11-29-2017 11:00 AM

A comparison of three prediction modelling approaches for clustered survival data with application to Lynch Syndrome Family

Bing Yu, *The University of Western Ontario*

Supervisor: Yun-Hee Choi, *The University of Western Ontario*

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

Abstract

The purpose of this study was to compare shared frailty model, joint frailty model and joint nested frailty model in terms of model fitting and prediction accuracy, as applied to Lynch Syndrome family data. The specific question we wanted to address was how the intervals between screening visits affect the risk of developing colorectal cancer among Lynch Syndrome family members. We also addressed questions on how the screening process has an effect on mortality and risks of developing different stages of colorectal cancer. Results from the models show that joint nested frailty model is preferable. This model improves the prediction accuracy by jointly modeling recurrent screenings and terminal event at the same time accounts for both individual and familial correlation.

Keywords: Joint modeling; frailty; cancer screening; dynamic prediction; recurrent data; time-to-event data; family data.

Acknowlegements

The completion of this thesis would not have been possible without the support of many people. First and foremost, I thank my supervisor Dr. Yun-Hee Choi of the Department of Epidemiology and Biostatistics at Western University. The door to Dr. Choi's office was always open whenever I ran into a bug in my codes or had questions about my research or writing. She allowed me the room to work on the field of my interest and steered me in the right direction whenever she thought I needed it. One could not wish for a better or friendlier supervisor.

I am also indebted to Dr. Laurent Biollais of Lunenfeld-Tanenbaum Research Institute at Mount Sinai Hospital, who deserves the credit for initiating the project. Dr. Briollais aided his effort in enlightening me with new ideas and developing the models. His encouragement and insights have been invaluable. Special thanks to him for giving me the access to Mount Sinai Hospital data.

Dr. Guangyong (GY) Zou, of the Department of Epidemiology and Biostatistics at Western University, was my supervisory committee member. He guided me throughout the grammatical edit of my thesis. I appreciate the valuable feedback offered by him.

I thank Mount Sinai Hospital for permission to use Lynch Syndrome family data for my analysis.

Finally, I must express my very profound gratitude to my parents for supporting and encouraging me wholeheartedly throughout all my studies at University. This accomplishment would not have been possible without them.

Contents

List of Figures

List of Tables

List of Abbreviations

LS	Lynch Syndrome
CRC	Colorectal Cancer
MMR	mismatch repair
JNFM	Joint nested frailty model
MSE	mean square error
se	standard error
sd	standard deviation
HR	hazard ratio
FGICR	Familial Gastrointestinal Cancer Registry
ANOVA	analysis of variance
HPV	human papiloma virus
PCF	piecewise constant functions
EM	Expectation-Maximization
ROC	Receiver Operating Characteristics
AUC	Area Under the Curve
BS	Brier Score
IPCWE	inverse probability of censoring weighted error
CIs	confidence intervals
LCV	likelihood cross-validation criterion

Chapter 1

Introduction

Lynch Syndrome (LS) is a hereditary cancer caused by mutations in DNA mismatch repair (MMR) genes such as MLH1, MSH2 and MSH6, and predisposes carriers mainly to colorectal cancer (CRC) and other extra-colonic cancers (Lynch et al., 2009). Patients with LS have a noticeable increased chance (70%-80%) of developing CRC in their lifetime, often at a young age, usually under age 50 (American Society of Clinical Oncology, 2017). Not all mutation carriers will develop a cancer. Carriers have variable expressivity, leading to phenotypic heterogeneity. In Canada, CRC is the second most commonly diagnosed cancer (excluding non-melanoma skin cancers) and the second leading cause of death from cancer in men and the third leading cause of death from cancer in women (Canadian Cancer Society, 2016). Although environment factors are the dominant reasons that cause CRC, hereditary factors have been found to account for 10%-15% of all cases (Vasen, Mslein, Alonso and et al., 2010). Three percent of CRC incidence is estimated to account for by LS (Hampel et al., 2008). However, CRC is a highly treatable cancer if it is detected early and it is up to 90 percent preventable with opportune and thorough CRC screenings (Smith et al., 2001). Unfortunately, nearly half of those diagnoses find out too late (Colon Cancer Canada, 2013). Colonoscopy screenings are recommended for individuals in LS family every 1-2 years starting at 20-25 years of age (Stuckless, 2012; Vasen et al., 2007; Lindor et al., 2006; Canadian Cancer Society, 2016). These recommendations are mostly based on crude analyses, without considering the complexity of the screening visits and disease process, by comparing the incidence and the mortality of CRC in screening group and non-screening group (Jrvinen et al., 2000; Stupart et al., 2009; Vasen, Abdirahman, Brohet and et al., 2010). The screening visit process can have an effect on the disease risk and mortality; meanwhile, disease risk can also influence the visit process. In other words, those with high risk of CRC tend to have more frequent screening visits. This motivates us to use joint modeling of the screening visits and the disease occurence. In addition, some unmeasured but influential factors induce correlations within individuals and within families. For example, a positive screening test in an individual with LS may increase the number of further screening visits for this individual and also encourage the screening visits for other individuals in the same family (Thomas, 2017).

In this thesis, our primary interest was to evaluate the screening efficiency on the risk of developing CRC using three modeling approaches – joint frailty, joint nested frailty and shared frailty models – by taking the complexity of screening visits and the information (adenoma or other polyp detection, visit age etc.) obtained from screening visits into account. We aimed to evaluate the prediction abilities of the three models for detecting CRC depending on various factors such as screening intervals, first visit age, gender, and adenoma detection and removal status. Adenomas, also called adenomatous polyps, are found commonly at colonoscopy. They are removed after detection because of their tendency to become malignant and lead to CRC.

We employed three approaches for modeling screening visits and times to CRC and compared their prediction ability. The evaluations of the three models are applied based on LS family data collected by the Familial Gastrointestinal Cancer Registry (FGICR) in the Zane Cohen Centre at Mount Sinai hospital. The FGICR is a family study center located at Mount Sinai Hospital in Toronto that focuses on families affected by rare familial gastrointestinal cancer syndromes to provide a secondary prevention of cancer through early diagnosis and treatment. Lynch Syndrome is one of their main focuses. They used the following criteria to identify LS: three family members, two of whom must be first-degree relatives, in two successive generations, must have colorectal cancer or colorectal cancer and a combination of gynecologic, genitourinary, or other gastrointestinal cancer (http://www.zanecohencentre.com/

gi-cancers/fgicr/about-us). In addition, anyone diagnosed with CRC before age 35, or diagnosed with colon cancer and one other LS-related cancer, was included into the study regardless of their family history.

From FGICR, 423 LS families with 1068 unique patients have been identified and offered with DNA testing. The data consist of the following: demographic information including age at study entry, and gender; familial information such as familiy relation, proband (the first affected individual in a family) indicator, and type of mutation genes; and medical information including cancer stage, surgery information of eligible individuals. We also obtained the data at each screening visit, including ages at screening visit, detection of adenoma and other polyps, detection of CRC, cancer stage, proportion of removed colon and so on. Screening intervals vary within and between individuals. For modeling gap times between screening visits, we included individuals with at least two screening visits in our analysis, which led to 242 LS families with 422 individuals. Main variables we considered for data analysis were gender, screening visit ages, age at death, type of mutation genes, site of detected cancer, cancer stage, and polyp information.

1.1 Objectives

The objectives of our study were as follows:

1. Compare and evaluate the impact of screening visits (screening intervals) on the risk of developing a first CRC for LS families, using three modeling approaches: a joint frailty model considering individual dependence, a joint nested frailty model considering both individual and familial dependences, and a shared frailty model for terminal event only, taking time-dependent and time-independent covariates of interest into account.

2. Provide dynamic predictions of CRC risks for individuals in LS families based on individual's visit history and the visit and disease histories of other family members. A valid assessment of predictive accuracy for joint nested frailty model is provided with

comparisons to the joint frailty model and the shared frailty model.

3. Assess mutation type specific risks of developing CRC and associated effects of covariates.

4. Identify and examine prognostic factors for individuals with CRC using survival time after the first CRC. Covariates are considered including detection of adenoma before CRC, age at CRC, average gap time of screening visits, cumulative number of visits after detection of CRC, cancer stage and proportion of colon removed, and gender.

5. Assess the effect of screening process on the risk to detect low stage cancer and high stage cancer.

1.2 Organization of the thesis

The remainder of the thesis is structured as follow. Chapter 2 presents a literature review on joint frailty model, joint nested frailty model, maximum penalized likelihood estimation, and dynamic prediction. Chapter 3 describes the three statistical models and their dynamic predictions formulations. Our models and dynamic predictions are applied to LS family data in Chapter 4. Chapter 5 evaluates the performance of predictions under the three statistical models. Future research and some discussion are presented in Chapter 6.

Chapter 2

Literature Review

2.1 Cancer Screening

Cancer screening, the routine testing of asymptomatic individuals without a history of the disease of interest, is an important aspect of cancer prevention and control. By the time symptoms appear, cancer may have begun to spread and be harder to treat or cure. Early detection of abnormal tissue or cancer can increase the chance of successful treatment. Regular screening visits benefit individuals at high risk of cancer from early cancer detection and prevention.

Many authors have already reported evidence in support of surveillance colonoscopy on cancer risks mostly based on observational studies (Jrvinen et al., 2000; Stupart et al., 2009; Vasen, Abdirahman, Brohet and et al., 2010). Jrvinen et al. (2000) compared incidence of CRC and survival rates in two cohorts of 22 families with LS (133 subjects with screening and 119 subjects without). Colonoscopy screening at 3-year intervals reduced the risk of CRC by 62%, prevented CRC deaths, and decreased overall mortality by about 65% in LS families. Ladabaum and Song (2005) estimated that with screening uptake of 75%, the incidence of CRC could decrease by 17% to 54% and CRC deaths could decrease by 28% to 60%. Based on a 5-year prospective cohort study specific for 178 MLH1 gene carriers, Stupart et al. (2009) showed that 11% in screening group developed CRC, comparing to 27% in non-screening group, and 2% in screening group died from CRC, comparing to 12% in non-screening group. The average survival ages from birth for the two groups are 78 years for screening group and 55 years for non-screening group. Some randomized trails were also conducted to illustrate the effect of cancer screening. With screening intervention, a noticeable decrease in the amount of patients diagnosed with CRC was observed (Atkin et al., 2010; Segnan et al., 2011).

Mutation carriers of DNA mismatch repair (MMR) genes including MLH1, MSH2, MSH6 are at high risk of developing CRC, yet different kinds of mutated genes may have varying effects on CRC risks and mortalities. Some researches were done specifically for several main MMR genes. Mean onset ages of colorectal cancer were 44 and 46 years for MLH1 and MSH2, respectively, comparing to 69 years in the general population (Lin et al., 1998). The estimated cumulative risks of CRC by age 70 years were found significantly different: 41% for MLH1 mutation carriers, 48% for MSH2, and 12% for MSH6 (Bonadona et al., 2011). Further considering gender difference, average CRC cumulative risks at the age of 70 years for MLH1 and MSH2 mutation carriers, respectively, were estimated to be 34% and 47% for male carriers and 36% and 37% for female carriers (Dowty et al., 2013). The risk for CRC was significantly lower in MSH6 than in MLH1 or MSH2 mutation carriers (Hendriks et al., 2004; Bonadona et al., 2011), and age of CRC onset for MSH6 carriers was around ten years later than the ones for MLH1 and MSH2 gene carriers (Plaschke et al., 2004).

Screening interval also matters; more frequent screening visits can reduce the risk of developing a CRC and the mortality from CRC. A 1-2 years screening interval of colonoscopy screening is recommended starting at age 20-25 (Mandel et al., 1999; Vasen et al., 2007; Jørgensen et al., 2002; Vasen, Abdirahman, Brohet and et al., 2010; Stuckless, 2012; Canadian Cancer Society, 2016). Risk of developing CRC for members of LS families would be reduced with screening intervals of 1-2 years than with screening intervals of 2-3 years; 6% with intervals of 1-2 years, comparing to 10% with intervals of 2-3 years (Vasen, Abdirahman, Brohet and et al., 2010). Stuckless (2012) showed that colonoscopy screening at 1 to 2-year interval delayed the onset age of CRC by more than 10 years and lead to 4 to 15-year improvement in life expectancy. Cumulative 18-year CRC mortality rate was 33% lower in the annual screening group and 21% lower in the biennial screening group, comparing to the non-screening group (Mandel et al., 1999).

Several modeling strategies were considered in literature for evaluating screening effect on risk of CRC. First, screening visit process was incorporated as a covariate in a Cox regression model. Analysis using Cox regression was carried out to identify risk factors independently associated with CRC risk during screenings, followed by stratified analyses. The median age at start of follow-up was considered as one of the risk factors; carriers aged 40 years or older at the start of the screenings had a higher risk than carriers younger than 40 years of age (Vasen, Abdirahman, Brohet and et al., 2010).

Second, screening process was jointly modeled with a detection of disease through a joint frailty model. Katki et al. (2015) fitted a joint frailty model for time to clearance of human papiloma virus (HPV) and time to a cervical cancer and provided suggestions on optimal screening intervals in screening guidelines. Lee et al. (2013) investigated a pooled time lag to benefit across trials by fitting a joint random effects (frailty) model and showed that CRC mortality decreased steadily with longer follow-up.

Finally, screening process was evaluated by a simulation study. Gunsoy et al. (2014) demonstrated that mortality reduction was highly dependent on screening frequency, age range, and uptake, based on a 13-state Markov simulation model including 13 states from healthy to different status of in-situ and further to death. Different screening strategies were also considered, including triennial screening in women aged 47-73 and annual screening in women aged 47-73 in their simulation study. The results showed that predicted breast cancer mortality reduction due to screening ranged from 15.9% to 36.7% from age 40 to 85 years for different scenarios. Thomas (2017) studied the effect of screening on cancer risk adjusting for screening behavior variables using simulation study. A conceptual model was simulated accounting for the interplay between screening behavior and the cancer process, with two measured covariates and two unobserved frailties (one individual frailty and one frailty among members of the same sibship). Ages at polyps were simulated from a Weibull distribution. Growth rate of polyps and age at the tumor diagnosis were also generated. In screening process, age at the first screening was generated by a lognormal distribution. Following screening times were generated from a lognormal distribution with different intercept from the previous one, as well more covariates were considered including total number of polyps on the last screening, an indicator of whether any of the individuals siblings had at least one screening by the then, an indicator of whether any polyp had been found in siblings, and an indicator of whether any of siblings had a cancer diagnosis.

2.2 Shared Frailty Models

A frailty model is an extension of the proportional hazard model by including a random component, to account for the heterogeneity caused by unmeasured covariates (Rondeau et al., 2012). The choice of frailty distribution is an important issue when using frailty models. Two main frailty distributions are the gamma distribution (Clayton, 1978; Vaupel et al., 1979) and the log-normal distribution (McGilchrist and Aisbett, 1991). Gamma frailty is commonly used because of its straightforward interpretation of correlation parameter, robustness to misspecification for both regression coefficients and hazard functions, and mathematically convenience (Dixon et al., 2011). Frailty models are used for explaining heterogeneity in individuals or within families, for example, correlated survival times for individuals, like twins or family members, and repeated events for the same individual (Hougaard, 1995).

Shared frailty model is the simplest form of frailty model and usually used for clustered data. Joint frailty model is in the joint modeling context, by adding an shared individual uncertainty in the model to jointly model the hazard of recurrent events and the hazard of a terminal event. Joint nested frailty model is a joint frailty model specifically for hierarchically clustered data. Two frailties, which account for random effects within an individual and within a cluster, are included in the joint nested frailty model when jointly models for recurrent events with a terminal event.

The shared frailty model was introduced by Clayton (1978) and extensively studied in

Hougaard (2000). A frailty term in the model represents an unobserved random effect that modifies the hazard function, multiplicatively (Hougaard, 1995). Shared frailty models are usually applied to correlated cluster data for example, family data and recurrent events. In addition, shared frailty model commonly uses recurrent events (cancer relapse or visiting process). The survival times are conditionally independent with respect to the shared frailty (Wienke, 2014). Hougaard (1995) used shared frailty model for multivariate failure times as the conditional of independent times given the random effect.

2.3 Joint Frailty Models

Joint frailty models are usually used to account for the dependence between recurrent events and time-to-event data by sharing a random effect. For modeling the recurrent events, we can consider two time-scales; one focuses on the times between two successive events (i.e. gap times) while the other focuses on the time to events (i.e. calendar times) (Duchateau et al., 2003). Choice of time scale can change the interpretation of the time evolution entirely. In our study, we chose gap time scale because our interest lies in screening intervals between two successive screenings.

There are several choices for modeling the baseline hazard function, including cubic M-splines, piecewise constant functions (PCF) or the Weibull distribution. PCF requires to choose an appropriate number of the intervals to capture enough the flexibility of the true hazard function. For Weibull distribution, a small number of parameters are estimated but resulting estimated functions are monotone, which might be too limiting in some cases (Król et al., 2017). Cubic M-splines can be used for estimating baseline functions. They are non-negative and easy to integrate (Belot et al., 2014).

2.3.1 Joint Nested Frailty Model

Nested frailties can be used for modeling the hierarchical clustering of the data by including two nested random effects. It is appropriate and necessary when data are clustered at several hierarchical levels. Taking into account of these two kinds of random effects could lead to more accurate estimates of parameters of interest (Rondeau et al., 2006). Sastry (1997) applied a multivariate hazard model with family and community frailties, estimated the model using the EM algorithm, to the survey data collected via a hierarchically clustered sampling scheme. Manda (2001) fit a Cox proportional hazard model with two random frailties, family and community frailties, acting on the hazard rate, and examined the results by using Gibbs sampler and the EM algorithm. Rondeau et al. (2006) applied a multilevel proportional hazards model with two frailties, cluster-level random effect and sub-cluster random effect, multiplicatively affecting the hazard function. A semi-parametric penalized likelihood approach was used for estimations of parameters in the models. Results showed that if the hierarchical structures of the data were ignored, the variances of the random effects were overestimated. Underestimation of the regression coefficients in the shared frailty model was also found under a large intra-subgroup correlation.

The nested frailties were applied to joint modeling to account for the hierarchical structure of family data (Choi et al., 2017). This joint nested frailty model (JNFM) extends to a more complete case which underlines two frailties, individual frailty and familial frailty, for studying the impact of recurrent process on the terminal event. In this thesis, both individual and familial frailties were included for modeling the recurrent screening visits and the time to a first CRC as a function of screening history and covariates of interest.

2.3.2 Maximum Penalized Likelihood Estimation

In the construction of likelihood for frailty models, penalized likelihood is considered to account for model complexity when estimating parameters in the models. In the general likelihood estimation, adding more parameters into models would increase the likelihood and provide a better fit to the data, since more aspects of data are taken into account. However, this could lead to a more complex model. It is not practical to always choose a sophisticated model. Penalized likelihood is developed by adding a penalty term in log-likelihood to balance the complexity of a model and the goodness of fit. A penalty can be viewed as a tolerable degree of bias in exchange for reduction in the variability of parameter estimates (Rothman et al., 2008). The smoothing parameter associated with the penalty controls the trade-off between data fit and smoothness of the baselines functions, thus determines how much the data are smoothed to produce the estimates.

Leroux and Puterman (1992) implemented maximum penalized likelihood method when estimating the parameters in independent and Markov-dependent mixture models, in a study of breathing and body movements in fetal lambs. A potential saving in computation results from the observation that whenever the addition of a component fails to produce an increase in likelihood, the maximum value of the likelihood has been found. Rondeau et al. (2003) applied maximum penalized likelihood to nonparametric estimation of a continuous hazard function in a shared gamma-frailty model with right-censored and left-truncated data. Cole et al. (2014) illustrated the mechanics of maximum penalized likelihood estimation and described extensions which are better suited to observational health research. In epidemiological studies, data can present different problems including small sample size or sparse data. Application of penalization would reduce bias and further reduce mean square error (MSE).

2.3.3 Likelihood cross-validation criterion

The likelihood cross-validation (LCV) was proposed for choosing the smoothing parameter in the penalized likelihood and also for choosing between different semi-parametric models (Joly et al., 1998). Joly et al. (2002) applied LCV to interval censored data in an illness-death model, in order to choose smoothing parameters in penalized likelihood simultaneously. Commenges and Gegout-Petit (2007) argued that LCV could be used for choosing between different semi-parametric models, such as stratified and non-stratified proportional hazard survival model. LCV is adopted to guide the choice of the model structures used in the analysis based on penalized likelihoods. The LCV criteria is utilized by default in R package `frailtypack` (Rondeau et al., 2017) and is defined as:

$$LCV = \frac{1}{n}(tr(\hat{H}_{pl}^{-1}\hat{H}_l) - l(\hat{\psi})),$$

where $\hat{\psi}$ is the maximum penalized likelihood esetimator, H_{pl} is the converged hessian of the penalized log-likelihood, H_l is the converged hessian of the log-likelihood, and $l(\cdot)$ is the full log-likelihood at $\hat{\psi}$. Lower values of LCV indicate a better fitting model.

2.4 Dynamic Prediction

In longitudinal study, the prediction of risk of a terminal event is always of interest. One would have access to all the available longitudinal information up to a certain time point in order to predict the risk of event after that time point. By dynamic prediction one can update the prediction using longitudinal measurements. A more reliable prediction risk would be obtained when updating the prediction values conditioning on accumulated information up to date. For patients' disease prognosis, dynamic predictions are widely studied. It requires making updated predictions as time goes by and more data are observed. There are two approaches for dynamic prediction, which are joint modeling of longitudinal biomarkers and survival data, and landmark analysis.

Dafni (2011) explained the concepts of landmark method and applied the method in observational study cases. In landmark analysis, a time point is selected as a landmark time. Responders and non-responders are separated based on the landmark time; only individuals who are alive at the landmark time are included in the analysis. The landmark method ignores the responses after the fixed landmark time and the events before the same landmark time. van Houwelingen (2007) first came up the idea of applying land-marking for dynamic prediction. This approach would dynamically adjust predictive models during the follow-up time, by fitting models for individuals who were still at risk at the landmark point. van Houwelingen and Putter (2011) applied different models in different cases, such as dynamic prognostic models for survival data using time-dependent information, and dynamic prediction based on genomic data. Rizopoulos et al. (2013) presented and compared landmark analysis and joint models, which provided dynamic estimates of survival probabilities, for longitudinal and time-to-event data. van Houwelingen and Putter (2008) compared landmark model and multi-state model on the dynamic prediction for 5-year failure free survival after bone marrow transplantation in acute lymphoid leukemia patients. The results from the two methods were similar; however, landmark model had the advantage that had easy prediction rules due to the simplicity of the model. However, it did not provide the insight in the biological process which could be obtained from multi-state model. Longitudinal biomarkers are used for predictions of prognosis. However, biomarkers usually have nonlinear or not monotone trajectories, which makes the fitting of parametric models computationally difficult. Huang et al. (2016) proposed an approach for dynamic prediction which assumed that the biomarker effects on the risk of disease recurrence are smooth functions over time, and compared this method to the other two traditional methods, which are joint modeling and landmark method. Mauguen et al. (2013) proposed dynamic prediction tools under three different prediction scenarios, which considers the exact recurrence history, observed recurrence history and no recurrence history of an individual. The proposed tools were tested by using observational invasive breast cancer data.

