} e^{\beta_1 G_{fp}}}{k}\right]^{-k}, \end{aligned}$$

where a_{fp} is the age at examination of the proband, G_{fp} is the genotype of the proband and $S_1(t_1|G_{fp})$ is the marginal survivor function for T_1 obtained from the bivariate survival function provided in equation (2.2).

Clinic-based design

For the clinic-based design, the ascertainment scheme involves the proband, his/her parents and siblings. The ascertainment probability for family f is the probability of observing the disease statuses of those who were involved in the ascertainment process

at their ages at examination and can be expressed as:

$$P(Y_f, Asc_f) = P(T_1 < a_{fp}|G_{fp}) \sum_{\omega\in\Omega} \Bigg[P(T_1 < a_{fd}|G_{\omega fd})^{\delta_{fd1}} P(T_1 \geq a_{fd}|G_{\omega fd})^{1-\delta_{fd1}} \times$$
$$P(T_1 < a_{fm}|G_{\omega fm})^{\delta_{fm1}} P(T_1 \geq a_{fm}|G_{\omega fm})^{1-\delta_{fm1}} P(G_{\omega fd}, G_{\omega fm}|G_{fp}) \times$$
$$\prod_{fs} \left\{ P(T_1 < a_{fs}|G_{\omega fs})^{\delta_{fs1}} P(T_1 \geq a_{fs}|G_{\omega fs})^{1-\delta_{fs1}} P(G_{\omega fs}|G_{\omega fd}, G_{\omega fm}) \right\} \Bigg],$$

where a_{fp},a_{fd},a_{fm}, and a_{fs} are the ages at examination of the proband, father, mother, and sibling, respectively. Similarly, $G_{\omega fp}$,$G_{\omega fd}$,$G_{\omega fm}$, and $G_{\omega fs}$ are the genotypes of the proband, father, mother, and sibling belonging to the genotypic configuration, ω, and δ_{fp1}, δ_{fd1}, δ_{fm1}, and δ_{fs1} are their corresponding censoring indicators for event 1. The ascertainment probability is obtained by summing over possible genotypic configurations of the father, mother, and sibling. This probability involves the computation of the conditional genotypes for the parents and sibling given the genotype of the proband. The conditional genotype probabilities are derived in the Appendix A.2.

Now, putting every pieces together, we write the ascertainment corrected likelihood for the family data arising from n families, each with n_f family members as

$$L(\theta) = \prod_{f=1}^{n} \frac{\prod_{i=1}^{n_f} N_{fi}}{A_f},$$

where the numerator consists of the contribution of the i^{th} member in f^{th} family and the denominator is the ascertainment probability of family f. Then the corresponding log-likelihood is given by

$$l(\theta) = \sum_{f=1}^{n} \sum_{i=1}^{n_f} \log N_{fi} - \sum_{f=1}^{n} \log A_f. \tag{2.5}$$

By maximizing the ascertainment corrected log-likelihood in equation (2.5), we can obtain the maximum likelihood estimates of the parameters $\theta = (\nu_1, \varphi_1, \nu_2, \varphi_2, k, \beta_1^\top, \beta_2^\top)$ in the model. The maximum likelihood estimate, $\widehat{\theta}$, asymptotically follows a normal

distribution with mean θ and variance Σ. In the following section, we derive the robust variance estimators (White, 1982a) for the relative and absolute risk estimates of the mutated gene in order to handle the model mis-specification for not modeling the dependence among family members.

2.5 Variance estimation

To incorporate possible residual familial correlation, we first derive the robust variance estimator for the estimated parameters in the model, following which, we obtain the robust variance estimators for the estimated penetrance functions.

2.5.1 Robust variance estimator for parameter estimates

Let $\widehat{\theta} = (\widehat{\nu}_1, \widehat{\varphi}_1, \widehat{\nu}_2, \widehat{\varphi}_2, \widehat{k}, \widehat{\beta}_1^\top, \widehat{\beta}_2^\top)^\top$ denote the maximum likelihood estimate of the vector of parameters in our shared frailty model. The robust variance estimator (White, 1982a) for $\widehat{\theta}$ is expressed as

$$Var(\widehat{\theta}) = H^{-1}(\theta)V(\theta)H^{-1}(\theta),$$

where $H(\theta)$ is the Fisher information matrix consisting of the second order derivatives of the log-likelihood function in equation (2.5) with respect to the parameters, θ, and $V(\theta)$ is the variance of the score vector; they have the following forms,

$$\begin{aligned}
H(\theta) &= -E\left[\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^\top}\right] = -E\left[\sum_{f=1}^{n}\sum_{i=1}^{n_f}\frac{\partial^2}{\partial\theta\partial\theta^\top}\log N_{fi} - \sum_{f=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta^\top}\log A_f\right] \text{ and} \\
V(\theta) &= Var\left[\frac{\partial l(\theta)}{\partial\theta}\right] \\
&= \sum_f Var\left[\sum_{i=1}^{n_f}\frac{\partial}{\partial\theta}\log N_{fi} - \frac{\partial}{\partial\theta}\log A_f\right] \\
&= \sum_f E\left[\left\{\sum_{i=1}^{n_f}\frac{\partial}{\partial\theta}\log N_{fi} - \frac{\partial}{\partial\theta}\log A_f\right\}\left\{\sum_{i=1}^{n_f}\frac{\partial}{\partial\theta}\log N_{fi} - \frac{\partial}{\partial\theta}\log A_f\right\}^\top\right].
\end{aligned}$$

Then we estimate $H(\theta)$ and $V(\theta)$ empirically as follows

$$\begin{aligned}\widehat{H}(\theta) &= -\frac{1}{n}l''(\theta) \\ \widehat{V}(\theta) &= \frac{1}{n}\sum_{f=1}^{n} l_f{}'(\theta) l_f{}'(\theta)^\top\end{aligned}$$

where $l''(\theta)$ is the second order derivatives of the log likelihood function with respect to the parameter θ, and $l'_f(\theta)$ is the contribution of f^{th} family ($f = 1, \ldots, n$) to the score function.

Therefore, the robust variance estimate of the estimated parameters is obtained from the following variance-covariance matrix,

$$\widehat{Var}(\widehat{\theta}) = \widehat{H}(\widehat{\theta})^{-1}\widehat{V}(\widehat{\theta})\widehat{H}(\widehat{\theta})^{-1}. \tag{2.6}$$

2.5.2 Robust variance estimator for penetrance estimates

The robust variance of the penetrance function is obtained by using the Delta method. Penetrance is defined as the cumulative probability of failure such that $F_1(t_1;\theta) = 1 - S_1(t_1;\theta)$, where $S_1(t_1;\theta)$ is the marginal survival function of the first event at time t_1, given by

$$S_1(t_1;\theta) = \left(\frac{k + \nu_1 t_1{}^{\varphi_1} e^{\beta_1^\top x_1}}{k}\right)^{-k}.$$

We derive the variance of the penetrance function in two steps in order to (log) transform the skewed distribution of the survival function, as the robust variance estimator assumes a normal distribution. First, we compute the variance of the cumulative hazard function, $Var\{\Lambda_1(t_1;\theta)\}$, using the Delta method where

$$\begin{aligned}\Lambda_1(t_1;\theta) &= -\log S_1(t_1;\theta) \\ &= k\log\left(k + \nu_1 t_1{}^{\varphi_1} e^{\beta_1^\top x_1}\right) - k\log k.\end{aligned} \tag{2.7}$$

Next, using the relation in equation (2.7), we compute the variance of the penetrance function, $Var\{F_1(t_1;\widehat{\theta})\}$, by using the Delta method again.

Step 1: The variance estimator of the cumulative hazard function has the form:

$$Var\{\Lambda_1(t_1;\widehat{\theta})\} = D_\theta(t_1)^\top \Sigma D_\theta(t_1),$$

where Σ is the robust variance-covariance matrix of the parameters obtained using equation (2.6) and $D_\theta(t_1)$ is the vector of partial derivatives of $\Lambda_1(t_1;\theta)$ with respect to each parameter, such that

$$D_\theta(t_1) = \frac{\partial \Lambda_1(t_1;\theta)}{\partial \theta} = \left(\frac{\partial \Lambda_1(t_1;\theta)}{\partial \nu_1}, \frac{\partial \Lambda_1(t_1;\theta)}{\partial \varphi_1}, \frac{\partial \Lambda_1(t_1;\theta)}{\partial k}, \frac{\partial \Lambda_1(t_1;\theta)}{\partial \beta_1}\right)^\top.$$

Step 2: The robust variance estimator for the penetrance of developing the first event by the age t_1 can be expressed as

$$\begin{aligned} Var\{F_1(t_1;\widehat{\theta})\} &= Var\{S_1(t_1;\widehat{\theta})\} \\ &= Var\{e^{-\Lambda_1(t_1;\widehat{\theta})}\} \\ &= \left[-e^{-\Lambda_1(t_1;\widehat{\theta})}\right]^2 Var\{\Lambda_1(t_1;\widehat{\theta})\} \\ &= \left\{e^{-k\log(k+\nu_1 t_1{}^{\varphi_1} e^{\beta_1^\top x_1)+k\log k}}\right\}^2 Var\{\Lambda_1(t_1;\widehat{\theta})\}. \end{aligned}$$

Similarly, the robust variance estimator for the penetrance function of event 2 at time t_2, after the first event, can be obtained as

$$Var\{F_2(t_2;\widehat{\theta})\} = \left\{e^{-k\log(k+\nu_2 t_2{}^{\varphi_2} e^{\beta_2^\top x_2)+k\log k}}\right\}^2 Var\{\Lambda_2(t_2;\widehat{\theta})\}.$$

2.6 Summary

Using a shared frailty model, we modeled the bivariate event times to account for the dependence between two sequential events. For the data that arise from a family-based study design, we incorporated a retrospective likelihood to account for the study design in our analysis. The retrospective likelihood possesses the advantage to implicitly correct for complex ascertainment schemes and is capable of providing

population-based inference. To account for the familial correlation, the robust variance estimators of the penetrance functions and the relative risks for the first and second events were derived using the Delta method.

Chapter 3

SIMULATION STUDY

A simulation study was conducted to evaluate the performance of our proposed method in a large sample setting. We compared the precision and accuracy of the disease risks derived from our frailty model, that accounts for the dependency between two sequential event times, to those from an independent model that ignores it.

In the following section, we define the different parameter combinations considered in our simulation studies and their corresponding parameter values. Then, we provide details about family data generation involving the bivariate event times for each family member. Finally, we provide the simulation results and conclude the chapter with a discussion of the results.

3.1 Parameter combinations

For our simulation studies, we considered the families to arise from a population-based study design, where each family was selected through an affected mutation carrier proband. These probands were assumed to be randomly sampled from a population-based disease registry, for example, a provincial cancer registry. We considered the following models for the two sequential event times T_1 and T_2,

$$\begin{aligned} h_1(t_1|X_1, X_2, Z) &= Z\nu_1\varphi_1(t_1 - 20)^{\varphi_1 - 1}e^{\beta_1 X_1 + \beta_2 X_2} \\ h_2(t_2|X_2, Z) &= Z\nu_2\varphi_2 t_2^{\varphi_2 - 1}e^{\beta_3 X_2}, \end{aligned}$$

where t_1 is the age-at-onset of the first event with a minimum age of onset as 20 years, t_2 is the time of occurrence of the second event since the first event (in years), and Z is the frailty variable following the gamma distribution with mean 1 and variance $1/k$. For the first event, we adjusted for a gender effect, X_1, and a gene mutation effect, X_2, and for the second event, we adjusted only for the mutation effect, X_2. The parameters involved in our shared frailty model are $\theta = (\nu_1, \varphi_1, \nu_2, \varphi_2, k, \beta_1, \beta_2, \beta_3)$, where ($\nu_1$,$\varphi_1$) and ($\nu_2$,$\varphi_2$) are the scale and shape parameters of the baseline hazards for the first and second events, respectively, k is the frailty parameter, and β_1, β_2, and β_3 are the regression coefficients (log relative risks) of X_1, X_2, and X_2, respectively.

Our parameter combinations involved different genetic models and diverse penetrance levels for the mutated gene. In what follows, we explain each of the parameter combination and how we chose the appropriate parameter values. We also present all the parameter combinations considered for in simulation study in Figure 3.1. The simulation study was designed according to the guidelines provided by Burton et al. (2006).

3.1.1 Genetic models

We considered two genetic models − dominant and recessive models. In the dominant model, at least one copy (one inherited from each parent) of the mutant allele is enough for an individual to be at risk for the disease, whereas for the recessive model, two copies are required to be at risk. We also varied the prevalence of the mutated gene for each of the genetic models; for the dominant model, we considered a rare variant with allele frequency, $q = 2\%$, and for the recessive model, we considered a common variant with, $q = 30\%$. Therefore, we considered these two genetic models: dominant model with a rare gene and recessive model with a common gene. The allele frequencies for these two models were set based on the work of Choi et al. (2008).

3.1.2 Parameter combinations for the first event

For the first event, we fixed the baseline parameters at ($\nu_1 = 5.35\times10^{-6}$, $\varphi_1 = 2.33$) and the log-relative risk of the gender effect at $\beta_1 = 1.19$. These values provided the penetrances of first event by the age of 70 years to be 15% and 5% for the male and female non-mutation carriers, respectively (baseline population). On the other hand, for the mutation carriers, we considered two penetrance levels − high and low penetrances, with the log-relative risk of mutated gene, β_2, set at 2.5 and 1.55, respectively. When $k = 10$, the high penetrance model ($\beta_2 = 2.5$), corresponded to a penetrance for the first event by the age of 70 years to be 83% and 44% for the male and female mutation carriers, respectively, and the low penetrance model ($\beta_2 = 1.55$), corresponded to 52% and 20% of penetrances for male and female carriers, respectively. The penetrance values covered when $k = 1$ and 3 are provided in Figure 3.1. The true values for the log-relative risks were decided based on our analyses of family data from Newfoundland in Chapter 4.

3.1.3 Parameter combinations for the second event

As for the second event, we investigated two distinct baseline settings − low baseline ($\nu_2 = 0.00724$, $\varphi_2 = 1.14$) and high baseline ($\nu_2 = 0.00324$, $\varphi_2 = 1.84$). The corresponding penetrance values for developing a second event in 10 years after the first event among those baseline populations (non-mutation carriers) were 10% and 23%, respectively. We used two baselines because the risk of successive cancers following a primary cancer is predominently low in the general population. However, for Lynch syndrome families, there exists a very high risk for cancer recurrence. Therefore, in order to study this diversity in risk, we considered low and high baselines for T_2. For the mutation carriers, in addition to the two baseline settings, we also considered two penetrance models − high and low penetrance, where their log-relative risks for the mutated gene (β_3) were fixed at 0.75 and 0.3, respectively. Consequently, the fol-

lowing four scenarios were resulted for the second event penetrance: (i) low baseline with low penetrance, (ii) low baseline with high penetrance, (iii) high baseline with low penetrance, and (iv) high baseline with high penetrance. Correspondingly, the cumulative risks of developing a second event in 10 years after the first event for a mutation carrier were 13%, 19%, 26%, and 37% for scenarios (i), (ii), (iii), and (iv), when $k = 10$.

We decided these penetrance values for the second event based on extensive literature search on the occurrence of second primary colorectal cancer (CRC). The penetrance values for mutation carriers were determined by considering articles which estimated the cumulative risk of second CRC using kindred studies. For example, Aarnio et al. (1995) and Parry et al. (2011) estimated the cumulative risk of second CRC after 10 years of follow-up for mutation carriers to be close to 16%, Mecklin and Jarvïnen (1986) estimated the penetrance of second CRC to be approximately 40% for 10 years using 22 Finnish CRC kindreds, and Fitzgibbons Jr. et al. (1987) sampled 10 kindreds and estimated the penetrance to be 40% for 10 years after the first CRC. For the non-carrier penetrances, Myrhøj et al. (1997) estimated the age-specific cumulative risks using sporadic cases as 10%.

3.1.4 Dependence levels, family sizes, and simulation runs

Finally, we considered three dependence levels between the successive events − high dependence ($k = 1$), moderate dependence ($k = 2$), and small dependence ($k = 10$). We recall that the dependence between the successive events is measured by the variance of the frailty distribution. For a gamma frailty, the variance is provided by the inverse of the frailty parameter, var(Z)=$1/k$. Hence, as the value of k increases, the dependence between the successive events decreases. The variance of the frailty distribution, used as a measure of dependence between the event times, can be related to Kendall's tau using the relation: $\tau = \dfrac{1}{1+2k}$.

We generated the family data for our simulation studies with two varying sample

sizes − 100 families and 200 families. Under a population-based study design, the aforementioned sample sizes are readily achievable due to access to large family registries such as the NCI funded Breast and Colon Cancer Familial Registries (CFR) (http://www.cfr.epi.uci.edu/).

In total, our simulation studies considered 96 parameter combinations (2 penetrance models for first event, 4 penetrance models for second event, 3 dependence levels, 2 genetic models, and 2 sample sizes) and for each parameter combination, we performed 500 replications. Figure 3.1 summarizes all the parameter values considered and their corresponding penetrance values for the first and second event.

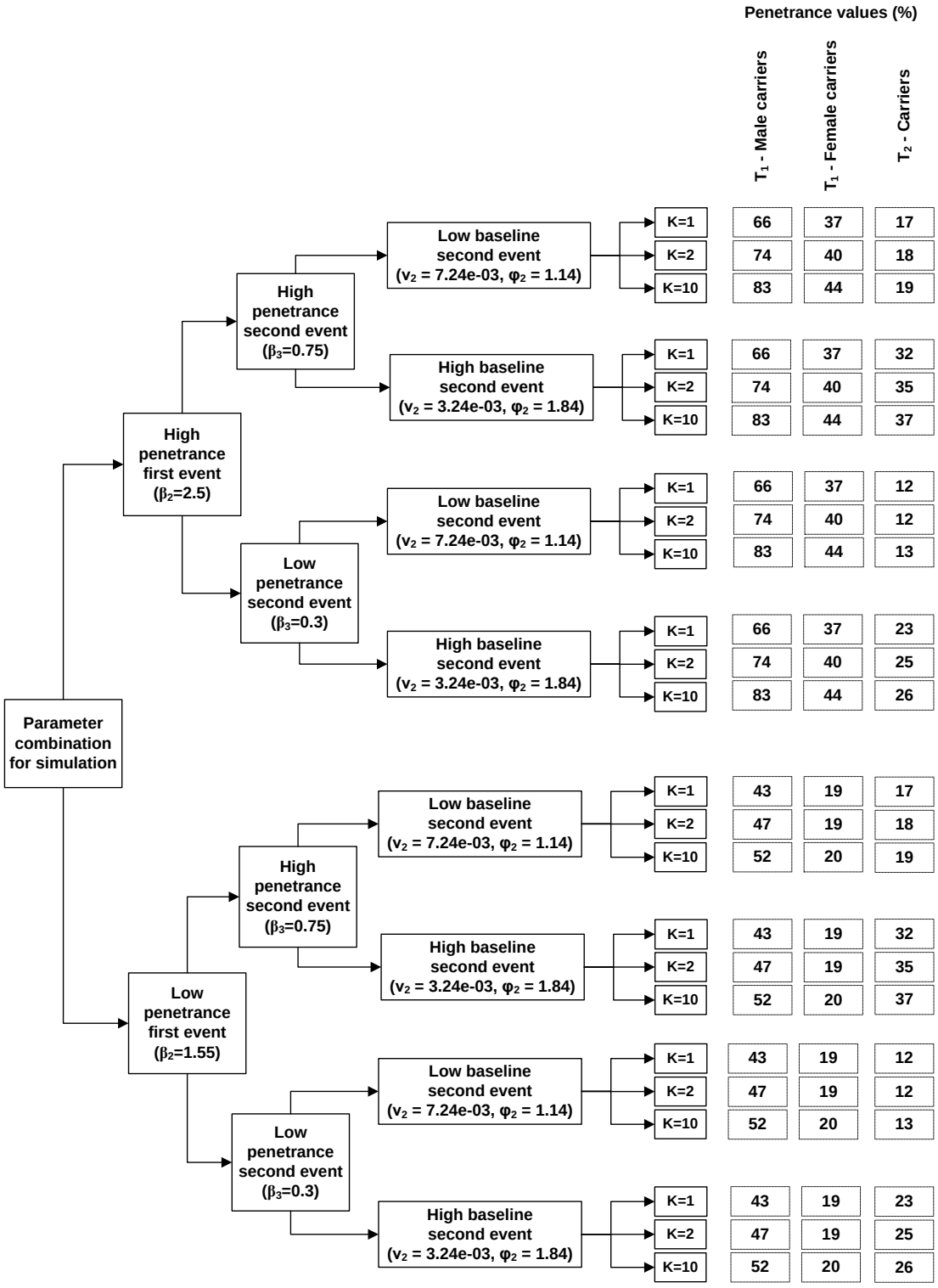

Figure 3.1: Parameter values chosen for our simulation study and the corresponding penetrance values for the first (T_1) and second (T_2) event.

3.2 Pedigree generation

We generated the pedigrees for our simulation studies based on the ideas of Gauderman (1995). Each simulated family consists of three generations of family members - two parents, their offspring (one of whom is a proband) and each offspring has a spouse and children. The number of siblings in the second generation as well as their offspring were varied between two and five using a truncated negative binomial distribution. Figure 3.2 illustrates the pedigree structure that we used for our simulation study.

For each family member, we generated their gender, age at examination, mutation status, and bivariate event times. We began by simulating family members with equal probabilities of male and female and their ages at examination using a normal distribution with the mean age as 45 for the members belonging to the first and second generations and 20 years for the members of the third generation. The variances were fixed at 2.5 years for the first two generations and 1 year for the third generation. We assumed the minimum age-at-onset for the first event as 20 years and the maximum

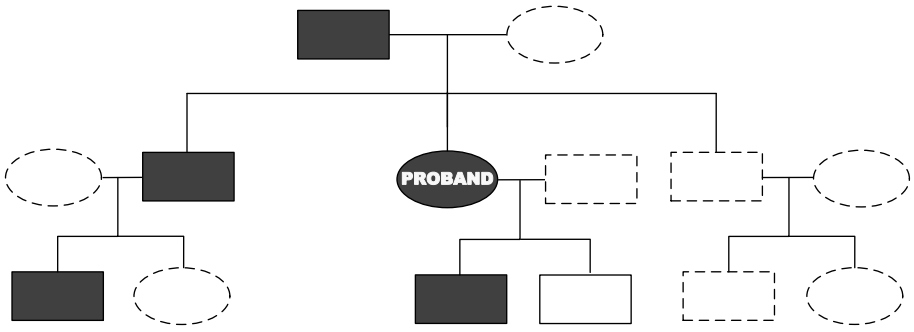

Figure 3.2: A simulated family with the proband including two parents, two siblings and each having two children. Males are displayed in rectangles and females in ovals. Solid and dashed outlines represent mutation carriers and non-carriers, respectively, and shaded if affected.

age at examination as 90 years.

In the next step, we generated the genotype of the proband conditioning on his/her age at examination and gender. We assumed the proband to be affected by the first event before his/her age at examination and to carry a disease mutation. Conditioning on the proband's genotype, the genotypes of other family members were generated either using Hardy Weinberg equilibrium or Mendelian transmission probabilities. Now, given the genotype and gender for each family member, we generated their bivariate event times. Details of generating the bivariate event times are presented in the following section.

3.2.1 Simulation of bivariate event times

For our simulation studies, we simulated two sequential times using the following bivariate distribution,

$$S(t_1, t_2|X_1, X_2) = \left[\frac{k + \nu_1(t_1 - 20)^{\varphi_1} e^{\beta_1 X_1 + \beta_2 X_2} + \nu_2 {t_2}^{\varphi_2} e^{\beta_3 X_2}}{k}\right]^{-k}. \tag{3.1}$$

We began by generating the age-at-onset for first event from the corresponding marginal distribution and then generated time to second event from the conditional distribution of second event time given the value of the first event time. The detailed procedure of simulating the two sequential event times is as follows:

1. Generate two independent random variables, u and v from the Uniform distribution, $U(0, 1)$, respectively.
2. Using equation (3.1), derive the marginal survival function for the first event, $S_1(t_1|X_1, X_2)$, where X_1 and X_2 are the gender and genotype of the simulated family member.
3. Now, set $u = S_1(t_1|X_1, X_2)$ and solve for t_1 such that

$$t_1 = \left[\frac{k u^{\frac{-1}{k}} - k}{\nu_1 e^{\beta_1 X_1 + \beta_2 X_2}}\right]^{\frac{1}{\varphi_1}} + 20.$$

4. Using the conditional survival distribution for the second event conditioning on $T_1 = t_1$, obtain t_2 by setting $v = S_2(t_2|t_1, X_2)$, such that

$$t_2 = \left[\frac{v^{\frac{-1}{k+1}}(k + \nu_1(t_1 - 20)^{\varphi_1} e^{\beta_1 X_1 + \beta_2 X_2}) - k - \nu_1(t_1 - 20)^{\varphi_1} e^{\beta_1 X_1 + \beta_2 X_2}}{\nu_2 e^{\beta_3 X_2}} \right]^{\frac{1}{\varphi_2}}$$

5. Thus, (t_1, t_2) are the bivariate event times generated from the bivariate distribution in equation (3.1).

Finally, the censoring indicators were derived for the first event as $\delta_1 = 1$ if $t_1 < a$ and 0, otherwise, where a is the age at examination and for the second event, $\delta_2 = 1$ if $(t_1 + t_2) < a$ and 0, otherwise. We set $T_1 = t_1$ if $\delta_1 = 1$ and $T_1 = a$ if $\delta_1 = 0$. Correspondingly, $T_2 = t_2$ if $\delta_2 = 1$ and $T_2 = a - t_1$ if $\delta_2 = 0$. Since the proband was assumed to be affected by the first event ($\delta_1 = 1$), his/her time-to-onset for the first event was generated to be less than his/her age at examination.

3.3 Evaluation criteria

We compared the estimates of disease risks (penetrance and relative risks) obtained from our frailty-based approach to those from an independent model, which ignored the dependence between the events. For the latter, we assumed two independent Weibull models for the first and second events. We evaluated the accuracy and precision of the disease risk estimates using the following characteristics:

Median bias

The bias of an estimate was computed as the difference between the estimate and the true value. The median bias was reported due to the presence of few extreme values in some settings of our simulation study. In addition, we reported the first and third quartiles.

Standard error

For each simulation, we computed the model-based robust standard errors (SE) of the estimates. The median SE estimate from the 500 simulations was reported along

with their first and third quartiles.