2.4.1 Prediction Accuracy

There are several indices to quantify the accuracy of prediction models. The well-known indices are the area under Receiver Operating Characteristics (ROC) curve and Brier Score (BS). The ROC curve is created by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity). It shows the ability of the forecast to discriminate between event of interest and non-event. The corresponding Area Under the Curve (AUC) represents the probability that a person who experienced an event has a higher risk score than a person who did not experience any event (Chambless et al., 2011). ROC AUC has a drawback that it treats sensitivity and specificity of the same importance; however, they are probably not equivalent (Halligan et al., 2015). In terms of CRC, poor sensitivity could mean missed cancer, delayed treatment or even death, whereas poor specificity could just mean unnecessary colonoscopy. ROC AUC ignores clinical differences in misclassification cost so that it is not sensitive to bias, which means that a biased prediction can also provide a good resolution. The BS is defined as mean squared probability difference between the predicted values and the true statuses at the certain prediction time points. It measures the calibration of a set of probabilistic predictions or the magnitude of the prediction error. Choi et al. (2017) used BS as the criteria to compare the dynamic prediction accuracies of joint frailty model and joint nested frailty model.

Proust-Lima and Taylor (2009) validated their prediction based on joint modeling, by focusing on the prediction error. They calculated the expectation of loss function of the difference between actual values of prediction with the observed data. Rizopoulos (2011) assessed prediction through how well their marker is capable of discriminating between subjects who have the event within a medically meaningful time range from subjects who do not. They validated their prediction based on joint model by ROC AUC. In the literature, there is no general agreement about which measure should be preferred when validating and comparing predictive rules in a survival case (Pencina et al., 2008). Ikeda et al. (2002) derived a theoretical functional relationship between ROC and BS, assuming that the calibration in the observer's probability estimate is perfect. They noticed a theoretical monotone relationship between BS and the area under the ROC curve. When the area under the ROC curve increases, the BS value monotonically decreases.

Chapter 3

Statistical Models

This chapter describes the statistical models and methods used for the analysis of the LS data set. Section 3.1 to 3.3 describe shared frailty model for terminal event with time-dependent covariates, joint frailty model and joint nested frailty model, respectively. The maximum penalized likelihood estimation and the dynamic prediction with prediction accuracy are included under sections of different models.

Notation used in this thesis are following by Choi et al. (2017). We use two superscript R and D to distinguish two processes, where R represents for recurrent process and D for a terminal event. The parameter estimation in our models are based on maximizing the penalized marginal likelihood obtained from integrating the random variable(s) over its(their) distribution(s).

3.1 Shared frailty model for the terminal event

Shared frailty model for the terminal event is developed here to incorporate time-dependent variables and the familial random variable. Let f indicate for family, $f = 1, \ldots, n$, where n is the total number of families, i for individual, $i = 1, \ldots, m_f$, where m_f is the size of family

f. For subject i in family f, $\tilde{T}^D_{fi}$ is denoted by the true time to a terminal event, C^D_{fi} by the censoring time, $T^D_{fi} = min(\tilde{T}^D_{fi}, C^D_{fi})$ by the observed time to event. δ^D_{fi} is denoted as an event indicator, which takes value 1 if $T^D_{fi} = \tilde{T}^D_{fi}$, 0 otherwise. Time-dependent covariates, which are represented as $X^D_{fi}(t)$, indicate that the values of covariates of interest are not necessarily constant over time. $r(t)$ and $\lambda(t)$ represent the hazard functions for the visit process and the terminal event, respectively, with $r_0(t)$ and $\lambda_0(t)$ as their baseline hazard functions. Let the family-specific random effect w_f follow $\Gamma(1/\eta, 1/\eta)$ with mean 1 and variance η and the density function denoted by $g_w(w)$. Let $\psi = (\lambda_0(t), \gamma, \eta)$ represent the vector of the parameters involved in the shared frailty model with time-dependent covariates.

The hazard function for a terminal event for individual i in family f is,

$$\lambda_{fi}(t_{fi}; X^D_{fi}(t_{fi}), w_f) = \lambda_0(t_{fi}) w_f \exp(X^D_{fi}(t_{fi})\gamma),$$

where $t_{fi} \in [0, T^D_{fi}]$ and $T^D_{fi} = min(\tilde{T}^D_{fi}, C^D_{fi})$, w_f is the family-specific random effect.

The corresponding survival function for individual i in family f is,

$$S_{fi}(t_{fi}; X^D_{fi}(t_{fi}), w_f) = e^{-\int_0^{t_{fi}} \lambda_{fi}(s; X^D_{fi}(s), w_f) ds} = e^{-\int_0^{t_{fi}} \lambda_0(s) w_f \exp(X^D_{fi}(s)\gamma) ds}.$$

By integrating over the familial frailty distribution,

$$\begin{aligned} S_{fi}(t_{fi}; X^D_{fi}(t_{fi})) &= \int_{w_f} S_{fi}(t_{fi}; X^D_{fi}(t_{fi}), w_f) g_w(w_f) dw_f \\ &= \int_{w_f} e^{-\int_0^{t_{fi}} \lambda_0(s) w_f \exp(X^D_{fi}(s)\gamma) ds} g_w(w_f) dw_f. \end{aligned}$$

The conditional likelihood given the random effect w_f can be expressed as

$$L^C(\psi) = \prod_{f=1}^{n} \prod_{i=1}^{m_f} \lambda_0(t_{fi}) w_f \exp(X^D_{fi}(t_{fi})\gamma).$$

However, the conditional likelihood cannot be used directly for statistical inference since the frailty is unobservable. We will need to integrate over the frailty distribution in order to discover the observable consequences of the model. By integrating over the familial frailty distribution, the marginal likelihood is obtained as:

$$L(\psi) = \prod_{f=1}^{n} \int_{w_f} {w_f}^{m_f} g_w(w_f) dw_f \times \prod_{f=1}^{n} \prod_{i=1}^{m_f} \lambda_0(t_{fi}) \exp(X^D_{fi}(t_{fi})\gamma).$$

A predictive probability of event between t and horizon time $t+s$ for individual i in family f can be caluculated:

$$\begin{aligned}P(t,s) &= P(\tilde{T}_{fi}^{D} < t+s | \tilde{T}_{fi}^{D} > t, X_{fi}^{D}(t), \psi)\\ &= \frac{S_{fi}(t;X_{fi}^{D}(t)) - S_{fi}(t+s;X_{fi}^{D}(t))}{S_{fi}(t;X_{fi}^{D}(t))}.\end{aligned}$$

3.2 Joint frailty model

Screening visit process is considered as recurrent events and the first CRC is considered as a terminal event in our case. The screening visit process covers all the colonoscopy visits for each individual up to the last follow-up time, which is the disease time if one had a CRC during the study time. Joint frailty model jointly models the gap time between two successive screening visits and the time to a first CRC, while taking into account the individual random effects.

In terms of the recurrent process for subject i, let T_{ij}^{R} be the gap time between visit $j-1$ and j, and δ_{ij}^{R} be a visit indicator, which takes value 1 if $j = 1,\ldots,n_i$ (visits), and 0 if $j = n_i + 1$ (terminal event or the last follow-up). The gap time between the last screening visit and the terminal event or last follow-up is treated as censored. The hazard function for the recurrent event is noted as $r(t)$ with the baseline hazard as $r_0(t)$. X^R and β represent the covariates and corresponding vector of coefficients in the reccurent event model. Joint frailty model by Rondeau et al. (2007) is employed in this section, and applied to model screening visits.

Individual i is redefined as $i = 1,\ldots,m$ where m is the total number of individuals in the study. Let j indicate for visit, $j = 1,\ldots,n_i + 1$, where n_i is the total number of screening visits until the last follow-up time for individual i and the $(n_i + 1)^{th}$ visit refers to the terminal event or the last follow-up as cencored. We denote u_i as an individual-specific random effect, following a gamma distribution $\Gamma(1/\theta, 1/\theta)$ with mean 1 and variance θ and probability density function denoted by $g_u(u)$.

The hazard function for the recurrent events is written as:

$$r_{ij}(t_{ij};X_{ij}^R,u_i) = r_0(t_{ij})u_i\exp(X_{ij}^R\beta),$$

where X_{ij}^R is the vector of covariates measured at visit $j-1$ in the visit process, β is the vector of regression coefficients.

The corresponding survival function is,

$$S_{ij}(t_{ij};X_{ij}^R,u_i) = e^{-\int_0^{t_{ij}} r_{ij}(s;X_{ij}^R,u_i)ds} = e^{-\int_0^{t_{ij}} r_0(s)u_i\exp(X_{ij}^R\beta)ds}.$$

In terms of the terminal event of disease for subject i, $\tilde{T}_i^D$ is the true time to a terminal event, δ_i^D is an event indicator. The hazard function for the terminal event is noted as $\lambda(t)$ with the baseline hazard as $\lambda_0(t)$. X^D and γ represent the covariates and corresponding vector of coefficients in the terminal event model. Especially in Cox model with time-dependent covariates, $X_{fi}^D(t)$ are used to represent the time dependent covariates in the model.

The hazard function for the terminal event is written as:

$$\lambda_i(t;X_i^D,u_i) = \lambda_0(t)u_i^\alpha\exp(X_i^D\gamma),$$

where X_i^D is the vector of covariates considered in the terminal event, γ is the vector of regression coefficients in joint frailty model.

The corresponding survival function is:

$$S_i(t_i;X_i^D,u_i) = e^{-\int_0^{t_i}\lambda_i(s;X_i^D,u_i)ds} = e^{-\int_0^{t_i}\lambda_0(s)u_i^\alpha\exp(X_i^D\gamma)ds}.$$

Let $\psi = (r_0(t),\lambda_0(t),\beta,\gamma,\theta,\alpha)$ represent the vector of the parameters involved in the joint frailty model. According to Rondeau et al. (2012), the marginal likelihood function of joint frailty model can be expressed as:

$$L(\psi) = \prod_{i=1}^m \int_{u_i} u_i^{n_i+\alpha\delta_i^D}\exp\left\{-u_i\sum_{j=1}^{n_i+1} R_0(t_{ij}^R;X_{ij}^R)e^{X_{ij}^R\beta} - u_i^\alpha\Lambda_0(t_i^D;X_i^D)e^{X_i^D\gamma}\right\}g(u_i)du_i\times$$

$$\prod_{i=1}^m\prod_{j=1}^{n_i+1}\{r_0(t_{ij}^R;X_{ij}^R)e^{X_{ij}^R\beta}\}^{\delta_{ij}^R}\{\lambda_0(t_i^D;t_i^D)e^{X_i^D\gamma}\}^{\delta_i^D},$$

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

The baseline hazard functions are approximated on the basis of splines to allow flexible functions, as we do not have any a priori on the shape of these risk functions. Cubic M-splines (polynomial functions of 3^{rd} order that are combined linearly to approximate a function in an interval) are used. Five knots were used in our applications in order to get a close approximation to the true hazard function.

The penalized likelihood for joint modeling is defined by Rondeau et al. (2007):

$$pl(\psi) = \log L(\psi) - \kappa_1 \int_0^\infty r_0''(t)^2 dt - \kappa_2 \int_0^\infty \lambda_0''(t)^2 dt, \tag{3.1}$$

where κ_1 and κ_2 (κ_1 for hazard function of recurrent events and κ_2 for hazard function of terminal event) are two positive smoothing parameters to control the trade-off between the data fit and the smoothness of the baseline functions. The smoothing parameters can be fixed as some certain values or evaluated by maximization of a likelihood cross-validation criterion. Values can be obtained by separately fitting two models, one is a shared frailty model for only recurrent events with an individual frailty and the other is a shared frailty model for only terminal event with a familial frailty. The two obtained smoothing parameters κ_1 and κ_2 are used in the joint frailty model as two fixed values.

Dynamic prediction for joint frailty model is to predict the probability of disease in a specific time window given the history of patient i before the time of prediction t. Individual's visit history of subject i in family f before time t is:

$$Y_{fi}^R(t) = \left\{T_{fij}^R, \forall j \in \{1, \ldots, j_i^*\}\right\},$$

which includes all visits and corresponding covariates information until the last visit of subject i before time t, noted as visit j_i^*.

A prediction of disease between t and $t+s$ given that the individual had j recurrent events before time t is written as follow. The prediction probability is obtained marginally by integrating over the individual frailty distribution (Rondeau et al., 2012):

$$
\begin{aligned}
P(t,s) &= P(\tilde{T}_i^D \leq t+s|\tilde{T}_i^D > t, X_{ij}^R(t), X_i^D(t), Y_i^R(t), \psi) \\
&= \int_{u_i} P(\tilde{T}_i^D < t+s|\tilde{T}_i^D > t, Y_i^R(t), X_{ij}^R(t), X_i^D(t), u_i, \psi) g_u(u_i) du_i,
\end{aligned}
$$

where $g_u(u_i)$ is used to represent $g_u(u_i|\tilde{T}_i^D > t, X_{ij}^R(t), X_i^D(t), Y_i^R(t), \psi)$ for simplicity.

We note that alternatively the prediction probability can be obtained conditionally on the estimated frailty values as proposed by Mauguen et al. (2013):

$$
\begin{aligned}
P^{cond}(t,s) &= P(\tilde{T}_i^D \leq t+s|\tilde{T}_i^D > t, X_i^D, \hat{u}_i, \hat{\psi}) \\
&= \frac{S_i^D(t|X_i^D, \hat{u}_i, \hat{\psi}) - S_i^D(t+s|X_i^D, \hat{u}_i, \hat{\psi})}{S_i^D(t|X_i^D, \hat{u}_i, \hat{\psi})} \\
&= 1 - \left(\frac{S_0^D(t+s)}{S_0^D(t)} \right)^{\hat{u}_i^{\hat{\alpha}} exp(\hat{\beta}' X_i^D)},
\end{aligned}
$$

where u_i is the individual frailty for the patient i and $S_0^D(.)$ is the baseline survival function. The $\hat{u}_i$ are obtained from the posterior distribution of the u_i conditional on the observed data, knowing the estimated values of the regression parameters. Thus conditional prediction is only possible for the patients we already know what happened in the recurrent events. In practice, estimating $\hat{u}_i$ each time for different t and s is not easy. Therefore, we used the marginal approach for both joint frailty model and joint nested frailty model in our application.

3.3 Joint nested frailty model

We denote u_{fi} as an individual-specific random effect, and w_f is a family-specific random effect; u_{fi} follows a gamma distribution $\Gamma(1/\theta, 1/\theta)$ with mean 1 and variance θ and density function denoted by $g_u(u)$, and w_f follows $\Gamma(1/\eta, 1/\eta)$ with mean 1 and variance η and density function denoted by $g_w(w)$. In addition, α and ξ in joint nested frailty model represent the associations between two processes at individual level and at familial level, respectively.

As proposed by Choi et al. (2017), the hazard function for the recurrent events is written as:

$$
r_{fij}(t; X_{fij}^R, u_{fi}, w_f) = r_0(t) u_{fi} w_f^{\xi} \exp(X_{fij}^R \beta),
$$

where X_{fij}^R is the vector of covariates measured at visit $j-1$, β is the vector of regression coefficients in the recurrent event model.
The hazard function for the terminal event is written as:

$$\lambda_{fi}(t;X_{fi}^D,u_{fi},w_f) = \lambda_0(t)u_{fi}^{\alpha}w_f\exp(X_{fi}^D\gamma),$$

where X_{fi}^D is the vector of covariates, γ is the vector of regression coefficients in the termianl event model.

Two frailties shared by two processes, u_{fi} and w_f, represent the unobserved or unmeasured random effects which are not explained by the observed covariates in the model. θ and η are their variations, respectively. Especially, a larger η implies a greater heterogeneity in frailty across families and a greater correlation of the survival times for individuals that belong to the same family. In addition, α and ξ allow possible associations between the two processes; $\alpha > 0$ (or $\alpha < 0$) represents the positive (or negative) association between the two processes due to the individual random effects. A zero value of α means that u_{fi} affects only on the recurrent events and the dependence between two processes can be fully explained by the observed covariates. Similarly, $\xi > 0$ (or $\xi < 0$) represents the positive (or negative) association between two processes due to unknown familial effects. A zero value of ξ means that the family-specific random variable w_f affects only the terminal event.

Let $\psi = (\lambda_0(t), r_0(t), \beta, \gamma, \theta, \eta, \alpha, \xi)$ represent the vector of the parameters involved in the joint nested frailty model. The full marginal likelihood function is obtained by integrating the likelihood over the frailty distributions,

$$\begin{aligned} L(\psi) &= \prod_{f=1}^{n}\int_{w_f} w_f^{v_f\xi+d_f}\left[\prod_{i=1}^{m_f}\int_{u_{fi}} u_{fi}^{n_{fi}+\alpha\delta_{fi}^D}\exp\{-u_{fi}w_f^{\xi}\sum_{j=1}^{n_{fi}+1}R_0(t_{fij})e^{X_{fij}^R\beta}\right. \\ &\quad \left. -u_{fi}^{\alpha}w_f\Lambda_0(t_{fi}^D)e^{X_{fi}^D\gamma}\}g_u(u_{fi})du_{fi}\right]g_w(w_f)dw_f \times \\ &\quad \prod_{f=1}^{n}\prod_{i=1}^{m_f}\left\{\prod_{j=1}^{n_{fi}}r_0(t_{fij})e^{X_{fij}^R\beta}\right\}\lambda_0(t_{fi})^{\delta_{fi}^D}e^{\delta_{fi}^D X_{fi}^D\gamma}, \end{aligned}$$

where $d_f = \sum_{i=1}^{m_f}\delta_{fi}^D$ is the number of events in family f, $v_f = \sum_{i=1}^{m_f} n_{fi}$ is the total number of visits in family f and at $(n_{fi}+1)^{th}$ visit, $t_{fi(n_{fi}+1)}^R = t_{fi}^D - \sum_{j=1}^{n_{fi}} t_{fij}^R$. $R_{ij}(t)$ is the cumulative hazard function for the recurrent event, and $\Lambda_i(t)$ is the cumulative hazard function for the terminal event.

The penalized likelihood (equation 3.1) is used for estimating the parameters. The two smoothing parameters κ_1 and κ_2 are obtained by a shared frailty model for only recurrent events with an individual frailty and a shared frailty model for only terminal event with a familial frailty, and used in the joint nested frailty model.

Following the dynamic prediction for the joint nested frailty model (Choi et al., 2017), we obtain the probability of developing a first CRC event between time t and $t+s$ for subject i who survived by time t conditioning on this individual's own visit history and the family histories observed by time t.

Let T^R_{fij} be the gap time between visit $j-1$ and j, $Y^R_{fi}(t)$ be the individual's visit history of subject i in family f before time t. The individual's visit history, denoted by $Y^R_{fi}(t)$, is defined as:

$$Y^R_{fi}(t) = \left\{T^R_{fij}, \forall j \in \{1,\ldots,j^*_i\}\right\},$$

which includes all visits and corresponding covariates information until the last visit of subject i before time t, noted as visit j^*_i.

Let $T^D_{fi}(t)$ be the observed time to an event before t. $\delta^D_{fi}(t)$ is the disease indicator by time t, which takes value 1 if $T^D_{fi}(t) = \tilde{T}^D_{fi}$, and 0 otherwise. Family history, including visit and disease history, of all individuals but not individual i in family f is defined as:

$$H_{f(-i)}(t) = \left\{Y^R_{fl}(t), T^D_{fl}(t), \delta^D_{fl}(t), \forall l \in \{1,\ldots,i-1,i+1,\ldots,m_f\}\right\}.$$

The prediction probability is obtained by integrating over the two frailty distributions. The probability of developing a first CRC between t and $t+s$ for subject i is specified as:

$$\begin{aligned}P(t,s) &= P(\tilde{T}^D_{fi} < t+s | \tilde{T}^D_{fi} > t, Y^R_{fi}(t), H_{f(-i)}(t), X^R_{fi}(t), X^D_{fi}(t), \psi)\\ &= \int_{u_{fi}}\int_{w_f} P(\tilde{T}^D_{fi} < t+s | \tilde{T}^D_{fi} > t, Y^R_{fi}(t), H_{f(-i)}(t), X^R_{fi}(t), X^D_{fi}(t), u_{fi}, w_f, \psi) P(u_{fi}, w_f) du_{fi} dw_f,\end{aligned}$$

where $P(u_{fi}, w_f)$ is used to represent $P(u_{fi}, w_f|\tilde{T}^D_{fi} > t, Y^R_{fi}(t), H_{f(-i)}(t), X^R_{fi}(t), X^D_{fi}(t), \psi)$ for simplicity.

This is developed by conditioning on 1) individual survived by time t, 2) individual's visit history, $Y^R_{fi}(t)$ up to time t, 3) other family members' visit and disease history by time t, $H_{f(-i)}(t)$, and 4) individual's covariate information observed upto time t, $X^R_{fi}(t) = \{X^R_{fij}, \forall j \in \{1, \ldots, j^*_i\}\}$ for recurrent events and $X^D_{fi}(t)$ for the terminal event.

3.3.1 Evaluation of predictive accuracy

Dynamic prediction is an important feature of joint modeling and is increasingly popular in applied bioscience and biostatistics field. Dynamic predictions for an individual of interest can be computed after fitting joint models. We compare the predictive accuracies of the dynamic predictions among the shared frailty model for terminal event, the joint frailty model and the joint nested frailty model.

A 10-fold cross-validation is used to evaluate the prediction accuracy. The original data set is randomly partitioned into ten equal sized subsets. One of the ten subsets is used for testing the model while the remaining nine subsets are used as training data. By using each of the ten subsets once for validation, the cross-validation process is repeated ten times. We divided data based on families rather than individuals, since it is crucial to keep the family structure when fitting joint nested frailty model and computing dynamic predictions.

The predictive accuracy of the proposed model is evaluated using Brier Score, a quadratic prediction error. The difference between the predicted values of developing CRC at certain time points from the model and the true observed individual's status is measured with BS. Due to the existence of censoring, the true status of CRC is unknown and cannot be computed. However, we considered everyone in the model, treating the censored individuals as no cancer; this may lead to bias in the prediction, since those censored individuals may develop cancer after the last follow-up time. To make the right censoring into account, a weighted BS is calculated. Inverse probability of censoring weighted error (IPCWE) is applied to account for right censoring (Blanche et al., 2015). $\hat{G}_n(t)$ represents Kaplan-Meier estimate of the population censoring distribution. Let $W_i(t,s,\hat{G}_n)$ be the weight in IPCWE, which has the following form:

$$W_i(t,s,\hat{G}_n) = \frac{I(T_i^D \leq t+s)\delta_i^D}{\hat{G}_n(T_i^D)/\hat{G}_n(t)} + \frac{I(T_i^D > t+s)}{\hat{G}_n(t+s)/\hat{G}_n(t)}.$$

The weighted BS is defined as:

$$\hat{BS}(t,s) = \frac{1}{n\hat{S}_{\tilde{T}}(t)} \sum_{t=1}^{n} \hat{W}_i(t,s)\left(\tilde{D}_i(t,s) - P_i(t,s)\right)^2,$$

where $\hat{S}_{\tilde{T}}(t) = \frac{1}{n}\sum_{i=1}^{n} I(\tilde{T}_i > t)$ estimates the probability of observing a subject at risk (i.e. alive and uncensored) at the prediction time point t. $\tilde{D}_i(t,s) = I(t < \tilde{T}_i \leq t+s, \tilde{\delta}_i = 1)$ is an indicator which equals to 1 when an subject i developed a cancer in the time range $(t, t+s]$ and equals to 0 otherwise. $P_i(t,s)$ represents the prediction probability.