Coverage probability

We also presented the coverage probability (CP) based on the estimated model-based SE. It was calculated as the proportion of times the 95% confidence interval of the estimates included the true value. Burton et al. (2006) suggests that the CP for an estimate must approximately lie within the two SEs of the nominal coverage probability. The SE for a 95% nominal CP with 500 simulations is given by $\sqrt{\frac{0.95(1-0.95)}{500}} = 0.0097$ and hence the acceptable coverage boundary is $(0.95 - (2 \times 0.0097) = 0.93,\ 0.95 + (2 \times 0.0097) = 0.97)$. Over-coverage (CP $> 95\%$) may lead to an increase in type II error, whereas an under-coverage (CP $< 95\%$) may increase the type I error rate.

3.4 Simulation results

We performed the simulation studies using the statistical software, R (R Development Core Team, 2011). We summarize the simulation results in Tables 3.1 - 3.5 for the dominant models and Tables 3.6 - 3.10 for the recessive models. In addition, Figures 3.3 - 3.7 graphically display the bias of the estimate and its 95% confidence interval (CI) for the dominant model and Figures 3.8 - 3.12 display the same for the recessive model. The 95% CIs were computed using the median model-based SEs obtained from the simulations. Each Table and Figure also contain the results from the independent model. In this Chapter, we present the simulation results based on 200 simulated families. The results from 100 families are presented in Appendix B.

3.4.1 Estimation of log genetic relative risks

Log genetic relative risk for the first event (β_2)

The simulation results for the estimation of log genetic relative risk of the first event (β_2) using the dominant and recessive models are presented in Tables 3.1 and 3.6,

respectively and also graphically displayed in Figures 3.3 and 3.8 in terms of the bias and its 95% confidence interval (CI). In the presence of the dominant model (see Table 3.1 and Figure 3.3), our frailty based approach produced almost unbiased estimates with an absolute value of the bias less than 0.062. The biases were predominantly positive and slightly increased with the value of k, i.e. when the dependency between the events reduced. But they were not significantly different from zero as their 95% CIs covered zero. On the other hand, the model-based standard errors (SEs) ranged between 0.215 and 0.258 over all different combinations of the first and second event penetrances and they tended to decrease slightly as the value of k increased. We also noticed that high penetrance yielded higher SEs than low penetrance setting for both first and second events. The coverage probabilities (CPs) were close to the prescribed nominal probability of 0.95 in all 24 parametric combinations; therefore, the type I error rate for testing the null hypothesis, $\beta_2 = 0$, is under control, i.e. $\alpha = 5\%$. Similarly, for the recessive model (see Table 3.6 and Figure 3.8), the bias was also negligible but appeared slightly higher compared to the dominant model in the presence of the low penetrance for the first event. The SEs and CPs remained almost same for the dominant and recessive models.

Log genetic relative risk for the second event (β_3)

We tabulate the simulation results for the estimation of log relative risk of mutated gene for the second event (β_3) in Tables 3.2 and 3.7 using the dominant and recessive models, respectively, and also graphically display its bias and the 95% CI in Figures 3.4 and 3.9. In the estimation of β_3 using the two genetic models, the bias appeared to be negligible and ranged between -0.040 and 0.068. The 95% CI for the bias covered zero in all parameter settings. The model-based SEs for $\hat{\beta}_3$ ranged between 0.379 and 0.615 and they tended to increase with the value of k in the low baseline (LBL) setting of the second event but they decreased in most of the high baseline (HBL) setting. The SEs from the high penetrance for the first event appeared lower than

those from the low penetrance of the first event and a similar tendency was observed for the high and low penetrance settings for the second event. The SEs for $\hat{\beta}_3$ were twice as large as those obtained for $\hat{\beta}_2$ due to relatively smaller number of second events. The CPs were mostly close to the prescribed 95% CP.

We also compared the performance of our frailty approach to that of a model that ignores the dependence between the events. The genetic relative risk estimates for different dependence levels ($k = 1,\ 2,$ and 10) were compared to those from the independent model. Tables 3.1, 3.2, 3.6, and 3.7 clearly suggested that in the presence of a high ($k = 1$) or moderate ($k = 2$) dependency between the successive events, the independent model resulted in negative bias in the estimates of log genetic relative risks for both the first event, β_2, as well as for the second event, β_3. The median bias from the independent model can be as large as -0.254 in the estimates of β_2 and as large as -0.215 in the estimates of β_3. However, as shown in Figure 3.3, 3.4, 3.8, and 3.9, the 95% CIs for the biases of $\hat{\beta}_2$ and $\hat{\beta}_3$ from the independent model included zero irrespective of the genetic model and the value of k. In spite of that, the CPs for the independent model were far below the prescribed 95% nominal level for $k = 1$ and 2, especially for the estimation of β_2 under the high penetrance for the first event. This suggested that the independent model increased the type I error rates in the test of null hypothesis of the relative risks. Overall, regardless of the genetic model and parameter combinations, a model which incorrectly assumed independence between two dependent events resulted in some sinvalid estimates of the genetic relative risks.

3.4.2 Estimation of penetrances

In our simulation studies, we estimated the penetrance of developing a first event by the age of 70 years, separately for male and female mutation carriers and the penetrance of developing a second event in 10 years after the first event among carriers of the mutated gene.

Penetrance estimation of first event for male carriers

The simulation results obtained for the penetrance estimation of first event among male mutation carriers are presented in Tables 3.3 and 3.8 for the dominant and recessive models, respectively. We also graphically display the bias and its 95% CI in Figures 3.5 and 3.10, using the two genetic models. Regardless of the genetic model, the biases of the penetrance estimates obtained from our frailty approach estimate appeared to be negligible as the highest absolute value was only 0.014 and their robust SEs ranged between 0.037 and 0.053. The resulting 95% CIs of the bias included zero in all settings. Moreover, there was a very minimal decrease in SEs with value of k and the values of SE was the smallest when k=10, irrespective of parameter combinations. The bias and SE of the estimate were somewhat larger in the presence of the high penetrance model for T_1 compared to the low penetrance model. Finally, the CPs were all close to 95%.

Penetrance estimation of first event for female carriers

As for the penetrance estimation for female mutation carriers, we present the simulation results in Tables 3.4 and 3.9 for the dominant and recessive models, respectively. We also graphically display the bias of the estimate and its 95% CI in Figures 3.6 and 3.11. For the dominant model, the biases and SEs remained negligible but slightly smaller compared to those obtained for male carriers. As shown in Table 3.4, the model-based SEs tended to increase with the value of k and the SEs for the high penetrance model for T_1 were slightly larger (ranged between 0.040 and 0.045) than those corresponding to the low penetrance model (ranged between 0.032 and 0.036). The CPs were similar to those obtained for the male counterparts, all close to the 95% nominal level. However, the SEs for the female carriers seemed to be slightly smaller than those for the male carriers as shown in Figures 3.6 and 3.11. The bias, SE and CP remained almost same for the dominant and recessive models.

Irrespective of the genetic model, our frailty based approach was less biased com-

pared to the independent model in the penetrance estimation of the first event. Surprisingly, the independent model yielded almost unbiased estimates in the penetrance estimation for female mutation carriers (Figures 3.6 and 3.11), in addition to the CPs close to 95% (Tables 3.4 and 3.9). We investigated the reason behind this peculiar situation and found that the independent model overestimated the scale parameter (ν_1), which would lead to overestimation of the penetrance but the underestimated log relative risk of mutated gene (β_2) neutralized its effect, so resulting in an unbiased penetrance estimates for female carrier. We also observed that the robust SEs for the independent model seemed slightly higher than those obtained from our frailty model in the estimation of penetrance values for female carries.

Penetrance estimation of second event for mutation carriers

The simulation results for the dominant model in the estimation of penetrance of developing a second event in 10 years after the first event among mutation carriers are presented in Table 3.5 and graphically in Figure 3.7. And the recessive model's results are presented in Table 3.10 and Figure 3.12. In the estimation of penetrance for second event, the absolute values of the bias for both the genetic models remained less than 0.020. As Tables 3.5 and 3.10 show, it is evident that there was no impact of different genetic models in the estimation of penetrance function as their SEs and CPs were almost similar. Irrespective of the genetic model, the SEs were larger under the high baseline for T_2 (ranged between 0.034 and 0.085) compared to the corresponding low baseline (ranged between 0.022 and 0.055). The SEs tended to decrease as the value of k increased; the SEs of the estimates were the lowest when $k = 10$, regardless of the parameter combinations and genetic models. Lastly, the CPs for the two genetic models ranged between 0.71 and 0.94, which were far below the 95% nominal level. This could possibly be due to the limited number of second event occurrences in these settings. With an idea to improve the CPs, we simulated 500 families (results not shown) and found the CPs to be better than those obtained using 200 families

(above 75%), but still slightly less than the considered 95% probability.

The independent model clearly produced biased penetrance estimates compared to our frailty model which accounted for the dependence between events. The estimates were upwardly biased in the presence of a high or moderate dependence, i.e. when $k = 1$ or 2 (see Figures 3.7 and 3.12) and the biases were significantly different from zero. This was supported by the presence of very low CP. However, when $k = 10$, the penetrance estimates from the independent model were close to those obtained from our frailty approach as expected and it can also be noted that the SEs from the independent model were always smaller compared to those from our frailty approach for any value of k.

3.5 Summary

Using a population-based study design, we demonstrated that our proposed method provided more consistent and reliable estimates for both relative risk and penetrance compared to the independent model. The two genetic models − dominant and recessive models, had some impact on the estimation of genetic relative risks of first event where the recessive model produced slightly larger bias than the dominant model. However, in the penetrance estimation there was no noticeable difference in our evaluation criteria using these two genetic models.

On the other hand, the independent model yielded unreliable estimates in the presence of high or medium dependence as their CPs often did not achieve the desired coverage level. In the estimation of penetrance for the first event, the estimates were biased for male carriers, however, the estimates were unbiased for female carriers. In the estimation of penetrance for a second event, the estimates were highly biased from an independent model in the presence of high dependence ($k = 1$). The CPs under the independent model were predominantly lower than the prescribed 95% nominal probability and especially on the estimation of penetrance of second event for

mutation carriers where the CPs were less than 50% in the presence of high or moderate dependence between the sequential events. Even though the precision from the independent model seemed to be slightly better (smaller SEs) than our frailty model, the largely biased estimates produced by the independent model counterbalanced the benefit, especially in the presence of high dependence between the two events. When there is low dependence between the event times, the independent model provides better CP values than our frailty based approach, i.e. closer to the prescribed 95% nominal probability. This could possibly be due to the additional burden in estimating the frailty parameter (k) using our approach, in the absence of a substantial dependence between the events. We note that the median model-based SEs of the estimates were close to the simulation-based SEs. The average censoring rate for the first event in the simulated families was close to 90% and the censoring rate slightly increased as the dependence between the event times increased and hence can be viewed as a possible reason behind the under coverage of the CPs.

We also varied the sample size of our simulation from $n = 200$ families to $n = 100$ families (results presented in Appendix B) and found the inferences were similar to those obtained with 200 samples with the CPs close to the prescribed 95%. Using the $n = 100$, the biases appeared slightly greater for the estimation of genetic relative risks, β_2 and β_3, but were not statistically significant. It was also expected to see that the model based SEs with the 100 families were about 1.5 fold larger than those with 200 families on the estimation of genetic relative risks and penetrance functions for the first and second events.

Table 3.1: Estimation of log relative genetic risk (β_2) of developing the first event under the dominant genetic model with rare allele frequency ($q = 2\%$) using 200 simulated families.

Parameters				Frailty model							Independent model						
T_1	T_2		k	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	-0.006	-0.16	0.17	0.242	0.23	0.26	0.97	-0.254	-0.39	-0.12	0.206	0.20	0.22	0.74
			2	0.016	-0.13	0.16	0.232	0.22	0.25	0.95	-0.137	-0.28	0.00	0.205	0.19	0.22	0.87
			10	0.035	-0.11	0.18	0.222	0.20	0.24	0.94	-0.039	-0.17	0.11	0.202	0.19	0.22	0.95
		HBL	1	0.033	-0.13	0.20	0.239	0.22	0.26	0.94	-0.242	-0.39	-0.08	0.207	0.20	0.22	0.75
			2	-0.017	-0.15	0.14	0.227	0.21	0.24	0.95	-0.147	-0.27	-0.02	0.203	0.19	0.22	0.88
			10	0.014	-0.12	0.18	0.215	0.20	0.23	0.93	-0.032	-0.17	0.12	0.202	0.19	0.21	0.94
	LP^2	LBL	1	0.028	-0.13	0.18	0.248	0.23	0.26	0.96	-0.239	-0.38	-0.11	0.209	0.20	0.22	0.76
			2	0.040	-0.14	0.22	0.238	0.22	0.26	0.95	-0.140	-0.29	0.02	0.206	0.19	0.22	0.86
			10	0.042	-0.08	0.18	0.220	0.20	0.24	0.95	-0.021	-0.15	0.11	0.202	0.19	0.21	0.96
		HBL	1	0.040	-0.13	0.19	0.245	0.23	0.26	0.95	-0.242	-0.38	-0.10	0.209	0.20	0.22	0.76
			2	0.027	-0.13	0.18	0.232	0.22	0.25	0.94	-0.132	-0.27	0.01	0.205	0.19	0.22	0.88
			10	0.023	-0.14	0.18	0.215	0.20	0.23	0.92	-0.038	-0.17	0.12	0.199	0.19	0.21	0.92
LP^1	HP^2	LBL	1	-0.005	-0.16	0.18	0.255	0.23	0.28	0.95	-0.131	-0.26	0.01	0.223	0.21	0.23	0.90
			2	0.033	-0.13	0.20	0.244	0.23	0.27	0.92	-0.048	-0.19	0.09	0.217	0.21	0.23	0.93
			10	0.055	-0.11	0.20	0.239	0.22	0.26	0.92	0.005	-0.17	0.13	0.218	0.21	0.23	0.94
		HBL	1	0.021	-0.15	0.20	0.253	0.23	0.27	0.94	-0.116	-0.27	0.05	0.223	0.21	0.23	0.91
			2	0.012	-0.15	0.17	0.240	0.22	0.26	0.95	-0.067	-0.23	0.09	0.220	0.21	0.23	0.93
			10	0.040	-0.12	0.20	0.228	0.21	0.25	0.92	-0.009	-0.14	0.15	0.218	0.21	0.23	0.95
	LP^2	LBL	1	0.011	-0.15	0.18	0.258	0.24	0.28	0.95	-0.121	-0.26	0.03	0.223	0.21	0.24	0.92
			2	0.045	-0.11	0.21	0.249	0.23	0.28	0.93	-0.036	-0.20	0.11	0.221	0.21	0.23	0.94
			10	0.062	-0.10	0.24	0.242	0.22	0.26	0.91	0.008	-0.14	0.16	0.219	0.21	0.23	0.94
		HBL	1	0.004	-0.15	0.16	0.256	0.23	0.28	0.96	-0.135	-0.26	0.01	0.223	0.21	0.23	0.93
			2	0.032	-0.13	0.20	0.241	0.22	0.26	0.93	-0.054	-0.20	0.11	0.219	0.21	0.23	0.93
			10	0.041	-0.09	0.20	0.233	0.21	0.26	0.93	0.003	-0.14	0.16	0.218	0.21	0.23	0.95

Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low (ν_2=0.00724,φ_2=1.14) and high (ν_2=0.00324,φ_2=1.84) baselines for second event, respectively.

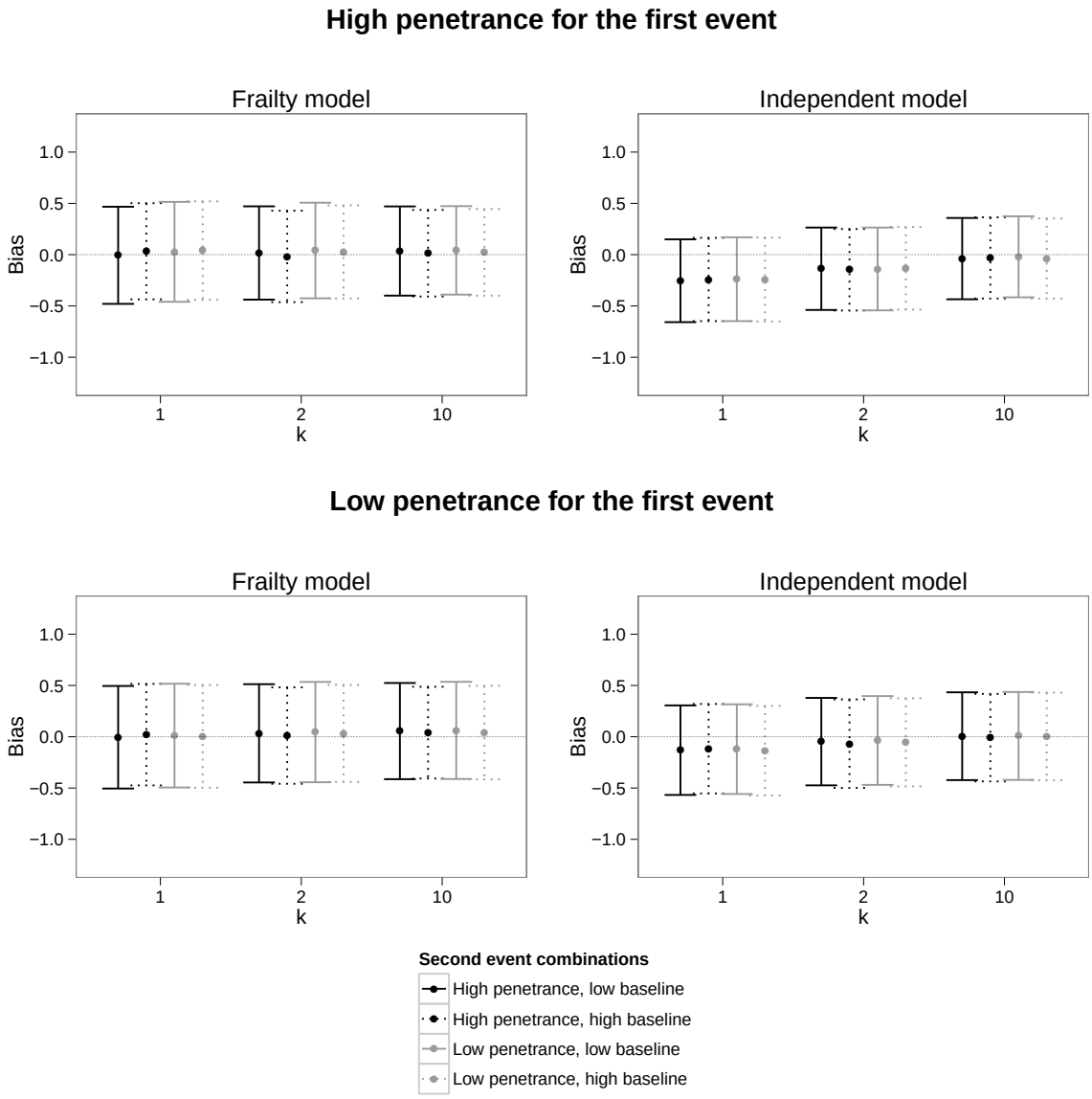

Figure 3.3: Bias and its 95% confidence interval in the log genetic relative risk estimation of the first event (β_2) from frailty model (left) and independent model (right), respectively, under high penetrance (top) and low penetrance (bottom) for the first event using the dominant genetic model with rare allele frequency (2%) with a sample size of 200 families.

Table 3.2: Estimation of log relative genetic risk (β_3) of developing the second event under the dominant genetic model with rare allele frequency ($q = 2\%$) using 200 simulated families.

Parameters				Frailty model							Independent model						
T_1	T_2		k	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	0.028	-0.31	0.44	0.502	0.44	0.59	0.96	-0.151	-0.43	0.25	0.464	0.40	0.56	0.92
			2	0.033	-0.33	0.46	0.546	0.46	0.64	0.95	-0.074	-0.42	0.34	0.519	0.44	0.61	0.94
			10	0.013	-0.33	0.44	0.573	0.49	0.73	0.95	-0.010	-0.37	0.41	0.562	0.48	0.71	0.97
		HBL	1	-0.002	-0.27	0.36	0.405	0.35	0.47	0.93	-0.215	-0.45	0.07	0.332	0.29	0.39	0.84
			2	0.001	-0.25	0.30	0.396	0.34	0.45	0.93	-0.143	-0.38	0.12	0.350	0.30	0.40	0.89
			10	-0.016	-0.29	0.29	0.387	0.33	0.44	0.90	-0.080	-0.34	0.22	0.367	0.31	0.42	0.91
	LP^2	LBL	1	-0.024	-0.36	0.34	0.498	0.44	0.58	0.95	-0.159	-0.44	0.21	0.460	0.40	0.54	0.93
			2	0.029	-0.26	0.46	0.549	0.47	0.65	0.96	-0.046	-0.34	0.39	0.523	0.45	0.62	0.95
			10	0.017	-0.32	0.46	0.585	0.49	0.73	0.92	0.005	-0.35	0.41	0.579	0.49	0.72	0.95
		HBL	1	0.024	-0.26	0.30	0.396	0.36	0.45	0.92	-0.123	-0.37	0.11	0.336	0.30	0.39	0.89
			2	0.015	-0.26	0.31	0.389	0.35	0.45	0.94	-0.093	-0.32	0.19	0.347	0.31	0.40	0.91
			10	0.012	-0.25	0.26	0.386	0.34	0.46	0.94	-0.032	-0.27	0.22	0.364	0.32	0.42	0.95
LP^1	HP^2	LBL	1	0.046	-0.24	0.43	0.512	0.44	0.59	0.95	-0.047	-0.33	0.30	0.470	0.41	0.55	0.95
			2	0.062	-0.29	0.48	0.541	0.47	0.65	0.93	-0.021	-0.32	0.43	0.517	0.45	0.61	0.95
			10	0.022	-0.32	0.56	0.586	0.49	0.73	0.92	-0.003	-0.34	0.51	0.581	0.49	0.72	0.95
		HBL	1	-0.023	-0.32	0.25	0.402	0.35	0.46	0.93	-0.157	-0.40	0.12	0.335	0.30	0.39	0.88
			2	-0.040	-0.29	0.30	0.402	0.35	0.47	0.96	-0.133	-0.36	0.15	0.350	0.31	0.41	0.93
			10	0.014	-0.24	0.31	0.399	0.35	0.47	0.92	-0.030	-0.28	0.26	0.372	0.33	0.43	0.94
	LP^2	LBL	1	0.002	-0.31	0.36	0.505	0.44	0.57	0.94	-0.040	-0.33	0.29	0.465	0.41	0.54	0.94
			2	0.040	-0.33	0.46	0.545	0.48	0.69	0.94	-0.003	-0.35	0.44	0.528	0.46	0.64	0.95
			10	0.036	-0.32	0.49	0.615	0.52	0.76	0.93	0.016	-0.32	0.45	0.602	0.51	0.74	0.95
		HBL	1	-0.020	-0.28	0.27	0.401	0.35	0.47	0.94	-0.085	-0.30	0.16	0.348	0.31	0.40	0.93
			2	-0.000	-0.29	0.32	0.405	0.35	0.47	0.93	-0.047	-0.31	0.23	0.360	0.31	0.42	0.92
			10	-0.008	-0.29	0.31	0.407	0.35	0.48	0.92	-0.023	-0.31	0.28	0.376	0.33	0.44	0.93

Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low (ν_2=0.00724,φ_2=1.14) and high (ν_2=0.00324,φ_2=1.84) baselines for second event, respectively.

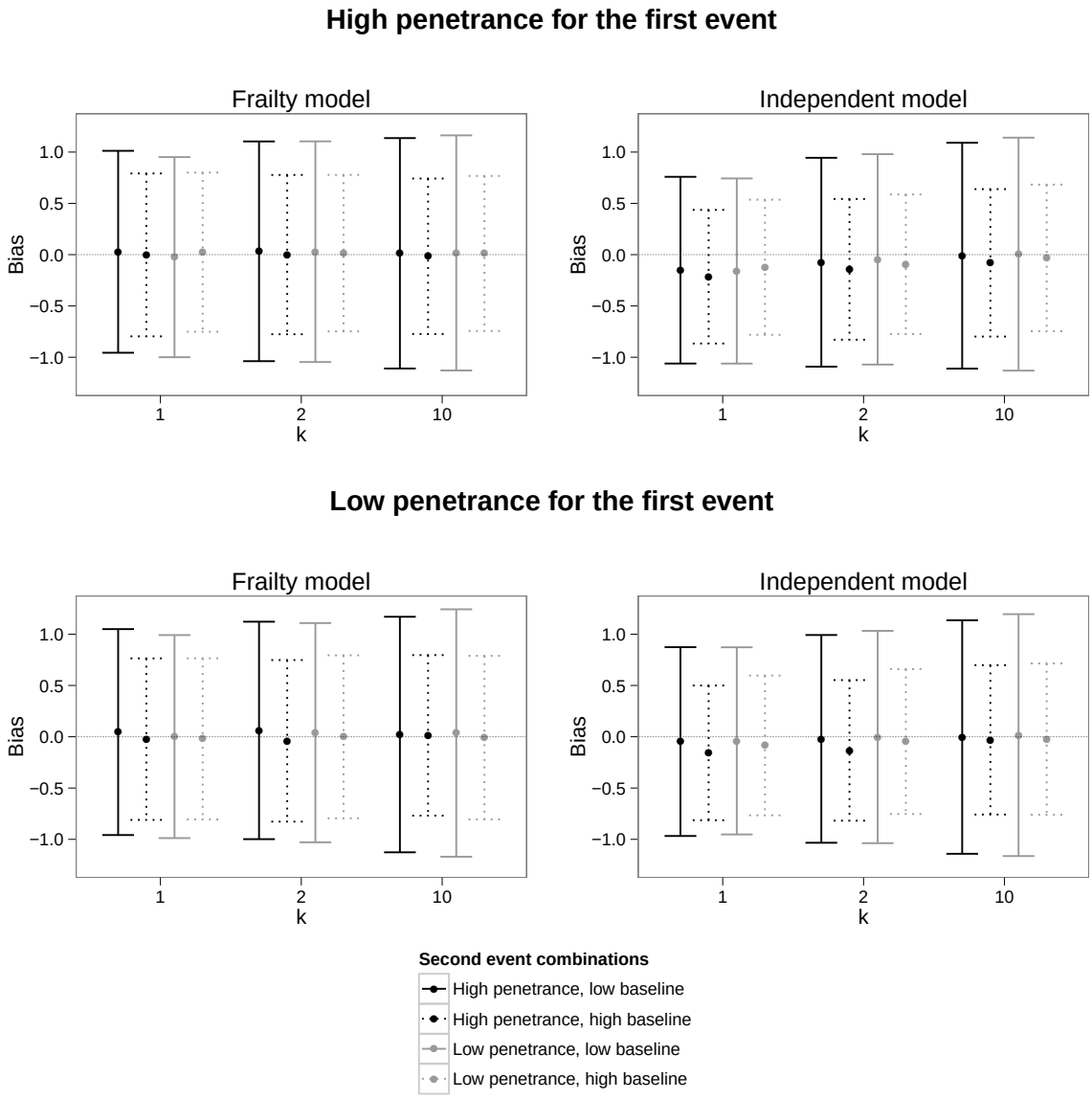

Figure 3.4: Bias and 95% confidence interval of the bias in the log genetic relative risk estimation of the second event (β_3) from frailty model (left) and independent model (right), respectively, under high penetrance (top) and low penetrance (bottom) for the first event using the dominant genetic model with rare allele frequency (2%) with a sample size of 200 families.