3.4 Advantages and disadvantages of three models

One of our main goals is to incorporate the screening visits in prediction. Three models, shared frailty model, joint frailty model and joint nested failty model, were compared in terms of prediction. The shared frailty model is the easiest one to apply among the three models due to its simplicity. It accounts for the hierarchical structure but only models the terminal event by using time-dependent covariates to represent the recurrent event process; it ignores the interplay between two processes. The joint nested frailty model extends to a more complex case which underlines two frailties, individual frailty and familial frailty, for studying the impact of recurrent process on the terminal event. However, sometimes it can be difficult to fit the joint nested frailty model because of the complexity of the model. In terms of prediction, the joint nested frailty model is supposed to have better prediction accuracy among three models, since it reflects how the screening visits affect the risk of CRC in LS families. It is crucial to consider the hierarchical structure in the analysis of family data. The joint frailty model does not reflect the family structure so its prediction may be less accurate in the presence of strong familial correlation as shown in our LS families. The shared frailty model only for terminal event can provide a decent prediction accuracy, since it accounts for the family structure and takes screening process into account via time-dependent covariates.

Chapter 4

Application to Lynch Syndrome Family

This chapter presents the analysis of the Lynch Syndrome family data. Section 1 introduces the data and provides descriptive statistics. Section 2 describes the specifications of the fitted models. Section 3 evaluates the impact of screening on risk of CRC, including estimations of coefficients and parameters from the models. The chapter finishes with Section 4, providing dynamic predictions of the risk of CRC. Fitting the models and estimating the predictions were based on the R package `frailtypack` (Rondeau et al., 2016). `frailtypack` fits several classes of frailty models with gamma or normal random effects. Time-varying effect covariates can be considered in Cox, shared and joint frailty models.

4.1 Lynch Syndrome Family data

The observational LS data set we are using in this study was provided by Familial Gastrointestinal Cancer Registry (FGICR) (`http://www.zanecohencentre.com/gi-cancers/fgicr/about-us`). FGICR is a family study center at Mount Sinai Hospital with focus on families affected with rare familial gastrointestinal cancer syndromes, in order to give a secondary prevention of cancer through early diagnosis and treatment. Medical and surgical cares are maintained for patients and their family members. Genetic testing may be given through their service and research molecular laboratory program, for those with specific forms of familial colorectal, gastric and pancreatic cancer.

The FGICR includes 423 LS families with 1068 unique patients in total. The main aim of our study is to evaluate the screening efficiency on the risk of developing CRC and provide dynamic prediction based on screening and disease history. Since we are interested in association between the intervals of screening visits and CRC occurrence, we included individuals who had at least two screening visits with no detected CRC at the first visit. Thus, 242 LS families with 422 individuals (152 males and 270 females) were identified for analysis. Table 4.1 summarizes the distribution of different mutation types included in the data set. MLH1, MSH2 and MSH6 are three main mutation types commonly observed in the data, so that we include analyses based on these three types in Section 4.3. Basic information and characteristics of the LS family data are summarized in Tables 4.2, 4.3 and 4.4. Of 422 individuals, 103 developed CRC at an average age of 50.69 (standard deviation (sd) = 14.75) years and the average gap time between first screening and actual diagnosis of CRC was 8.18 (sd = 10.14) years for these CRC patients. The remaining 319 individuals who did not develop any CRC during the study time have been followed up with screenings on average every 8.46 (sd = 6.46) years. Additionally, 135 individuals in the data set had at least one adenoma detection and removal during their screening visits. The average gap time between two successive screening visits was 3.39 (sd = 4.48) years. Individuals in the study had 3.90 (sd = 2.76) screening visits on average before any CRC detected. Probands were kept in the data set in order to avoid losing too much information. We used probands' CRC ages as a covariate in the model in order to incorporate ascertainment of families into study.

While exploring the differences among mutation types, some distinct patterns can be found in Table 4.4. MLH1 patients start their screening visits at the earliest age, of age 38.5 (sd = 12.97), among all the five mutation types patients, but also have a longer visiting interval. Proband's age of CRC is noticable older among MSH6 patients. MSH6 patients are the last to start screening visits and have the oldest age of onset for CRC. For a better comparison, analysis of variance (ANOVA) was used to test differences among means of the main covariates of five different mutation types. P-values are listed for explanation. The differences among the group means of gap year between visits, and the ones of age at the first CRC are not statistically significant, with p-values equal to 0.151 and 0.742, respectively. For other variables, not all of group means among five mutation types can be treated as equal.

Table 4.1: Distribution of mutation types

	Number (%)					
	Total	MLH1	MSH2	MSH6	PMS2	EPCAM
Number of individuals	422 (100%)	146 (34.60%)	194 (45.97%)	59 (13.98%)	18 (4.27%)	5 (1.18%)
Number of families	242 (100%)	83 (34.30%)	97 (40.08%)	40 (16.53%)	17 (7.02%)	5 (2.07%)

Table 4.2: Basic information of Lynch Syndrome family data

		gender		**adenoma**	
	Total	Males	Females	yes	no
Number of individuals	422	152 (36%)	270 (64%)	135 (32%)	287 (68%)
CRC	103	47 (45.6%)	56 (54.4%)	18 (17.5%)	85 (82.5%)
MLH1	146	63 (43.15%)	83 (56.85%)	47 (32.19%)	99 (67.81%)
MSH2	194	68 (35.05%)	126(46.29%)	67 (34.54%)	127 (65.46%)
MSH6	59	14 (23.73%)	45(76.27%)	24 (40.68%)	35 (59.32%)
PMS2	18	5 (27.78%)	13 (72.22%)	5 (27.78%)	13 (72.22%)
EPCAM	5	2 (40%)	3 (60%)	1 (20%)	4 (80%)

Table 4.3: The characteristics of Lynch Syndrome family data

	whole data set		
	(min,max)	mean (sd)	median
Number of visits per person	(1,16)	3.898 (2.762)	3
Age(years) at the first visit	(3,78)	40.470 (13.295)	40
Gap(years) between visits	(0.2,34)	3.389 (4.476)	2
Proband's age at CRC	(12,81)	42.400 (13.154)	42
For those who had CRC			
Age at the first CRC	(20,88)	50.690 (14.746)	50
Gap year of CRC from the first visit	(0,37)	8.175 (10.141)	5
For those who had no CRC			
Follow-up time	(0.2,38)	8.464 (6.456)	6

Table 4.4: The characteristics of Lynch Syndrome family data – comparing five mutation types; p-values are for assessing the null hypothesis: the means of a certain variable for five mutation types are all equal.

	MLH1	**MSH2**	**MSH6**	**PMS2**	**EPCAM**	**ANOVA**
	mean (sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)	p-value
Number of visits	3.77 (2.76)	4.36 (2.99)	3.29 (2.05)	2.33 (1.09)	2.80 (0.84)	<0.001
Age at the first visit	38.55 (12.97)	39.77 (13.44)	46.78 (10.60)	42.28 (15.48)	42.60 (19.42)	0.001
Gap year between visits	4.09 (5.53)	3.02 (3.91)	3.00 (3.76)	3.33 (2.40)	2.13 (1.28)	0.151
Proband's age at CRC	41.12 (13.82)	40.88 (12.06)	51.83 (10.69)	38.78 (15.52)	40.60 (7.09)	<0.001
For those who had CRC						
Age at the first CRC	49.38 (15.12)	51.24 (15.34)	55.50 (11.20)	50.38 (13.70)	no CRC	0.742
Gap year of CRC from the first visit	9.89 (11.02)	7.60 (10.16)	2.25 (4.10)	7.50 (7.21)	no CRC	0.015
For those who had no CRC						
Follow-up time	11.41 (9.42)	10.28 (8.25)	6.41 (5.19)	8.19 (5.62)	4.20 (3.03)	<0.001

4.2 Model Description

In the analysis of the LS data, screening visits were used as recurrent events and CRC was used as a terminal event. Three models, including the shared frailty model for terminal event, the joint frailty model and the joint nested frailty model, were fitted into the LS data. A gap time scale was used to model the duration between two recurrent events.

While modeling the screening visits, the first visit age, the age at the previous visit, indicator of any detection and removal of adenoma at the previous visit, indicator of any detection and removal of other polyps at the previous visit, proband's CRC age and gender (1 for male and 0 for female) were included in the screening visit process. While modeling the terminal event, the first visit age, indicator of detection and removal of adenoma before CRC, indicator of detection and removal of other polyps before CRC, proband's CRC age and gender (1 for male and 0 for female) were included. Three age-related variables, the first visit age, age at the previous visit and the proband's CRC age, were centered at their median ages (40, 46 and 40, respectively) and divided by 10, for a more meaningful intercept. The corresponding parameter estimates represented the increase or decrease in the log hazard scale for every 10-year increase in age.

Splines functions with 5 knots were used as baseline hazard functions in the three models. For shared frailty model, the smoothing parameter κ was set as 10 as a starting point for cross valiation, in order to add a penalty to the likelihood. For two joint models, the smoothing parameters κ_1 and κ_2 in the penalized log-likelihood function were chosen using the cross-validation from the two marginal models, i.e. shared frailty models for screening visit process only, and for the terminal event only.

4.2.1 Shared frailty model for terminal event

For the terminal event (CRC) only, shared frailty model was fitted. The first visit age, detection and removal of adenoma before CRC, detection and removal of other polyps before CRC, proband's CRC age and gender (1 for male and 0 for female) were included in the model as covariates. We considered the cumulative number of visits as a time-dependent covariate to reflect the nature pattern of screening visits. Starting time points and ending time points were specified to incorporate the time-dependent covariate.

An example of R code for fitting the shared frailty model for terminal event are shown as follow. Splines function using percentile intervals with 5 knots is used as baseline hazard function. Positive smoothing parameter in the penalized likelihood estimation is chosen as 10 in this example to give a penalty on the likelihood. Cross validation procedure is used for estimating smoothing parameter.

```
mod.cox.gap <- frailtyPenal(Surv(t.start,t.stop,CRC)~
  cluster(famid) + x1 + x2,
  data = data, hazard = "Splines-per",
  n.knots = 5, kappa = 10, cross.validation = TRUE)
```

4.2.2 Joint frailty model

The screening visits and the terminal event were jointly modeled by sharing an individual random effect between the two processes. An example of R code for fitting the joint frailty model are shown as follow. Splines function with 5 knots is used as baseline hazard function. Positive smoothing parameters in the penalized likelihood estimation are obtained by fitting the corresponding shared frailty models, for visiting process and terminal event separately, using cross validation.

```
modJoint.gap <- frailtyPenal(Surv(gaptime,visit)~
  cluster(idnew) + x1 + x2 + terminal(CRC),
  formula.terminalEvent = ~ x1 + x3,
  data = data, n.knots = 5, kappa = c(k1, k2))
```

4.2.3 Joint nested frailty model

We assumed the same familial frailty effect for both the recurrent and terminal events for simplicity by fixing $\xi = 1$; the association between two processes mainly comes from individual frailty. An example of R code for fitting the joint nested frailty model are shown as follow. Splines function with 5 knots is used as baseline hazard function. Positive smoothing parameters in the penalized likelihood estimation are obtained by fitting the corresponding shared frailty models using cross validation. `initialize = TRUE` indicates fitting an appropriate joint frailty model without group effect to provide initial values for the joint nested model.

```
modJointnested.gap <- frailtyPenal(Surv(gaptime,visit)~
  subcluster(idnew)+ cluster(famid)+
  x1 + x2 + terminal(CRC),
  formula.terminalEvent = ~ x1 + x3,
  data = data, n.knots = 5, kappa = c(k1, k2), initialize = TRUE)
```

4.3 Analysis of CRC risks

Results from the three models are summarized in Table 4.5. Results for mutation-specific models are included in Appendix A.

4.3.1 Impact of screening visits on the risk of CRC

Based on the joint nested frailty model, the effects of age at the first screening visit, detection and removal of adenoma and gender are significantly associated with the risk of developing CRC. Those individuals who ever been detected and removed adenoma before CRC have exp(-0.915) = 0.4 ($p = 0.001$) times the risk of developing CRC than those who did not have adenoma before CRC, and the 10-year increase in the first screening age increases the risk of CRC by exp(0.335) = 1.4 ($p = 0.001$) times. Males have exp(0.711) = 2.04 ($p = 0.004$) times the risk of getting CRC than females. The older the proband's CRC age, the slightly higher in the risk of developing CRC, and as well those individuals who ever detected and removed other polyps have lower risk of developing CRC than those who did not have other polyps before, but these effects were not significant.

Similar patterns were found with the joint frailty model taking only individual frailty into account; the same variables, age at the first screening, detection and removal of adenoma before

CRC and gender, were also significantly associated with the risk of developing CRC. For the shared frailty model, the cumulative number of visits was considered as a time-dependent covariate only in the shared frailty model to account for the visit process in the model. The hazard ratio (HR) of the cumulative number of visits (exp(0.206) = 1.23, $p = 0.010$) was significant, indicating that every additional visit would increase the risk of developing CRC by 23%. Some different patterns appeared in the shared frailty model; the effect of adenoma detection became not significant. The coefficient for the proband's CRC age was found negative in shared frailty model for terminal event but not significant. Effects for other covariates remained similar as the ones from the other two models.

The intervals between screening visits were modeling in joint models. We found several variables associated with the screening gap times, i.e. visit frequency. In the joint nested frailty model, first visit age, age at previous visit, detection and removal of adenoma at the previous visit, detection of serrated polyps at the previous visit were marginally significantly associated with visiting frequency. Among them, for first visit age, the older the first visit age the lesser the visit frequency, with HR exp(-0.68) = 0.505 ($p < 0.001$). With HR exp(0.65) = 1.91 ($p < 0.001$) for age at previous visit, a 10-year-increase on age at previous visit would increase the frequency of visiting by around 91%. Detection and removal of adenoma at the previous visit would lead to a exp(0.36) = 1.43-fold ($p < 0.001$) increase in frequency of visits compared to the situation with no detection. Similarly, detection of a serrated polyp at the previous visit would lead to a exp(0.78) = 2.19-fold ($p = 0.005$) increase in frequency of visits compared to the situation with no detection. Proband's CRC age and gender were not significantly related to visit frequency according to our results. The joint frailty model provided consistent patterns as the joint nested frailty models.

We have conducted analyses by three mutation types, MLH1, MSH2, and MSH6, using the three models. The results of those mutation-specific models are summarized in Table A.1, A.2 and A.3 in Appendix A. MLH1-specific model showed slightly different patterns from the ones in general model for combined data. In terms of the CRC occurance in joint nested frailty model, only the detection and removal of adenoma and gender were found to be significant, with the HRs of exp(-1.003) = 0.367 ($p = 0.029$) and exp(0.77) = 2.16 ($p = 0.045$), respectively. Thus in families with MLH1 mutation, the individuals who had adenoma detected and removed during the follow-up time would have 33% lower risk of developing CRC than those who did not. Besides, males in MLH1 families had distinctly higher risk (116% higher risk) to develop CRC than females. In visiting process, MLH1-specific joint nested frailty model, the detection and removal of adenoma is not significant any more. In terms of MSH2-specific model, age at the first screening visit is the only covariate found to be significant in disease process, with the HR of exp(0.411) = 1.51 ($p = 0.005$). Thus in families with MSH2 mutation, 10-year increase in age at the first screening visit would lead to 51% higher risk of developing CRC. In visiting process, MSH2-specific joint nested frailty model showed a same pattern as the general joint nested frailty model. In MSH6-specific model, only the age at the first visit and gender were kept in the model to achieve convergence, due to the small number of MSH6 families. Gender is found to be significant with regression coefficient estimate = 3.615 ($se = 1.191$, $p = 0.002$), which means that males in MSH6 families have significantly higher risk to develop CRC than females.

However, limited amount of individuals in each group may lead to some unreliable model results. Thus, mutation types are considered as covariates in the general models additionally, with results shown in Table B.1 in Appendix B. MSH2 is chosen to be the reference category. EPCAM patients are removed from the data set since there only a small number of patients with the mutation. Estimates and significances are consistent with the ones in Table 4.5. MLH1 is significantly diverse from MSH2 in visiting process, and PMS2 is significantly different from MSH2 in both visiting process and terminal event. MSH2 patients have the most screening visits among the four mutation types we considerd, but have relatively smaller chance to detect CRC; only MSH6 patients have lower risk to detect cancer than MSH2 patients.

Table 4.5: Comparisons of three models–joint nested frailty, joint frailty, and shared frailty models–for analyzing LS family data

Visiting Process

Variables	Joint nested frailty			Joint frailty		
	estimate	se	p-value	estimate	se	p-value
age at the first visit	-0.683	0.081	<0.001	-0.629	0.079	<0.001
age at the previous visit	0.649	0.060	<0.001	0.645	0.060	<0.001
detection of adenoma						
at the previous visit	0.361	0.097	<0.001	0.367	0.097	<0.001
detection of serrated						
at the previous visit	0.782	0.280	0.005	0.782	0.280	0.005
proband's CRC age	0.020	0.047	0.665	0.018	0.046	0.689
gender	-0.053	0.125	0.672	-0.070	0.125	0.573

Terminal event (CRC)

Variables	Joint nested frailty			Joint frailty			Shared frailty		
	estimate	se	p-value	estimate	se	p-value	estimate	se	p-value
age at the first visit	0.335	0.102	0.001	0.288	0.086	0.001	0.416	0.125	0.001
detection of adenoma	-0.915	0.276	0.001	-0.767	0.244	0.002	-0.563	0.316	0.075
detection of other polyps	-0.128	0.278	0.645	0.031	0.245	0.899	0.212	0.346	0.540
cumulative number of visits †	-	-	-	-	-	-	0.206	0.080	0.010
proband's CRC age	0.010	0.102	0.919	0.023	0.077	0.770	-0.091	0.149	0.541
gender	0.711	0.248	0.004	0.608	0.209	0.004	0.649	0.266	0.015

† time-dependent covariate

Frailty Parameters

	Joint nested frailty			Joint frailty			Shared frailty		
	estimate	se	p-value	estimate	se	p-value	estimate	se	p-value
θ	0.649	0.064	-	0.657	0.065	-	0.488	0.161	-
α	0.731	0.105	<0.001	0.386	0.227	0.089			
η	0.737	0.181	-						
ξ	-	-	-						
penalized marginal									
log-likelihood	-2299.04			-2542.04			-315.72		
LCV	1.754			1.937			0.249		

θ denotes the variation of individual frailty,

α denotes the association between two processes at the individual level,

η denotes the variation of familial frailty,

ξ denotes the association between two processes at the familial level.

4.3.2 Impact of frailties

The effects of the frailties in the joint nested frailty model were estimated via four parameters (see Table 4.5); θ represents the variance of individual frailties, η represents the variance of familial frailties, α represents the association between the screening visits and CRC caused by individual random effects, and ξ represents the association between the screening visits and CRC that are due to unknown familial effects.

The specification of frailty models typically requires the heterogeneity parameters to be positive or, in case of homogeneity, to be zero. Therefore under the null hypothesis, the parameter is at the boundary of the parameter space which is $[0,\infty)$, and alternative hypothesis is one-sided which contains an inequality constraint. Maller and Zhou (2003) focused on a case of no covariates and obtained an asymptotic null distribution to the likelihood ratio statistic, which is an equal mixture of a point mass at zero and a chi-square distribution with one degree of freedom, denoted as $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. Claeskens et al. (2008) further studied the asymptotic distribution of the likelihood ratio test for the one-sided testing problem with covariates. The null distribution of the likelihood ratio statistic for testing the one-sided heterogeneity hypothesis in the shared gamma frailty model with Weibull baseline hazard is $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$.

Followed the mixture method, we calculate p-values for likelihood ratio tests for heterogeneity manually. The test statistic is twice the log of the likelihoods ratio, $2\times$ [log likelihood for alternative model - log likelihood for null model]. For testing $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$ while keeping the other parameters unrestricted in the joint nested frailty model, a new joint model with only familial frailty was fit as null model. For η in the joint nested frailty model, joint frailty model was used as null model. For θ in joint frailty model, two separate models for two processes, without frailty, were fit. The log-likelihood for the null model was obtained as the sum of the log-likelihoods from the two separate models. For the shared frailty model, a model without familial frailty was considered as null model. Mixture distributions for likelihood ratio tests for heterogeneity are shown in Table 4.6.

Table 4.6: Mixture distributions for likelihood ratio tests for heterogeneity

	Joint nested frailty			Joint frailty			Shared frailty		
	H_0	H_1	mixed distribution	H_0	H_1	mixed distribution	H_0	H_1	mixed distribution
θ	$\theta = 0, \eta > 0, \xi = 1$	$\theta > 0, \eta > 0, \xi = 1$	$\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$	$\theta = 0$	$\theta > 0$	$\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$	$\theta = 0$	$\theta > 0$	$\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$
η	$\eta = 0, \theta > 0$	$\eta > 0, \theta > 0, \xi = 1$	$\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$	-	-	-	-	-	-

Results are shown in Table 4.7 and 4.8 as follow:

Table 4.7: P-values for likelihood ratio tests for heterogeneity by mixed method

	Joint nested frailty			
	l_{alt}	l_{null}	test statistic	**p-value**
θ	-2299.040	-2670.780	743.480	<0.001
η	-2299.040	-2542.040	486	<0.001

Table 4.8: P-values for likelihood ratio tests for heterogeneity by mixed method

	Joint frailty				Shared frailty			
	l_{alt}	l_{null}	test statistic	**p-value**	l_{alt}	l_{null}	test statistic	**p-value**
θ	-2542.040	-2734.150	384.220	<0.001	-315.720	-331.730	32.020	7.63×10^{-9}

At the individual level, we found that frailty variance (estimate and p-value) and the associated power term of the frailty (estimate and p-value) were significant in the joint nested frailty model (significant frailty variance $\hat{\theta} = 0.65$ ($p < 0.001$) and significant power of the frailty $\hat{\alpha} = 0.731$ ($p < 0.001$)). This indicates obvious and positive residual link between the visiting process and the risk of CRC at the individual level. At the family level, the power (ξ) of the familial frailty was fixed as one. The variance of the familial frailty was significantly different from zero ($\hat{\eta} = 0.74$, $p < 0.001$) indicating positive residual intra-family correlation appears between the visiting process and the risk of CRC. In joint frailty model, the variance of individual frailty was significant (estimate, p-value) with its estimate $\hat{\theta} = 0.66$ ($p < 0.001$). However, α was not significant ($\hat{\alpha} = 0.39$, $p = 0.09$), which indicates no obvious residual link between the visiting process and the risk of CRC at the individual level based on joint frailty model. In shared frailty model only for the terminal event, the frailty parameter θ was significant, indicating significant residual correlation among the times to CRC within families ($\hat{\theta} = 0.49$, $p = 0.001$).

We also compared the estimates of frailty parameters from mutation-specific models (Tables A.1-A.3). The frailty parameters are all significant across mutation-specific models. MLH1 families showed stronger associations ($\hat{\theta} = 0.737$, comparing to $\hat{\theta} = 0.649$ in general model) between the recurrent and disease processes caused by individual random frailty. Estimates of frailty parameters in MSH2-specific model are very similar to the ones in general model. MSH6 families show a stronger association between two processes at the individual level ($\hat{\alpha} = 1.483$, comparing to $\hat{\alpha} = 0.731$ in general model).

4.3.3 Comparison of model fitting

Comparison of joint nested frailty model and joint frailty model can be guided by the values of likelihood cross-validation criterion (LCV) (Rondeau et al., 2017). Lower values of LCV indicate a better fitting model. The gain of using a joint nested frailty model instead of a joint frailty model can be evaluated by comparing the LCV (joint frailty model) = 1.94 to LCV (joint nested frailty model) = 1.75. Since the LCV is 10% lower for the joint nested frailty model, it would seem that the joint nested frailty model fits better than the joint frailty model as it accounts for the familial dependence.