Table 3.3: Penetrance estimation of male mutation carriers for the first event by the age of 70 years under the dominant genetic model with rare allele frequency ($q = 2\%$) using 200 simulated families.

Parameters				Pen	Frailty model							Independent model						
T_1	T_2		k	(%)	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	66	0.005	-0.03	0.03	0.046	0.04	0.05	0.94	0.063	0.04	0.10	0.040	0.04	0.04	0.60
			2	74	-0.001	-0.03	0.03	0.044	0.04	0.05	0.93	0.042	0.02	0.07	0.036	0.03	0.04	0.74
			10	83	-0.006	-0.03	0.02	0.039	0.03	0.04	0.94	0.011	-0.01	0.03	0.031	0.03	0.03	0.93
		HBL	1	66	-0.000	-0.03	0.04	0.044	0.04	0.05	0.94	0.064	0.04	0.09	0.040	0.04	0.04	0.63
			2	74	0.001	-0.02	0.03	0.042	0.04	0.04	0.93	0.039	0.02	0.07	0.036	0.03	0.04	0.77
			10	83	-0.005	-0.03	0.02	0.037	0.03	0.04	0.95	0.007	-0.01	0.03	0.031	0.03	0.03	0.92
	LP^2	LBL	1	66	-0.000	-0.03	0.03	0.046	0.04	0.05	0.93	0.061	0.04	0.09	0.040	0.04	0.04	0.64
			2	74	-0.001	-0.03	0.03	0.045	0.04	0.05	0.92	0.039	0.02	0.07	0.036	0.03	0.04	0.76
			10	83	-0.003	-0.02	0.02	0.040	0.03	0.04	0.94	0.013	-0.00	0.03	0.030	0.03	0.03	0.91
		HBL	1	66	0.002	-0.03	0.03	0.045	0.04	0.05	0.95	0.062	0.04	0.09	0.040	0.04	0.04	0.65
			2	74	0.002	-0.02	0.03	0.043	0.04	0.04	0.94	0.040	0.02	0.07	0.036	0.03	0.04	0.76
			10	83	-0.006	-0.03	0.02	0.037	0.03	0.04	0.94	0.011	-0.01	0.03	0.030	0.03	0.03	0.90
LP^1	HP^2	LBL	1	43	-0.008	-0.04	0.03	0.049	0.05	0.05	0.95	0.025	-0.00	0.06	0.046	0.04	0.05	0.91
			2	47	-0.005	-0.04	0.03	0.050	0.05	0.05	0.94	0.019	-0.01	0.05	0.047	0.04	0.05	0.92
			10	52	-0.014	-0.05	0.02	0.051	0.05	0.06	0.91	0.003	-0.04	0.03	0.047	0.04	0.05	0.91
		HBL	1	43	-0.002	-0.04	0.03	0.047	0.04	0.05	0.93	0.031	-0.01	0.06	0.046	0.04	0.05	0.88
			2	47	-0.008	-0.04	0.03	0.049	0.05	0.05	0.94	0.011	-0.01	0.05	0.047	0.04	0.05	0.94
			10	52	-0.003	-0.04	0.03	0.049	0.05	0.05	0.92	0.008	-0.03	0.04	0.047	0.04	0.05	0.94
	LP^2	LBL	1	43	-0.003	-0.04	0.04	0.049	0.05	0.05	0.94	0.027	-0.00	0.06	0.046	0.04	0.05	0.87
			2	47	-0.002	-0.04	0.04	0.051	0.05	0.06	0.92	0.021	-0.01	0.05	0.047	0.04	0.05	0.92
			10	52	-0.012	-0.05	0.02	0.052	0.05	0.06	0.94	0.001	-0.03	0.03	0.047	0.04	0.05	0.94
		HBL	1	43	-0.008	-0.04	0.03	0.048	0.05	0.05	0.96	0.024	-0.01	0.06	0.046	0.04	0.05	0.92
			2	47	-0.002	-0.04	0.03	0.049	0.05	0.05	0.93	0.022	-0.01	0.06	0.047	0.04	0.05	0.94
			10	52	-0.004	-0.04	0.03	0.049	0.05	0.05	0.92	0.009	-0.03	0.04	0.047	0.04	0.05	0.92

Pen - penetrance; Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low (ν_2=0.00724,φ_2=1.14) and high (ν_2=0.00324,φ_2=1.84) baselines for second event, respectively.

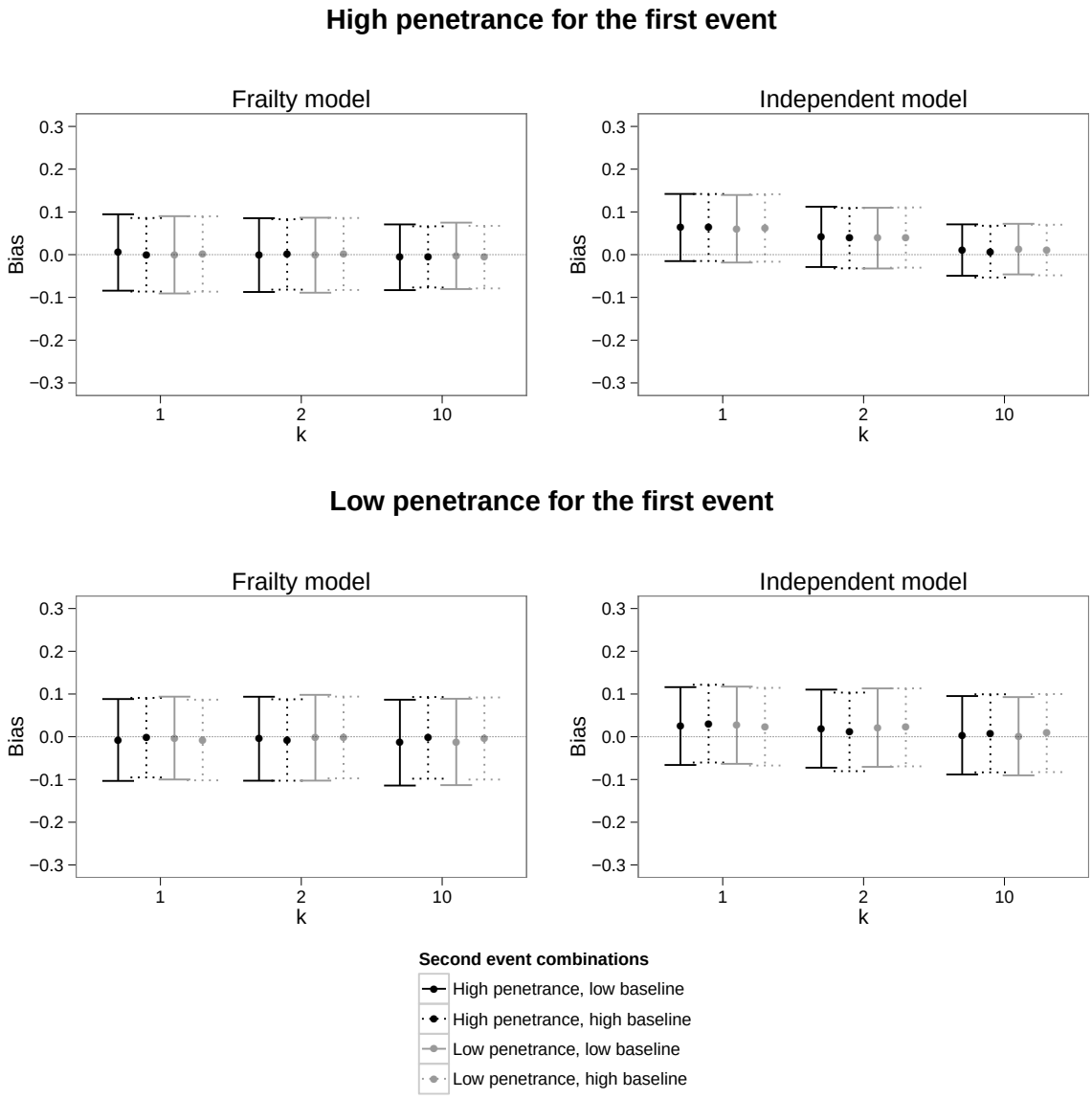

Figure 3.5: Bias and its 95% confidence interval in the first event penetrance estimation at age 70 years for male mutation carriers from frailty model (left) and independent model (right), respectively, under high penetrance (top) and low penetrance (bottom) for the first event using the dominant genetic model with rare allele frequency (2%) with a sample size of 200 families.

Table 3.4: Penetrance estimation of female mutation carriers for the first event by the age of 70 years under the dominant genetic model with rare allele frequency ($q = 2\%$) using 200 simulated families.

Parameters				Pen	Frailty model							Independent model						
T_1	T_2		k	(%)	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	37	-0.003	-0.02	0.03	0.040	0.04	0.04	0.95	-0.002	-0.03	0.04	0.044	0.04	0.04	0.95
			2	40	-0.005	-0.02	0.03	0.043	0.04	0.04	0.95	-0.003	-0.02	0.03	0.045	0.04	0.05	0.95
			10	44	-0.005	-0.03	0.03	0.045	0.04	0.05	0.95	-0.006	-0.03	0.03	0.046	0.04	0.05	0.96
		HBL	1	37	0.003	-0.03	0.03	0.040	0.04	0.04	0.95	0.003	-0.03	0.04	0.044	0.04	0.05	0.95
			2	40	-0.004	-0.03	0.03	0.043	0.04	0.04	0.92	-0.003	-0.03	0.03	0.045	0.04	0.05	0.92
			10	44	-0.003	-0.03	0.02	0.045	0.04	0.05	0.92	-0.003	-0.03	0.02	0.046	0.04	0.05	0.94
	LP^2	LBL	1	37	-0.002	-0.03	0.02	0.040	0.04	0.04	0.94	-0.001	-0.02	0.03	0.044	0.04	0.04	0.95
			2	40	-0.002	-0.03	0.03	0.043	0.04	0.04	0.94	-0.001	-0.03	0.03	0.045	0.04	0.05	0.93
			10	44	0.001	-0.03	0.02	0.045	0.04	0.05	0.96	0.002	-0.03	0.02	0.045	0.04	0.05	0.96
		HBL	1	37	-0.001	-0.03	0.03	0.040	0.04	0.04	0.95	-0.001	-0.03	0.03	0.044	0.04	0.04	0.95
			2	40	-0.000	-0.02	0.03	0.043	0.04	0.04	0.95	-0.000	-0.02	0.03	0.045	0.04	0.05	0.95
			10	44	0.003	-0.04	0.03	0.045	0.04	0.05	0.92	0.002	-0.04	0.03	0.045	0.04	0.05	0.93
LP^1	HP^2	LBL	1	19	-0.002	-0.03	0.01	0.032	0.03	0.04	0.94	-0.003	-0.03	0.01	0.034	0.03	0.04	0.93
			2	19	-0.001	-0.02	0.03	0.034	0.03	0.04	0.94	-0.002	-0.02	0.03	0.035	0.03	0.04	0.93
			10	20	-0.000	-0.02	0.03	0.035	0.03	0.04	0.95	-0.000	-0.02	0.03	0.036	0.03	0.04	0.96
		HBL	1	19	0.001	-0.03	0.02	0.033	0.03	0.04	0.94	0.001	-0.03	0.02	0.035	0.03	0.04	0.94
			2	19	-0.004	-0.02	0.03	0.034	0.03	0.04	0.95	-0.006	-0.02	0.03	0.035	0.03	0.04	0.95
			10	20	-0.002	-0.02	0.03	0.036	0.03	0.04	0.92	-0.003	-0.02	0.03	0.036	0.03	0.04	0.93
	LP^2	LBL	1	19	-0.003	-0.03	0.02	0.033	0.03	0.04	0.94	-0.004	-0.03	0.02	0.034	0.03	0.04	0.94
			2	19	0.001	-0.02	0.03	0.034	0.03	0.04	0.93	0.000	-0.02	0.03	0.035	0.03	0.04	0.93
			10	20	0.000	-0.02	0.03	0.036	0.03	0.04	0.93	-0.001	-0.02	0.03	0.036	0.03	0.04	0.94
		HBL	1	19	-0.002	-0.03	0.02	0.033	0.03	0.04	0.95	-0.005	-0.03	0.02	0.034	0.03	0.04	0.95
			2	19	0.002	-0.01	0.03	0.034	0.03	0.04	0.95	0.001	-0.02	0.03	0.035	0.03	0.04	0.96
			10	20	-0.004	-0.02	0.03	0.036	0.03	0.04	0.91	-0.005	-0.03	0.03	0.036	0.03	0.04	0.92

Pen - penetrance; Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low (ν_2=0.00724,φ_2=1.14) and high (ν_2=0.00324,φ_2=1.84) baselines for second event, respectively.

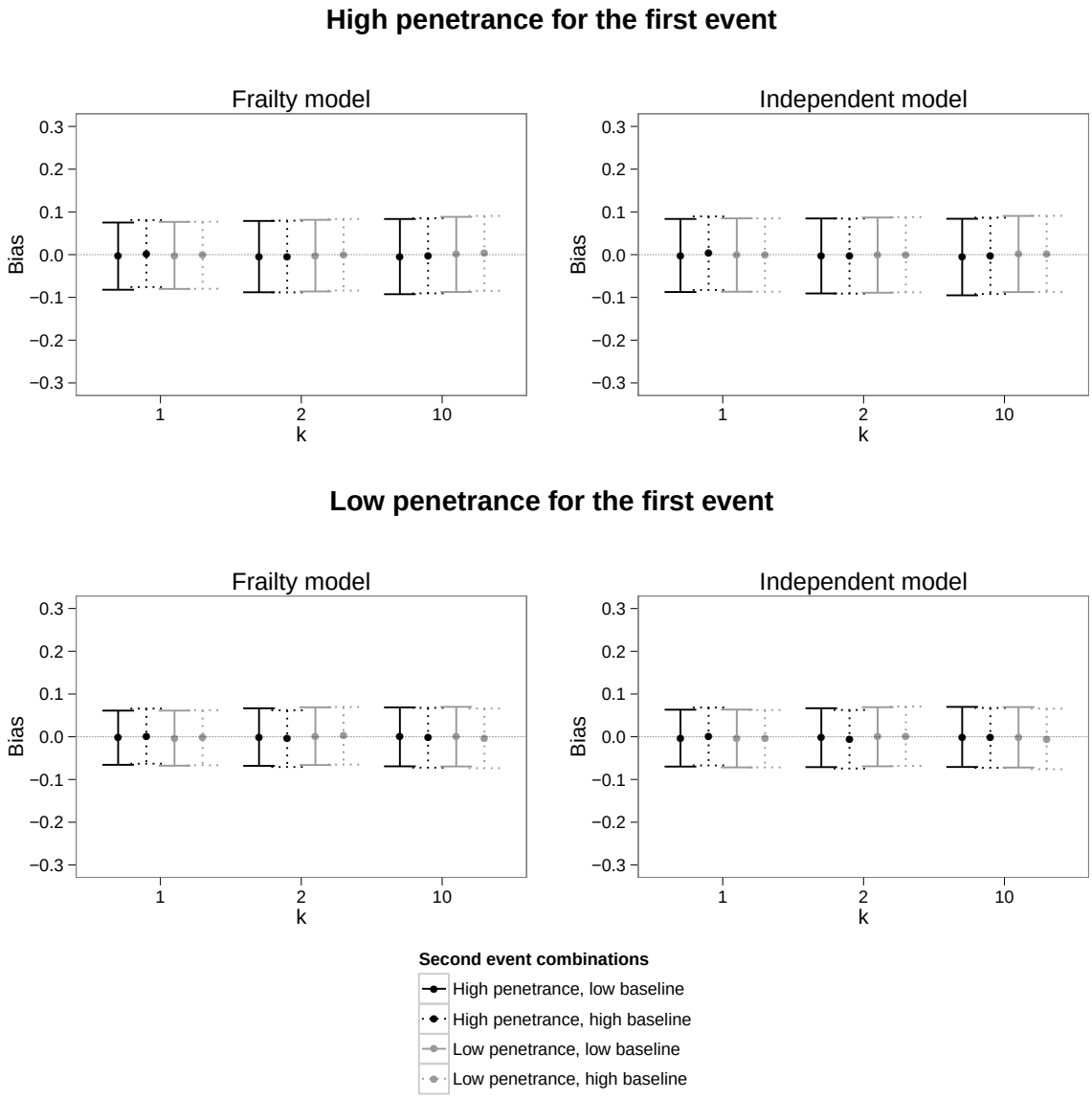

Figure 3.6: Bias and its 95% confidence interval in the first event penetrance estimation at age 70 years for female mutation carriers from frailty model (left) and independent model (right), respectively, under high penetrance (top) and low penetrance (bottom) for the first event using the dominant genetic model with rare allele frequency (2%) with a sample size of 200 families.

Table 3.5: Penetrance estimation of mutation carriers for developing the second event in 10 years after the first event under the dominant genetic model with rare allele frequency ($q = 2\%$) using 200 simulated families.

Parameters				Pen	Frailty model							Independent model						
T_1	T_2		k	(%)	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	17	0.002	-0.02	0.03	0.033	0.03	0.04	0.93	0.105	0.09	0.13	0.025	0.02	0.03	0.01
			2	18	-0.002	-0.02	0.03	0.034	0.03	0.04	0.93	0.057	0.05	0.08	0.024	0.02	0.02	0.28
			10	19	-0.005	-0.03	0.01	0.029	0.02	0.04	0.91	0.011	-0.00	0.02	0.022	0.02	0.02	0.94
		HBL	1	32	0.001	-0.03	0.04	0.048	0.04	0.06	0.92	0.157	0.14	0.18	0.029	0.03	0.03	0.00
			2	35	0.003	-0.03	0.04	0.047	0.04	0.05	0.91	0.088	0.07	0.10	0.028	0.03	0.03	0.11
			10	37	-0.005	-0.03	0.02	0.041	0.03	0.05	0.94	0.019	0.00	0.04	0.027	0.03	0.03	0.92
	LP^2	LBL	1	12	-0.002	-0.02	0.02	0.024	0.02	0.03	0.92	0.077	0.06	0.09	0.022	0.02	0.02	0.04
			2	12	-0.002	-0.01	0.02	0.025	0.02	0.03	0.92	0.043	0.03	0.06	0.020	0.02	0.02	0.44
			10	13	-0.006	-0.02	0.00	0.022	0.02	0.03	0.92	0.009	-0.01	0.02	0.018	0.02	0.02	0.95
		HBL	1	23	-0.002	-0.03	0.03	0.038	0.03	0.05	0.93	0.125	0.11	0.15	0.028	0.03	0.03	0.00
			2	25	-0.001	-0.03	0.02	0.038	0.03	0.04	0.90	0.070	0.04	0.08	0.026	0.02	0.03	0.26
			10	26	-0.004	-0.03	0.01	0.034	0.03	0.04	0.94	0.016	-0.00	0.03	0.024	0.02	0.02	0.91
LP^1	HP^2	LBL	1	17	-0.002	-0.04	0.07	0.053	0.04	0.08	0.76	0.128	0.11	0.15	0.031	0.03	0.03	0.01
			2	18	-0.003	-0.05	0.05	0.049	0.03	0.07	0.73	0.067	0.05	0.09	0.029	0.03	0.03	0.35
			10	19	-0.020	-0.06	0.01	0.041	0.03	0.06	0.83	0.014	-0.01	0.03	0.026	0.02	0.03	0.93
		HBL	1	32	0.007	-0.07	0.08	0.081	0.06	0.10	0.80	0.187	0.16	0.21	0.035	0.03	0.04	0.00
			2	35	-0.006	-0.08	0.07	0.075	0.06	0.10	0.75	0.103	0.08	0.12	0.034	0.03	0.04	0.15
			10	37	-0.004	-0.05	0.04	0.058	0.03	0.08	0.87	0.025	0.00	0.05	0.032	0.03	0.03	0.87
	LP^2	LBL	1	12	0.001	-0.03	0.05	0.040	0.03	0.06	0.74	0.093	0.07	0.11	0.027	0.03	0.03	0.05
			2	12	0.001	-0.03	0.04	0.034	0.02	0.05	0.72	0.048	0.03	0.07	0.025	0.02	0.03	0.50
			10	13	-0.015	-0.05	0.00	0.028	0.02	0.05	0.84	0.009	-0.01	0.02	0.022	0.02	0.02	0.93
		HBL	1	23	-0.010	-0.05	0.07	0.066	0.05	0.09	0.82	0.148	0.13	0.17	0.033	0.03	0.04	0.00
			2	25	-0.004	-0.06	0.06	0.058	0.04	0.08	0.72	0.083	0.06	0.10	0.031	0.03	0.03	0.25
			10	26	-0.008	-0.06	0.02	0.047	0.03	0.07	0.85	0.017	-0.00	0.04	0.029	0.03	0.03	0.93

Pen - penetrance; Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low (ν_2=0.00724,φ_2=1.14) and high (ν_2=0.00324,φ_2=1.84) baselines for second event, respectively.

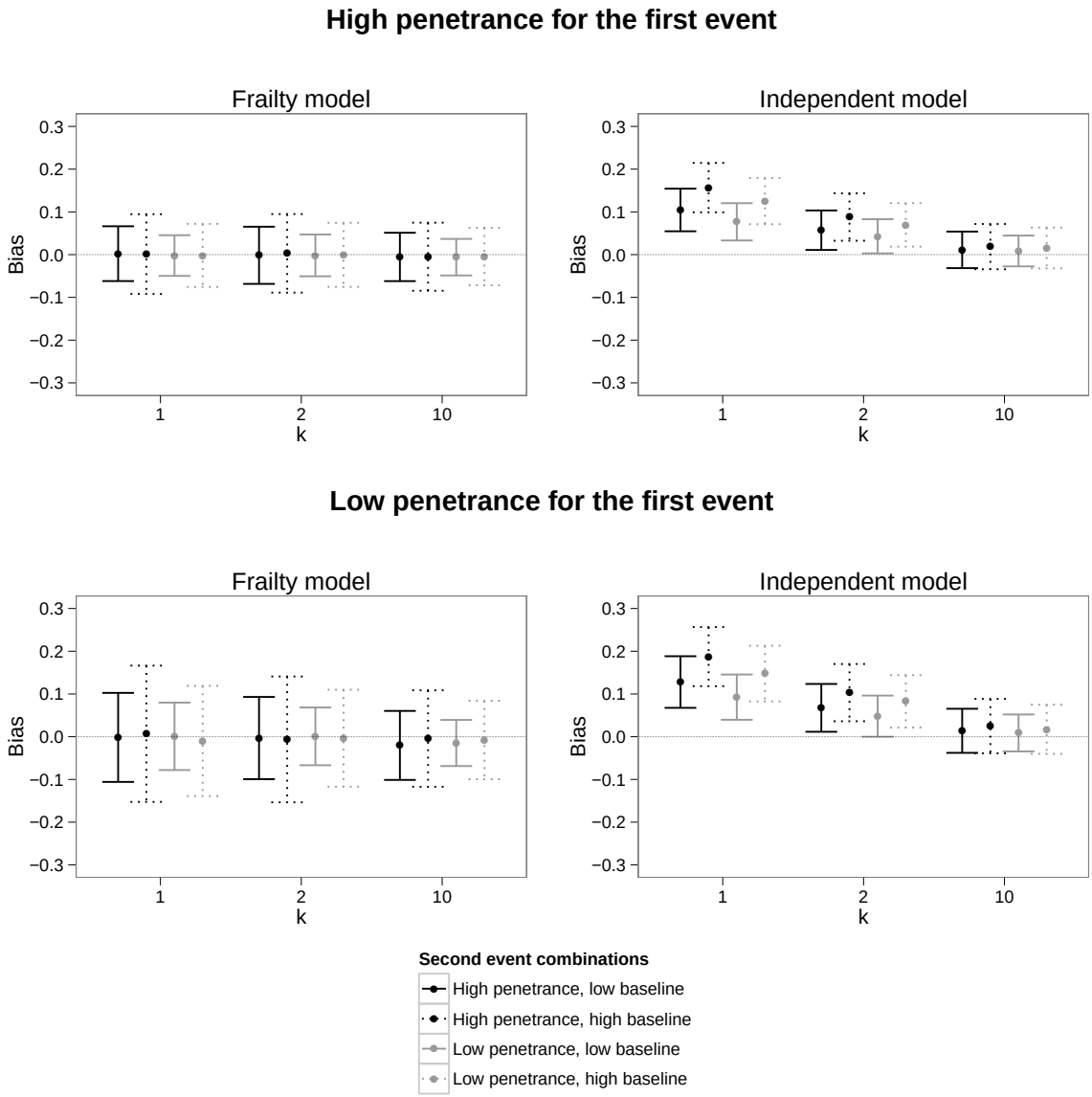

Figure 3.7: Bias and its 95% confidence interval in the 10-year penetrance estimation of the second event for mutation carriers from frailty model (left) and independent model (right), respectively, under high penetrance (top) and low penetrance (bottom) for the first event using the dominant genetic model with rare allele frequency (2%) with a sample size of 200 families.

Table 3.6: Estimation of log relative genetic risk (β_2) of developing the first event under the recessive genetic model with common allele frequency ($q = 30\%$) using 200 simulated families.