4.4 Dynamic prediction of CRC risk

Dynamic prediction can be provided based on joint modeling. The dynamic prediction $P(t,s)$ is estimated as the probability of developing a terminal event CRC for subjects who survived to a certain time point, based on their visiting and disease histories and those of other family members observed by that time point. When comparing different levels for a certain covariate, controlling other covariates are necessary. In our data set, it is difficult to find several individuals who have similar screening patterns but differ in the covariate of interest. To better illustrate the effect of the specific covariate while controlling for the other covariates, we created a virtual individual for prediction whose covariates are controlled under different situations. Then, this individual was plugged into a family whose proband had first CRC at age 39. Several scenarios were under consideration. We evaluated the following covariates and their effects on the dynamic predictions: gap time between visits, age at the first visit, detection and removal of adenoma, gender, family history, and mutation types.

Starting prediction points were fixed at prediction time $t = 0$, 2, 5 and 10 years with prediction window (s) increasing by one year until $t + s = 15$ years. At each prediction time point, we predicted the risk every year until reaching 15 years from the first visit.

The 95% confidence intervals (CIs) of the prediction probability of developing a terminal event between t and $t + s$ were obtained from Monte Carlo simulations of the parameter values. Calculation of the prediction probabilities is based on the parameter values $\hat{\psi} = (\hat{r}_0(t), \hat{\lambda}_0(t), \hat{\beta}, \hat{\gamma}, \hat{\theta}, \hat{\alpha})$, drawn from the Multivariate Gaussian distribution $MN(\hat{\psi}, \hat{\sum}_{\psi})$. The CIs are the 2.5^{th} and 97.5^{th} percentiles of the prediction probabilities estimated from the n simulated values. Five hundreds sets of parameters were generated to obtain CIs of the predictions in our application.

4.4.1 Effect of gap time between visits on CRC

For comparing the effects of different gap times between successive screening visits, this person was defined as a male who started his first screening visit at age 25 and detected adenoma at the first screening visit. Three gap times were applied to this person which were one-, two-, and three-year gap between two successive visits. Figure 4.1 displays the dynamic prediction for this person of the cumulative risk of developing CRC. It shows that smaller the gap time, higher the probability for this individual to detect a CRC. If this person had no cancer by year 5 and has screening visits every year, then about 40% of chance to detect by year 10 but every 2 years visit would decrease the chance of detecting CRC by 20%.

4.4.2 Effect of age at the first visit on CRC

For comparing different ages at the first visit, this person was defined as a male who had been screened every year, was detected of adenoma at the first visit. Three different ages at the first visit were considered: 20, 25 and 40. These three ages were chosen grounded on the clinical suggestions on the age when a LS individual should start a screening visit. Figure 4.2 displays the dynamic prediction for this person with different first visit ages. It is consistently shown that starting screening visit at an older age would increase the risk of developing a CRC. Figure C.1 includes two more ages at the first visit, 30 and 35, to make the comparison more general, where consistant conclusion can be reached.

4.4.3 Effect of adenoma on CRC

For comparing the effect of detection and removal of adenoma among visits, this person was defined as a male who started his first screening visit at age 25, had been screened every year. Situation with detection of adenoma was compared to the one without detection of adenoma. Figure 4.3 displays the dynamic prediction for this person depending on adenoma detection status. If one who had adenoma detected and removed during screening visit, the risk of CRC would be lower compared to the one with no adenoma detected. It is consistently shown that no detection and removal of adenoma among screenings would lead to a higher the risk of developing a CRC.

4.4.4 Effect of gender on CRC

Similarly, gender effect was compared while fixing the other covariates. Figure 4.4 shows that males have higher risk of developing CRC than females. The results remain consistent as the estimates from the joint nested frailty model.

4.4.5 Effect of family history on CRC

To demonstrate the effect of family history on the dynamic predictions, we chose two families whose family histories are different: high risk family and low risk family. The high risk family we chose contains six individuals with 22 visits including four CRCs in total, and the low risk family contains only one individual with two visits but no CRC. The Figure 4.5 shows that the high risk family with more visits and CRC history generally has a higher risk of developing a CRC than one with less visits without CRC history at all windows starting from prediction time 2.

4.4.6 Effects of mutation types on CRC

We evaluated the effects of different gap times specific to mutation type on dynamic predictions. We chose MLH1 and MSH2 mutations for comparison. The mutation-specfic models (Tables A.1 and A.2) were used for calculating dynamic predictions. Figure 4.6 shows that individuals in MLH1 families have higher risk of CRC than those in MSH2 families.

Besides, based on the model with mutation types as covariates (Tables B.1), dynamic prediction was conducted as well. Under each assigned visiting interval, MLH1, MSH2, MSH6 and PMS2 were compared. Other covariates were controlled at the same levels while comparing different mutation types. Figure C.2 to Figure C.5 in Appendix C shows that individuals in PMS2 families have much higher risk of CRC than those of any other mutation type; given one-year gap time for example, PMS2 patients have approximately 30% higher risk than MLH1 patients after ten years since the first visit. However, the differences decrease with longer gap time. In addition, more frequent screening visits would lead to higher risk to detect a CRC, applied to all kinds of mutation types.

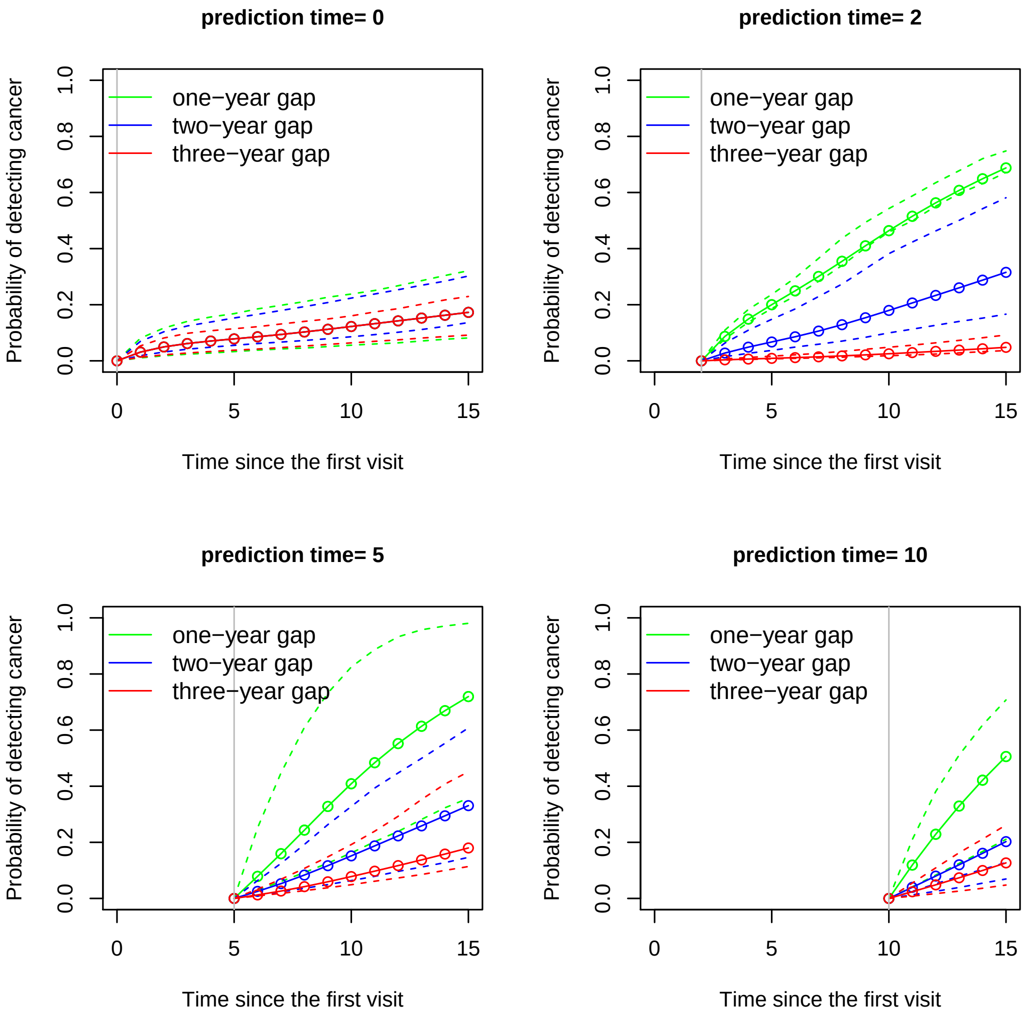

Figure 4.1: Comparison of the effects of screening intervals on dynamic prediction of CRC, $P(t,s)$, at fixed prediction time $t = 0,2,5$ and 10 years with s increasing by one until $t+s =$ 15 years, for comparing three different screening intervals. The grey vertical line represents the time of prediction. The dashed lines represent 95% confidence intervals of the prediction probabilities of developing a terminal event between t and $t+s$. A virtual individual was defined as a male who started his first screening visit at age 25, was detected of adenoma and presented CRC during the follow-up time. Situations of this person visited every one year, every two years, and every three years were compared.

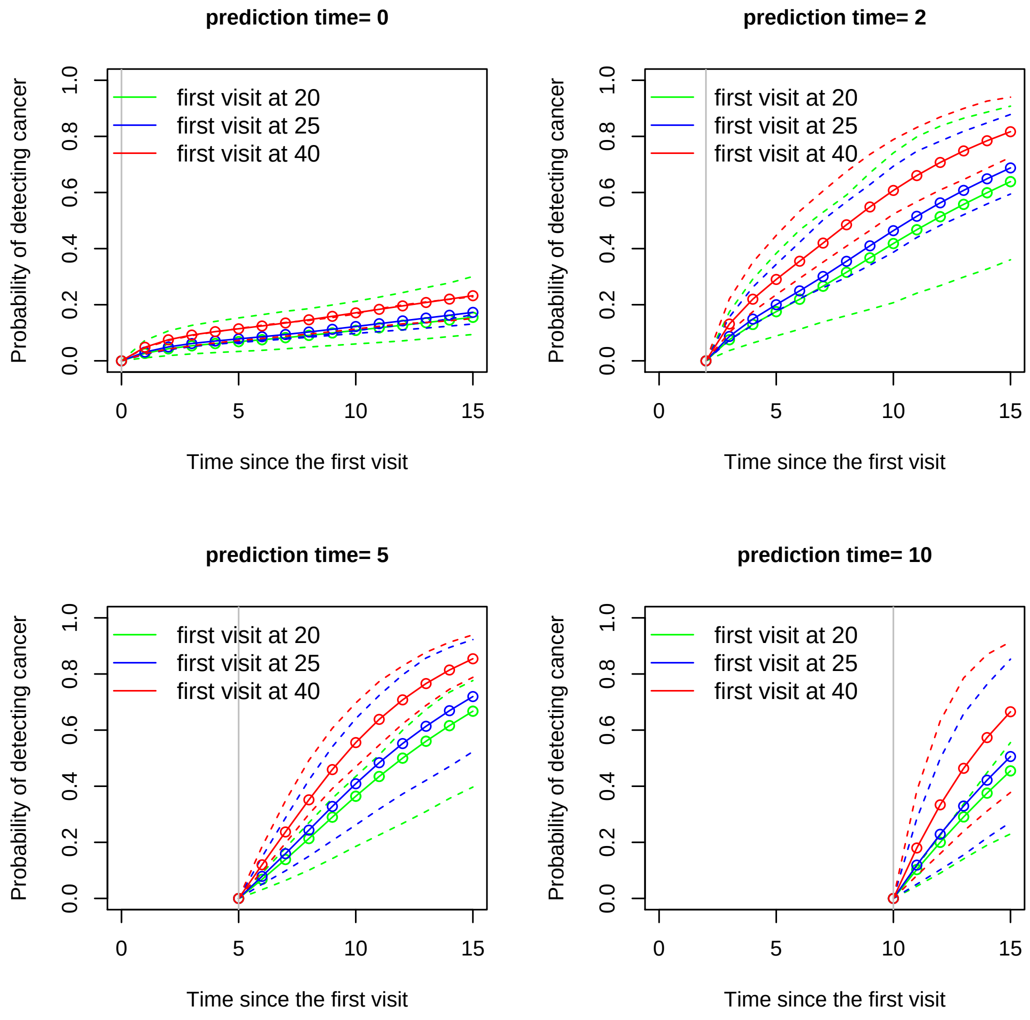

Figure 4.2: Comparison of the effects of the first visit ages on dynamic prediction of CRC, $P(t,s)$, at fixed prediction time $t = 0,2,5$ and 10 years with s increasing by one until $t+s = 15$ years, for comparing different ages at the first visit. The grey vertical line represents the time of prediction. The dashed lines represent 95% confidence intervals of the prediction probabilities of developing a terminal event between t and $t+s$. A virtual individual was defined as a male who had been screened every year, was detected of adenoma and presented CRC during the follow-up time. Situations of this person first visited at age of 20, 25 and 40 were compared.

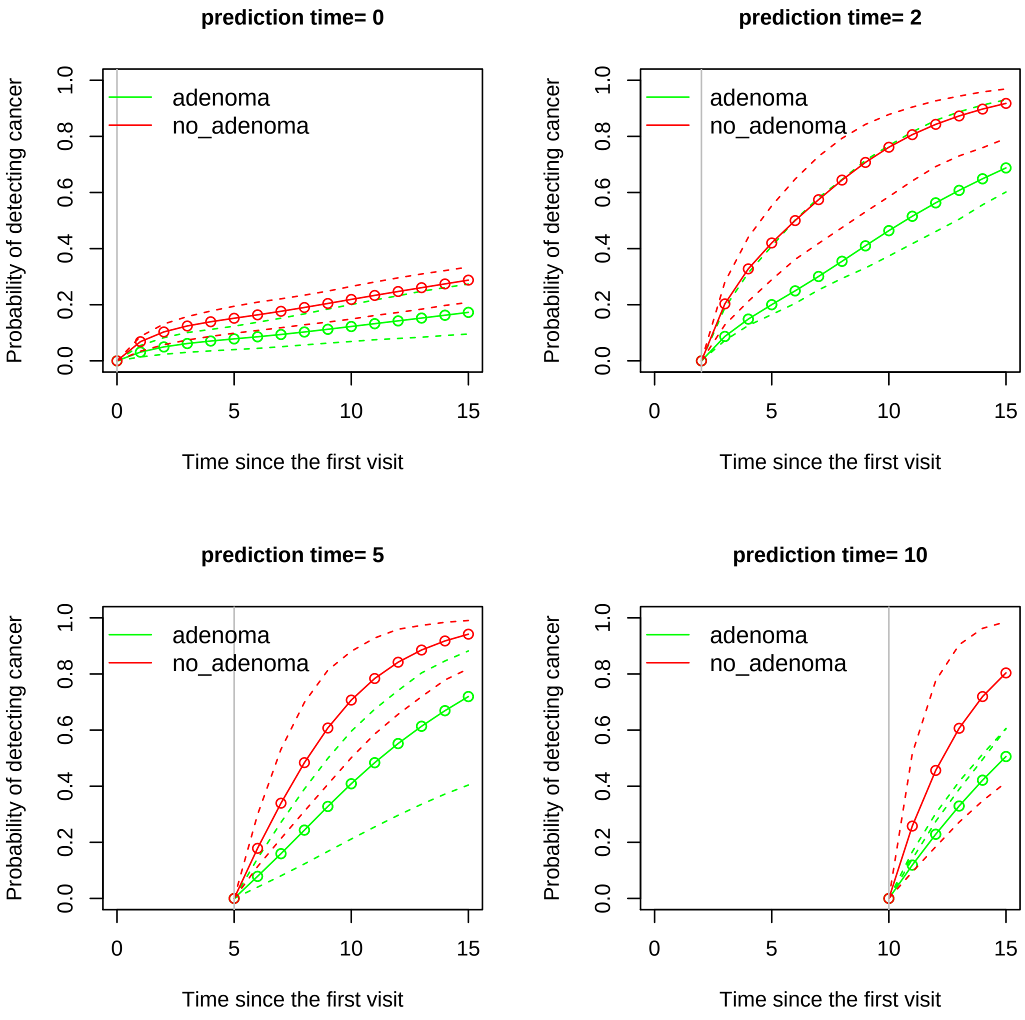

Figure 4.3: Comparison of adenoma effects on dynamic prediction of CRC, $P(t,s)$, at fixed prediction time $t = 0,2,5$ and 10 years with s increasing by one until $t+s = 15$ years, for comparing effect of adenoma among visits. The grey vertical line represents the time of prediction. The dashed lines represent 95% confidence intervals of the prediction probabilities of developing a terminal event between t and $t+s$. A virtual individual was defined as a male who started his first screening visit at age 25, had been screened every year and presented CRC during the follow-up time. Situations of when this person had adenomas before CRC and no adenomas were compared.

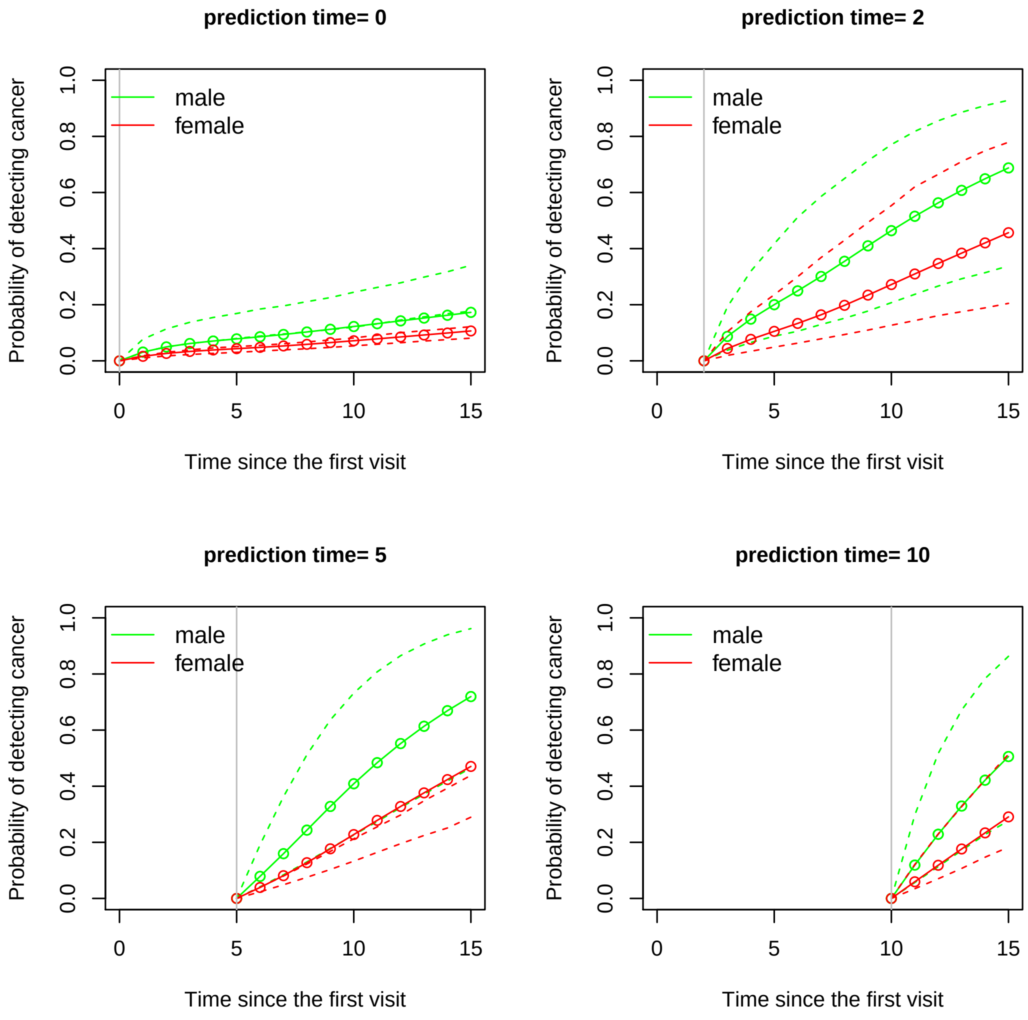

Figure 4.4: Comparison of gender effects on dynamic prediction of CRC, $P(t,s)$, at fixed prediction time $t = 0,2,5$ and 10 years with s increasing by one until $t+s = 15$ years, for comparing effect of gender. The grey vertical line represents the time of prediction. The dashed lines represent 95% confidence intervals of the prediction probabilities of developing a terminal event between t and $t+s$. A virtual individual was defined as a person who started his first screening visit at age 25, had been screened every year, was detected of adenoma, and presented CRC during the follow-up time. Situations of when this person was a male and a female were compared.

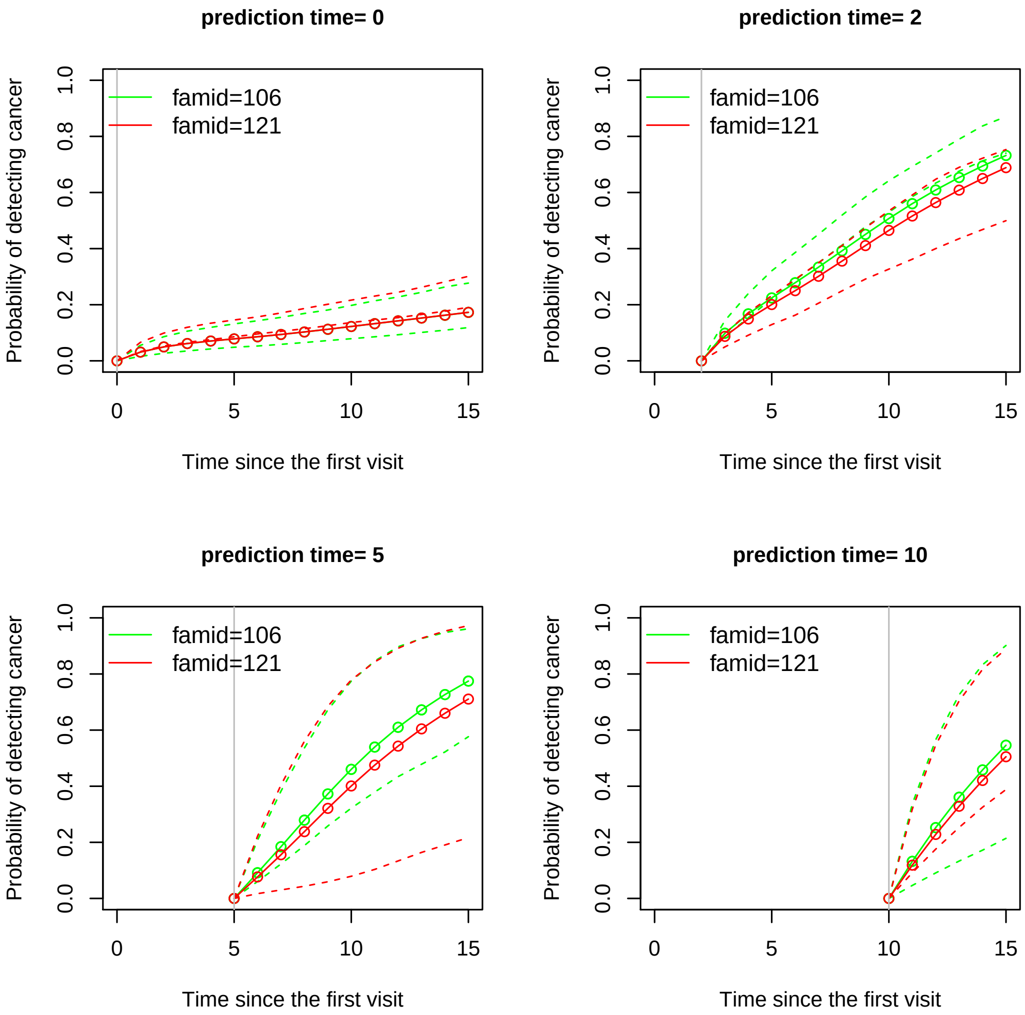

Figure 4.5: Comparison of the effects of family histories on dynamic prediction of CRC, $P(t,s)$, at fixed prediction time $t = 0{,}2{,}5$ and 10 years with s increasing by one until $t+s = 15$ years, for comparing different family histories. The grey vertical line represents the time of prediction. The dashed lines represent 95% confidence intervals of the prediction probabilities of developing a terminal event between t and $t+s$. A virtual individual was defined as a male who started his first screening visit at age 25, had been screened every year, was detected of adenoma, and presented CRC during the follow-up time. This individual was assigned into two families with different histories. High risk (family 106): six individuals with 22 visits including four CRCs in total; low risk (family 121): one individual with two visits but no CRC.