Parameters				Frailty model							Independent model						
T_1	T_2		k	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	0.039	-0.16	0.19	0.241	0.23	0.26	0.95	-0.249	-0.39	-0.10	0.204	0.19	0.22	0.73
			2	0.022	-0.15	0.20	0.232	0.22	0.25	0.93	-0.140	-0.28	0.02	0.203	0.19	0.21	0.86
			10	0.033	-0.11	0.19	0.220	0.20	0.24	0.95	-0.044	-0.16	0.12	0.201	0.19	0.21	0.94
		HBL	1	0.025	-0.13	0.18	0.240	0.23	0.25	0.94	-0.234	-0.37	-0.09	0.205	0.20	0.22	0.76
			2	0.017	-0.13	0.16	0.223	0.21	0.24	0.94	-0.146	-0.26	0.00	0.201	0.19	0.21	0.89
			10	0.021	-0.11	0.16	0.214	0.20	0.23	0.95	-0.024	-0.15	0.11	0.200	0.19	0.21	0.96
	LP^2	LBL	1	0.014	-0.14	0.20	0.247	0.23	0.26	0.96	-0.247	-0.39	-0.10	0.206	0.19	0.22	0.75
			2	0.021	-0.14	0.17	0.233	0.22	0.25	0.96	-0.143	-0.29	-0.01	0.203	0.19	0.21	0.88
			10	0.021	-0.12	0.18	0.221	0.20	0.24	0.92	-0.039	-0.17	0.09	0.200	0.19	0.21	0.94
		HBL	1	0.002	-0.15	0.19	0.244	0.23	0.26	0.95	-0.240	-0.40	-0.09	0.205	0.19	0.22	0.73
			2	-0.000	-0.13	0.15	0.226	0.21	0.24	0.96	-0.142	-0.28	-0.02	0.201	0.19	0.21	0.88
			10	0.019	-0.11	0.16	0.213	0.20	0.23	0.94	-0.045	-0.17	0.10	0.198	0.19	0.21	0.95
LP^1	HP^2	LBL	1	0.029	-0.13	0.19	0.259	0.24	0.28	0.94	-0.107	-0.24	0.04	0.224	0.21	0.23	0.94
			2	0.021	-0.14	0.18	0.244	0.22	0.27	0.93	-0.062	-0.22	0.07	0.219	0.21	0.23	0.92
			10	0.039	-0.09	0.20	0.237	0.21	0.26	0.91	-0.008	-0.14	0.13	0.216	0.20	0.23	0.95
		HBL	1	0.047	-0.11	0.20	0.251	0.23	0.27	0.94	-0.088	-0.23	0.06	0.224	0.21	0.23	0.93
			2	0.008	-0.13	0.16	0.242	0.22	0.26	0.94	-0.081	-0.20	0.06	0.218	0.21	0.23	0.93
			10	0.034	-0.12	0.22	0.231	0.21	0.25	0.93	0.002	-0.15	0.15	0.216	0.20	0.23	0.94
	LP^2	LBL	1	0.020	-0.16	0.19	0.260	0.23	0.28	0.94	-0.115	-0.24	0.04	0.222	0.21	0.24	0.92
			2	0.039	-0.12	0.20	0.246	0.23	0.27	0.92	-0.053	-0.20	0.11	0.219	0.21	0.23	0.94
			10	0.062	-0.10	0.21	0.240	0.22	0.26	0.94	-0.005	-0.14	0.15	0.217	0.21	0.23	0.96
		HBL	1	0.013	-0.12	0.18	0.253	0.23	0.28	0.94	-0.119	-0.26	0.03	0.222	0.21	0.23	0.93
			2	0.033	-0.15	0.20	0.241	0.22	0.26	0.94	-0.045	-0.21	0.10	0.219	0.21	0.23	0.94
			10	0.014	-0.13	0.19	0.229	0.21	0.25	0.92	-0.026	-0.16	0.13	0.214	0.20	0.23	0.94

Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low (ν_2=0.00724,φ_2=1.14) and high (ν_2=0.00324,φ_2=1.84) baselines for second event, respectively.

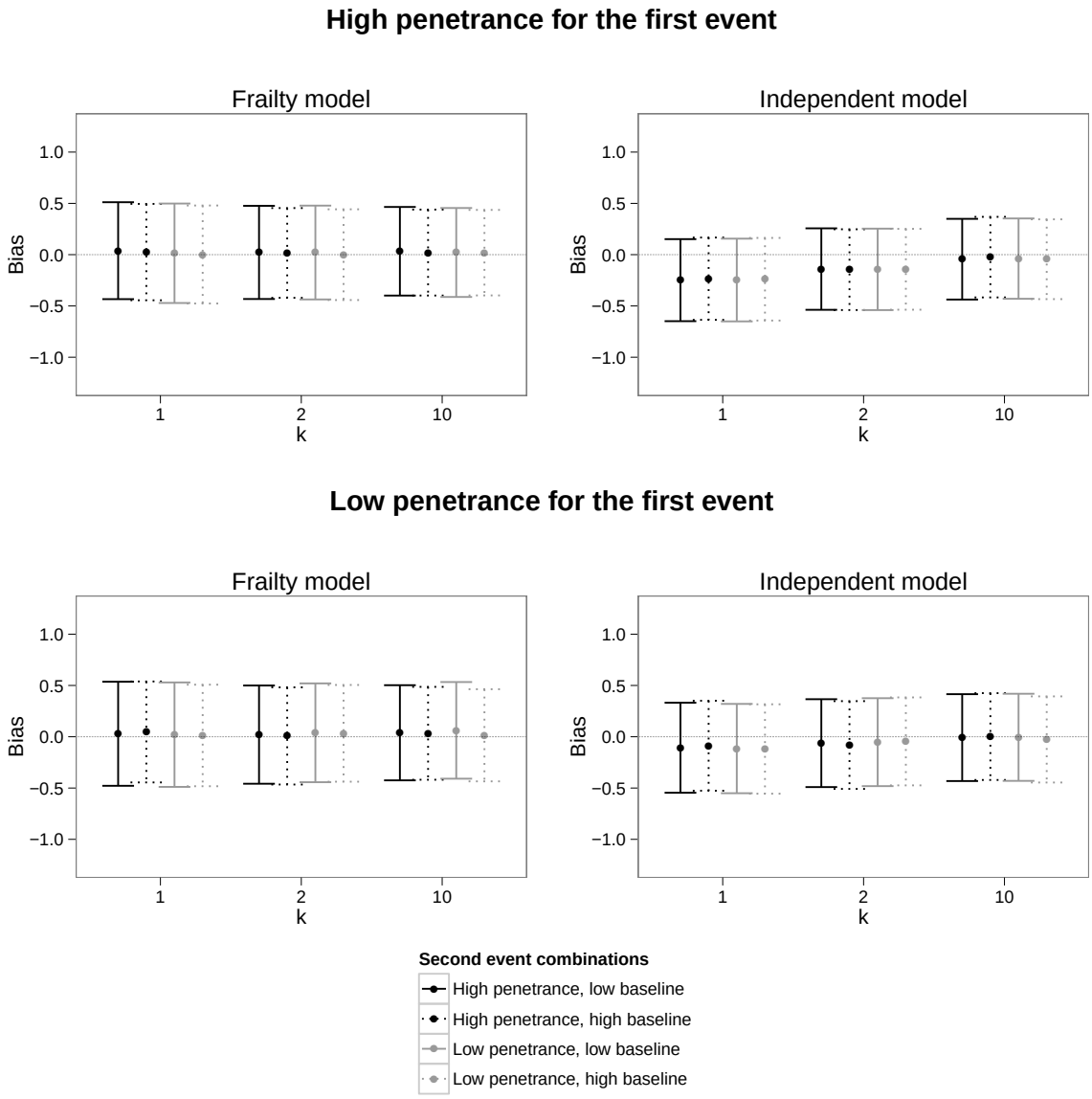

Figure 3.8: Bias and its 95% confidence interval in the log genetic relative risk estimation of the first event (β_2) from frailty model (left) and independent model (right), respectively, under high penetrance (top) and low penetrance (bottom) for the first event using the recessive genetic model with common allele frequency (30%) with a sample size of 200 families.

Table 3.7: Estimation of log relative genetic risk (β_3) of developing the second event under the recessive genetic model with common allele frequency ($q = 30\%$) using 200 simulated families.

Parameters				Frailty model							Independent model						
T_1	T_2		k	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	0.067	-0.26	0.41	0.496	0.43	0.58	0.95	-0.106	-0.40	0.21	0.449	0.40	0.53	0.93
			2	0.068	-0.29	0.46	0.538	0.46	0.65	0.96	-0.033	-0.37	0.34	0.508	0.44	0.61	0.96
			10	0.056	-0.33	0.43	0.555	0.47	0.69	0.93	0.006	-0.37	0.38	0.549	0.47	0.69	0.95
		HBL	1	-0.003	-0.24	0.33	0.399	0.35	0.45	0.95	-0.211	-0.43	0.06	0.332	0.29	0.38	0.88
			2	-0.011	-0.27	0.22	0.392	0.34	0.44	0.94	-0.156	-0.40	0.08	0.344	0.29	0.39	0.90
			10	-0.013	-0.26	0.26	0.383	0.33	0.45	0.90	-0.050	-0.30	0.21	0.358	0.32	0.42	0.92
	LP^2	LBL	1	0.010	-0.29	0.39	0.504	0.44	0.58	0.98	-0.120	-0.41	0.25	0.466	0.40	0.55	0.94
			2	0.052	-0.28	0.42	0.534	0.47	0.62	0.96	-0.013	-0.34	0.32	0.510	0.44	0.59	0.96
			10	0.110	-0.29	0.51	0.580	0.49	0.73	0.94	0.082	-0.30	0.49	0.569	0.48	0.71	0.96
		HBL	1	-0.018	-0.25	0.29	0.379	0.34	0.44	0.94	-0.157	-0.35	0.09	0.324	0.29	0.38	0.91
			2	0.006	-0.24	0.29	0.390	0.34	0.46	0.93	-0.088	-0.30	0.18	0.352	0.31	0.41	0.91
			10	0.005	-0.26	0.28	0.383	0.33	0.45	0.93	-0.019	-0.27	0.23	0.360	0.32	0.42	0.94
LP^1	HP^2	LBL	1	0.021	-0.31	0.39	0.500	0.43	0.58	0.94	-0.072	-0.37	0.26	0.462	0.40	0.54	0.94
			2	0.009	-0.31	0.43	0.525	0.46	0.62	0.95	-0.043	-0.34	0.36	0.496	0.43	0.60	0.95
			10	0.033	-0.34	0.58	0.576	0.48	0.73	0.88	0.005	-0.35	0.53	0.570	0.48	0.72	0.94
		HBL	1	0.021	-0.27	0.29	0.401	0.35	0.47	0.94	-0.132	-0.36	0.13	0.335	0.30	0.39	0.90
			2	-0.001	-0.24	0.28	0.401	0.35	0.45	0.92	-0.093	-0.31	0.17	0.346	0.30	0.39	0.92
			10	0.045	-0.26	0.33	0.403	0.34	0.47	0.91	-0.019	-0.29	0.26	0.370	0.32	0.43	0.93
	LP^2	LBL	1	-0.018	-0.31	0.36	0.496	0.43	0.59	0.94	-0.054	-0.34	0.29	0.463	0.40	0.54	0.95
			2	-0.012	-0.34	0.41	0.530	0.46	0.62	0.94	-0.031	-0.36	0.36	0.514	0.44	0.60	0.96
			10	0.065	-0.34	0.51	0.590	0.50	0.73	0.95	0.037	-0.35	0.47	0.577	0.49	0.72	0.98
		HBL	1	-0.013	-0.24	0.29	0.398	0.35	0.45	0.94	-0.057	-0.28	0.18	0.339	0.30	0.38	0.92
			2	-0.004	-0.26	0.28	0.399	0.35	0.46	0.94	-0.032	-0.25	0.22	0.358	0.32	0.40	0.95
			10	0.010	-0.27	0.33	0.403	0.35	0.47	0.92	-0.007	-0.28	0.28	0.372	0.33	0.43	0.94

Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low (ν_2=0.00724,φ_2=1.14) and high (ν_2=0.00324,φ_2=1.84) baselines for second event, respectively.

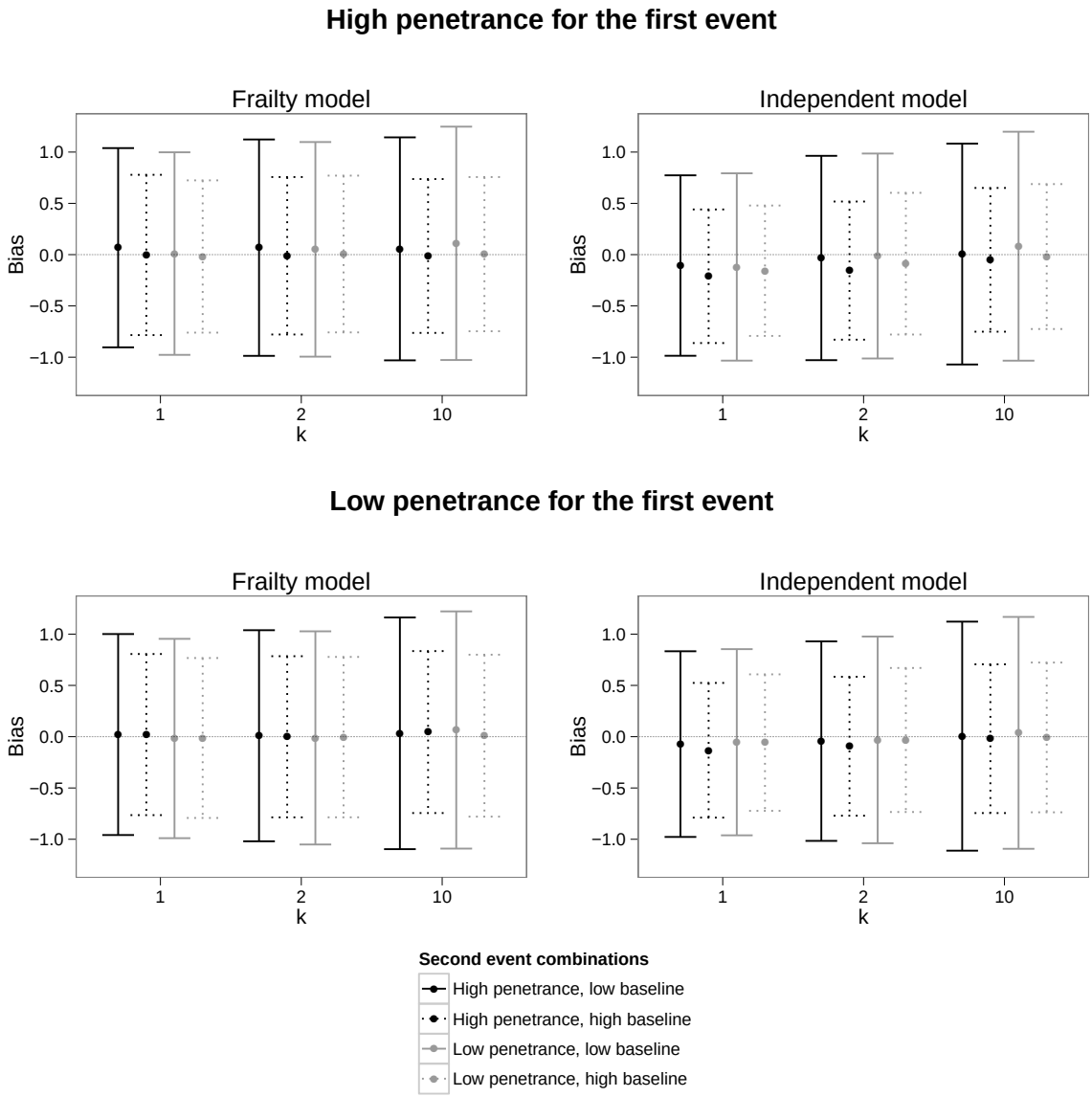

Figure 3.9: Bias and its 95% confidence interval in the log genetic relative risk estimation of the second event (β_3) from frailty model (left) and independent model (right), respectively, under high penetrance (top) and low penetrance (bottom) for the first event using the recessive genetic model with common allele frequency (30%) with a sample size of 200 families.

Table 3.8: Penetrance estimation of male mutation carriers for the first event by the age of 70 years under the recessive genetic model with common allele frequency ($q = 30\%$) using 200 simulated families.

Parameters				Pen	Frailty model							Independent model						
T_1	T_2		k	(%)	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	66	-0.000	-0.03	0.04	0.047	0.04	0.05	0.94	0.067	0.04	0.10	0.041	0.04	0.04	0.60
			2	74	0.001	-0.03	0.03	0.046	0.04	0.05	0.95	0.046	0.02	0.07	0.037	0.03	0.04	0.75
			10	83	-0.008	-0.03	0.02	0.041	0.04	0.04	0.94	0.010	-0.01	0.03	0.031	0.03	0.03	0.93
		HBL	1	66	0.003	-0.03	0.03	0.045	0.04	0.05	0.93	0.070	0.04	0.10	0.040	0.04	0.04	0.58
			2	74	-0.001	-0.02	0.04	0.043	0.04	0.05	0.93	0.044	0.02	0.07	0.036	0.03	0.04	0.74
			10	83	-0.005	-0.03	0.02	0.038	0.03	0.04	0.94	0.011	-0.01	0.03	0.031	0.03	0.03	0.93
	LP^2	LBL	1	66	-0.003	-0.04	0.02	0.048	0.04	0.05	0.97	0.066	0.04	0.09	0.041	0.04	0.04	0.62
			2	74	-0.002	-0.03	0.03	0.047	0.04	0.05	0.94	0.042	0.02	0.07	0.037	0.03	0.04	0.75
			10	83	-0.006	-0.03	0.02	0.042	0.04	0.05	0.94	0.010	-0.01	0.03	0.031	0.03	0.03	0.91
		HBL	1	66	0.003	-0.03	0.03	0.046	0.04	0.05	0.93	0.067	0.04	0.10	0.040	0.04	0.04	0.62
			2	74	0.003	-0.02	0.03	0.044	0.04	0.05	0.93	0.046	0.02	0.07	0.036	0.03	0.04	0.73
			10	83	-0.003	-0.03	0.02	0.038	0.03	0.04	0.94	0.011	-0.01	0.03	0.031	0.03	0.03	0.91
LP^1	HP^2	LBL	1	43	-0.001	-0.04	0.03	0.051	0.05	0.06	0.94	0.031	0.00	0.06	0.048	0.05	0.05	0.90
			2	47	-0.005	-0.04	0.04	0.052	0.05	0.06	0.92	0.021	-0.01	0.06	0.048	0.05	0.05	0.92
			10	52	-0.012	-0.05	0.02	0.052	0.05	0.06	0.92	0.004	-0.04	0.03	0.047	0.05	0.05	0.95
		HBL	1	43	0.006	-0.03	0.03	0.049	0.05	0.05	0.94	0.035	0.01	0.07	0.047	0.04	0.05	0.90
			2	47	-0.010	-0.04	0.02	0.050	0.05	0.05	0.95	0.015	-0.01	0.04	0.048	0.05	0.05	0.94
			10	52	-0.006	-0.05	0.02	0.050	0.05	0.05	0.94	0.007	-0.03	0.04	0.048	0.05	0.05	0.93
	LP^2	LBL	1	43	-0.006	-0.04	0.03	0.052	0.05	0.06	0.93	0.031	-0.01	0.06	0.047	0.05	0.05	0.90
			2	47	-0.003	-0.03	0.03	0.053	0.05	0.06	0.93	0.023	-0.01	0.05	0.048	0.05	0.05	0.94
			10	52	-0.012	-0.05	0.01	0.054	0.05	0.06	0.94	0.002	-0.03	0.03	0.048	0.05	0.05	0.96
		HBL	1	43	-0.001	-0.04	0.03	0.050	0.05	0.05	0.93	0.034	0.00	0.06	0.048	0.05	0.05	0.88
			2	47	-0.001	-0.04	0.03	0.051	0.05	0.05	0.94	0.018	-0.02	0.05	0.048	0.05	0.05	0.93
			10	52	-0.010	-0.05	0.02	0.051	0.05	0.06	0.94	0.005	-0.03	0.03	0.048	0.05	0.05	0.94

Pen - penetrance; Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low (ν_2=0.00724,φ_2=1.14) and high (ν_2=0.00324,φ_2=1.84) baselines for second event, respectively.

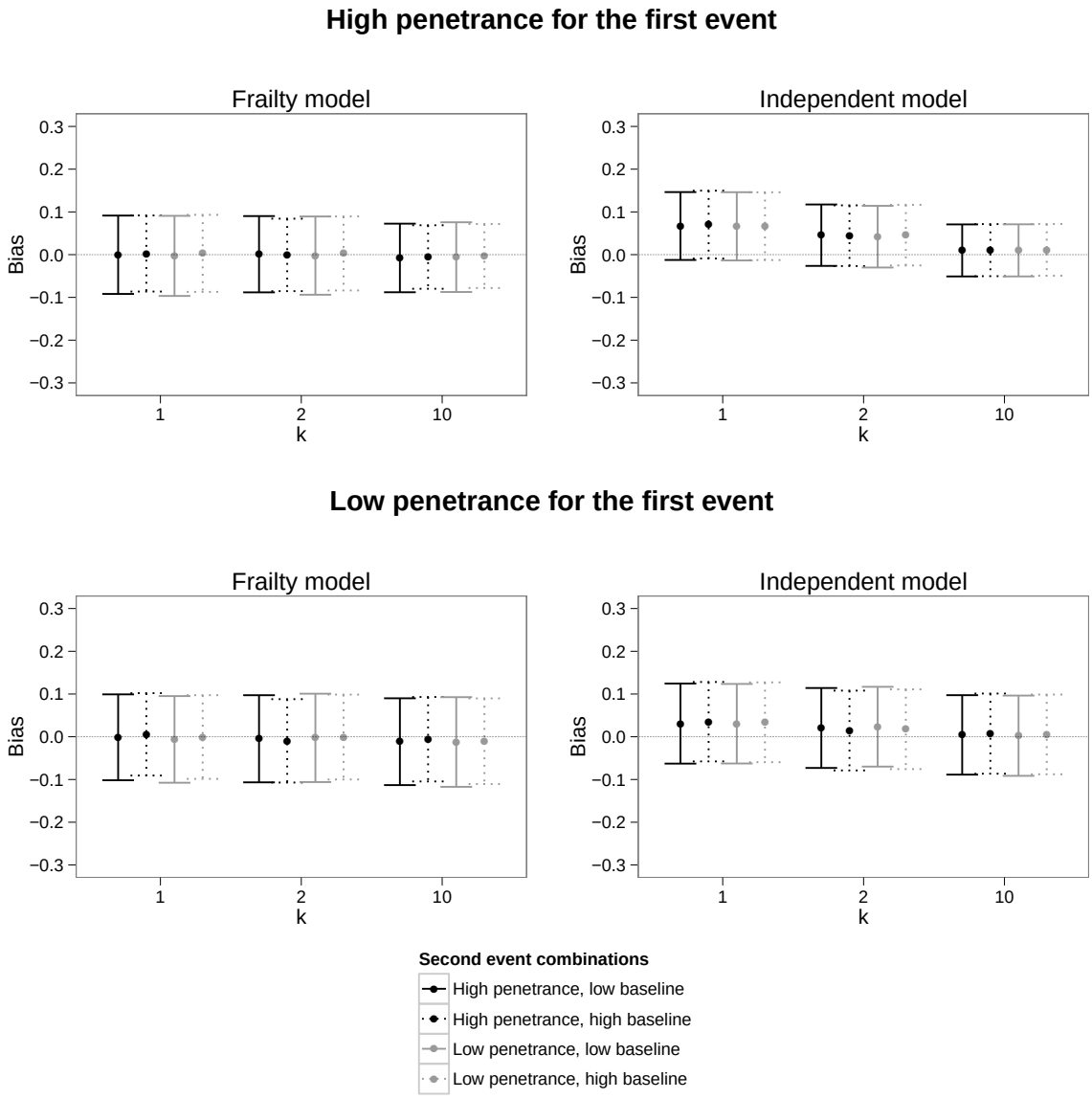

Figure 3.10: Bias and its 95% confidence interval in the first event penetrance estimation at age 70 years for male mutation carriers from frailty model (left) and independent model (right), respectively, under high penetrance (top) and low penetrance (bottom) for the first event using the recessive genetic model with common allele frequency (30%) with a sample size of 200 families.

Table 3.9: Penetrance estimation of female mutation carriers for the first event by the age of 70 years under the recessive genetic model with common allele frequency ($q = 30\%$) using 200 simulated families.

Parameters				Pen	Frailty model							Independent model						
T_1	T_2		k	(%)	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	37	0.002	-0.02	0.03	0.041	0.04	0.04	0.97	0.003	-0.02	0.04	0.045	0.04	0.05	0.96
			2	40	-0.001	-0.03	0.03	0.043	0.04	0.04	0.95	0.000	-0.03	0.04	0.046	0.04	0.05	0.95
			10	44	-0.004	-0.03	0.02	0.045	0.04	0.05	0.95	-0.004	-0.03	0.02	0.046	0.04	0.05	0.95
		HBL	1	37	0.004	-0.02	0.04	0.041	0.04	0.04	0.94	0.007	-0.02	0.04	0.045	0.04	0.05	0.94
			2	40	-0.002	-0.03	0.03	0.043	0.04	0.04	0.95	-0.001	-0.03	0.04	0.046	0.04	0.05	0.95
			10	44	-0.002	-0.04	0.02	0.046	0.04	0.05	0.92	-0.002	-0.04	0.02	0.047	0.04	0.05	0.92
	LP^2	LBL	1	37	-0.003	-0.03	0.02	0.041	0.04	0.04	0.97	-0.000	-0.02	0.03	0.045	0.04	0.05	0.98
			2	40	-0.004	-0.03	0.03	0.043	0.04	0.04	0.94	-0.002	-0.03	0.03	0.046	0.04	0.05	0.93
			10	44	-0.002	-0.04	0.03	0.046	0.04	0.05	0.93	-0.003	-0.04	0.03	0.046	0.04	0.05	0.93
		HBL	1	37	-0.003	-0.02	0.03	0.041	0.04	0.04	0.95	0.001	-0.02	0.04	0.045	0.04	0.05	0.95
			2	40	-0.005	-0.03	0.03	0.043	0.04	0.04	0.96	-0.004	-0.03	0.03	0.045	0.04	0.05	0.95
			10	44	-0.003	-0.04	0.03	0.046	0.04	0.05	0.94	-0.003	-0.04	0.03	0.046	0.04	0.05	0.94
LP^1	HP^2	LBL	1	19	-0.004	-0.03	0.02	0.033	0.03	0.04	0.94	-0.005	-0.03	0.02	0.035	0.03	0.04	0.93
			2	19	0.000	-0.02	0.03	0.035	0.03	0.04	0.95	-0.000	-0.02	0.03	0.036	0.03	0.04	0.95
			10	20	-0.005	-0.03	0.02	0.036	0.03	0.04	0.92	-0.006	-0.03	0.02	0.036	0.03	0.04	0.94
		HBL	1	19	-0.002	-0.03	0.02	0.034	0.03	0.04	0.93	-0.003	-0.03	0.02	0.035	0.03	0.04	0.92
			2	19	-0.001	-0.02	0.03	0.035	0.03	0.04	0.95	-0.003	-0.02	0.03	0.036	0.03	0.04	0.94
			10	20	-0.000	-0.02	0.03	0.037	0.03	0.04	0.93	-0.001	-0.02	0.03	0.037	0.03	0.04	0.92
	LP^2	LBL	1	19	-0.002	-0.02	0.02	0.033	0.03	0.04	0.95	-0.002	-0.03	0.02	0.035	0.03	0.04	0.96
			2	19	-0.002	-0.02	0.03	0.035	0.03	0.04	0.94	-0.002	-0.02	0.03	0.036	0.03	0.04	0.95
			10	20	0.000	-0.02	0.03	0.036	0.03	0.04	0.95	0.000	-0.02	0.03	0.037	0.03	0.04	0.94
		HBL	1	19	0.001	-0.03	0.02	0.033	0.03	0.04	0.94	-0.001	-0.03	0.02	0.035	0.03	0.04	0.94
			2	19	-0.002	-0.02	0.03	0.035	0.03	0.04	0.93	-0.003	-0.02	0.03	0.036	0.03	0.04	0.94
			10	20	-0.004	-0.02	0.03	0.036	0.03	0.04	0.95	-0.005	-0.02	0.03	0.036	0.03	0.04	0.96

Pen - penetrance; Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low (ν_2=0.00724,φ_2=1.14) and high (ν_2=0.00324,φ_2=1.84) baselines for second event, respectively.