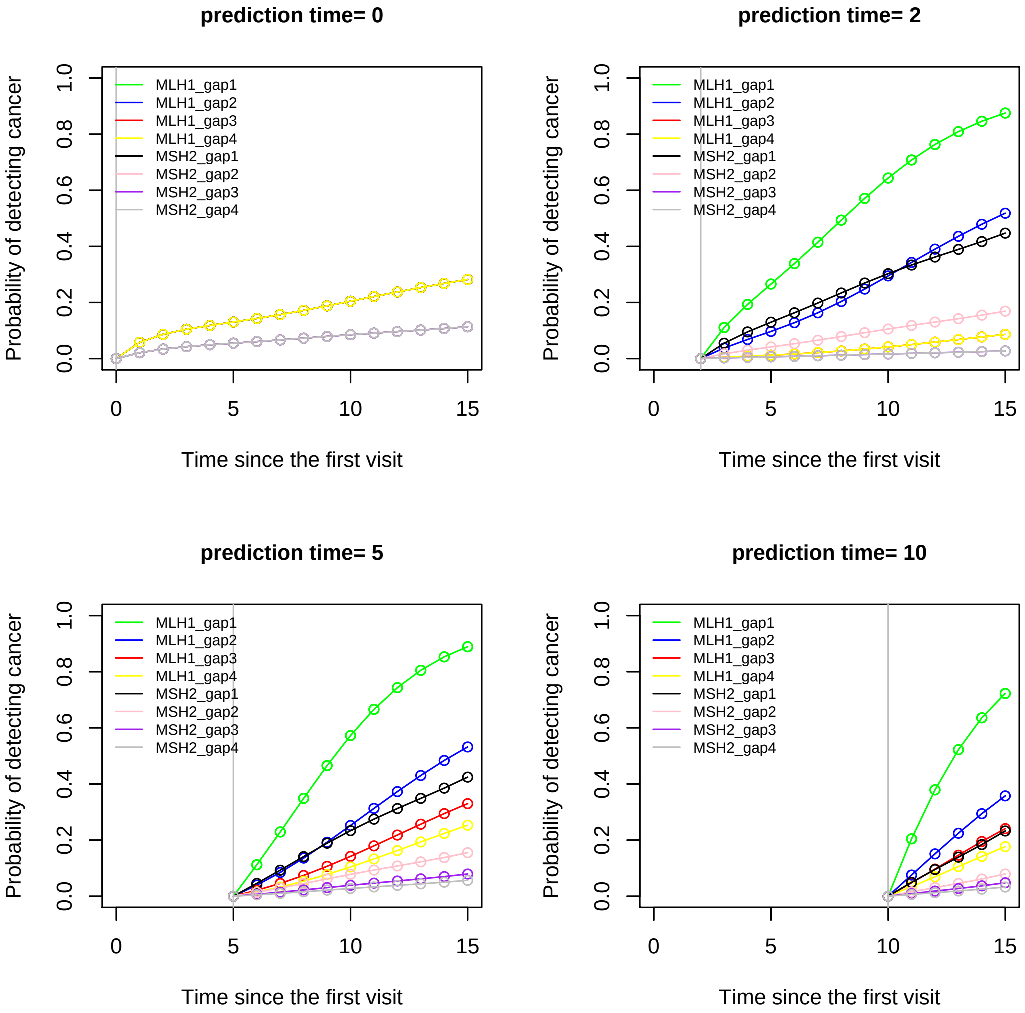

Figure 4.6: Comparison of the effects of gap times specific to mutation type on dynamic prediction of CRC, $P(t,s)$, at fixed prediction time $t = 0,2,5$ and 10 years with s increasing by one until $t+s =15$ years. The grey vertical line represents the time of prediction. A virtual individual was defined as a male who started his first screening visit at age 25, had been screened every year, was detected of adenoma, and presented CRC during the follow-up time. Two mutation types, MLH1 and MSH2 were assigned to this individual.

4.5 Comparison of prediction accuracies

Main purpose of our study was to compare the three models used in our analysis in terms of their prediction accuracies. We evaluated the prediction accuracy of the dynamic predictions by 10-fold cross-validated Brier Scores. For k-fold cross validation, while there is no overlap between the test sets on which the models are evaluated, there is overlap between the training sets for cross validation with $k > 2$. The overlap is the largest for leave-one-out cross validation, which leads to the correlation among learned models and increase of the variance with the amount of covariance. Therefore, leave-one-out cross validation has larger variance in comparison to k-fold cross validation with smaller k. The general suggestion of Kohavi (1995) to use 10-fold cross validation has been widely accepted. It is crucial to keep the family structure intact with familial frailty models. We therefore divided data based on families for cross-validation. Predictions were done for all the individuals in the first fold at prediction times $t = 0,2,5$ and 10 up to 15 years based on the model fitted using training data. Brier Score was calculated every time after prediction. This procedure was looped for all of the ten folds. At given prediction time t and prediction window s, the weighted Brier Scores accounting for right censoring was calculated and the results across all t and s are presented in Figure 4.7.

The joint nested frailty model consistently provided lower prediction error than the shared frailty model. Figure 4.7 clearly shows that BS values of joint nested frailty model (indicated as blue line) are the lowest compared to the ones from the other models at all windows at prediction time 0 and 5. The shared frailty model (indicated in black) has lower BS values than joint nested frailty model's at the last two prediction windows at prediction time 2; before prediction window of twelve years at prediction time point 2, the contrary is the case. At prediction time point 10, the joint nested frailty model shows worse prediction accuracy than the other two models. In general, joint nested frailty model appears to have better dynamic prediction accuracy than the other models as it accounts for familial history and correlation.

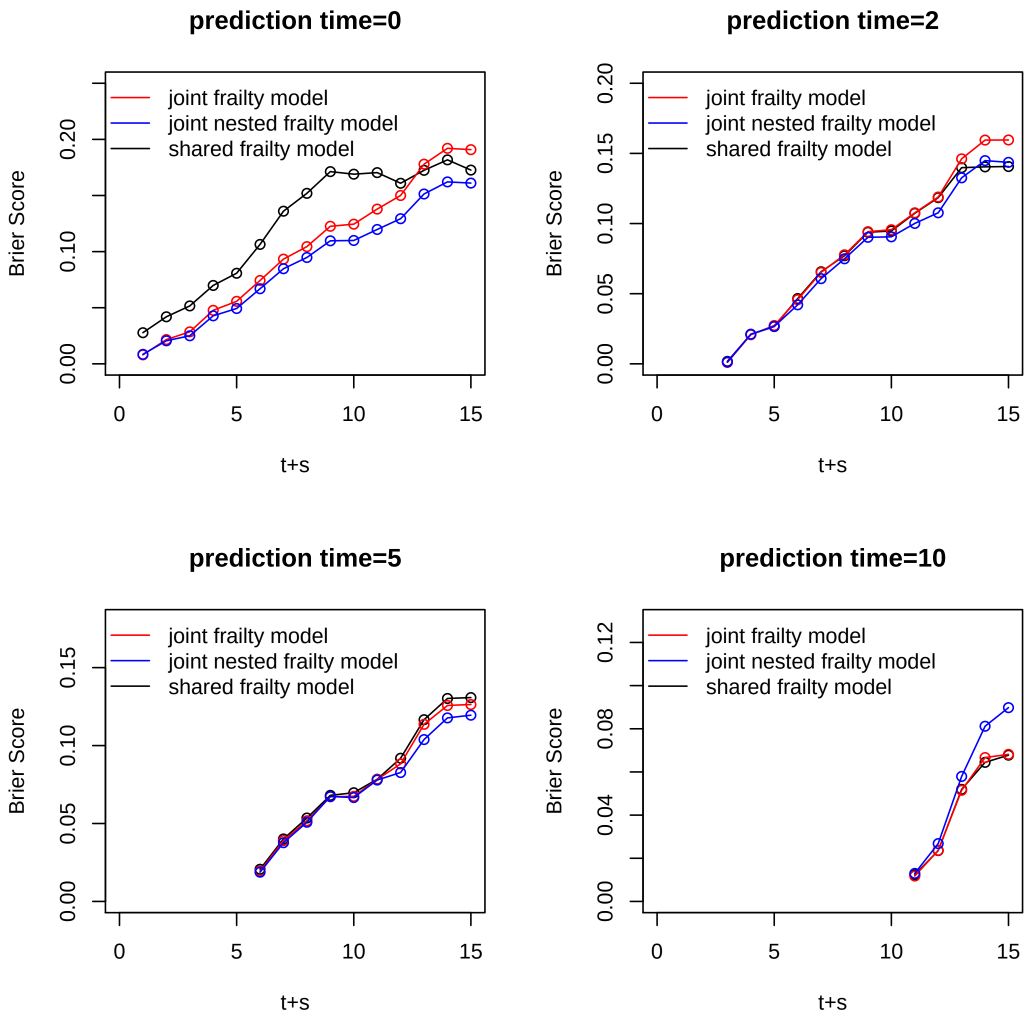

Figure 4.7: Brier Score using 10-folds cross validation from three models at different time points ($t = 0, 2, 5, 10$) and varying windows ($s = 1, \ldots, 15$).

4.6 Analysis of the screening effect on mortality

Literature has shown that screening process not only affects the risk of developing CRC, but also impacts the mortality from CRC. We therefore set out to evaluate the impact of the screening on death after CRC using the same data set. We considered death as a terminal event in the shared frailty model with familial frailty.

For the analysis of death from CRC, we included 103 individuals who had CRC during the follow-up time. Ten of 103 individuals were excluded from the analysis due to missing information on proportion of removed colon or cancer stage. Out of the remaining 93 individuals who detected CRC during the study time, only 65 followed up screening visits after the first CRC; of 93, 16 died and 77 survived. The remaining 28 individuals did not have any screening visit after CRC; of 28, five died and 23 survived, whom we treated as censored.

The time to death from CRC was analyzed using the shared frailty model with covariates of interest. In terms of screening related covariates, we considered average gaptime before CRC, the cumulative number of visits after CRC as a time-dependent covariate, adenomas detection status before CRC, and the first visit age. Related to cancer, we considered age at the first CRC, cancer stage (0 for low stage, and 1 for high stage) and the proportion of colon removed in the model.

Results from the shared frailty model for death from CRC are summarized in Table 4.9. P-value by mixed method is calculated and shown in Table 4.10. In this model, all covariates are significant except for the cumulative number of visits after CRC; age at CRC, adenoma detection and removal, average gap time, cancer stage, proportion of colon removed and gender were all significantly associated with the mortality after CRC. Those who are one-year older to develop a CRC are at a exp(0.064) = 1.066 ($p < 0.001$) times mortality. One-year increase in gap time between two screening visits can lead to a exp(0.120) = 1.127-fold ($p = 0.003$) risk of death. Patients with high cancer stage CRC have dramatically higher mortalities (exp(0.861) = 2.365, $p = 0.039$) than those with low cancer stage CRC. Proportion of removal of colon is also shown to play an vital role on risk of death; patients removed larger proportion of colon have dramatically lower mortalities (83.7% lower, $p = 0.019$) than those removed smaller proportion of colon. Detection and removal of adenoma before CRC would significantly lower the mortality (exp(-3.108) = 0.045, $p = 0.001$). Estimates of two indicators of mutation types show that MSH2 patients have the highest risk of death from CRC among the three types under consideration. MLH1 patients have much lower risk of death than MSH2 patients; the difference is significant. Familial random effect is significant in this model ($p = 0.038$).

To better illustrate CRC mortality associated with cancer stage and average screening gap time, Figure 4.8 and Figure 4.9 displays survival functions specific to cancer stage and mutation type across different average gap times when other covariates are fixed. Two cancer stages are shown in separate plots for comparison. Low stage cancer is more likely to survive from death than high stage cancer. For low stage CRC patients, after 20 years from detection of CRC, the survival probability decreases to around 90% to 95%, while it takes less than 10 years for high stage CRC patients. Screening interval plays an important role in survival probability. More frequent screenings significantly leads to a higher survival probability, working for both of the two cancer stages. Consistently found in the figures, MLH1 patients have the highest survival probability, followed by MSH6, and MSH2 patients have the lowest chance to survive from CRC, applied to both cancer stages and all the four different gap times. Comparison focusing on the effect of gap times can be found in the Figure C.6 to Figure C.8 in Appendix C. From these plots, the effect of gap time on the survival probability is obvious; more frquent screening visits can increase the survival probability, for all the three mutation types patients and also for both two cancer stages. The results from this model is more persuasive for giving suggestions to patients on frequent screening visits.

Results from the mutation type specific shared frailty models for death from CRC are summarized in Table A.4, Table A.5 and Table A.6. Proportion of removed colon and gender are no longer significant in MLH1- and MSH6-specific models, but remain significant in MSH2-specific model. Cumulative number of visits after CRC is significant for MSH2-specific model, meaning that for MSH2 patients, one more visit after detection of CRC would lead to 33.5% ($p = 0.004$) higher risk of death. However, detection and removal of adenoma and age at CRC have no significant impact on MSH2 patients' mortality. For MSH6 patients, only age at the first visit is significant to their risk of death. Results are just for reference due to limited amount of individuals in each group.

Table 4.9: Shared frailty model with Splines baseline hazard using death from CRC as the terminal event with mutation type as covariates (indicators)

Variables	estimate	se	p-value
detection of adenoma before crc	-3.108	0.947	0.001
age at crc	0.064	0.017	<0.001
average gap time between visits before crc	0.120	0.041	0.003
cumulative number of visit after crc [†]	0.059	0.068	0.387
MLH1 compared to MSH2	-1.093	0.454	0.016
MSH6 compared to MSH2	-0.431	0.556	0.438
cancer stage	0.861	0.417	0.039
proportion of removed colon	-1.813	0.770	0.019
gender	1.434	0.434	0.001

[†] time-dependent covariate

Frailty parameters			
	estimate	se	p-value
θ	0.002	0.002	-
penalized marginal log-likelihood	-157.280		
LCV	0.450		

Table 4.10: P-value for likelihood ratio test for heterogeneity by mixed method – death from CRC as the terminal event

	Shared frailty			
	l_{alt}	l_{null}	test statistic	**p-value**
θ	-157.280	-158.860	3.160	0.038

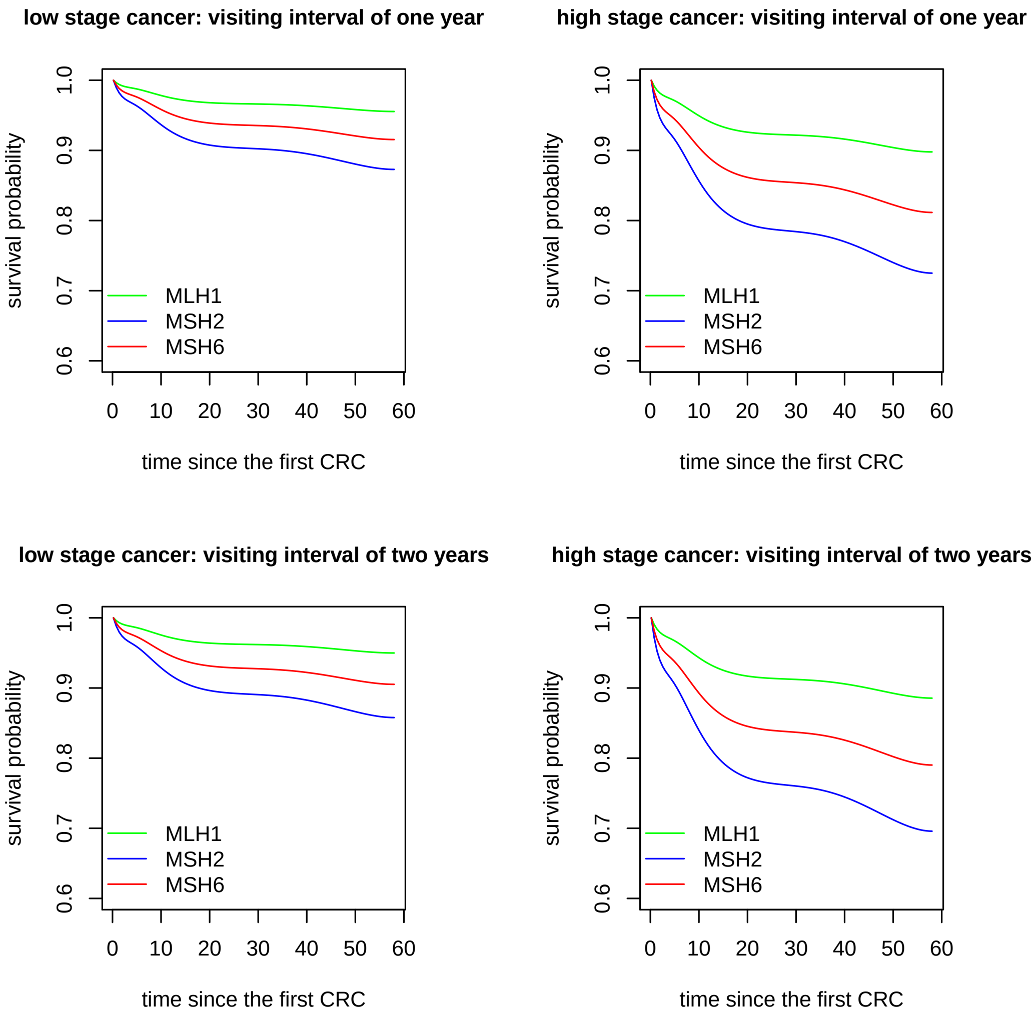

Figure 4.8: Survival probabilities after the first CRC, specific to cancer stage, screening gap times, and mutation types. Gap times are assigned as one year and two years. Under each assigned visiting interval, MLH1, MSH2 and MSH6 patients are compared in two separate plots of two different stage of CRC. The x-axis is the time in years since the detection of the first CRC. The y-axis is the survival probability from CRC death.

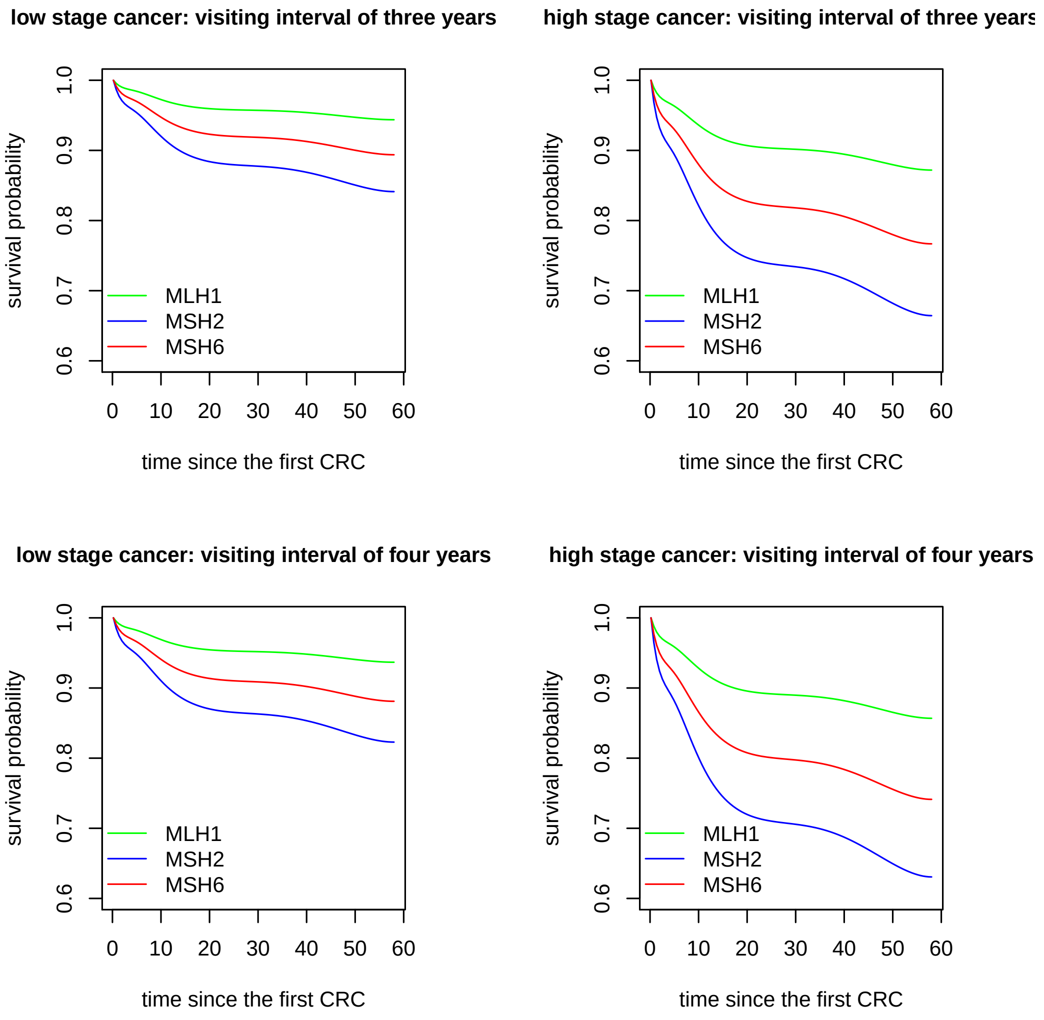

Figure 4.9: Survival probabilities after the first CRC, specific to cancer stage, screening gap times, and mutation types. Gap times are assigned as three years and four years. Under each assigned visiting interval, MLH1, MSH2 and MSH6 patients are compared in two separate plots of two different stage of CRC. The x-axis is the time in years since the detection of the first CRC. The y-axis is the survival probability from CRC death.

4.7 Analysis of screening effect on cancer stage

The cancer stage information is an important information for treatment plan and prognosis. Generally, the higher the number, the more the cancer has spread. We considered cancer stage, 0, 1 and 2 as low stage cancer and 3 and 4 as high stage cancer. Physicians may be more interested in questions such as, what factors differentiate the two levels of stage. Among those patients who had CRC, cancer stage was considered as a terminal event.

The gap times from the first visit to specific cancer stages were analyzed using the shared frailty model with covariates of interest. For modeling the time to low stage cancer, individuals with high stage cancer were removed from the data. Similarly, for modeling the time to a high stage cancer, individuals with low stage cancer were removed. As a result, 47 patients were considered as low stage cancer and 49 individuals were considered as high stage cancer.

Covariates considered for both models are: the first visit age, detection and removal of adenoma before CRC, detection and removal of other polyps before CRC, the screening frequency or average gaptime before CRC, proband's CRC age and gender (1 for male and 0 for female). Some interactions between covariates were considered but found not significant so they were not included in the final model. The parameter estimates are summarized in Table 4.11, separately for the low stage cancer and the high stage cancer. P-values by mixed method are shown in Table 4.12. Mutation specific estimates can be found in Appendix A in Table A.7, A.8 and A.9.