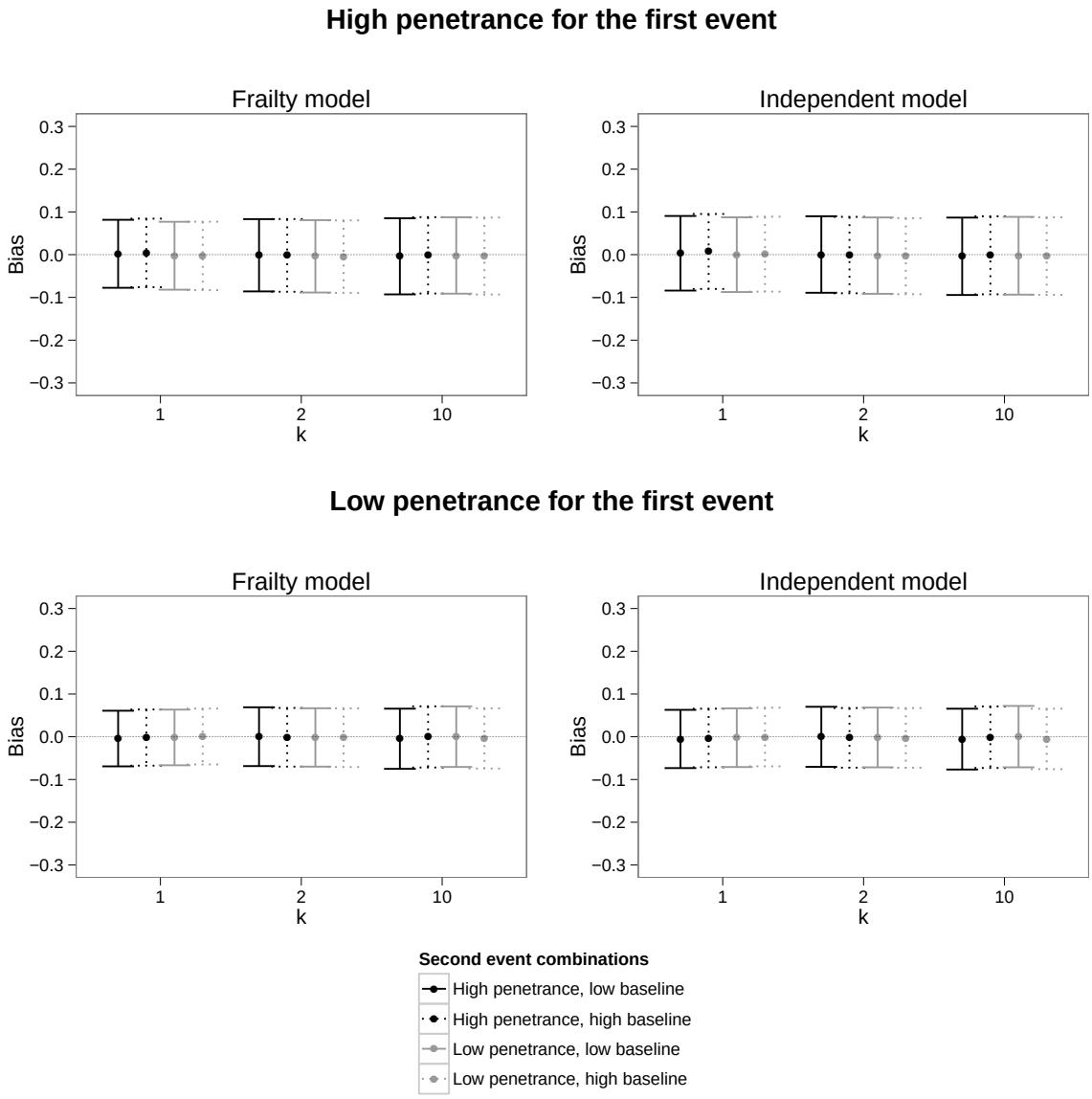

Figure 3.11: Bias and its 95% confidence interval in the first event penetrance estimation at age 70 years for female mutation carriers from frailty model (left) and independent model (right), respectively, under high penetrance (top) and low penetrance (bottom) for the first event using the recessive genetic model with common allele frequency (30%) with a sample size of 200 families.

Table 3.10: Penetrance estimation of mutation carriers for developing the second event in 10 years after the first event under the recessive genetic model with common allele frequency ($q = 30\%$) using 200 simulated families.

Parameters				Pen	Frailty model							Independent model						
T_1	T_2		k	(%)	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	17	0.002	-0.02	0.03	0.034	0.03	0.04	0.92	0.107	0.10	0.13	0.026	0.02	0.03	0.01
			2	18	-0.001	-0.02	0.03	0.034	0.03	0.04	0.89	0.060	0.05	0.08	0.024	0.02	0.02	0.28
			10	19	-0.009	-0.03	0.01	0.030	0.02	0.04	0.91	0.011	-0.01	0.03	0.022	0.02	0.02	0.92
		HBL	1	32	0.001	-0.03	0.04	0.050	0.04	0.06	0.94	0.161	0.14	0.18	0.030	0.03	0.03	0.00
			2	35	0.002	-0.04	0.04	0.048	0.04	0.06	0.91	0.087	0.07	0.10	0.029	0.03	0.03	0.13
			10	37	-0.007	-0.04	0.02	0.041	0.03	0.05	0.91	0.018	-0.00	0.04	0.027	0.03	0.03	0.87
	LP^2	LBL	1	12	-0.002	-0.02	0.01	0.025	0.02	0.03	0.93	0.079	0.06	0.09	0.023	0.02	0.02	0.05
			2	12	-0.001	-0.02	0.02	0.026	0.02	0.03	0.87	0.043	0.03	0.06	0.021	0.02	0.02	0.47
			10	13	-0.003	-0.02	0.01	0.023	0.02	0.03	0.90	0.010	-0.01	0.02	0.019	0.02	0.02	0.92
		HBL	1	23	0.002	-0.02	0.04	0.040	0.03	0.05	0.90	0.129	0.11	0.15	0.028	0.03	0.03	0.00
			2	25	0.000	-0.04	0.03	0.040	0.03	0.05	0.89	0.071	0.05	0.08	0.026	0.02	0.03	0.25
			10	26	-0.004	-0.03	0.01	0.035	0.03	0.04	0.92	0.016	-0.00	0.03	0.025	0.02	0.03	0.90
LP^1	HP^2	LBL	1	17	-0.006	-0.04	0.06	0.055	0.04	0.08	0.79	0.125	0.11	0.15	0.031	0.03	0.03	0.02
			2	18	0.007	-0.04	0.06	0.052	0.03	0.08	0.71	0.070	0.06	0.09	0.029	0.03	0.03	0.29
			10	19	-0.017	-0.06	0.01	0.039	0.03	0.06	0.85	0.015	-0.00	0.03	0.027	0.02	0.03	0.92
		HBL	1	32	0.002	-0.06	0.09	0.085	0.06	0.11	0.82	0.181	0.16	0.21	0.036	0.03	0.04	0.00
			2	35	-0.011	-0.08	0.07	0.080	0.06	0.10	0.77	0.101	0.08	0.12	0.034	0.03	0.04	0.15
			10	37	-0.007	-0.07	0.03	0.060	0.03	0.09	0.87	0.030	0.00	0.05	0.033	0.03	0.03	0.88
	LP^2	LBL	1	12	-0.000	-0.03	0.05	0.040	0.03	0.06	0.74	0.092	0.07	0.11	0.027	0.03	0.03	0.06
			2	12	-0.001	-0.03	0.04	0.035	0.02	0.06	0.74	0.050	0.03	0.07	0.025	0.02	0.03	0.49
			10	13	-0.013	-0.05	0.00	0.029	0.02	0.05	0.84	0.010	-0.01	0.02	0.022	0.02	0.02	0.95
		HBL	1	23	-0.007	-0.05	0.06	0.067	0.05	0.09	0.82	0.150	0.13	0.17	0.034	0.03	0.04	0.00
			2	25	0.007	-0.06	0.06	0.061	0.04	0.08	0.74	0.082	0.06	0.10	0.032	0.03	0.03	0.25
			10	26	-0.010	-0.06	0.02	0.048	0.03	0.07	0.85	0.017	-0.01	0.04	0.030	0.03	0.03	0.92

Pen - penetrance; Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low (ν_2=0.00724,φ_2=1.14) and high (ν_2=0.00324,φ_2=1.84) baselines for second event, respectively.

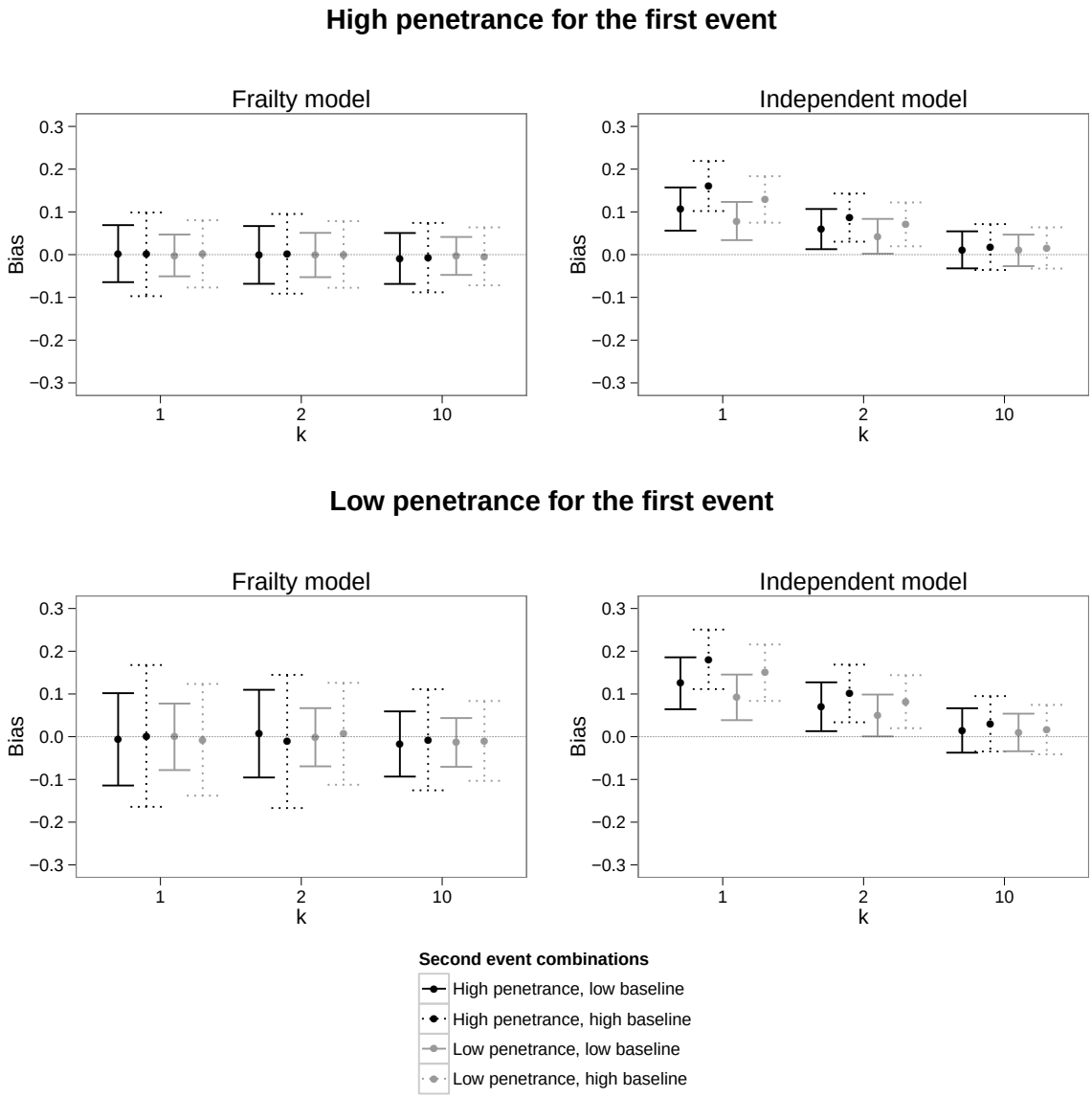

Figure 3.12: Bias and its 95% confidence interval in the 10-year penetrance estimation of the second event for mutation carriers from frailty model (left) and independent model (right), respectively, under high penetrance (top) and low penetrance (bottom) for the first event using the recessive genetic model with common allele frequency (30%) with a sample size of 200 families.

Chapter 4

AN APPLICATION TO LYNCH SYNDROME FAMILIES

We applied our proposed frailty model to 12 large Lynch syndrome families sampled from Newfoundland. As discussed in Chapter 1, Lynch syndrome is a genetic condition that has a high risk of early-onset colorectal cancer (CRC), predominantly associated with MLH1 and MSH2 genes. Individuals with Lynch syndrome are also susceptible to successive cancers of the colon, stomach, ovary, endometrium, etc. (Lynch et al., 1977). The families considered for our analyses share a founder mutation in MSH2 gene and some family members have experienced multiple cancers. The two sequential event times observed from these Lynch syndrome families are modeled using our frailty model approach based on the ascertainment corrected retrospective likelihood. Thus, we provide the age-specific penetrance and genetic relative risks associated with the mutated gene for both first and second occurrence of colorectal cancer.

4.1 Data description

The data consist of a cohort of 12 high-risk Lynch syndrome families that were found to segregate the mutant MSH2 gene. These high-risk families were identified through affected mutation carrying probands, along with their highly vulnerable relatives from the Medical Genetics Clinic at Memorial University, St.John's, Newfoundland, Canada. Information on the disease history and mutation status among their relatives was gathered and the data were collected retrospectively (Kopciuk et al., 2009).

The sampled families consist of 343 individuals spread across two to five generations. The number of members in each family ranges between 5 and 54. For each family member, we have information on their age-at-onset of first CRC and time of second CRC since the first cancer (in years) with their corresponding censoring indicators, i.e. $(T_1, \delta_1), (T_2, \delta_2)$, gender, mutation status, age at examination (in years), and relationship to the proband. Of the 343 individuals, mutation status was available for 260 individuals. We excluded individuals with missing observations and used only complete cases for our analysis.

Figure 4.1 provides a schematic representation of the distribution of events among the 12 families from Newfoundland. The data contain equal proportions of males and females and the number of mutation carriers (161) is almost twice that of non-mutation carriers (99). Among the mutation carriers (79 males and 82 females), 40 males and 28 females experienced a first CRC. Of those mutation carriers, 13 males and 8 females experienced a second CRC. There were no CRC events among non-carrier males and females, except one male who had a first CRC. These numbers clearly exhibit the underlying effect of genetic mutation in the occurrence of CRC among these Lynch syndrome families. The Kaplan-Meier (K-M) estimates of the cumulative distribution function for the first and second event times are plotted in Figures 4.2 and 4.3, respectively. The probands were excluded for these K-M estimates. It can also be noted from these plots that males tended to have a higher risk than females for both first CRC and second CRC. Similarly, mutation carriers were largely at risk compared to non-carriers. The PH assumption fails in the cumulative hazard function of the second event (Figure 4.3) with a p-value for the log rank test as 0.709. Nevertheless, we considered the PH assumption for mathematical simplicity.

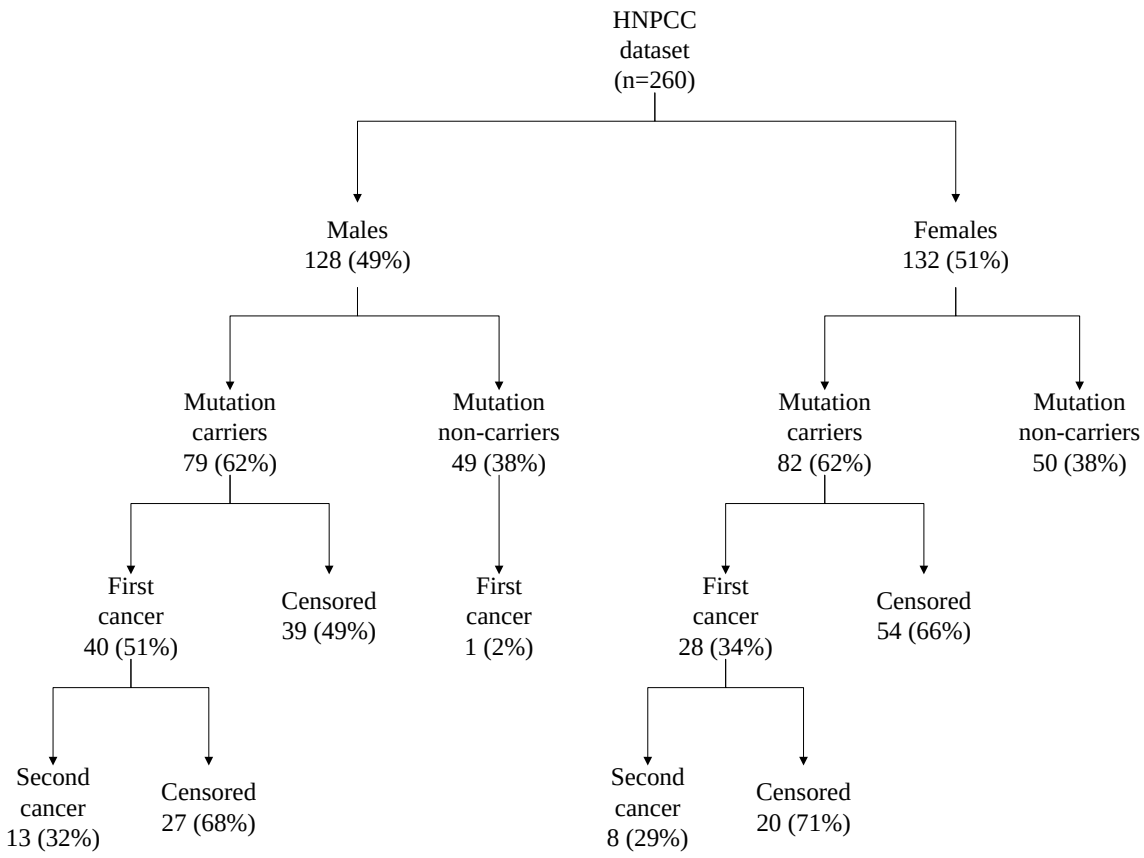

Figure 4.1: Distribution of colorectal cancer occurrences among males and females in 12 Lynch syndrome families from Newfoundland.

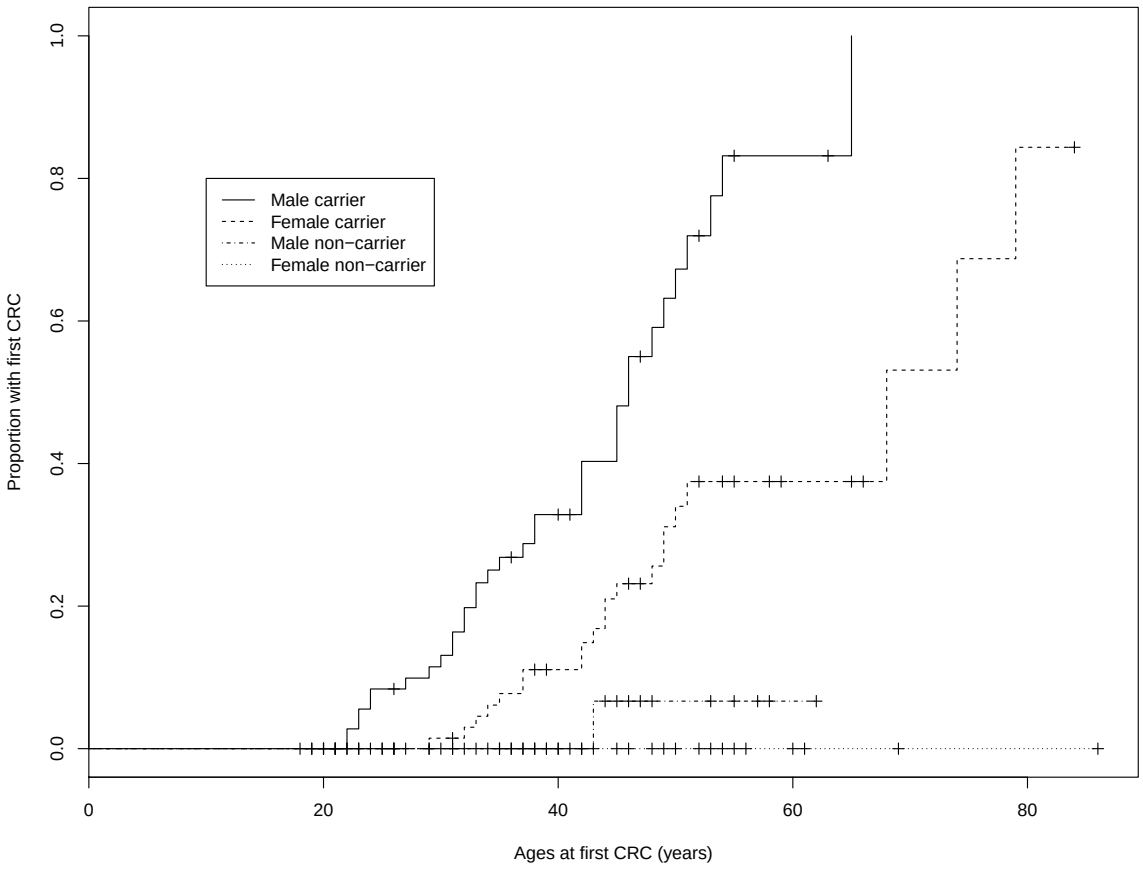

Figure 4.2: Kaplan-Meier curve of the cumulative distribution function for the time to first colorectal cancer among 12 Lynch syndrome families from Newfoundland.

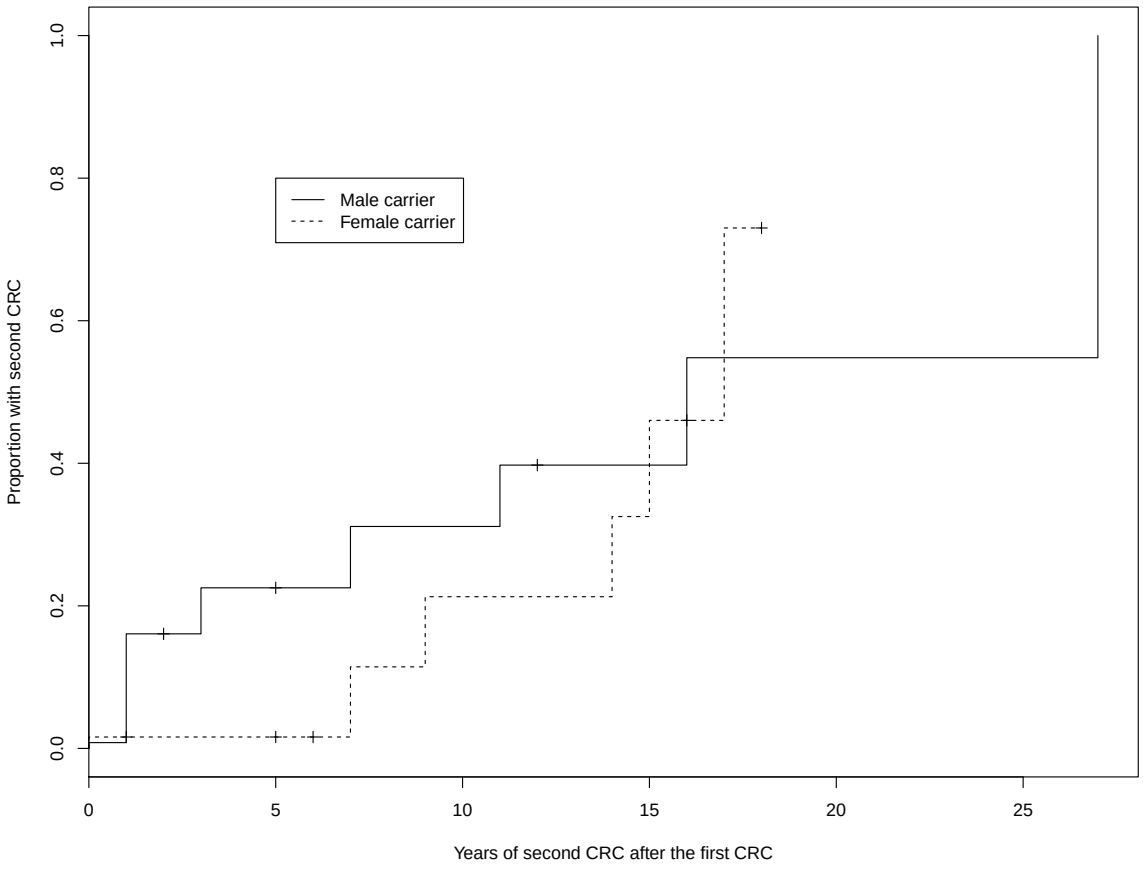

Figure 4.3: Kaplan-Meier curve of the cumulative distribution function for the time to a second colorectal cancer after the first cancer, among 12 Lynch syndrome families from Newfoundland.