In the model with low stage cancer, age at the first screening visit and average gap time between visits were significant; ten year older at the first screening visit increases the risk of developing a low stage cancer by 72.4% ($p = 0.025$), and one year increase in average gap time between visit significantly ($p = 0.033$) decreases the risk of low stage cancer by 12.2%. MSH6 patients have much lower risk to develop a low stage CRC than MSH2 patients, while MLH1 and PMS2 have higher risks of low stage CRC. However, none of the comparisons in terms of mutation types are statistically significant. Significant familial frailty parameter estimate ($\hat{\theta}$ = 4.508, $p = 0.001$) indicates positive familial correlation in the times to low stage cancer.

In the model with high stage cancer, detection and removal of adenoma was found significant, meaning that thsoe individuals who were detected and removed adenoma during the screenings have exp(-1.103) = 0.332 times of risk to develop a high stage cancer than those without any detection of adenoma. Gender is significant as well with *p*-value of 0.048, indicating males have higher risks of developing a high stage CRC than females. MSH2 patients have the lowest risks among all four mutation types under consideration. Compared to MSH2 patients, MLH1 patients have exp(0.712) = 2.038-fold risk of high stage CRC and PMS2 patients have exp(1.360) = 3.8977-fold risk of high stage CRC, with *p*-values 0.026 and 0.009, respectively. Familial frailty becomes not significant in model with high stage cancer. However, due to the limited amount of data in both groups, significance is for reference only.

Cumulative hazard plots specific to two stages of CRC are shown in Figure 4.10 and Figure 4.11. It is more likely to develop a low stage CRC than a high stage CRC in general. Applied to every assigned gap time, MLH1 and PMS2 patients have similar and the highest risk to develop a low stage CRC; after five years since the first visit, they are under a 80% hazard of low stage CRC, while MSH6 patients have less than 20% risk. For high stage CRC, PMS2 patients are under the highest risk as well, while MSH2 patients have the lowest hazard. The pattern remains the same across four visiting intervals.

Comparison focusing on the effect of gap times can be found in the Figure C.9 to Figure C.12 in Appendix C. From these plots, more frquent screening visits can increase the probability of detecting either low stage CRC or high stage CRC, for all the four mutation types patients. The differences among four gap times are more visible in detecting a low stage CRC than the ones in high stage CRC, which is consistent with the model results shown in Table 4.11.

Table 4.11: Shared frailty models with Weibull baseline hazard: cancer stages as the terminal events with mutation type as covariates (indicators)

	Low stage cancer			**High stage cancer**		
Variables	estimate	se	p-value	estimate	se	p-value
age at the first visit	0.544	0.243	0.025	0.099	0.114	0.388
detection of adenoma before crc	0.132	0.450	0.770	-1.103	0.408	0.007
detection of other polyps before crc	0.373	0.504	0.460	0.334	0.343	0.329
average gap time between visits	-0.131	0.061	0.033	-0.029	0.025	0.237
proband's crc age	0.089	0.214	0.678	0.025	0.119	0.831
gender	0.437	0.470	0.353	0.564	0.285	0.048
MLH1 compared to MSH2	0.676	0.582	0.245	0.712	0.320	0.026
MSH6 compared to MSH2	-1.759	1.041	0.091	0.433	0.483	0.371
PMS2 compared to MSH2	0.616	1.263	0.626	1.360	0.519	0.009
Frailty parameters						
θ	4.508	2.319	-	<0.001	0.010	-
marginal log-likelihood	-181.530			-177.750		
AIC	0.536			0.523		

Table 4.12: P-values for likelihood ratio tests for heterogeneity by mixed method – cancer stages as the terminal events

	Low stage CRC				High stage CRC			
	l_{alt}	l_{null}	test statistic	**p-value**	l_{alt}	l_{null}	test statistic	**p-value**
θ	-181.530	-186.460	9.860	0.001	-177.750	-177.760	0.020	>0.999

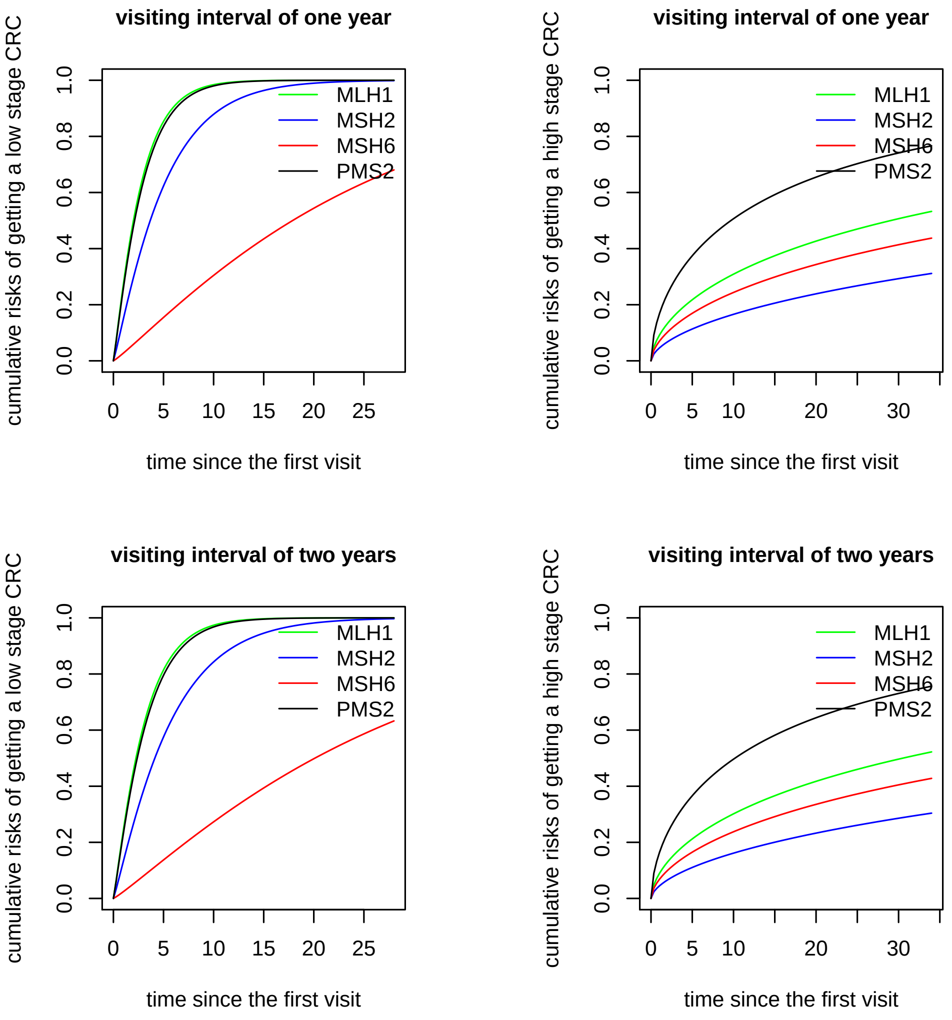

Figure 4.10: Cumulative hazard for low stage cancer (left) and high stage cancer (right), specific to screening gap times of one year and two years. Four mutation types are under comparison. The x-axis is the time in years since the first visit. The y-axis is the cumulative hazard of developing cancers in two stages.

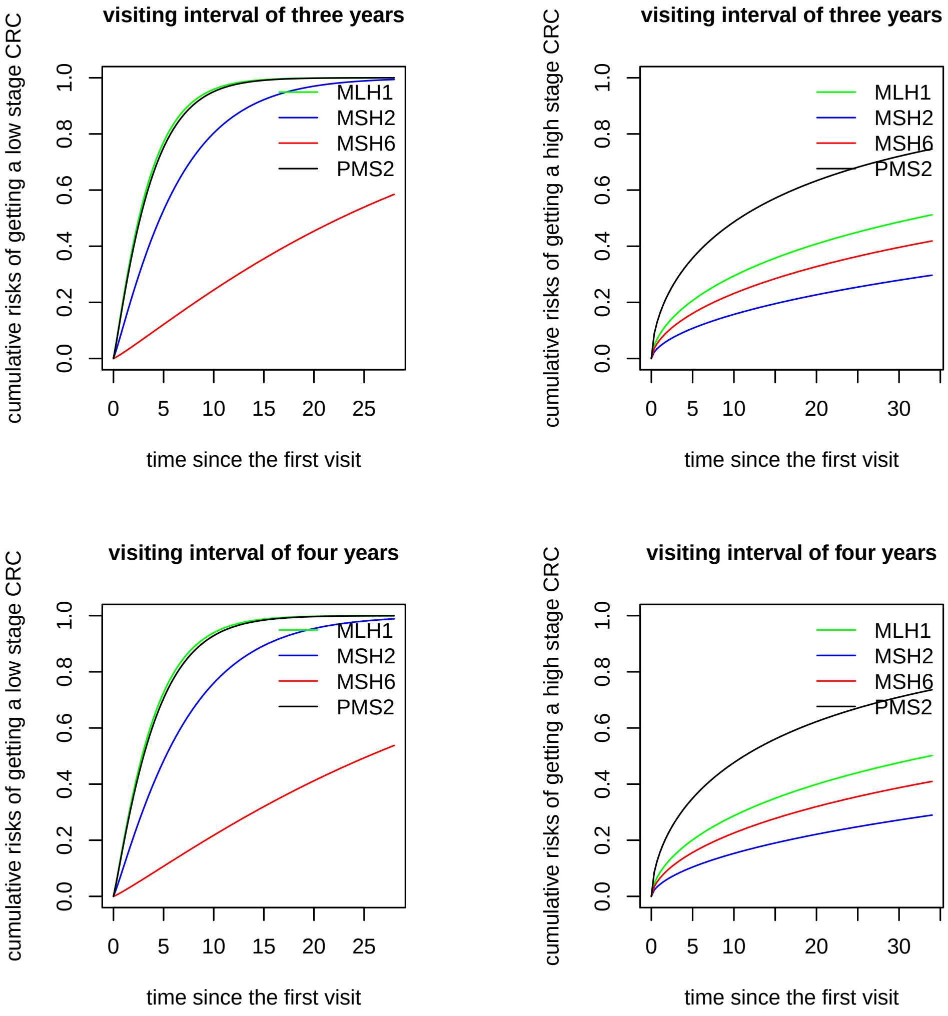

Figure 4.11: Cumulative hazard for low stage cancer (left) and high stage cancer (right), specific to screening gap times of three years and four years. Four mutation types are under comparison. The x-axis is the time in years since the first visit. The y-axis is the cumulative hazard of developing cancers in two stages.

Chapter 5

Discussion

This thesis set out to evaluate and compare three statistical models for evaluating the screening efficiency on the risk of disease or mortality for clustered survival data. We applied shared frailty model, joint frailty model and joint nested frailty model to LS family data to predict the risk of CRC. The impact of screening visits on the risk of developing a first CRC for LS families was assessed, based on the three models. Mutation type-specific CRC risks were estimated in the analysis. We provided dynamic predictions and assessed their prediction accuracies across the three models. To better understand the screening efficiency, we further analyzed the effects of screening visits associated with CRC mortality and also with cancer stages.

We addressed the following challenges throughout the thesis: 1) A simplest way to incorporate the screening visits is as a time-dependent covariate in the disease model. We treated the cumulative number of visits as a time-dependent covariate in the shared frailty model for CRC occurrence. This modeling approach was comparable to complicated joint modeling as it still captured important covariate effects while incorporating time-dependent screening visits in the model. 2) For analysis of clustered data, we have incorporated familial correlation into our modeling of screening visits with cancer occurrence and also with mortality. In addition, our data include information from multiple screening visits for each individual, leading to a complex nested structure; multiple visits observed within individuals and individuals are clustered in family. We employed the joint nested frailty model for better incorporating the complex data structure. The familial frailty in our joint nested frailty model, which takes the unmeasured factors within a family into account, was considered as well in shared frailty model. 3) One of the main objectives was to investigate mutation-specific risks of developing CRC and associated effects of covariates. There exists a potential problem in terms of model fitting, since sample size shrinks when separating the whole data set into subsets by mutation types. Choice of covariates matters for model fitting.

This study analyzed LS family data from Mount Sinai Hospital. Among the 242 LS families of 422 individuals included in the analysis, age at the first visit, age at the previous visit, detection of adenoma at the previous visit and detection of serrated polyps at the previous visit were all significantly related to the the frequency of screening visits ($p < 0.001$, $p < 0.001$, $p < 0.001$ and $p = 0.005$, respectively). Several covariates were found to be associated with the risk of developing CRC, which included age at the first visit, any detection of adenoma before CRC and gender. Those individuals who had adenoma before CRC decreased the risk of developing CRC compared to those who did not. A 10-year increase in the age at the first visit age would increase the risk of developing CRC by 40%. Besides, males have a 2.04-fold higher risk of getting CRC versus to females. The time-dependent covariate in proportional hazard model, cumulative number of visits before CRC, is found significantly influential on the risk of CRC, with exp(0.206) = 1.23-fold (23% higher) risk when cumulative number of visits increase one at a certain visit. These results uniformly show that the screening process is considerable when estimating the parameters in the terminal event.

Frailty parameters, θ and α, were found significant in joint nested frailty model. It is worthwhile to consider inter-individual and inter-family variation. A significant positive familial correlation is observed in proportional hazard model as well. Joint nested frailty model was revealed to be better than joint frailty model in terms of model fitting, according to LCV and penalized marginal log-likelihood. Our observations are consistent with Choi et al. (2017), who demonstrated the importance of screening process when estimating the parameters related to the terminal event, and omitting residual familial correlations could have impact on the screening effects.

In mutation type specific models, although magnitudes of estimates varied slightly, the directions of the effect kept consistent. This is the first study, to our knowledge, to examine the screening effect on risk of developing CRC in mutation-type-specific joint nested frailty model. In addition, there are few results on how the screening visit intervals would affect the risk of detecting low stage cancer and high stage cancer.

The study has a number of possible limitations:

1. Firstly, goodness of fit was only judged by the comparison of LCV and penalized marginal log-likelihood between candidate models. Simulation evaluation is needed to provide a thorough examination.

2. Secondly, the familial frailty is assumed to be shared over all family members. This assumption may not be true in reality. Siblings could have a stronger correlation than the correlation between kids and grandparents. The Kinship matrix may better accommodate familial relatedness. Kinship coefficient matrix is the matrix of probability that randomly selected allele from two individuals are identical by decent (Lange, 2003), which is widely used to measure the relatedness within a cluster. Instead of considering the relationship of any two individuals in the family, we could consider different frailties among different generations or any pairs of individuals within families.

3. Thirdly we only focused on the first CRC but successive CRCs may arise the problem of competing risk. Individuals may experience several stages, such as death before any CRC, developing a first CRC, death after CRC and survival upon the first CRC but getting a second CRC. However, the successive CRCs after the first CRC were not allowed in our model.

4. Finally, the occurrence of polyps over time could be added to our joint model as an additional potential recurrent process. This recurrent polyps process could be correlated with our screening visit process and terminal event and subject to the individual and familial frailties as well.

The above limitations are research avenue for future research. Kinship matrix can be considered to describe the relatedness among family members. Competing risk which appears when there exists successive CRCs, can also be a great extension of joint frailty model and joint nested frailty model.

Bibliography

American Society of Clinical Oncology (2017), 'Lynch syndrome', https://www.asco.org/practice-guidelines/cancer-care-initiatives/genetics-toolkit/management-individuals-increased.

Atkin, W. S., Edwards, R., Kralj-Hans, I., Wooldrage, K., Hart, A. R., Northover, J. M., Parkin, D. M., Wardle, J., Duffy, S. W. and Cuzick, J. (2010), 'Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial', *The Lancet* **375**(9726), 1624 – 1633.

Belot, A., Rondeau, V., Remontet, L., Giorgi, R. and the CENSUR working survival group (2014), 'A joint frailty model to estimate the recurrence process and the disease-specific mortality process without needing the cause of death', *Statistics in Medicine* **33**(18), 3147–3166.

Blanche, P., Proust-Lima, C., Loubre, L., Berr, C., Dartigues, J.-F. and Jacqmin-Gadda, H. (2015), 'Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks', *Biometrics* **71**(1), 102–113.

Bonadona, V., Bonati, B., Olschwang, S. and et al (2011), 'Cancer risks associated with germline mutations in mlh1, msh2, and msh6 genes in lynch syndrome', *JAMA* **305**(22), 2304–2310.

Canadian Cancer Society (2016), 'Colorectal cancer', http://www.cancer.ca/en/cancer-information/cancer-type/colorectal/colorectal-cancer/?region=on.

Chambless, L. E., Cummiskey, C. P. and Cui, G. (2011), ‘Several methods to assess improvement in risk prediction models: Extension to survival analysis’, *Statistics in Medicine* **30**(1), 22–38.

Choi, Y.-H., Jacqmin-Gadda, H., Krol, A., Parfrey, P., Briollais, L. and Rondeau, V. (2017), Joint nested frailty models for screening visit and disease processes in lynch syndrome families. manuscript under review.

Claeskens, G., Nguti, R. and Janssen, P. (2008), ‘One-sided tests in shared frailty models’, *TEST* **17**(1), 69–82.

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

Cole, S. R., Chu, H. and Greenland, S. (2014), ‘Maximum likelihood, profile likelihood, and penalized likelihood: A primer’, *American Journal of Epidemiology* **179**(2), 252–260.

Colon Cancer Canada (2013), ‘Fast facts on colorectal cancer (crc)’, `http://coloncancercanada.ca/fast-facts-on-colorectal-cancer-crc/`.

Commenges, D. and Gegout-Petit, A. (2007), ‘Likelihood for generally coarsened observations from multistate or counting process models’, *Scandinavian Journal of Statistics* **34**(2), 432–450.

Dafni, U. (2011), ‘Landmark analysis at the 25-year landmark point’, *Circulation: Cardiovascular Quality and Outcomes* **4**(3), 363–371.

Dixon, S. N., Darlington, G. A. and Desmond, A. F. (2011), ‘A competing risks model for correlated data based on the subdistribution hazard’, *Lifetime Data Analysis* **17**, 472–495.

Dowty, J. G., Win, A. K., Buchanan, D. D., Lindor, N. M., Macrae, F. A., Clendenning, M., Antill, Y. C., Thibodeau, S. N., Casey, G., Gallinger, S., Marchand, L. L., Newcomb, P. A., Haile, R. W., Young, G. P., James, P. A., Giles, G. G., Gunawardena, S. R., Leggett, B. A., Gattas, M., Boussioutas, A., Ahnen, D. J., Baron, J. A., Parry, S., Goldblatt, J., Young, J. P., Hopper, J. L. and Jenkins, M. A. (2013), 'Cancer risks for mlh1 and msh2 mutation carriers', *Human Mutation* **34**(3), 490–497.

Duchateau, L., Janssen, P., Kezic, I. and Fortpied, C. (2003), 'Evolution of recurrent asthma event rate over time in frailty models', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **52**(3), 355–363.

Gunsoy, N., Garcia-Closas, M. and Moss, S. (2014), 'Estimating breast cancer mortality reduction and overdiagnosis due to screening for different strategies in the united kingdom', *British Journal of Cancer* **110**, 2412–2419.

Halligan, S., Altman, D. G. and Mallett, S. (2015), 'Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach', *European Radiology* **25**(4), 932–939.

Hampel, H., Frankel, W. L., Martin, E., Arnold, M., Khanduja, K., Kuebler, P., Clendenning, M., Sotamaa, K., Prior, T., Westman, J. A., Panescu, J., Fix, D., Lockman, J., LaJeunesse, J., Comeras, I. and de la Chapelle, A. (2008), 'Feasibility of screening for lynch syndrome among patients with colorectal cancer', *Journal of Clinical Oncology* **26**(35), 5783–5788. PMID: 18809606.

Hendriks, Y. M., Wagner, A., Morreau, H., Menko, F., Stormorken, A., Quehenberger, F., Sandkuijl, L., Mller, P., Genuardi, M., van Houwelingen, H., Tops, C., van Puijenbroek, M., Verkuijlen, P., Kenter, G., van Mil, A., Meijers-Heijboer, H., B., T. G., Breuning, M. H., Fodde, R., Winjen, J. T., Brcker-Vriends, A. H. and Vasen, H. (2004), 'Cancer risk in hereditary nonpolyposis colorectal cancer due to msh6 mutations: impact on counseling and surveillance', *Gastroenterology* **127**(1), 17 – 25.

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

Hougaard, P. (2000), *Analysis of Multivariate Survival Data*, Statistics for Biology and Health, Springer New York, New York.

Huang, X., Yan, F., Ning, J., Feng, Z., Choi, S. and Cortes, J. (2016), 'A two-stage approach for dynamic prediction of time-to-event distributions', *Statistics in Medicine* **35**(13), 2167–2182.

Ikeda, M., Ishigaki, T. and Yamauchi, K. (2002), 'Relationship between brier score and area under the binormal roc curve', *Computer Methods and Programs in Biomedicine* **67**(3), 187 – 194.

Joly, P., Commenges, D., Helmer, C. and Letenneur, L. (2002), 'A penalized likelihood approach for an illnessdeath model with intervalcensored data: application to agespecific incidence of dementia', *Biostatistics* **3**(3), 433–443.

Joly, P., Commenges, D. and Letenneur, L. (1998), 'A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia', *Biometrics* **54**(1), 185–194.

Jørgensen, O. D., Kronborg, O. and Fenger, C. (2002), 'A randomised study of screening for colorectal cancer using faecal occult blood testing: results after 13 years and seven biennial screening rounds', *Gut* **50**(1), 29–32.

Jrvinen, H. J., Aarnio, M., Mustonen, H., AktanCollan, K., Aaltonen, L. A., Peltomki, P., De La Chapelle, A. and Mecklin, J. (2000), 'Controlled 15-year trial on screening for colorectal cancer in families with hereditary nonpolyposis colorectal cancer', *Gastroenterology* **118**(5), 829 – 834.

Katki, H. A., Cheung, L. C., Fetterman, B., Castle, P. E. and Sundaram, R. (2015), 'A joint model of persistent human papilloma virus infection and cervical cancer risk: implications for cervical cancer screening', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **178**(4), 903–923.

Kohavi, R. (1995), 'A study of cross-validation and bootstrap for accuracy estimation and model selection'.

Król, A., Mauguen, A., Mazroui, Y., Laurent, A., Michiels, S. and Rondeau, V. (2017), 'Tutorial in Joint Modeling and Prediction: a Statistical Software for Correlated Longitudinal Outcomes, Recurrent Events and a Terminal Event', *ArXiv e-prints* .

Ladabaum, U. and Song, K. (2005), ‘Projected national impact of colorectal cancer screening on clinical and economic outcomes and health services demand’, *Gastroenterology* **129**(4), 1151 – 1162.

Lange, K. (2003), *Mathematical and statistical methods for genetic analysis*, Springer Science & Business Media, New York.

Lee, S. J., Boscardin, W. J., Stijacic-Cenzer, I., Conell-Price, J., O’Brien, S. and Walter, L. C. (2013), ‘Time lag to benefit after screening for breast and colorectal cancer: meta-analysis of survival data from the united states, sweden, united kingdom, and denmark’, *BMJ* **346**.

Leroux, B. G. and Puterman, M. L. (1992), ‘Maximum-penalized-likelihood estimation for independent and markov- dependent mixture models’, *Biometrics* **48**(2), 545–558.

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 & Rectum* **41**(4), 428–433.

Lindor, N., Petersen, G., Hadley, D. and et al (2006), ‘Recommendations for the care of individuals with an inherited predisposition to lynch syndrome: A systematic review’, *JAMA* **296**(12), 1507–1517.

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

Maller, R. and Zhou, X. (2003), ‘Testing for individual heterogeneity in parametric models for event-history data’, *Mathematical Methods of Statistics* **12**(3), 276–304.

Manda, S. O. (2001), ‘A comparison of methods for analysing a nested frailty model to child survival in malawi’, *Australian and New Zealand Journal of Statistics* **43**(1), 7–16.

Mandel, J. S., Church, T. R., Ederer, F. and Bond, J. H. (1999), ‘Colorectal cancer mortality: Effectiveness of biennial screening for fecal occult blood’, *JNCI: Journal of the National Cancer Institute* **91**(5), 434–437.

Mauguen, A., Rachet, B., Mathoulin-Plissier, S., MacGrogan, G., Laurent, A. and Rondeau, V. (2013), 'Dynamic prediction of risk of death using history of cancer recurrences in joint frailty models', *Statistics in Medicine* **32**(30), 5366–5380.

McGilchrist, C. A. and Aisbett, C. W. (1991), 'Regression with frailty in survival analysis', *Biometrics* **47**(2), 461–466.

Pencina, M. J., D' Agostino, R. B., D' Agostino, R. B. and Vasan, R. S. (2008), 'Evaluating the added predictive ability of a new marker: From area under the roc curve to reclassification and beyond', *Statistics in Medicine* **27**(2), 157–172.

Plaschke, J., Engel, C., Krger, S., Holinski-Feder, E., Pagenstecher, C., Mangold, E., Moeslein, G., Schulmann, K., Gebert, J., Doeberitz, M. v. K., Rschoff, J., Loeffler, M. and Schackert, H. K. (2004), 'Lower incidence of colorectal cancer and later age of disease onset in 27 families with pathogenic msh6 germline mutations compared with families with mlh1 or msh2 mutations: The german hereditary nonpolyposis colorectal cancer consortium', *Journal of Clinical Oncology* **22**(22), 4486–4494. PMID: 15483016.

Proust-Lima, C. and Taylor, J. M. G. (2009), 'Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach', *Biostatistics* **10**(3), 535–549.

Rizopoulos, D. (2011), 'Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data', *Biometrics* **67**(3), 819–829.

Rizopoulos, D., Murawska, M., Andrinopoulou, E.-R., Molenberghs, G., Takkenberg, J. J. M. and Lesaffre, E. (2013), 'Dynamic Predictions with Time-Dependent Covariates in Survival Analysis using Joint Modeling and Landmarking', *ArXiv e-prints* .

Rondeau, V., Commenges, D. and Joly, P. (2003), 'Maximum penalized likelihood estimation in a gamma-frailty model', *Lifetime Data Analysis* **9**(2), 139–153.

Rondeau, V., Filleul, L. and Joly, P. (2006), 'Nested frailty models using maximum penalized likelihood estimation', *Statistics in Medicine* **25**(23), 4036–4052.

Rondeau, V., Gonzalez, J. R., Mazroui, Y., Mauguen, A., Krol, A., Diakite, A. and Laurent, A. (2016), *frailtypack: General Frailty models: shared, joint and nested frailty models with prediction*. R package version 2.8.3.

Rondeau, V., Gonzalez, J. R., Mazroui, Y., Mauguen, A., Krol, A., Diakite, A., Laurent, A. and Lopez, M. (2017), 'General frailty models: Shared, joint and nested frailty models with prediction'. Package 'frailtypack'. R package version 2.12.3.

Rondeau, V., Marzrou, Y. and Gonzalez, J. (2012), 'frailtypack: An r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation', *Journal of Statistical Software, Articles* **47**(4), 1–28.

Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H., Brouste, V. and Soubeyran, P. (2007), 'Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events', *Biostatistics* **8**(4), 708–721.

Rothman, K., Greenland, S. and Lash, T. (2008), *Modern Epidemiology*, 3 edn, Lippincott-Raven Publishers, Philadelphia.

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.

Segnan, N., Armaroli, P., Bonelli, L., Risio, M., Sciallero, S., Zappa, M., Andreoni, B., Arrigoni, A., Bisanti, L., Casella, C., Crosta, C., Falcini, F., Ferrero, F., Giacomin, A., Giuliani, O., Santarelli, A., Visioli, C. B., Zanetti, R., Atkin, W. S. and Senore, C. (2011), 'Once-only sigmoidoscopy in colorectal cancer screening: Follow-up findings of the italian randomized controlled trialscore', *JNCI: Journal of the National Cancer Institute* **103**(17), 1310–1322.

Smith, R. A., von Eschenbach, A. C., Wender, R., Levin, B., Byers, T., Rothenberger, D., Brooks, D., Creasman, W., Cohen, C., Runowicz, C., Saslow, D., Cokkinides, V. and Eyre, H. (2001), 'American cancer society guidelines for the early detection of cancer: Update of early detection guidelines for prostate, colorectal, and endometrial cancers: Also: Update

2001testing for early lung cancer detection', *CA: A Cancer Journal for Clinicians* **51**(1), 38–75.

Stuckless, S. N. (2012), The impact of screening on the clinical course of lynch syndrome, Submission, Memorial University of Newfoundland, http://research.library.mun.ca/id/eprint/11487.

Stupart, D. A., Goldberg, P. A., Algar, U. and Ramesar, R. (2009), 'Surveillance colonoscopy improves survival in a cohort of subjects with a single mismatch repair gene mutation', *Colorectal Disease* **11**(2), 126–130.

Thomas, D. C. (2017), 'Estimating the effect of targeted screening strategies: An application to colonoscopy and colorectal cancer', *Epidemiology (Cambridge, Mass.)* **28**(4), 470478.

van Houwelingen, H. C. (2007), 'Dynamic prediction by landmarking in event history analysis', *Scandinavian Journal of Statistics* **34**(1), 70–85.

van Houwelingen, H. C. and Putter, H. (2008), 'Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data', *Lifetime Data Analysis* **14**(4), 447.

van Houwelingen, H. and Putter, H. (2011), *Dynamic Prediction in Clinical Survival Analysis*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press, Boca Raton.

Vasen, H. F. A., Möslein, G., Alonso, A., Bernstein, I., Bertario, L., Blanco, I., Burn, J., Capella, G., Engel, C., Frayling, I., Friedl, W., Hes, F. J., Hodgson, S., Mecklin, J.-P., Møller, P., Nagengast, F., Parc, Y., Renkonen-Sinisalo, L., Sampson, J. R., Stormorken, A. and Wijnen, J. (2007), 'Guidelines for the clinical management of lynch syndrome (hereditary non-polyposis cancer)', *Journal of Medical Genetics* **44**(6), 353–362.

Vasen, H. F. A., Mslein, G., Alonso, A. and et al. (2010), 'Recommendations to improve identification of hereditary and familial colorectal cancer in europe', *Familial Cancer* **9**(2), 109–115.

Vasen, H. F., Abdirahman, M., Brohet, R. and et al. (2010), 'One to 2-year surveillance intervals reduce risk of colorectal cancer in families with lynch syndrome', *Gastroenterology* **138**(7), 2300 – 2306.

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.

Wienke, A. (2014), *Frailty Models*, John Wiley and Sons, Ltd, Rostock.

Appendix A

Model results for different mutation types

Table A.1: Comparisons of three models–joint nested frailty, joint frailty, and shared frailty models–for analyzing MLH1 families

Visit Process

Variables	Joint nested frailty			Joint frailty		
	estimate	se	p-value	estimate	se	p-value
age at the first visit	-0.589	0.119	<0.001	-0.577	0.126	<0.001
age at the previous visit	0.662	0.097	<0.001	0.657	0.098	<0.001
detection of adenoma at the previous visit	0.231	0.189	0.221	0.249	0.190	0.190
detection of serrated at the previous visit	0.933	0.439	0.034	0.942	0.44	0.032

Terminal event (CRC)

Variables	Joint nested frailty			Joint frailty			Shared frailty		
	estimate	se	p-value	estimate	se	p-value	estimate	se	p-value
age at the first visit	0.217	0.171	0.205	0.175	0.138	0.207	0.136	0.202	0.500
detection of adenoma	-1.003	0.459	0.029	-0.839	0.396	0.034	-0.220	0.523	0.675
detection of others polyps	–	–	–	–	–	–	0.527	0.643	0.413
cumulative number of visits [†]							0.401	0.125	0.001
proband's CRC age	0.086	0.157	0.582	0.087	0.110	0.427	-0.046	0.207	0.823
gender	0.770	0.383	0.045	0.637	0.330	0.054	0.327	0.415	0.431

[†] time-dependent covariate

Frailty Parameters

	Joint nested frailty			Joint frailty			Shared frailty		
	estimate	se	p-value	estimate	se	p-value	estimate	se	p-value
θ	0.737	0.112	<0.001	0.740	0.112	<0.001	0.339	0.144	0.009
α	1.310	0.228	<0.001	0.825	0.368	0.025			
η	0.663	0.370	0.037						
ξ	–	–	–						
penalized marginal log-likelihood	-834.160			-916.970			-126.940		
LCV	1.913			2.096			0.314		

θ denotes the variation of individual frailty,

α denotes the association between two processes at the individual level,

η denotes the variation of familial frailty,

ξ denotes the association between two processes at the familial level.

Table A.2: Comparisons of three models–joint nested frailty, joint frailty, and shared frailty models–for analyzing MSH2 families

Visit Process

Variables	Joint nested frailty			Joint frailty		
	estimate	se	p-value	estimate	se	p-value
age at the first visit	-0.723	0.101	<0.001	-0.719	0.106	<0.001
age at the previous visit	0.676	0.079	<0.001	0.673	0.080	<0.001
detection of adenoma at the previous visit	0.387	0.129	0.003	0.388	0.13	0.003
detection of serrated at the previous visit	0.819	0.418	0.050	0.819	0.418	0.050

Terminal event (CRC)

Variables	Joint nested frailty			Joint frailty			Shared frailty		
	estimate	se	p-value	estimate	se	p-value	estimate	se	p-value
age at the first visit	0.411	0.148	0.005	0.376	0.132	0.004	0.556	0.171	0.001
detetction of adenoma	-0.743	0.402	0.064	-0.565	0.362	0.119	-0.424	0.480	0.377
detetction of other polyps	–	–	–	–	–	–	0.385	0.442	0.383
cumulatove number of visits †							0.132	0.117	0.260
proband's CRC age	-0.028	0.165	0.866	-0.038	0.131	0.770	-0.054	0.248	0.828
gender	0.333	0.374	0.373	0.367	0.336	0.275	0.24	0.434	0.580

† time-dependent covariate

Frailty Parameters

	Joint nested frailty			Joint frailty			Shared frailty		
	estimate	se	p-value	estimate	se	p-value	estimate	se	p-value
θ	0.636	0.087	<0.001	0.648	0.090	<0.001	0.518	0.262	0.024
α	0.626	0.146	<0.001	0.299	0.356	0.402			
η	0.781	0.258	0.001						
ξ	–	–	–						
penalized marginal log-likelihood	-1135.930			-1234.710			-136.580		
LCV	1.675			1.815			0.217		

θ denotes the variation of individual frailty,

α denotes the association between two processes at the individual level,

η denotes the variation of familial frailty,

ξ denotes the association between two processes at the familial level.

Table A.3: Comparisons of three models–joint nested frailty, joint frailty, and shared frailty models–for analyzing MSH6 families

Visit Process

Variables	Joint nested frailty			Joint frailty		
	estimate	se	p-value	estimate	se	p-value
age at the first visit	-0.331	0.249	0.181	-0.318	0.249	0.201
age at the previous visit	0.483	0.200	0.016	0.472	0.198	0.017
detection of adenoma at the previous visit	0.490	0.253	0.053	0.506	0.250	0.043

Terminal event (CRC)

Variables	Joint nested frailty			Joint frailty			Shared frailty		
	estimate	se	p-value	estimate	se	p-value	estimate	se	p-value
age at the first visit	1.249	0.696	0.073	1.316	0.701	0.061	2.225	1.336	0.096
detetction of adenoma	–	–	–	–	–	–	-1.896	1.783	0.288
detetction of other polyps	–	–	–	–	–	–	-2.337	3.780	0.536
cumulative number of visits[†]							0.019	0.404	0.962
proband's CRC age	–	–	–	–	–	–	0.022	1.404	0.987
gender	3.615	1.191	0.002	3.647	1.308	0.005	5.360	2.049	0.009

[†] time-dependent covariate

Frailty Parameters

	Joint nested frailty			Joint frailty			Shared frailty		
	estimate	se	p-value	estimate	se	p-value	estimate	se	p-value
θ	0.351	0.118	0.001	0.344	0.115	0.001	3.085	2.551	0.113
α	1.483	0.246	<0.001	2.052	1.204	0.088			
η	0.787	0.436	0.035						
ξ	–	–	–						
penalized marginal log-likelihood	-217.630			-257.700			-19.500 (marginal log-likelihood)		
LCV	1.674			1.947			0.1990 (AIC)		

θ denotes the variation of individual frailty,

α denotes the association between two processes at the individual level,

η denotes the variation of familial frailty,

ξ denotes the association between two processes at the familial level.

Table A.4: MLH1 families: Shared frailty model with Weibull baseline hazard using death from CRC as the terminal event with cumulative number of visit after CRC as a time-dependent covariate

MLH1

Variables	estimate	se	p-value
detection of adenoma before CRC	-4.425	2.160	0.041
age at CRC	0.086	0.019	<0.001
average gap time between visits before CRC	0.067	0.024	0.006
cumulative number of visit after CRC [†]	-0.109	0.130	0.404
cancer stage	2.098	0.770	0.006
proportion of removed colon	-2.180	1.926	0.258
gender	0.186	0.729	0.799

[†] time-dependent covariate

Frailty parameters

	estimate	se	p-value
θ	<0.001	<0.001	0.315
marginal log-likelihood	-52.750		
AIC	0.292		

Table A.5: MSH2 families: Shared frailty model with Weibull baseline hazard using death from CRC as the terminal event with cumulative number of visit after CRC as a time-dependent covariate

MSH2

Variables	estimate	se	p-value
detection of adenoma before CRC	-13.006	14.861	0.382
age at CRC	0.018	0.032	0.575
average gap time between visits before CRC	0.198	0.044	<0.001
cumulative number of visit after CRC †	0.289	0.100	0.004
cancer stage	-1.442	0.882	0.102
proportion of removed colon	-3.197	1.037	0.002
gender	3.757	0.711	<0.001

† time-dependent covariate

Frailty parameters

	estimate	se	p-value
θ	<0.001	<0.001	0.500
marginal log-likelihood	-50.730		
AIC	0.408		

Table A.6: MSH6 families: Shared frailty model with Weibull baseline hazard using death from CRC as the terminal event with cumulative number of visit after CRC as a time-dependent covariate

MSH6

Variables	estimate	se	p-value
age at the first visit	3.357	1.134	0.003
detection of adenoma before CRC	-2.969	1.764	0.092
cumulative number of visit after CRC †	-0.469	0.413	0.256
cancer stage	-1.127	0.756	0.136
proportion of removed colon	5.082	4.543	0.263
gender	0.510	2.324	0.826

† time-dependent covariate

Frailty parameters

	estimate	se	p-value
θ	<0.001	<0.001	0.500
marginal log-likelihood	-21.370		
AIC	1.320		

Table A.7: Shared frailty models for MLH1 families with Weibull baseline hazard: cancer stages as the terminal events with average screening gap time and other polyp

MLH1	**Low stage cancer**			**High stage cancer**		
Variables	estimate	se	p-value	estimate	se	p-value
age at the first visit	0.111	0.312	0.721	0.063	0.170	0.710
detection of adenoma	-0.114	0.724	0.875	-0.666	0.578	0.249
detection of other polyps	0.287	0.929	0.757	3.563	0.916	<0.001
average gap time between visits	-0.061	0.070	0.385	-0.007	0.030	0.807
proband's CRC age	0.384	0.299	0.199	-0.218	0.183	0.234
gender	0.974	0.812	0.230	-0.154	0.458	0.737
interaction between other polyps and average gap time	-	-	-	-1.394	0.540	0.010

Frailty parameters						
	estimate	se	p-value	estimate	se	p-value
θ	3.490	2.582	0.088	<0.001	<0.001	0.500
marginal log-likelihood	-72.650			-70.570		
AIC	0.675			0.655		

Table A.8: Shared frailty models for MSH2 families with Weibull baseline hazard: cancer stages as the terminal events with average screening gap time and other polyp

MSH2	**Low stage cancer**			**High stage cancer**		
Variables	estimate	se	p-value	estimate	se	p-value
age at the first visit	1.353	0.637	0.034	0.031	0.217	0.888
detection of adenoma	0.432	0.712	0.544	-1.252	0.769	0.104
detection of other polyps	0.700	0.780	0.370	0.136	0.594	0.819
average gap time between visits	-0.423	0.250	0.091	-0.003	0.041	0.939
proband's CRC age	-0.397	0.481	0.409	0.193	0.217	0.376
gender	0.329	0.869	0.705	0.596	0.506	0.239

Frailty parameters						
	estimate	se	p-value	estimate	se	p-value
θ	8.636	6.019	0.076	<0.001	<0.001	0.500
marginal log-likelihood	-82.770			-65.350		
AIC	0.527			0.443		

Table A.9: Shared frailty models for MSH6 families with Weibull baseline hazard: cancer stages as the terminal events with average screening gap time and other polyp

MSH6	**Low stage cancer**			**High stage cancer**		
Variables	estimate	se	p-value	estimate	se	p-value
age at the first visit	0.617	1.450	0.670	1.975	0.847	0.020
detection of adenoma	-	-	-	-4.352	1.913	0.023
detection of other polyps	-	-	-	2.375	1.967	0.227
average gap time between visits	-1.754	1.430	0.220	-1.704	0.909	0.061
proband's CRC age	-	-	-	0.105	0.564	0.852
gender	-	-	-	2.589	1.558	0.097

Frailty parameters						
	estimate	se	p-value	estimate	se	p-value
θ	0.250	7.012	0.486	<0.001	<0.001	0.500
marginal log-likelihood	-6.340			-7.250		
AIC	0.214			0.285		

Appendix B

Model results considering mutation types as covariates

Table B.1: Comparisons of joint nested frailty and joint frailty models, with mutation type as covariates (indicators), for analyzing LS family data

Variables	**Joint nested frailty model**			**Joint frailty model**		
	estimate	se	p-value	estimate	se	p-value
Visiting Process						
age at the first visit	-0.652	0.077	<0.001	-0.647	0.079	<0.001
age at the previous visit	0.658	0.059	<0.001	0.656	0.060	<0.001
detection of adenoma at the previous visit	0.359	0.097	<0.001	0.362	0.097	<0.001
detection of serrated polyp at the previous visit	0.800	0.277	0.004	0.805	0.277	0.004
MLH1 compared to MSH2	-0.300	0.116	0.010	-0.305	0.118	0.010
MSH6 compared to MSH2	-0.091	0.171	0.597	-0.100	0.171	0.559
PMS2 compared to MSH2	-0.863	0.335	0.010	-0.877	0.337	0.009
Disease Process						
age at the first visit	0.336	0.099	0.001	0.297	0.084	<0.001
detection of adenoma	-0.904	0.273	0.001	-0.763	0.244	0.002
MLH1 compared to MSH2	0.195	0.289	0.502	0.154	0.224	0.491
MSH6 compared to MSH2	-0.242	0.465	0.602	-0.224	0.399	0.574
PMS2 compared to MSH2	0.971	0.541	0.073	0.821	0.418	0.049
gender	0.721	0.241	0.003	0.636	0.210	0.002
Frailty Parameters						
θ	0.645	0.061	<0.001	0.648	0.062	<0.001
α	0.839	0.113	<0.001	0.572	0.235	0.015
η	0.709	0.191	<0.001	-	-	-
ξ	-	-	-	-	-	-
penalized marginal log-likelihood	-2278.170			-2515.930		
LCV	1.752			1.932		

Appendix C

Additional plots – prediction plots, survival plots and cumulative hazard plots

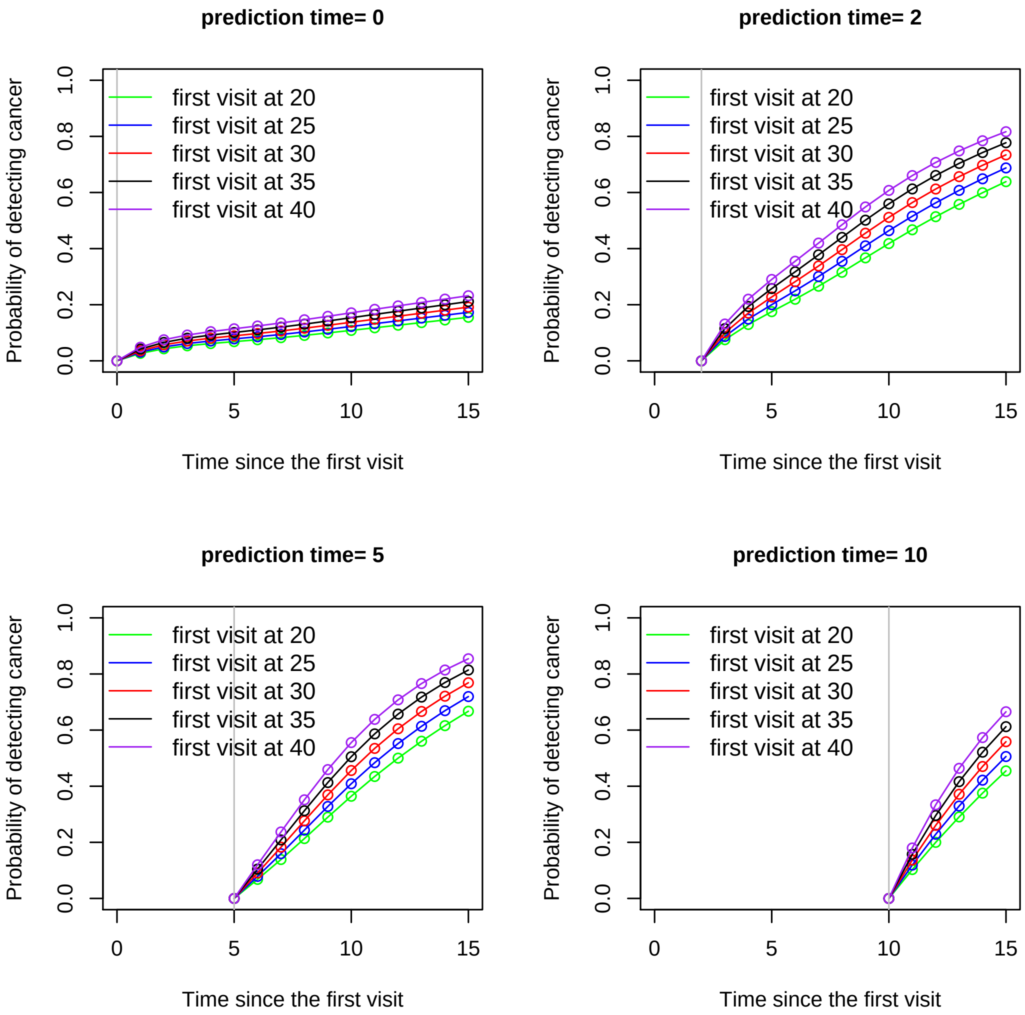

Figure C.1: Comparison of the effects of the first visit ages on dynamic prediction of CRC, $P(t,s)$, at fixed prediction time $t = 0,2,5$ and 10 years with s increasing by one until $t+s = 15$ years, for comparing different ages at the first visit. The grey vertical line represents the time of prediction. A virtual individual was defined as a male who had been screened every year, was detected of adenoma and presented CRC during the follow-up time. Situations of this person first visited at age of 20, 25, 30, 35 and 40 were compared.