4.2 Modeling sequential events

We considered the sequential event times in the occurrence of first and second CRC such that we define T_1 as the age-at-onset for first CRC and define T_2 as the time to second CRC after the occurrence of first cancer. We fitted a shared frailty model for the bivariate event times T_1 and T_2, observed from the 12 high-risk families, based on the following two hazard functions

$$\begin{aligned} \lambda_1(t_1|Z, X_1, X_2) &= Z\nu_1\varphi_1(t_1 - 20)^{\varphi_1 - 1}e^{\beta_1 X_1 + \beta_2 X_2} \\ \lambda_2(t_2|Z, X_1) &= Z\nu_2\varphi_2 {t_2}^{\varphi_2 - 1}e^{\beta_3 X_1}, \end{aligned}$$

where t_1 and t_2 are the event times for the first and second cancer, respectively, with the minimum age of onset for first cancer as 20 years and Z is the frailty variable that follows the gamma distribution with mean 1 and variance $1/k$. For the first cancer, we adjusted for the gender effect, X_1, (coded as 1-male, 0-female) and the mutation effect, X_2, (coded as 1-carrier, 0-non-carrier). For the second cancer, we adjusted only for the gender effect, X_1, as the non-mutation carriers remained free from the event of second CRC. We assumed Weibull distributions for the baseline hazard functions of T_1 and T_2. The frailties are considered to be time independent and are shared at the individual level in order to model the dependence between successive events of cancer.

The retrospective likelihood was applied to correct for the complex ascertainment process involved in sampling these high-risk families. The data used in our analyses arose from a clinic-based study design so that families were ascertained into the study based on multiple affected family members in addition to the probands. We fixed the allele frequency of the mutation gene as 2% (Lynch and Smyrk, 1996). We performed our analyses using the statistical software, R (R Development Core Team, 2011) and obtained the relative risk and penetrance estimates for the first and second CRC along with their robust standard errors.

4.2.1 Relative risks estimation

Given a value of frailty, the log relative risk of the first CRC for males compared to females was 1.34 (Standard Error, SE = 0.19) and was statistically significant (p-value < 0.001). The log relative risk of second CRC for males compared to females was 0.61 (SE = 0.39) and was not statistically significant (p-value = 0.11). Carriers of the MSH2 gene had a significantly higher risk of first CRC, 3.99 (SE = 0.96) compared to the non-carriers (p-value < 0.001). The dependence between the first and second CRCs was measured by the variance of the frailty distribution, i.e. the inverse of the estimated frailty parameter. From our analysis, we obtained a very high value for the frailty parameter estimate, 5.38×10^4 with a 95% confidence interval $(2.68 \times 10^4, 97.97 \times 10^7)$; therefore, we regard the events to be independent. This was also substantiated by fitting two separate models for the first and second events and the estimates from these models were identical to the one obtained using our frailty approach.

4.2.2 Penetrances estimation

The penetrance estimates (absolute risk) of the first and second CRC were obtained as a complement of their respective marginal survival functions. The penetrance (standard error) of first CRC by the age of 70 years among male mutation carriers was 94.20% (SE = 4%) and among female mutation carriers was 52.43% (SE = 12%). Figure 4.4 presents the estimated penetrance curves for the event of first CRC by the gender status and mutation carrier status. The male carriers tended to have higher risk compared to female carriers. It can also be seen that the life-time risk of first cancer among non-carriers was very low (less than 5% for both genders). However, for the mutation carriers, the cumulative risk seemed to increase rapidly between the ages of 30 and 70 years (especially for males) and gradually stabilized later.

The risks of developing a second CRC in 10 years after the first cancer were 49.15%

(SE = 11%) for male mutation carriers and 31% (SE = 6%) for female mutation carriers. Figure 4.5 presents the estimated penetrance curves of second CRC among males and females. Similar to the penetrance curves of the first cancer, males were at a higher risk compared to females and the difference in risks between them tended to widen over time.

4.3 Summary

In this chapter, we illustrated our proposed method using real data from Newfoundland. Using 12 very large Lynch syndrome families, we estimated the relative risk and penetrance function of the mutated gene. We considered only complete cases and excluded individuals with missing genotypes. This could possibly explain the large estimate obtained for the relative risk of the mutated gene on the first cancer. We found that gender and mutation effects were statistically significant on the occurrence of first CRC. However, the gender effect was observed to be statistically not significant for the second occurrence of CRC. Finally, we obtained a very large estimate for the frailty parameter and therefore, we conclude that the occurrence of first and second CRC is almost independent for this data. We strongly feel that this might be possibly due to the presence of fewer individuals who experienced a second event. Overall, we were able to model successive events in the occurrence of colorectal cancer and obtained the risk estimates.

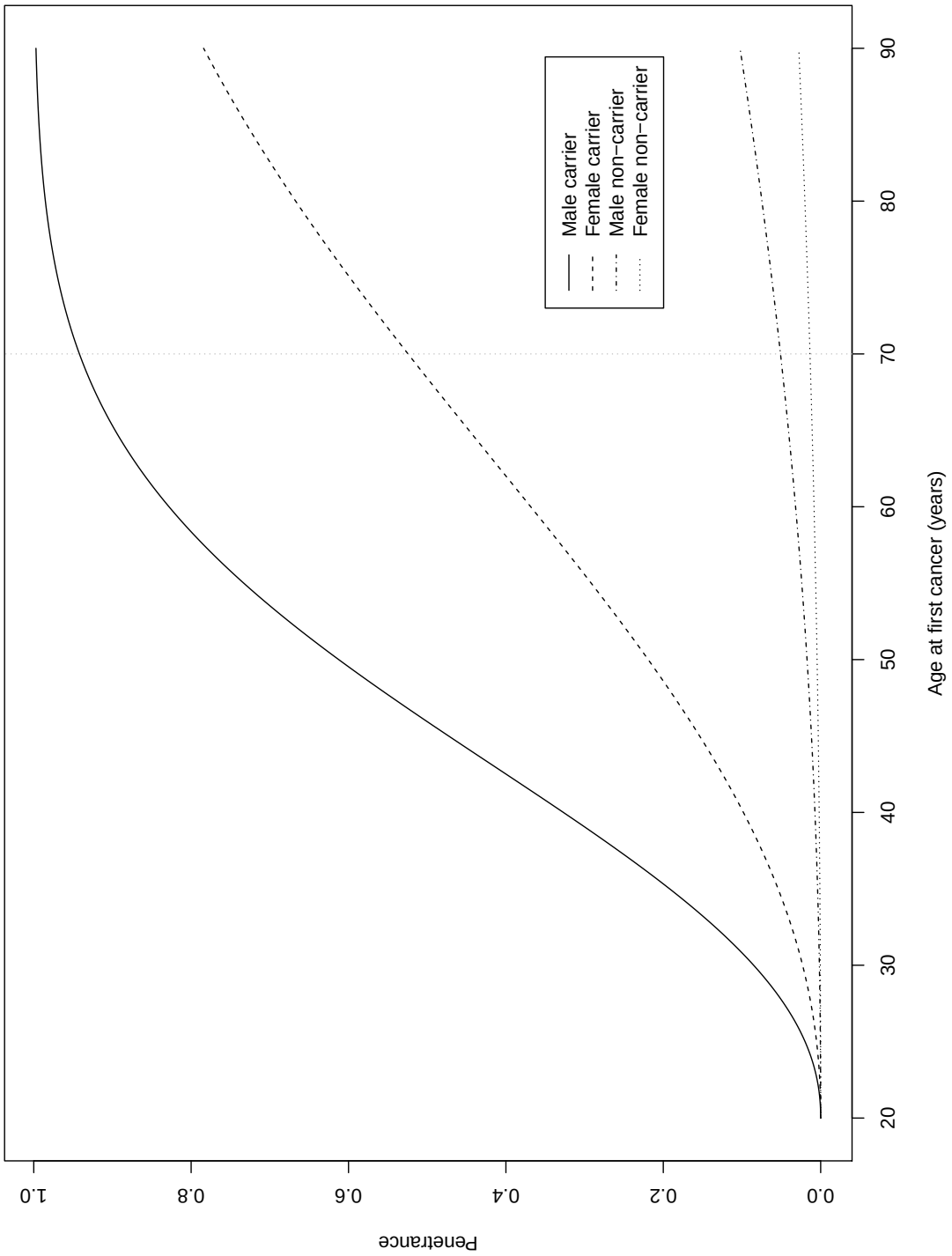

Figure 4.4: Estimated age-specific penetrance function of first colorectal cancer using 12 Lynch syndrome families from Newfoundland.

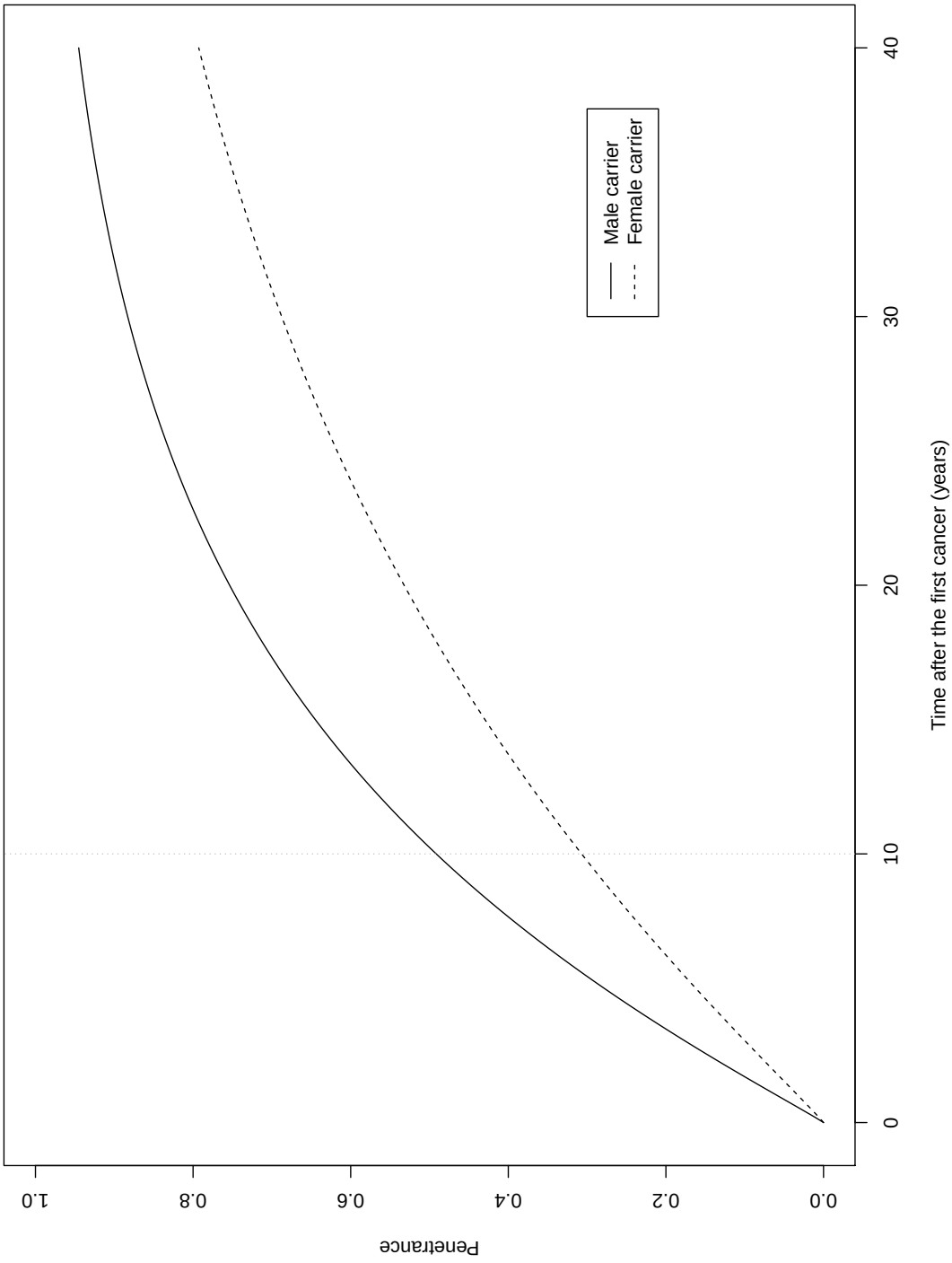

Figure 4.5: Estimated age-specific penetrance function of second colorectal cancer after the occurrence of first cancer, using 12 Lynch syndrome families from Newfoundland.

Chapter 5

DISCUSSION

Early-onset of disease and high chance for multiple events are the hallmarks of complex genetic mutations like Lynch syndrome. Knowledge of age-specific cumulative risk of disease and relative risk of the disease gene is highly valuable in the management of genetic diseases. In this thesis, we developed a statistical framework to model two sequential event times arising from two types of family designs − population- and clinic-based study designs and estimated the relative and absolute risks associated with a mutated gene. We modeled the dependence between the event times using a shared frailty model and incorporated an ascertainment corrected retrospective likelihood to account for the non-random sampling of families.

Using simulation studies, we demonstrated that our frailty approach can provide unbiased estimates of both relative and absolute risks of a mutated gene. Also, our model is capable of producing a valid estimate of the standard error such that the desired coverage probability is achievable. We strongly feel that an increase in sample size would prove helpful to achieve accurate coverage probability using our approach. We also investigated the effect of ignoring the dependence between the event times using an independent model. We conclude that the independent model would produce unreliable risk estimates, especially for the penetrance estimation of a second event in the presence of a high dependence between two events. The independent model would also underestimate the standard error, which in turn would result in a coverage probability far less than 0.95.

We illustrated our proposed method using a sample of 12 large Lynch syndrome families from Newfoundland and estimated the relative risk of the mutated gene and the age-specific penetrance for the occurrence of first and second colorectal cancer. The difference in sample sizes between our simulation studies (n=100 and 200) and our real data application (n=12) can be explained by the choice of study design, i.e. for the former we considered a population-based study design, and for the latter a clinic-based study design. The population-based design is highly capable of sampling a large number of families as it conditions only on the disease status of the proband, whereas the clinic-based design requires multiple affected individuals within a family. We are currently investigating the performance of our frailty-based approach for a clinic-based study design using simulation studies.

There were a few potential limitations to our study that are noteworthy. First, we assumed the conditional independence among family members given their genotypes with the rationale that the disease causing gene is the only source of familial correlation. However, most genetic disorders are highly complex and there may be a second gene or a modifier gene that may be associated with the disease outcome. Violation of the conditional independence assumption can lead to upwardly biased estimates (Gail et al., 2001). We did not explicitly model the residual familial correlation, but derived the robust variance estimates using the sandwich-variance estimator approach. Second, we considered the pedigree data to have complete genotype and phenotype (outcome) information for all sampled family members. In practice, this may not be possible because it is common to have missing information at several stages of data collection. Nevertheless, missing genotypes can be inferred using the Expectation-Maximization (EM) approach (Choi and Briollais, 2011). In this approach, the maximum likelihood estimates (MLEs) can be computed in the presence of missing informations using a two-step iterative procedure. The first step is the expectation step where the expectation of the log-likelihood for the complete data is taken with respect to the conditional distribution of missing genotypes given ob-

served genotype and phenotype information from the family members and current choice of parameter values. Then, in the maximization step, the parameter estimates are updated by maximizing the log-likelihood function using the estimate of missing data in the expectation step. These two steps iterate until convergence to obtain the maximum likelihood estimates. Third, we assumed independent censoring for the first event but this may not be true if death can occur due to other cancers. Such a situation may result in informative censoring for both first and second events and may alter the probabilities of the event of interest. This could possibly be averted by including an additional state in our three-state progressive model to account for death as a competing risk. Fourth, we considered proportional hazards (PH) assumption in our modeling of bivariate event times. However, if the assumption is violated, then a stratified PH model can be fit. Finally, we assumed parametric distributions for the baseline hazard functions and for the frailty. The consequence of misspecification of these distributions can affect the risk estimates. To obtain more robust estimates, non-parametric or piecewise constant approaches can be used to specify the baseline distributions but at the cost of intensive computing in the estimation of baseline parameters.

In future work, we plan to extend our approach using a nested frailty model. The nested frailty model (Sastry, 1997) is a popular way to model dependencies in the presence of multilevel, clustered time-to-event data. Using this model, we consider two frailty variables, Z_f and Z_{fi} where the frailty Z_f models the residual familial correlation among the members of family f ($f = 1, \ldots, n$) that is not explained by the observed risk factors, and the second frailty Z_{fj}, which is nested under Z_f, models the dependence between the two sequential event times experienced by an individual i ($i = 1, \ldots, n_f$). By considering a nested frailty model, we can overcome the limitations of the aforesaid conditional independence assumption.

Appendix A

CARRIER PROBABILITY

A.1 Transmission probabilities

The retrospective likelihood provided in equation (2.4) involves the calculation of individual genotype probabilities. Human beings are considered to be biallelic at an autosomal loci, and therefore there are three possible genotypic configurations: AA (homozygous dominant), Aa (heterozygous), and aa (homozygous recessive). If a dominant model is considered, a person is termed to be a carrier of the mutant gene if s/he possesses at least one mutant allele at a locus, i.e. belonging to the type AA or Aa. If a recessive model is considered, then a person must carry the mutant gene in both the alleles, i.e. AA.

For a founder, i.e. a person whose parents' genotypes are unknown, the genotype probabilities can be derived using the Hardy-Weinberg equilibrium with the knowledge of population allele frequency of the mutant gene. The Hardy-Weinberg equilibrium assumes the allele frequency in a population remains constant through several generations provided assumptions like random mating, no mutation, and large population size are met. Let the population allele frequency for the mutant allele, A, be q and the allele frequency for a be $1-q$. If the required assumptions are met, then the genotype frequencies for the types AA, Aa, and aa are $q^2, 2q(1-q)$, and $(1-q)^2$, respectively.

For a non-founder, the genotype probability can be calculated using the Mendelian transmission probability. The genotype of an offspring at a loci is formed due to the contribution of an allele from each parent. Therefore, given the parents‘ genotype,

Table A.1: Offspring's genotypic probabilities conditional on parent's genotype - Mendelian transmission probabilities.

Father's genotype	Mother's genotype	Offspring's genotype		
		aa	Aa	AA
aa	aa	1	0	0
	Aa	1/2	1/2	0
	AA	0	1	0
Aa	aa	1/2	1/2	0
	Aa	1/4	1/2	1/4
	AA	0	1/2	1/2
AA	aa	0	1	0
	Aa	0	1/2	1/2
	AA	0	0	1

the offspring's probability of carrying a disease allele can be computed. For instance, if we let the genotype of the father be Aa and the genotype of the mother be Aa then the offspring has the probabilities 1/4, 1/2, and 1/4 to be homozygous dominant AA, heterozygous Aa, and homozygous recessive aa, respectively. Table A.1 provides the transmission probabilities for different combinations of parental genotypes.

A.2 Conditional genotype probabilities for relatives

The genotype probability of the relatives conditional on the proband's genotype are summarized in Table A.2 (Thomas, 2004). Using these conditional probabilities, the conditional carrier probabilities can be derived either for a dominant or recessive model. For instance, the conditional probability that the mother (M) is a carrier

given the child (C) is a carrier can be derived for a dominant model as follows:

$$\begin{aligned} P[M=1|C=1] &= P[M=AA\ or\ Aa|C=AA\ or\ AA] \\ &= \frac{P[C=AA\ or\ Aa|M=AA]P[M=AA]}{P[C=AA\ or\ Aa]} \\ &\quad + \frac{P[C=AA\ or\ Aa|M=Aa]P[M=Aa]}{P[C=AA\ or\ Aa]} \\ &= \frac{1+q-q^2}{2-q} \end{aligned}$$

Table A.2: Relative's genotypic probabilities conditional on proband's genotype.

Proband's genotype	Relative's genotype		
	aa	Aa	aa
	Parents or offspring		
aa	$1-q$	q	0
Aa	$\frac{1-q}{2}$	$\frac{1}{2}$	$\frac{q}{2}$
AA	0	$1-q$	q
	Sibling		
aa	$1-q+\frac{q^2}{4}$	$q-\frac{q^2}{2}$	$\frac{q^2}{4}$
Aa	$\frac{1}{2}-\frac{3q}{4}+\frac{q^2}{4}$	$\frac{1}{2}+\frac{q}{2}-\frac{q^2}{2}$	$\frac{q}{4}+\frac{q^2}{4}$
AA	$\frac{(1-q)^2}{4}$	$\frac{1}{2}-\frac{q^2}{2}$	$\frac{1}{4}+\frac{q}{2}+\frac{q^2}{4}$

Table A.3 provides the carrier probabilities of the relatives given the carrier status of the proband for a dominant model. Similarly, Table A.4 provides the conditional carrier probabilities for the recessive model, where two copies of the mutant allele is required to cause disease.

Table A.3: Relative's carrier probabilities conditional on proband's carrier status for a dominant model.

Proband's	Relative's carrier probabilities	
carrier status	1	0
	Parents or offspring	
1	$\frac{1+q-q^2}{2-q}$	$\frac{1-2q+q^2}{2-q}$
0	q	$1-q$
	Sibling	
1	$\frac{4q+5q^2-6q^3+q^4}{4(2q-q^2)}$	$\frac{4q-9q^2+6q^3-q^4}{4(2q-q^2)}$
0	$q-\frac{q^2}{4}$	$1-q+\frac{q^2}{4}$

Table A.4: Relative's carrier probabilities conditional on proband's carrier status for a recessive model.

Proband's	Relative's carrier probabilities	
carrier status	1	0
	Parents or offspring	
1	q	$1-q$
0	$1-q$	q
	Sibling	
1	$\frac{(1+q)^2}{4}$	$1-\frac{(1+q)^2}{4}$
0	$\frac{3-2q-q^2}{4}$	$\frac{1+2q+q^2}{4}$

Appendix B

SIMULATION RESULTS USING 100 FAMILIES

In Chapter 3, we presented the simulation results obtained using 200 families to evaluate the performance of our frailty approach to model bivariate event times. Here, we present the results obtained using 100 families sampled using a population-based study design. The conclusions arrived at earlier using 200 families still hold for the sample size of 100 families. Tables B.1 − B.5 present the simulation results obtained in the presence of the dominant genetic model and Tables B.6 − B.10 provide the results obtained in the presence of the recessive genetic model.

In the estimation of genetic relative risk of the first and second events, the biases were slightly greater than those obtained using 200 random families. For instance, in the estimation of β_2 using the dominant model (Table B.1), the highest value of the absolute bias was 0.084 using 100 families, whereas using 200 families (Table 3.1) the value was 0.062 . Similarly, in the estimation of β_3 (Table B.2), the highest value of the absolute bias using the dominant model was 0.102 using 100 families compared to 0.062 using 200 families (Table 3.7). However, in the estimation of penetrance for both the events, the biases were almost similar among the two sample sizes (Tables B.3, B.4, B.5, B.8, B.9, and B.10).

The model-based standard errors using 100 families were relatively larger than (almost 1.5 times) those obtained using 200 families, irrespective of the disease risk being estimated and the genetic model considered. Nevertheless, the coverage probabilities in the estimation of genetic relative risks and penetrance functions were almost similar and close to the prescribed 95% probability, except in the estimation of penetrance for second event among mutation carriers in the presence of a low penetrance

setting for the first event. Lastly, the independent model (model that ignored the dependence between event times) provided largely biased estimates compared to our frailty approach in the estimation of both relative and absolute risks. In conclusion, our frailty approach produced unbiased estimates of the relative risks and penetrance functions compared to the independent model.

Table B.1: Estimation of log relative genetic risk (β_2) of developing the first event under the dominant genetic model with a rare allele frequency ($q = 2\%$) using 100 simulated families.

Parameters				Frailty model							Independent model						
T_1	T_2		k	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	-0.022	-0.24	0.23	0.345	0.31	0.38	0.95	-0.285	-0.46	-0.08	0.289	0.27	0.31	0.80
			2	0.042	-0.17	0.27	0.334	0.30	0.37	0.93	-0.129	-0.33	0.08	0.290	0.27	0.31	0.90
			10	0.062	-0.17	0.27	0.315	0.28	0.36	0.91	-0.010	-0.25	0.20	0.289	0.26	0.32	0.93
		HBL	1	0.020	-0.20	0.25	0.342	0.31	0.37	0.95	-0.253	-0.46	-0.03	0.291	0.27	0.32	0.81
			2	0.009	-0.20	0.30	0.328	0.30	0.36	0.94	-0.156	-0.35	0.12	0.288	0.27	0.32	0.92
			10	0.060	-0.13	0.28	0.311	0.28	0.35	0.92	-0.010	-0.18	0.21	0.288	0.26	0.31	0.96
	LP^2	LBL	1	0.042	-0.17	0.28	0.349	0.32	0.39	0.96	-0.236	-0.43	-0.04	0.294	0.27	0.32	0.84
			2	0.038	-0.15	0.29	0.333	0.30	0.37	0.94	-0.143	-0.32	0.06	0.286	0.27	0.31	0.90
			10	0.050	-0.15	0.27	0.316	0.28	0.36	0.92	-0.022	-0.21	0.17	0.283	0.26	0.31	0.94
		HBL	1	0.049	-0.18	0.30	0.343	0.32	0.38	0.94	-0.220	-0.43	-0.01	0.291	0.27	0.32	0.84
			2	0.065	-0.21	0.28	0.330	0.30	0.37	0.94	-0.122	-0.32	0.10	0.291	0.27	0.32	0.90
			10	0.048	-0.12	0.23	0.310	0.28	0.35	0.95	-0.020	-0.20	0.17	0.285	0.26	0.31	0.97
LP^1	HP^2	LBL	1	0.048	-0.19	0.30	0.368	0.32	0.41	0.92	-0.103	-0.32	0.12	0.314	0.29	0.34	0.92
			2	0.059	-0.14	0.31	0.348	0.31	0.40	0.95	-0.059	-0.25	0.16	0.311	0.29	0.33	0.96
			10	0.071	-0.13	0.30	0.337	0.30	0.38	0.93	-0.000	-0.20	0.21	0.309	0.29	0.33	0.95
		HBL	1	0.055	-0.20	0.29	0.359	0.32	0.40	0.94	-0.116	-0.33	0.11	0.317	0.29	0.35	0.93
			2	0.054	-0.18	0.28	0.342	0.30	0.39	0.93	-0.052	-0.27	0.14	0.312	0.29	0.34	0.94
			10	0.074	-0.12	0.32	0.332	0.30	0.38	0.93	0.016	-0.18	0.21	0.309	0.29	0.33	0.98
	LP^2	LBL	1	0.052	-0.19	0.31	0.366	0.33	0.42	0.95	-0.092	-0.31	0.12	0.317	0.29	0.34	0.91
			2	0.061	-0.16	0.33	0.356	0.32	0.40	0.93	-0.069	-0.26	0.17	0.315	0.29	0.34	0.95
			10	0.084	-0.12	0.34	0.346	0.30	0.40	0.92	0.005	-0.19	0.21	0.310	0.29	0.34	0.95
		HBL	1	0.020	-0.21	0.26	0.363	0.32	0.41	0.94	-0.090	-0.33	0.10	0.316	0.29	0.34	0.93
			2	0.038	-0.18	0.26	0.340	0.31	0.39	0.92	-0.054	-0.25	0.13	0.312	0.29	0.34	0.94
			10	0.070	-0.14	0.30	0.336	0.30	0.39	0.93	0.015	-0.18	0.21	0.309	0.29	0.33	0.97

Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low (ν_2=0.00724,φ_2=1.14) and high (ν_2=0.00324,φ_2=1.84) baselines for second event, respectively.

Table B.2: Estimation of log relative genetic risk (β_3) of developing the second event under the dominant genetic model with a rare allele frequency ($q = 2\%$) using 100 simulated families.

Parameters				Frailty model							Independent model						
T_1	T_2		k	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	0.030	-0.42	0.64	0.711	0.58	0.98	0.96	-0.131	-0.56	0.42	0.649	0.53	0.95	0.93
			2	0.102	-0.39	0.89	0.774	0.63	1.06	0.91	-0.003	-0.47	0.75	0.752	0.61	1.04	0.93
			10	0.007	-0.46	0.66	0.816	0.62	1.09	0.87	-0.033	-0.49	0.58	0.847	0.63	1.06	0.94
		HBL	1	0.042	-0.32	0.40	0.542	0.46	0.66	0.95	-0.212	-0.52	0.14	0.462	0.38	0.56	0.90
			2	0.051	-0.34	0.44	0.564	0.46	0.71	0.92	-0.102	-0.46	0.22	0.485	0.40	0.63	0.89
			10	0.018	-0.33	0.44	0.546	0.45	0.71	0.90	-0.040	-0.37	0.36	0.509	0.42	0.64	0.93
	LP^2	LBL	1	0.031	-0.44	0.57	0.725	0.59	0.93	0.95	-0.096	-0.54	0.40	0.672	0.55	0.92	0.94
			2	0.102	-0.48	0.61	0.772	0.64	1.02	0.93	0.012	-0.51	0.50	0.737	0.61	1.01	0.94
			10	0.080	-0.48	0.72	0.801	0.64	1.07	0.85	0.044	-0.52	0.68	0.809	0.66	1.06	0.94
		HBL	1	0.057	-0.39	0.48	0.554	0.46	0.68	0.92	-0.113	-0.51	0.26	0.459	0.39	0.58	0.88
			2	0.011	-0.36	0.44	0.551	0.45	0.67	0.90	-0.081	-0.44	0.30	0.490	0.40	0.60	0.89
			10	0.021	-0.39	0.46	0.544	0.44	0.69	0.89	-0.030	-0.42	0.39	0.505	0.42	0.62	0.91
LP^1	HP^2	LBL	1	0.055	-0.44	0.61	0.712	0.58	0.91	0.91	-0.052	-0.50	0.50	0.679	0.55	0.90	0.93
			2	-0.008	-0.48	0.66	0.753	0.61	0.98	0.90	-0.073	-0.54	0.56	0.717	0.59	0.99	0.94
			10	0.132	-0.43	0.85	0.859	0.69	1.12	0.85	0.054	-0.46	0.74	0.923	0.68	1.08	0.96
		HBL	1	0.064	-0.33	0.48	0.561	0.47	0.71	0.93	-0.080	-0.43	0.26	0.469	0.40	0.59	0.92
			2	0.071	-0.30	0.46	0.578	0.46	0.73	0.90	-0.053	-0.40	0.32	0.507	0.42	0.63	0.92
			10	0.034	-0.36	0.52	0.575	0.47	0.73	0.90	-0.028	-0.40	0.41	0.533	0.43	0.64	0.92
	LP^2	LBL	1	0.092	-0.43	0.65	0.741	0.60	1.02	0.94	0.020	-0.42	0.55	0.699	0.56	0.97	0.96
			2	0.089	-0.42	0.69	0.800	0.63	1.05	0.92	0.028	-0.46	0.62	0.756	0.61	1.03	0.94
			10	0.144	-0.40	0.79	0.873	0.70	1.11	0.87	0.091	-0.42	0.76	0.917	0.70	1.09	0.97
		HBL	1	0.006	-0.38	0.45	0.556	0.47	0.69	0.93	-0.055	-0.40	0.29	0.481	0.40	0.58	0.92
			2	0.026	-0.32	0.46	0.562	0.45	0.69	0.91	-0.017	-0.36	0.36	0.489	0.41	0.60	0.92
			10	0.021	-0.38	0.48	0.583	0.47	0.74	0.90	-0.029	-0.41	0.41	0.546	0.43	0.66	0.94

Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low (ν_2=0.00724,φ_2=1.14) and high (ν_2=0.00324,φ_2=1.84) baselines for second event, respectively.

Table B.3: Penetrance estimation of male mutation carriers for the first event by the age of 70 years under the dominant genetic model with a rare allele frequency ($q = 2\%$) using 100 simulated families.

Parameters				Pen	Frailty model							Independent model						
T_1	T_2		k	(%)	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	66	-0.002	-0.04	0.04	0.064	0.06	0.07	0.95	0.063	0.02	0.10	0.057	0.05	0.06	0.80
			2	74	-0.006	-0.04	0.04	0.062	0.06	0.07	0.94	0.037	0.01	0.07	0.051	0.05	0.06	0.87
			10	83	-0.008	-0.04	0.03	0.054	0.04	0.06	0.93	0.012	-0.02	0.04	0.043	0.04	0.05	0.91
		HBL	1	66	0.002	-0.04	0.05	0.062	0.06	0.07	0.93	0.063	0.02	0.10	0.057	0.05	0.06	0.77
			2	74	-0.003	-0.04	0.04	0.059	0.05	0.06	0.96	0.040	0.01	0.07	0.051	0.05	0.06	0.84
			10	83	-0.000	-0.03	0.03	0.050	0.04	0.06	0.95	0.013	-0.01	0.04	0.042	0.04	0.05	0.89
	LP^2	LBL	1	66	0.001	-0.04	0.05	0.064	0.06	0.07	0.94	0.065	0.03	0.10	0.057	0.05	0.06	0.77
			2	74	-0.000	-0.04	0.04	0.063	0.06	0.07	0.93	0.045	0.01	0.08	0.051	0.05	0.06	0.85
			10	83	-0.007	-0.04	0.03	0.055	0.04	0.06	0.93	0.013	-0.02	0.04	0.043	0.04	0.05	0.92
		HBL	1	66	0.003	-0.04	0.05	0.062	0.06	0.07	0.94	0.063	0.02	0.10	0.056	0.05	0.06	0.77
			2	74	0.003	-0.04	0.04	0.060	0.06	0.07	0.92	0.042	0.01	0.08	0.051	0.05	0.06	0.83
			10	83	-0.006	-0.04	0.03	0.052	0.04	0.06	0.94	0.010	-0.02	0.04	0.043	0.04	0.05	0.90
LP^1	HP^2	LBL	1	43	-0.004	-0.04	0.04	0.068	0.06	0.07	0.92	0.031	-0.01	0.07	0.065	0.06	0.07	0.91
			2	47	-0.007	-0.05	0.05	0.069	0.06	0.08	0.90	0.020	-0.02	0.07	0.066	0.06	0.07	0.90
			10	52	-0.019	-0.07	0.03	0.070	0.06	0.08	0.90	0.002	-0.05	0.04	0.065	0.06	0.07	0.92
		HBL	1	43	-0.010	-0.05	0.04	0.066	0.06	0.07	0.92	0.025	-0.01	0.07	0.065	0.06	0.07	0.92
			2	47	-0.006	-0.05	0.04	0.068	0.06	0.07	0.92	0.018	-0.02	0.07	0.066	0.06	0.07	0.92
			10	52	-0.014	-0.06	0.03	0.068	0.06	0.07	0.93	0.000	-0.04	0.04	0.065	0.06	0.07	0.95
	LP^2	LBL	1	43	-0.004	-0.05	0.04	0.068	0.06	0.07	0.91	0.031	-0.02	0.07	0.065	0.06	0.07	0.93
			2	47	-0.016	-0.06	0.04	0.070	0.06	0.08	0.93	0.011	-0.03	0.07	0.066	0.06	0.07	0.93
			10	52	-0.021	-0.07	0.03	0.071	0.06	0.08	0.92	0.002	-0.05	0.04	0.066	0.06	0.07	0.95
		HBL	1	43	-0.001	-0.05	0.04	0.066	0.06	0.07	0.92	0.030	-0.02	0.07	0.065	0.06	0.07	0.92
			2	47	-0.009	-0.06	0.04	0.068	0.06	0.07	0.92	0.019	-0.03	0.06	0.066	0.06	0.07	0.92
			10	52	-0.013	-0.06	0.03	0.069	0.06	0.08	0.94	0.009	-0.04	0.04	0.066	0.06	0.07	0.96

Pen - penetrance; Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low ($\nu_2 = 0.0072$, $\varphi_2 = 1.14$) and high ($\nu_2 = 0.0032$, $\varphi_2 = 1.84$) baselines for second event, respectively.

Table B.4: Penetrance estimation of female mutation carriers for the first event by the age of 70 years under the dominant genetic model with a rare allele frequency ($q = 2\%$) using 100 simulated families.

Parameters				Pen	Frailty model							Independent model						
T_1	T_2		k	(%)	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	37	-0.007	-0.05	0.03	0.056	0.05	0.06	0.96	-0.003	-0.05	0.04	0.061	0.06	0.06	0.95
			2	40	-0.000	-0.04	0.05	0.059	0.06	0.06	0.92	0.000	-0.04	0.05	0.063	0.06	0.07	0.93
			10	44	-0.000	-0.04	0.04	0.063	0.06	0.07	0.94	0.000	-0.04	0.04	0.065	0.06	0.07	0.95
		HBL	1	37	0.001	-0.04	0.04	0.056	0.05	0.06	0.94	0.004	-0.04	0.05	0.062	0.06	0.06	0.93
			2	40	-0.005	-0.04	0.04	0.060	0.06	0.06	0.95	-0.005	-0.04	0.04	0.063	0.06	0.07	0.95
			10	44	-0.005	-0.05	0.04	0.064	0.06	0.07	0.93	-0.004	-0.05	0.04	0.065	0.06	0.07	0.93
	LP^2	LBL	1	37	0.000	-0.03	0.04	0.056	0.05	0.06	0.94	0.003	-0.04	0.05	0.062	0.06	0.06	0.94
			2	40	0.002	-0.03	0.05	0.060	0.06	0.06	0.94	0.004	-0.03	0.06	0.063	0.06	0.07	0.94
			10	44	0.001	-0.05	0.04	0.063	0.06	0.07	0.93	-0.000	-0.05	0.04	0.064	0.06	0.07	0.94
		HBL	1	37	0.000	-0.04	0.04	0.056	0.05	0.06	0.93	0.005	-0.04	0.05	0.061	0.06	0.06	0.93
			2	40	-0.006	-0.04	0.04	0.060	0.06	0.06	0.94	-0.003	-0.04	0.05	0.063	0.06	0.07	0.94
			10	44	-0.004	-0.05	0.04	0.063	0.06	0.07	0.95	-0.004	-0.05	0.04	0.064	0.06	0.07	0.96
LP^1	HP^2	LBL	1	19	-0.003	-0.03	0.02	0.046	0.04	0.05	0.94	-0.004	-0.04	0.02	0.048	0.04	0.05	0.95
			2	19	-0.003	-0.03	0.04	0.048	0.04	0.05	0.94	-0.004	-0.03	0.04	0.050	0.04	0.05	0.95
			10	20	-0.002	-0.03	0.04	0.050	0.04	0.06	0.93	-0.003	-0.03	0.04	0.051	0.05	0.06	0.93
		HBL	1	19	-0.006	-0.04	0.02	0.046	0.04	0.05	0.95	-0.006	-0.04	0.03	0.048	0.04	0.05	0.95
			2	19	-0.004	-0.03	0.04	0.048	0.04	0.05	0.92	-0.004	-0.03	0.04	0.049	0.04	0.05	0.93
			10	20	0.000	-0.03	0.04	0.050	0.04	0.06	0.93	-0.000	-0.03	0.04	0.051	0.05	0.06	0.95
	LP^2	LBL	1	19	-0.002	-0.04	0.03	0.046	0.04	0.05	0.92	-0.003	-0.04	0.03	0.049	0.04	0.05	0.92
			2	19	-0.003	-0.03	0.03	0.048	0.04	0.05	0.92	-0.003	-0.03	0.03	0.049	0.04	0.05	0.92
			10	20	-0.003	-0.03	0.04	0.049	0.04	0.06	0.91	-0.004	-0.04	0.04	0.051	0.05	0.06	0.92
		HBL	1	19	-0.002	-0.04	0.03	0.047	0.04	0.05	0.91	-0.003	-0.04	0.03	0.048	0.04	0.05	0.91
			2	19	-0.004	-0.03	0.03	0.048	0.04	0.05	0.93	-0.006	-0.03	0.03	0.049	0.04	0.05	0.94
			10	20	-0.001	-0.03	0.04	0.050	0.04	0.06	0.91	-0.002	-0.03	0.04	0.050	0.05	0.06	0.93

Pen - penetrance; Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low ($\nu_2 = 0.0072$, $\varphi_2 = 1.14$) and high ($\nu_2 = 0.0032$, $\varphi_2 = 1.84$) baselines for second event, respectively.

Table B.5: Penetrance estimation of mutation carriers for developing the second event in 10 years after the first event under the dominant genetic model with a rare allele frequency ($q = 2\%$) using 100 simulated families.

Parameters				Pen	Frailty model							Independent model						
T_1	T_2		k	(%)	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	17	-0.002	-0.02	0.03	0.044	0.04	0.06	0.89	0.103	0.09	0.13	0.036	0.03	0.04	0.18
			2	18	-0.004	-0.03	0.03	0.041	0.03	0.05	0.83	0.057	0.04	0.08	0.034	0.03	0.04	0.63
			10	19	-0.011	-0.04	0.02	0.036	0.03	0.05	0.91	0.011	-0.01	0.03	0.031	0.03	0.03	0.95
		HBL	1	32	0.001	-0.04	0.05	0.065	0.06	0.08	0.92	0.159	0.14	0.19	0.042	0.04	0.04	0.04
			2	35	-0.004	-0.05	0.04	0.061	0.05	0.07	0.90	0.087	0.06	0.12	0.040	0.04	0.04	0.44
			10	37	-0.010	-0.05	0.03	0.052	0.04	0.06	0.89	0.021	-0.01	0.05	0.038	0.04	0.04	0.91
	LP^2	LBL	1	12	0.001	-0.02	0.02	0.033	0.03	0.04	0.87	0.074	0.06	0.09	0.031	0.03	0.03	0.35
			2	12	-0.003	-0.02	0.03	0.032	0.02	0.04	0.86	0.041	0.02	0.06	0.029	0.03	0.03	0.75
			10	13	-0.008	-0.03	0.01	0.029	0.02	0.04	0.89	0.009	-0.01	0.02	0.027	0.02	0.03	0.93
		HBL	1	23	0.003	-0.03	0.06	0.052	0.04	0.07	0.88	0.129	0.10	0.16	0.039	0.04	0.04	0.09
			2	25	-0.003	-0.04	0.04	0.050	0.04	0.06	0.89	0.068	0.04	0.09	0.037	0.04	0.04	0.58
			10	26	-0.007	-0.04	0.02	0.042	0.04	0.05	0.88	0.015	-0.01	0.04	0.034	0.03	0.04	0.91
LP^1	HP^2	LBL	1	17	-0.007	-0.05	0.08	0.056	0.04	0.09	0.70	0.121	0.10	0.16	0.043	0.04	0.05	0.23
			2	18	-0.011	-0.06	0.05	0.052	0.04	0.08	0.71	0.067	0.04	0.09	0.040	0.04	0.04	0.67
			10	19	-0.028	-0.08	0.01	0.042	0.04	0.07	0.78	0.012	-0.01	0.04	0.038	0.04	0.04	0.95
		HBL	1	32	-0.007	-0.08	0.11	0.097	0.07	0.14	0.75	0.182	0.15	0.22	0.050	0.05	0.05	0.06
			2	35	0.001	-0.10	0.08	0.086	0.05	0.12	0.74	0.103	0.07	0.14	0.048	0.04	0.05	0.43
			10	37	-0.023	-0.10	0.03	0.066	0.05	0.10	0.84	0.021	-0.01	0.05	0.046	0.04	0.05	0.92
	LP^2	LBL	1	12	-0.006	-0.05	0.05	0.043	0.03	0.07	0.73	0.087	0.06	0.11	0.038	0.04	0.04	0.39
			2	12	-0.011	-0.05	0.04	0.040	0.03	0.07	0.76	0.047	0.02	0.07	0.035	0.03	0.04	0.77
			10	13	-0.024	-0.06	0.00	0.035	0.03	0.06	0.80	0.005	-0.02	0.02	0.031	0.03	0.04	0.94
		HBL	1	23	0.001	-0.06	0.11	0.072	0.05	0.11	0.71	0.150	0.12	0.18	0.047	0.04	0.05	0.06
			2	25	0.004	-0.08	0.06	0.064	0.04	0.10	0.73	0.082	0.05	0.11	0.044	0.04	0.05	0.56
			10	26	-0.019	-0.09	0.02	0.049	0.04	0.08	0.81	0.019	-0.01	0.04	0.042	0.04	0.04	0.95

Pen - penetrance; Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low ($\nu_2 = 0.0072$, $\varphi_2 = 1.14$) and high ($\nu_2 = 0.0032$, $\varphi_2 = 1.84$) baselines for second event, respectively.

Table B.6: Estimation of log relative genetic risk (β_2) of developing the first event under the recessive genetic model with a common allele frequency ($q = 30\%$) using 100 simulated families.

Parameters				Frailty model							Independent model						
T_1	T_2		k	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	0.041	-0.16	0.29	0.346	0.32	0.38	0.94	-0.235	-0.44	-0.01	0.290	0.27	0.32	0.85
			2	0.064	-0.15	0.32	0.336	0.31	0.37	0.93	-0.121	-0.30	0.12	0.293	0.27	0.31	0.91
			10	0.085	-0.15	0.30	0.310	0.28	0.35	0.93	-0.018	-0.20	0.21	0.283	0.26	0.31	0.94
		HBL	1	0.040	-0.20	0.28	0.338	0.31	0.38	0.94	-0.219	-0.42	-0.02	0.292	0.27	0.32	0.85
			2	0.043	-0.16	0.30	0.322	0.29	0.36	0.94	-0.099	-0.31	0.10	0.286	0.27	0.31	0.92
			10	0.067	-0.14	0.26	0.307	0.28	0.34	0.94	-0.007	-0.20	0.18	0.283	0.26	0.31	0.96
	LP^2	LBL	1	0.008	-0.19	0.27	0.347	0.32	0.38	0.95	-0.243	-0.44	-0.04	0.288	0.27	0.31	0.84
			2	0.055	-0.20	0.27	0.335	0.30	0.37	0.94	-0.129	-0.31	0.08	0.287	0.26	0.31	0.90
			10	0.078	-0.11	0.30	0.320	0.28	0.36	0.90	0.001	-0.19	0.20	0.285	0.26	0.31	0.96
		HBL	1	0.042	-0.18	0.27	0.345	0.31	0.38	0.94	-0.219	-0.43	-0.04	0.290	0.27	0.32	0.85
			2	0.039	-0.18	0.29	0.328	0.30	0.37	0.94	-0.123	-0.32	0.10	0.286	0.27	0.31	0.92
			10	0.028	-0.16	0.23	0.309	0.28	0.35	0.93	-0.030	-0.22	0.17	0.282	0.26	0.31	0.94
LP^1	HP^2	LBL	1	0.049	-0.18	0.29	0.362	0.32	0.41	0.94	-0.103	-0.31	0.11	0.315	0.29	0.34	0.94
			2	0.063	-0.18	0.33	0.348	0.31	0.40	0.93	-0.043	-0.25	0.18	0.313	0.29	0.34	0.95
			10	0.071	-0.14	0.32	0.343	0.30	0.39	0.92	0.007	-0.20	0.22	0.307	0.29	0.33	0.95
		HBL	1	0.039	-0.18	0.31	0.357	0.32	0.40	0.94	-0.087	-0.29	0.13	0.317	0.29	0.34	0.94
			2	0.043	-0.19	0.28	0.341	0.31	0.39	0.91	-0.068	-0.27	0.17	0.313	0.29	0.34	0.93
			10	0.035	-0.22	0.27	0.327	0.30	0.37	0.92	-0.027	-0.24	0.20	0.308	0.29	0.33	0.95
	LP^2	LBL	1	0.077	-0.17	0.32	0.372	0.33	0.42	0.94	-0.095	-0.32	0.14	0.317	0.30	0.34	0.95
			2	0.061	-0.17	0.28	0.356	0.32	0.40	0.93	-0.054	-0.26	0.15	0.310	0.29	0.33	0.95
			10	0.068	-0.15	0.29	0.346	0.30	0.39	0.94	-0.024	-0.22	0.19	0.308	0.29	0.34	0.96
		HBL	1	0.028	-0.24	0.31	0.363	0.32	0.41	0.92	-0.117	-0.36	0.14	0.318	0.29	0.34	0.91
			2	0.023	-0.19	0.31	0.345	0.31	0.39	0.94	-0.065	-0.26	0.18	0.311	0.29	0.34	0.95
			10	0.047	-0.18	0.28	0.327	0.29	0.38	0.92	-0.012	-0.24	0.21	0.307	0.29	0.33	0.96

Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low (ν_2=0.00724,φ_2=1.14) and high (ν_2=0.00324,φ_2=1.84) baselines for second event, respectively.

Table B.7: Estimation of log relative genetic risk (β_3) of developing the second event under the recessive genetic model with a common allele frequency ($q = 30\%$) using 100 simulated families.

Parameters				Frailty model							Independent model						
T_1	T_2		k	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	0.052	-0.34	0.61	0.734	0.60	0.91	0.95	-0.097	-0.48	0.43	0.678	0.55	0.89	0.94
			2	0.050	-0.39	0.68	0.745	0.61	1.03	0.92	-0.042	-0.47	0.56	0.720	0.59	1.01	0.94
			10	0.088	-0.36	0.77	0.810	0.65	1.11	0.88	0.058	-0.40	0.74	0.906	0.66	1.07	0.96
		HBL	1	0.030	-0.27	0.43	0.533	0.45	0.66	0.93	-0.200	-0.48	0.15	0.445	0.37	0.56	0.88
			2	0.032	-0.38	0.46	0.550	0.46	0.68	0.92	-0.117	-0.48	0.24	0.478	0.40	0.61	0.90
			10	0.021	-0.35	0.47	0.547	0.44	0.68	0.92	-0.039	-0.38	0.37	0.504	0.42	0.61	0.94
	LP^2	LBL	1	0.039	-0.45	0.61	0.718	0.58	0.93	0.95	-0.096	-0.52	0.46	0.659	0.54	0.84	0.94
			2	0.002	-0.50	0.55	0.739	0.59	1.02	0.93	-0.085	-0.57	0.45	0.717	0.57	1.01	0.94
			10	0.121	-0.37	0.78	0.845	0.66	1.08	0.87	0.089	-0.40	0.73	0.882	0.68	1.07	0.96
		HBL	1	0.012	-0.35	0.44	0.540	0.46	0.64	0.92	-0.153	-0.47	0.24	0.459	0.38	0.55	0.90
			2	0.034	-0.37	0.39	0.541	0.45	0.65	0.92	-0.070	-0.40	0.26	0.479	0.40	0.59	0.91
			10	-0.026	-0.36	0.39	0.530	0.43	0.67	0.89	-0.078	-0.38	0.34	0.507	0.41	0.62	0.94
LP^1	HP^2	LBL	1	0.029	-0.40	0.60	0.707	0.59	0.89	0.93	-0.054	-0.44	0.46	0.640	0.55	0.81	0.95
			2	0.128	-0.46	0.66	0.752	0.61	1.05	0.87	0.025	-0.48	0.57	0.725	0.59	1.03	0.94
			10	0.221	-0.35	0.94	0.856	0.67	1.12	0.85	0.161	-0.38	0.88	0.919	0.67	1.09	0.95
		HBL	1	0.039	-0.36	0.54	0.583	0.48	0.71	0.94	-0.095	-0.47	0.29	0.480	0.40	0.58	0.91
			2	0.083	-0.36	0.48	0.553	0.47	0.68	0.91	-0.066	-0.42	0.35	0.481	0.41	0.58	0.90
			10	0.005	-0.36	0.46	0.558	0.45	0.70	0.89	-0.040	-0.38	0.34	0.521	0.43	0.63	0.93
	LP^2	LBL	1	0.182	-0.32	0.89	0.737	0.60	1.06	0.93	0.103	-0.33	0.77	0.696	0.56	1.03	0.96
			2	0.026	-0.48	0.63	0.752	0.62	1.02	0.90	-0.014	-0.49	0.58	0.719	0.59	1.01	0.93
			10	0.109	-0.41	0.91	0.817	0.67	1.11	0.85	0.034	-0.41	0.83	0.823	0.66	1.10	0.95
		HBL	1	-0.013	-0.43	0.42	0.560	0.47	0.69	0.91	-0.073	-0.42	0.31	0.485	0.40	0.59	0.90
			2	0.058	-0.34	0.49	0.575	0.47	0.71	0.89	-0.002	-0.37	0.38	0.503	0.41	0.61	0.90
			10	0.039	-0.38	0.52	0.570	0.46	0.73	0.91	0.004	-0.39	0.46	0.517	0.42	0.66	0.94

Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low (ν_2=0.00724,φ_2=1.14) and high (ν_2=0.00324,φ_2=1.84) baselines for second event, respectively.

Table B.8: Penetrance estimation of male mutation carriers for the first event by the age of 70 years under the recessive genetic model with a common allele frequency ($q = 30\%$) using 100 simulated families.

Parameters				Pen	Frailty model							Independent model						
T_1	T_2		k	(%)	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	66	0.001	-0.05	0.05	0.065	0.06	0.07	0.93	0.068	0.03	0.11	0.056	0.05	0.06	0.71
			2	74	0.001	-0.04	0.05	0.064	0.06	0.07	0.93	0.045	0.01	0.08	0.052	0.05	0.06	0.81
			10	83	-0.008	-0.04	0.03	0.057	0.05	0.07	0.93	0.012	-0.02	0.04	0.043	0.04	0.05	0.89
		HBL	1	66	0.006	-0.04	0.05	0.063	0.06	0.07	0.93	0.070	0.04	0.11	0.057	0.05	0.06	0.72
			2	74	0.001	-0.03	0.04	0.060	0.06	0.07	0.95	0.044	0.01	0.08	0.051	0.05	0.06	0.81
			10	83	-0.001	-0.04	0.03	0.052	0.04	0.06	0.96	0.015	-0.02	0.04	0.043	0.04	0.05	0.93
	LP^2	LBL	1	66	0.007	-0.05	0.05	0.067	0.06	0.07	0.92	0.076	0.04	0.11	0.057	0.05	0.06	0.72
			2	74	-0.003	-0.04	0.04	0.065	0.06	0.07	0.93	0.043	0.01	0.08	0.052	0.05	0.06	0.85
			10	83	-0.007	-0.04	0.02	0.058	0.05	0.07	0.95	0.015	-0.01	0.04	0.043	0.04	0.05	0.93
		HBL	1	66	0.005	-0.04	0.05	0.064	0.06	0.07	0.94	0.071	0.03	0.11	0.057	0.05	0.06	0.75
			2	74	0.001	-0.04	0.04	0.062	0.06	0.07	0.94	0.041	0.01	0.08	0.052	0.05	0.06	0.83
			10	83	-0.007	-0.04	0.03	0.054	0.04	0.06	0.92	0.009	-0.02	0.04	0.044	0.04	0.05	0.90
LP^1	HP^2	LBL	1	43	-0.012	-0.06	0.04	0.069	0.06	0.08	0.93	0.027	-0.02	0.08	0.066	0.06	0.07	0.91
			2	47	-0.006	-0.05	0.04	0.071	0.06	0.08	0.91	0.021	-0.02	0.07	0.067	0.06	0.07	0.92
			10	52	-0.019	-0.07	0.03	0.072	0.07	0.08	0.92	0.002	-0.05	0.04	0.067	0.06	0.07	0.93
		HBL	1	43	-0.006	-0.05	0.04	0.069	0.06	0.07	0.94	0.033	-0.01	0.07	0.067	0.06	0.07	0.92
			2	47	-0.006	-0.05	0.04	0.069	0.06	0.08	0.92	0.014	-0.02	0.07	0.067	0.06	0.07	0.93
			10	52	-0.017	-0.07	0.02	0.069	0.06	0.08	0.93	-0.003	-0.05	0.04	0.067	0.06	0.07	0.96
	LP^2	LBL	1	43	-0.010	-0.06	0.04	0.071	0.06	0.08	0.92	0.033	-0.01	0.07	0.067	0.06	0.07	0.93
			2	47	-0.009	-0.06	0.04	0.072	0.06	0.08	0.91	0.022	-0.02	0.07	0.067	0.06	0.07	0.91
			10	52	-0.023	-0.08	0.02	0.073	0.07	0.08	0.91	-0.000	-0.05	0.04	0.067	0.06	0.07	0.94
		HBL	1	43	-0.006	-0.06	0.04	0.068	0.06	0.08	0.92	0.025	-0.02	0.08	0.066	0.06	0.07	0.93
			2	47	-0.018	-0.06	0.04	0.070	0.06	0.08	0.92	0.012	-0.03	0.06	0.067	0.06	0.07	0.94
			10	52	-0.013	-0.07	0.03	0.071	0.06	0.08	0.91	0.002	-0.05	0.04	0.067	0.06	0.07	0.95

Pen - penetrance; Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low ($\nu_2 = 0.0072$, $\varphi_2 = 1.14$) and high ($\nu_2 = 0.0032$, $\varphi_2 = 1.84$) baselines for second event, respectively.

Table B.9: Penetrance estimation of female mutation carriers for the first event by the age of 70 years under the recessive genetic model with a common allele frequency ($q = 30\%$) using 100 simulated families.

Parameters				Pen	Frailty model							Independent model						
T_1	T_2		k	(%)	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	37	-0.002	-0.04	0.04	0.057	0.05	0.06	0.92	0.005	-0.04	0.05	0.063	0.06	0.07	0.93
			2	40	-0.004	-0.04	0.04	0.061	0.06	0.06	0.93	-0.003	-0.04	0.05	0.065	0.06	0.07	0.94
			10	44	-0.002	-0.05	0.04	0.064	0.06	0.07	0.91	-0.002	-0.05	0.04	0.066	0.06	0.07	0.93
		HBL	1	37	0.003	-0.04	0.04	0.057	0.05	0.06	0.95	0.008	-0.04	0.05	0.063	0.06	0.07	0.95
			2	40	-0.002	-0.04	0.04	0.061	0.06	0.06	0.95	0.000	-0.04	0.05	0.064	0.06	0.07	0.95
			10	44	0.002	-0.04	0.04	0.065	0.06	0.07	0.93	0.003	-0.05	0.04	0.065	0.06	0.07	0.93
	LP^2	LBL	1	37	-0.003	-0.04	0.04	0.057	0.05	0.06	0.93	-0.001	-0.04	0.05	0.062	0.06	0.07	0.93
			2	40	-0.003	-0.04	0.05	0.061	0.06	0.06	0.94	-0.004	-0.04	0.05	0.064	0.06	0.07	0.94
			10	44	-0.004	-0.05	0.04	0.064	0.06	0.07	0.93	-0.003	-0.05	0.04	0.065	0.06	0.07	0.94
		HBL	1	37	0.001	-0.04	0.04	0.057	0.05	0.06	0.94	0.003	-0.04	0.05	0.063	0.06	0.07	0.94
			2	40	-0.004	-0.04	0.04	0.061	0.06	0.06	0.92	-0.003	-0.04	0.04	0.064	0.06	0.07	0.91
			10	44	-0.008	-0.06	0.04	0.064	0.06	0.07	0.94	-0.008	-0.06	0.04	0.065	0.06	0.07	0.94
LP^1	HP^2	LBL	1	19	-0.001	-0.04	0.02	0.047	0.04	0.05	0.92	-0.000	-0.04	0.03	0.050	0.04	0.05	0.92
			2	19	-0.002	-0.03	0.03	0.048	0.04	0.05	0.93	-0.003	-0.03	0.03	0.050	0.05	0.06	0.94
			10	20	0.001	-0.03	0.03	0.051	0.05	0.06	0.94	0.001	-0.03	0.03	0.051	0.05	0.06	0.93
		HBL	1	19	-0.004	-0.04	0.03	0.046	0.04	0.05	0.91	-0.004	-0.04	0.03	0.048	0.04	0.05	0.91
			2	19	-0.004	-0.03	0.04	0.048	0.04	0.05	0.91	-0.005	-0.04	0.04	0.050	0.04	0.06	0.92
			10	20	-0.003	-0.03	0.04	0.050	0.04	0.06	0.91	-0.003	-0.03	0.03	0.051	0.05	0.06	0.91
	LP^2	LBL	1	19	-0.000	-0.04	0.03	0.046	0.04	0.05	0.94	0.001	-0.04	0.03	0.049	0.04	0.06	0.94
			2	19	0.000	-0.03	0.04	0.049	0.04	0.06	0.92	0.000	-0.03	0.04	0.051	0.04	0.06	0.92
			10	20	-0.005	-0.03	0.03	0.050	0.04	0.05	0.94	-0.006	-0.03	0.03	0.051	0.05	0.06	0.94
		HBL	1	19	-0.007	-0.04	0.02	0.047	0.04	0.05	0.93	-0.008	-0.04	0.03	0.049	0.04	0.05	0.92
			2	19	0.001	-0.03	0.03	0.049	0.04	0.05	0.93	0.002	-0.03	0.04	0.050	0.05	0.06	0.93
			10	20	-0.003	-0.04	0.04	0.051	0.04	0.06	0.90	-0.005	-0.04	0.04	0.051	0.05	0.06	0.91

Pen - penetrance; Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low ($\nu_2 = 0.0072$, $\varphi_2 = 1.14$) and high ($\nu_2 = 0.0032$, $\varphi_2 = 1.84$) baselines for second event, respectively.

Table B.10: Penetrance estimation of mutation carriers for developing the second event in 10 years after the first event under the recessive genetic model with a common allele frequency ($q = 30\%$) using 100 simulated families.

Parameters				Pen	Frailty model							Independent model						
T_1	T_2		k	(%)	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP	Bias	Q_1^*	Q_3^*	SE	$Q_1^\dagger$	$Q_3^\dagger$	CP
HP^1	HP^2	LBL	1	17	-0.002	-0.02	0.04	0.044	0.04	0.06	0.87	0.109	0.09	0.14	0.037	0.04	0.04	0.18
			2	18	0.006	-0.03	0.04	0.043	0.03	0.06	0.86	0.063	0.04	0.09	0.034	0.03	0.04	0.59
			10	19	-0.012	-0.04	0.01	0.037	0.03	0.05	0.89	0.009	-0.01	0.03	0.032	0.03	0.03	0.93
		HBL	1	32	0.001	-0.04	0.05	0.066	0.06	0.08	0.88	0.158	0.13	0.19	0.042	0.04	0.04	0.03
			2	35	-0.007	-0.05	0.05	0.063	0.05	0.07	0.86	0.090	0.06	0.11	0.040	0.04	0.04	0.39
			10	37	-0.011	-0.05	0.03	0.053	0.04	0.07	0.91	0.019	-0.01	0.05	0.039	0.04	0.04	0.92
	LP^2	LBL	1	12	0.003	-0.02	0.03	0.034	0.03	0.05	0.87	0.080	0.06	0.10	0.032	0.03	0.03	0.33
			2	12	-0.002	-0.02	0.02	0.033	0.03	0.04	0.85	0.040	0.02	0.06	0.029	0.03	0.03	0.76
			10	13	-0.010	-0.04	0.00	0.030	0.02	0.04	0.88	0.005	-0.02	0.02	0.026	0.02	0.03	0.91
		HBL	1	23	-0.001	-0.03	0.05	0.053	0.04	0.06	0.88	0.127	0.10	0.16	0.039	0.04	0.04	0.11
			2	25	-0.001	-0.04	0.04	0.052	0.04	0.06	0.87	0.070	0.04	0.09	0.037	0.04	0.04	0.54
			10	26	-0.007	-0.04	0.02	0.043	0.04	0.06	0.88	0.013	-0.01	0.03	0.035	0.03	0.04	0.93
LP^1	HP^2	LBL	1	17	-0.008	-0.06	0.08	0.052	0.04	0.09	0.66	0.126	0.10	0.16	0.044	0.04	0.05	0.19
			2	18	-0.005	-0.06	0.06	0.048	0.04	0.08	0.73	0.067	0.04	0.10	0.041	0.04	0.04	0.68
			10	19	-0.026	-0.07	0.01	0.043	0.04	0.07	0.82	0.010	-0.01	0.04	0.038	0.04	0.04	0.94
		HBL	1	32	0.005	-0.09	0.10	0.097	0.07	0.14	0.76	0.181	0.14	0.22	0.050	0.05	0.05	0.04
			2	35	0.006	-0.09	0.08	0.090	0.05	0.12	0.74	0.104	0.06	0.14	0.049	0.05	0.05	0.46
			10	37	-0.020	-0.09	0.03	0.059	0.05	0.10	0.83	0.020	-0.01	0.05	0.046	0.04	0.05	0.90
	LP^2	LBL	1	12	-0.010	-0.04	0.04	0.045	0.03	0.07	0.73	0.093	0.07	0.12	0.039	0.04	0.04	0.36
			2	12	-0.010	-0.04	0.04	0.039	0.03	0.06	0.77	0.048	0.03	0.08	0.035	0.03	0.04	0.74
			10	13	-0.025	-0.06	0.00	0.034	0.03	0.06	0.78	0.007	-0.02	0.02	0.032	0.03	0.04	0.93
		HBL	1	23	0.001	-0.07	0.10	0.076	0.05	0.10	0.71	0.149	0.12	0.18	0.047	0.04	0.05	0.13
			2	25	-0.001	-0.08	0.06	0.064	0.04	0.10	0.71	0.083	0.05	0.11	0.044	0.04	0.05	0.52
			10	26	-0.019	-0.09	0.02	0.049	0.04	0.07	0.80	0.018	-0.01	0.04	0.042	0.04	0.04	0.93

Pen - penetrance; Bias - median bias; SE - robust standard error; CP - coverage probability.

Q_1^* and Q_3^* - First and third quartiles of bias, respectively.

$Q_1^\dagger$ and $Q_3^\dagger$ - First and third quartiles of robust standard error, respectively.

HP^1 and LP^1 - high ($\beta_2 = 2.5$) and low ($\beta_2 = 1.55$) penetrances for the first event, respectively.

HP^2 and LP^2 - high ($\beta_3 = 0.75$) and low ($\beta_3 = 0.3$) penetrances for the second event, respectively.

LBL and HBL - low ($\nu_2 = 0.0072$, $\varphi_2 = 1.14$) and high ($\nu_2 = 0.0032$, $\varphi_2 = 1.84$) baselines for second event, respectively.

BIBLIOGRAPHY

Aalen, O. O. (1994). Effects of frailty in survival analysis. *Statistical methods in medical research*, 3(3):227–243.

Aalen, O. O. and Tretli, S. (1999). Analyzing incidence of testis cancer by means of a frailty model. *Cancer Causes and Control*, 10(4):285–292.

Aarnio, M., Mecklin, J.-P., Aaltonen, L. A., Nystrom-Lahti, M., and Jarvinen, H. J. (1995). Life-time risk of different cancers in hereditary non-polyposis colorectal cancer (HNPCC) syndrome. *International Journal of Cancer*, 64(6):430–433.

Alarcon, F., Bonaïti-Pellié, C., and Harari-Kermadec, H. (2009). A nonparametric method for penetrance function estimation. *Genetic epidemiology*, 33(1):38–44.

Alarcon, F., Bourgain, C., Gauthier-Villars, M., Planté-Bordeneuve, V., Stoppa-Lyonnet, D., and Bonaïti-Pellié, C. (2008). PEL: An unbiased method for estimating age-dependent genetic disease risk from pedigree data unselected for family history. *Genetic epidemiology*, 33(5):379–385.

Bonney, G. E. (1998). Ascertainment corrections based on smaller family units. *American Journal of Human Genetics*, 63(4):1202–1215.

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

Carayol, J. and Bonaïti-Pellié, C. (2004). Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset. *Genetic epidemiology*, 27(2):109–117.

Chatterjee, N. and Wacholder, S. (2001). A marginal likelihood approach for estimating penetrance from kin-cohort designs. *Biometrics*, 57(1):245–252.

Chen, L., Hsu, L., and Malone, K. (2009). A frailty-model-based approach to estimating the age-dependent penetrance function of candidate genes using population-based case-control study designs: An application to data on the BRCA1 gene. *Biometrics*, 65(4):1105–1114.

Choi, Y.-H. (2012). A Frailty-Model-Based Method for Estimating Age-Dependent Penetrance from Family Data. *J Biomet Biostat*, S4(001).

Choi, Y.-H. and Briollais, L. (2011). An EM composite likelihood approach for Multistage sampling of family data. *Statistica Sinica*, 21(1):231–253.

Choi, Y.-H., Kopciuk, K. A., and Briollais, L. (2008). Estimating disease risk associated with mutated genes in family-based designs. *Human heredity*, 66(4):238–251.

Clayton, D. (2003). Conditional likelihood inference under complex ascertainment using data augmentation. *Biometrika*, 90(4):976–981.

Cook, R. and Lawless, J. (2010). *The Statistical Analysis of Recurrent Events*. Statistics for Biology and Health. Springer.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B Methodological*, 34(2):187–220.

Easton, D. F., Ford, D., Bishop, D. T., Haites, N., Milner, B., Allan, L., Easton, D. F., Ponder, B. A. J., Peto, J., Smith, S., Ford, D., Stratton, M., Narod, S. A., Lenoir, G. M., Feunteun, J., Lynch, H., Arason, A., Barkardottir, R., and Egilsson, V. (1995). Breast and ovarian cancer incidence in BRCA1-mutation carriers. *American Journal of Human Genetics*, 56(1):265–271.

Elston, R. C. and Sobel, E. (1979). Sampling considerations in the gathering and analysis of pedigree data. *American Journal of Human Genetics*, 31(1):62–69.

Ewens, W. and Elston, R. C. (2012). *Correcting for ascertainment*, volume 850 of *Methods in Molecular Biology*.

Fisher, R. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, 6:13–25.

Fitzgibbons Jr., R. J., Lynch, H. T., Stanislav, G. V., Watson, P. A., Lanspa, S. J., Marcus, J. N., Smyrk, T., Kriegler, M. D., and Lynch, J. F. (1987). Recognition and treatment of patients with hereditary nonpolyposis colon cancer (Lynch syndromes I and II). *Annals of Surgery*, 206(3):289–295.

Fleming, T. R. and Lin, D. Y. (2000). Survival analysis in clinical trials: Past developments and future directions. *Biometrics*, 56(4):971–983.

Frydman, H. (1992). A non-parametric estimation procedure for a periodically observed three-state Markov process, with application to AIDS. *Journal of the Royal Statistical Society*, 54(Series B):853–866.

Frydman, H. (1995). Semiparametric estimation in a three-state duration-dependent Markov model from interval-censored observations with application to AIDS data. *Biometrics*, 51(2):502–511.

Gail, M. H., Pee, D., Benichou, J., and Carroll, R. (1999). Designing studies to estimate the penetrance of an identified autosomal dominant mutation: Cohort, case-control, and genotyped-proband designs. *Genetic epidemiology*, 16(1):15–39.

Gail, M. H., Pee, D., and Carroll, R. (2001). Effects of violations of assumptions on likelihood methods for estimating the penetrance of an autosomal dominant mutation from kin-cohort studies. *Journal of Statistical Planning and Inference*, 96(1):167–177.

Gauderman, W. J. (1995). A method for simulating familial disease data with variable age at onset and genetic and environmental effects. *Statistics and Computing*, 5(3):237–243.

Gauderman, W. J., Witte, J. S., and Thomas, D. C. (1999). Family-based association studies. *Journal of the National Cancer Institute.Monographs*, (26):31–37.

Goethals, K., Janssen, P., and Duchateau, L. (2008). Frailty models and copulas: Similarities and differences. *Journal of Applied Statistics*, 35(9):1071–1079.

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

He, W. and Lawless, J. F. (2003). Flexible Maximum Likelihood Methods for Bivariate Proportional Hazards Models. *Biometrics*, 59(4):837–848.

Hopper, J. L., Bishop, D. T., and Easton, D. F. (2005). Population-based family studies in genetic epidemiology. *Lancet*, 366(9494):1397–1406.

Horowitz, J. L. (1999). Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. *Econometrica*, 67(5):1001–1028.

Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, 71(1):75–83.

Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika*, 73(3):671–678.

Janssen, J. and Limnios, N. (1999). *Semi-Markov Models and Applications*. Kluwer Academic Publishers.

Joly, P. and Commenges, D. (1999). A penalized likelihood approach for a progressive three-state model with censored and truncated data: Application to AIDS. *Biometrics*, 55(3):887–890.

Kopciuk, K. A., Choi, Y.-H., Parkhomenko, E., Parfrey, 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).

Kraft, P. and Thomas, D. C. (2000). Bias and efficiency in family-based gene-characterization studies: Conditional, prospective, retrospective, and joint likelihoods. *American Journal of Human Genetics*, 66(3):1119–1131.

Laird, N. M. and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics*, 7(5):385–394.

Lawless, J. F. and Yilmaz, Y. E. (2011). Semiparametric estimation in copula models for bivariate sequential survival times. *Biometrical Journal*, 53(5):779–796.

Le Bihan, C., Moutou, C., Brugieres, L., Feunteun, J., and Bonaiti-Pellie, C. (1995). ARCAD: A method for estimating age-dependent disease risk associated with mutation carrier status from family data. *Genetic epidemiology*, 12(1):13–25.

Lin, K. M., Shashidharan, M., Ternent, C. A., Thorson, A. G., Blatchford, G. J., Christensen, M. A., Lanspa, S. J., Lemon, S. J., Watson, P., and Lynch, H. T. (1998). Colorectal and Extracolonic cancer variations in MLH1/MSH2 Hereditary nonpolyposis colorectal cancer kindreds and the general population. *Diseases of the colon and rectum*, 41(4):428–433.

Lynch, H. T., Harris, R. E., and Lynch, P. M. (1977). Role of heredity in multiple primary cancer. *Cancer*, 40(4 , Suppl.):1849–1854.

Lynch, H. T., Lynch, J. F., and Attard, T. A. (2009). Diagnosis and management of hereditary colorectal cancer syndromes: Lynch syndrome as a model. *CMAJ*, 181(5):273–280.

Lynch, H. T. and Smyrk, T. (1996). Hereditary nonpolyposis colorectal cancer (Lynch syndrome): An updated review. *Cancer*, 78(6):1149–1167.

Mecklin, J.-P. and Jarvïnen, H. J. (1986). Clinical features of colorectal carcinoma in cancer family syndrome. *Diseases of the colon and rectum*, 29(3):160–164.

Meira-Machado, L. F., de Uña-Álvarez, J., Cadarso-Suárez, C., and Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical methods in medical research*, 18(2):195–222.

Myrhøj, T., Bisgaard, M. L., Bernstein, I., Svendsen, L. B., Søndergaard, J. O., and Bülow, S. (1997). Hereditary non-polyposis colorectal cancer: Clinical features

and survival: results from the danish HNPCC register. *Scandinavian Journal of Gastroenterology*, 32(6):572–576.

Nelsen, R. (2010). *An Introduction to Copulas*. Springer Series in Statistics. Springer.

Parry, S., Win, A. K., Parry, B., Macrae, F. A., Gurrin, L. C., Church, J. M., Baron, J. A., Giles, G. G., Leggett, B. A., Winship, I., Lipton, L., Young, G. P., Young, J. P., Lodge, C. J., Southey, M. C., Newcomb, P. A., Le Marchand, L., Haile, R. W., Lindor, N. M., Gallinger, S., Hopper, J. L., and Jenkins, M. A. (2011). Metachronous colorectal cancer risk for mismatch repair gene mutation carriers: The advantage of more extensive colon surgery. *Gut*, 60(7):950–957.

Putter, H., Fiocco, M., and Gekus, R. B. (2007). Tutorial in biostatistics: Competing risk and multi-state models. *Statistics in medicine*, 26(11):2389–2430.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sastry, N. (1997). A nested frailty model for survival data, with an application to the study of child survival in Northeast Brazil. *Journal of the American Statistical Association*, 92(438):426–435.

Siegmund, K. D., Whittemore, A. S., and Thomas, D. C. (1999). Multistage sampling for disease family registries. *Journal of the National Cancer Institute.Monographs*, (26):43–48.

Struewing, J. P., Hartge, P., Wacholder, S., Baker, S. M., Berlin, M., McAdams, M., Timmerman, M. M., Brody, L. C., and Tucker, M. A. (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *New England Journal of Medicine*, 336(20):1401–1408.

Thomas, D. C. (1999). Design of gene characterization studies: an overview. *Journal of the National Cancer Institute.Monographs*, (26):17–23.

Thomas, D. C. (2004). *Statistical methods in genetic epidemiology*. Oxford University Press.

Umar, A., Boland, C. R., Terdiman, J. P., Syngal, S., de la Chapelle, A., Rüschoff, J., Fishel, R., Lindor, N. M., Burgart, L. J., Hamelin, R., Hamilton, S. R., Hiatt, R. A., Jass, J., Lindblom, A., Lynch, H. T., Peltomaki, P., Ramsey, S. D., Rodriguez-Bigas, M. A., Vasen, H. F. A., Hawk, E. T., Barrett, J. C., Freedman, A. N., and Srivastava, S. (2004). Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *Journal of the National Cancer Institute*, 96(4):261–268.

Vasen, H. F. A., Watson, P., Mecklin, J.-P., and Lynch, H. T. (1999). New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative Group on HNPCC. *Gastroenterology*, 116(6):1453–1456.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

Vaupel, J. W. and Yashin, A. I. (1983). The deviant dynamics of death in heterogeneous populations. *International Institute for Applied Systems Analysis, Research Report*, (RR-83-1).

Vaupel, J. W. and Yashin, A. I. (1985). Heterogeneity's ruses: some surprising effects of selection on population dynamics. *American Statistician*, 39(3):176–185.

Wacholder, S., Hartge, P., Struewing, J. P., Pee, D., McAdams, M., Brody, L., and Tucker, M. (1998). The kin-cohort study for estimating penetrance. *American Journal of Epidemiology*, 148(7):623–630.

Weinberg, W. (1912). Further contributions to the theory of heredity. Part 5. On the inheritance of the predisposition to blood disease with methodological supplements to my sibship method. *Cancer*, Arch fur Rassen und Gesellschaftsbiologie(9):694–709.

White, H. (1982a). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

White, J. E. (1982b). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115(1):119–128.

Whittemore, A. S. and Halpern, J. (1997). Multi-stage sampling in genetic epidemiology. *Statistics in medicine*, 16(1-3):153–167.

Wienke, A. (2009). *Frailty models in survival analysis*. Chapman & Hall/CRC biostatistics series. Taylor and Francis.

Vita

- **Name**
  Balakumar Swaminathan

- **Post-secondary Education and Degrees**

  - Loyola College, Chennai, India
    2005-2008, B.Sc. in Statistics
  - Loyola College, Chennai, India
    2008-2010, M.Sc. in Statistics
  - The University of Western Ontario, London, Canada
    2010-2012, M.Sc. in Epidemiology and Biostatistics

- **Work Experience**
  Research Assistant, 2010-2012, The University of Western Ontario, London, Canada

- **Scholarship**
  Western Graduate Research Scholarship, 2010-2012, The University of Western Ontario, London, Canada