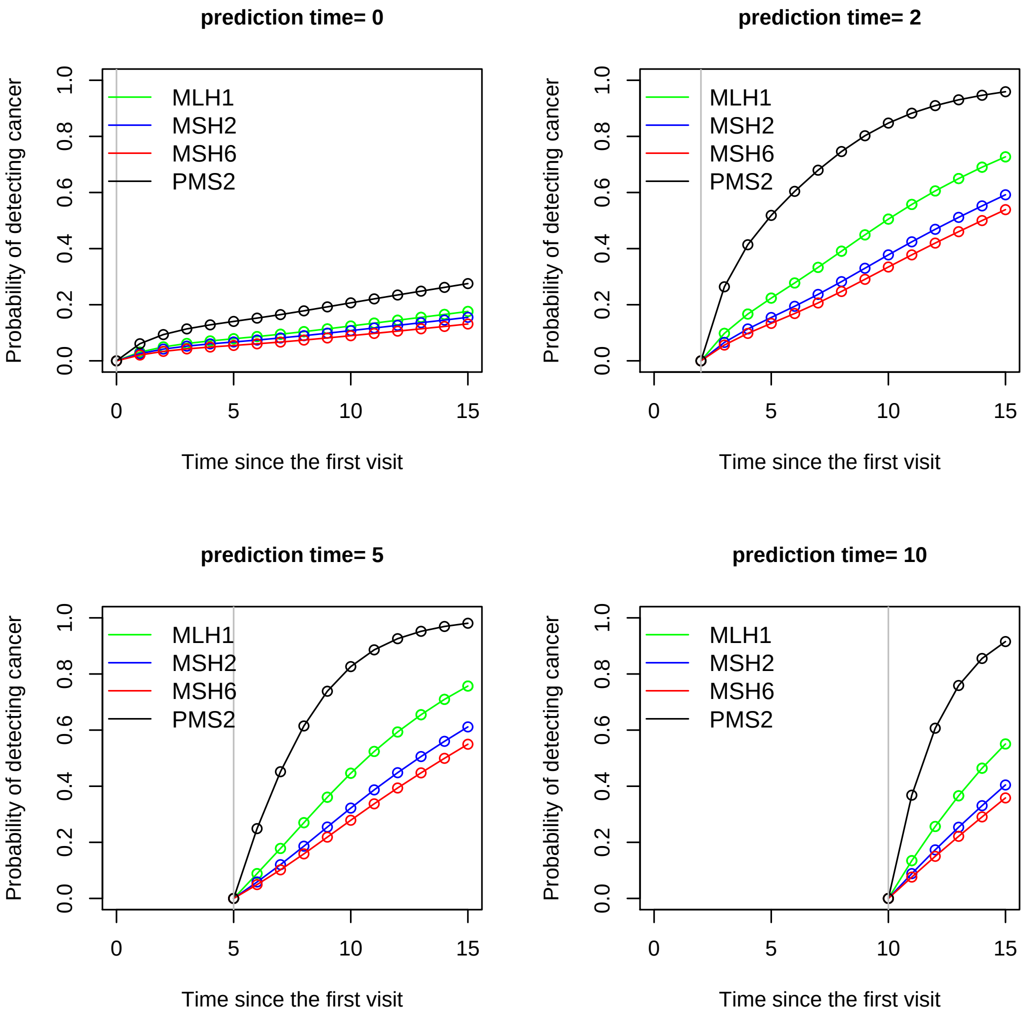

Figure C.2: One-year visiting interval: comparison of the effects mutation type on dynamic prediction of CRC, $P(t, t+s)$, at fixed prediction time t = 0,2,5 and 10 years with s increasing by one until $t+s$ =15 years. The grey vertical line represents the time of prediction. A virtual individual was defined as a male who started his first screening visit at age 25, had been screened every year, was detected of adenoma, and presented CRC during the follow-up time. Four mutation types, MLH1, MSH2, MSH6 and PMS2 were assigned to this individual for comparisons.

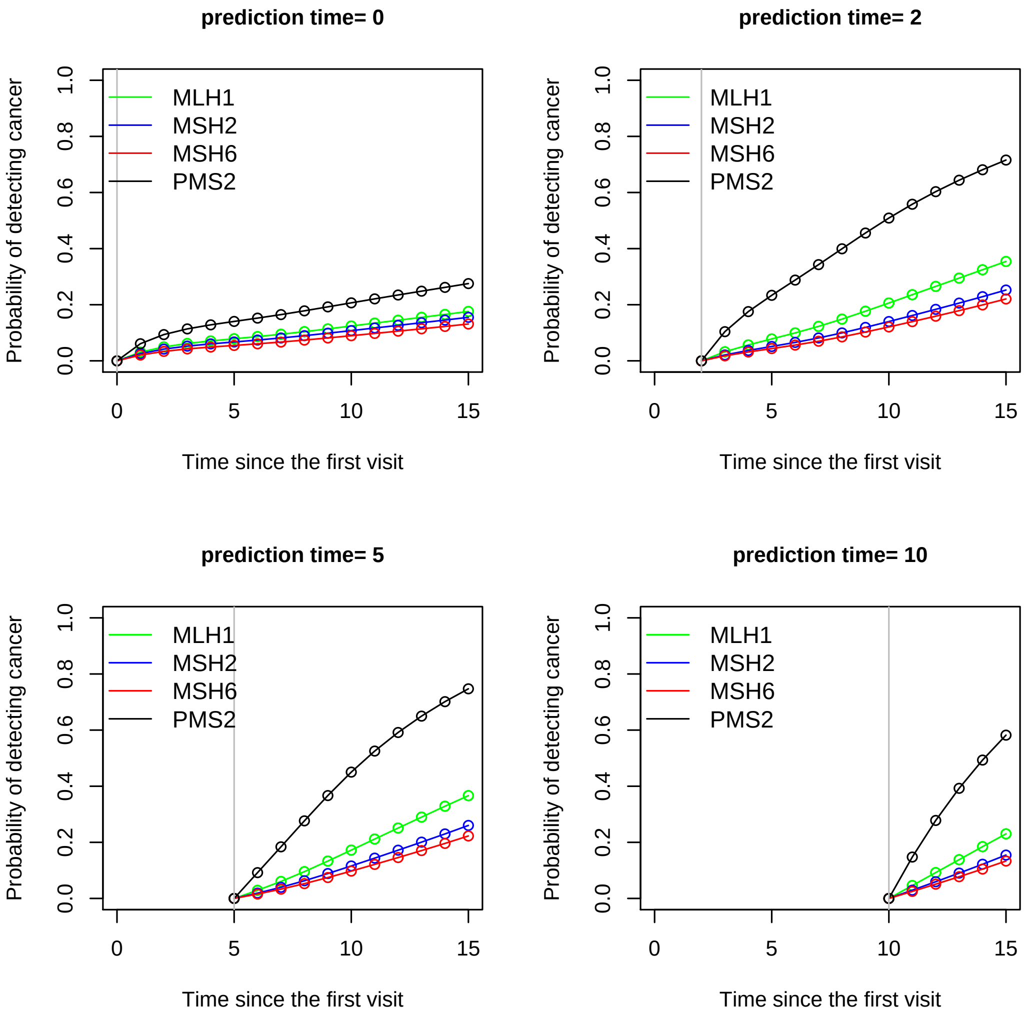

Figure C.3: Two-year visiting interval: comparison of the effects mutation type on dynamic prediction of CRC, $P(t,t+s)$, at fixed prediction time $t = 0{,}2{,}5$ and 10 years with s increasing by one until $t+s$ =15 years. The grey vertical line represents the time of prediction. A virtual individual was defined as a male who started his first screening visit at age 25, had been screened every two years, was detected of adenoma, and presented CRC during the follow-up time. Four mutation types, MLH1, MSH2, MSH6 and PMS2 were assigned to this individual for comparisons.

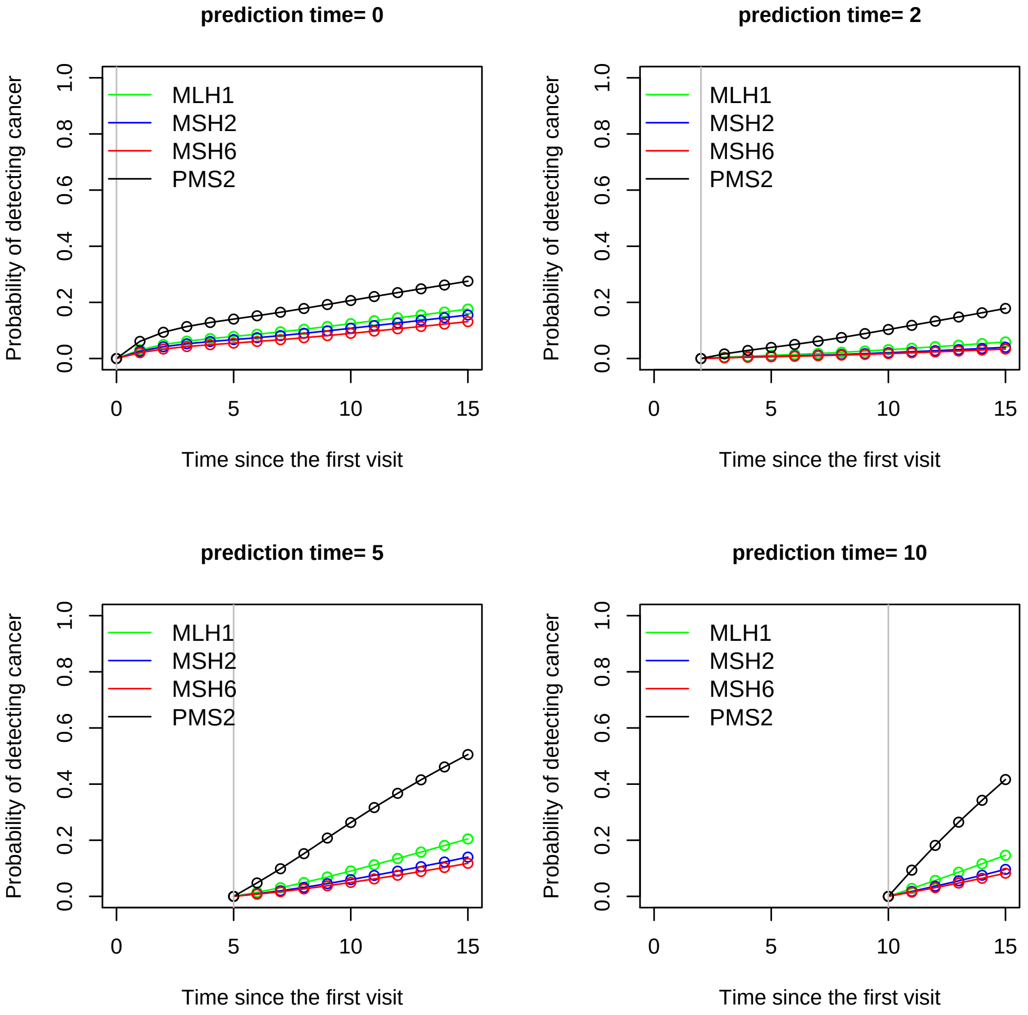

Figure C.4: Three-year visiting interval: comparison of the effects mutation type on dynamic prediction of CRC, $P(t,t+s)$, at fixed prediction time $t=$ 0,2,5 and 10 years with s increasing by one until $t+s$ =15 years. The grey vertical line represents the time of prediction. A virtual individual was defined as a male who started his first screening visit at age 25, had been screened every three years, was detected of adenoma, and presented CRC during the follow-up time. Four mutation types, MLH1, MSH2, MSH6 and PMS2 were assigned to this individual for comparisons.

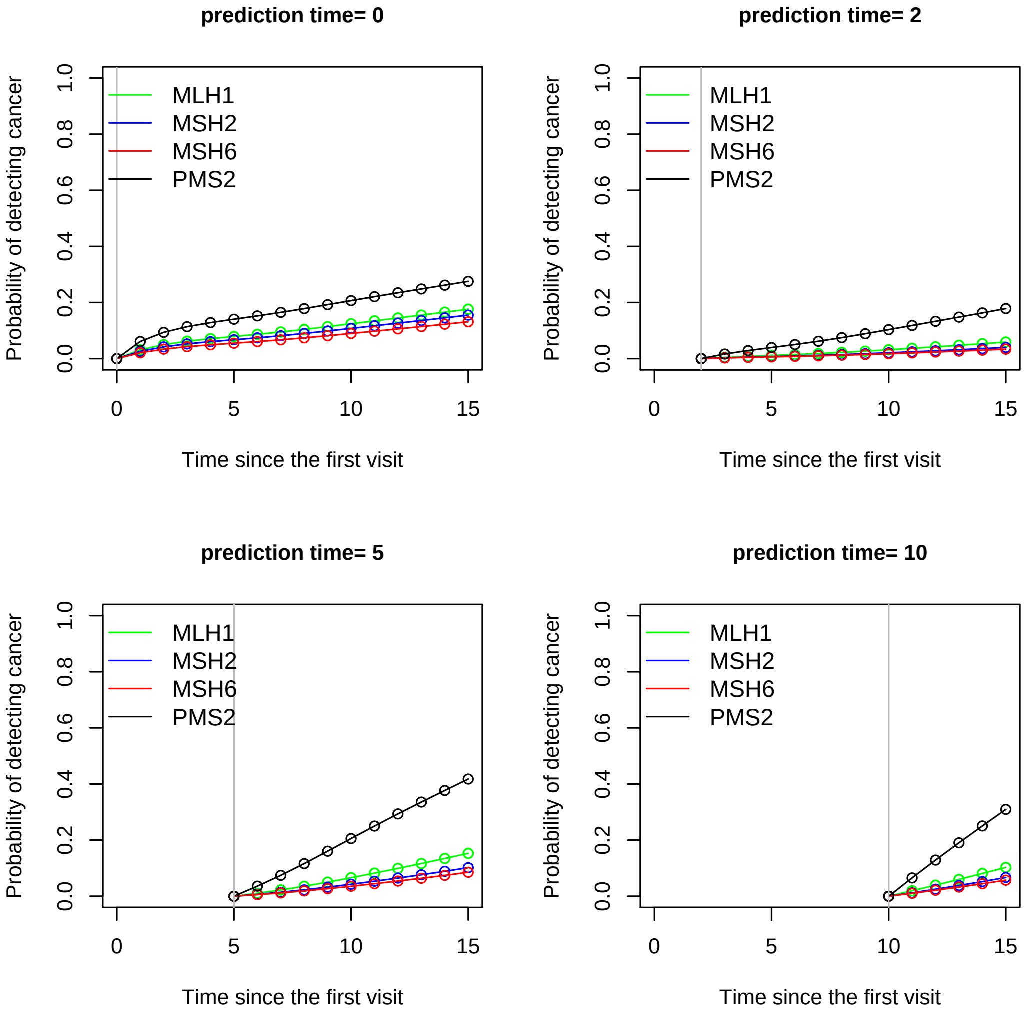

Figure C.5: Four-year visiting interval: comparison of the effects mutation type on dynamic prediction of CRC, $P(t,t+s)$, at fixed prediction time $t = 0,2,5$ and 10 years with s increasing by one until $t+s$ =15 years. The grey vertical line represents the time of prediction. A virtual individual was defined as a male who started his first screening visit at age 25, had been screened every four years, was detected of adenoma, and presented CRC during the follow-up time. Four mutation types, MLH1, MSH2, MSH6 and PMS2 were assigned to this individual for comparisons.

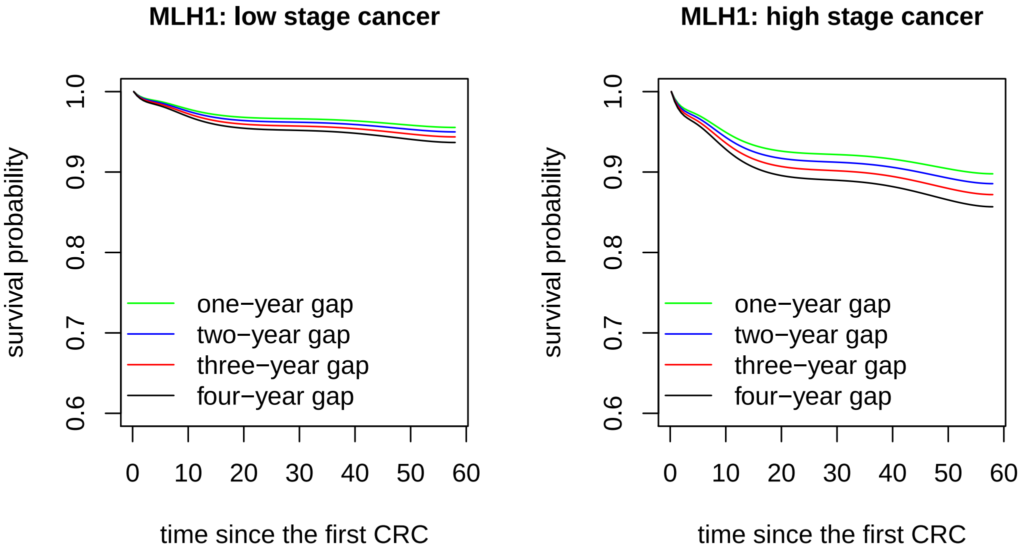

Figure C.6: MLH1: survival probabilities after the first CRC, specific to cancer stage, screening gap times. Gap times are assigned as one to four years. Under each cancer stage, four gap times are compared. The x-axis is the time in years since the detection of the first CRC. The y-axis is the survival probability from CRC death.

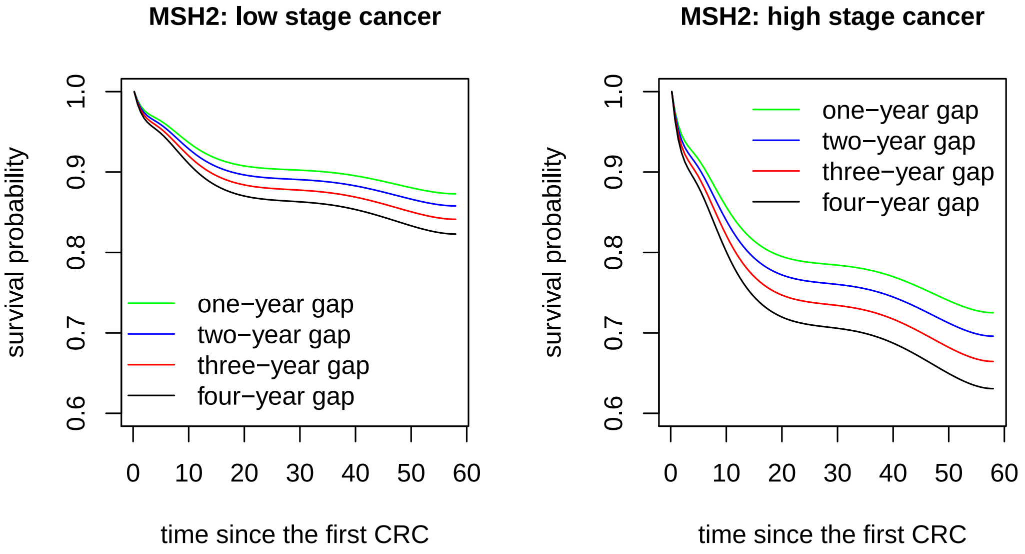

Figure C.7: MSH2: Survival probabilities after the first CRC, specific to cancer stage, screening gap times. Gap times are assigned as one to four years. Under each cancer stage, four gap times are compared. The x-axis is the time in years since the detection of the first CRC. The y-axis is the survival probability from CRC death.

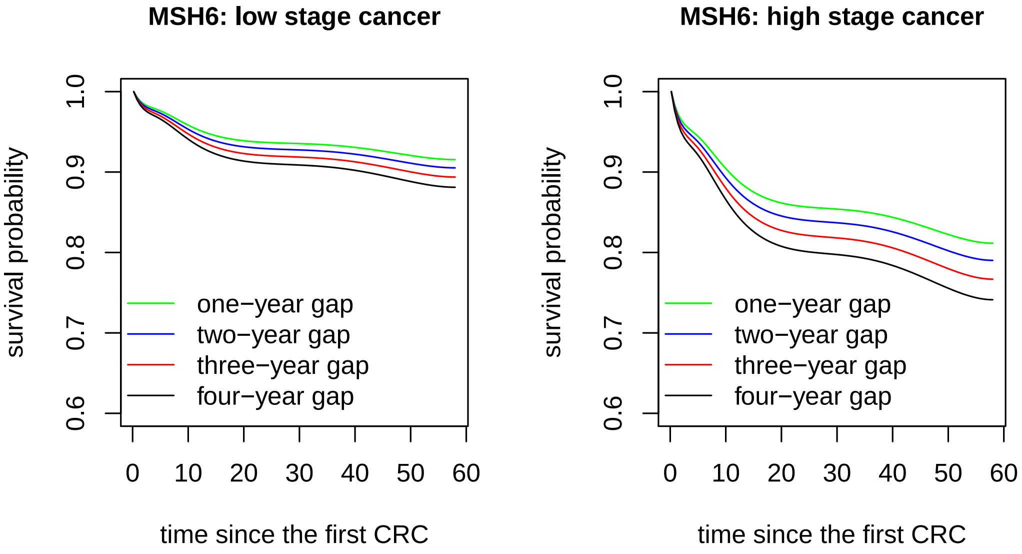

Figure C.8: MSH6: survival probabilities after the first CRC, specific to cancer stage, screening gap times. Gap times are assigned as one to four years. Under each cancer stage, four gap times are compared. The x-axis is the time in years since the detection of the first CRC. The y-axis is the survival probability from CRC death.

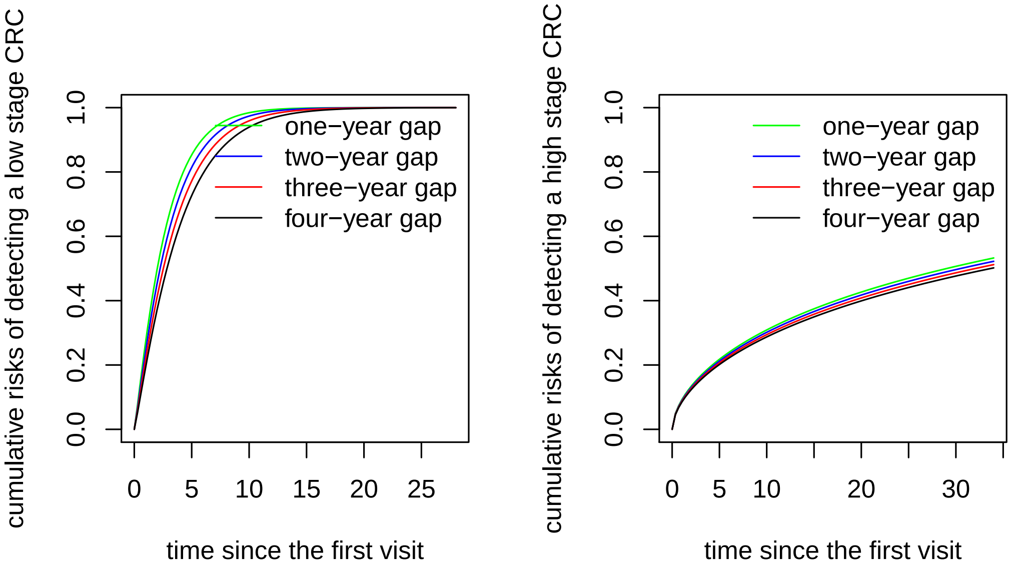

Figure C.9: MLH1: cumulative hazard for low stage cancer (left) and high stage cancer (right), specific to screening gap times. The x-axis is the time in years since the first visit. The y-axis is the cumulative hazard of developing cancers in two stages.

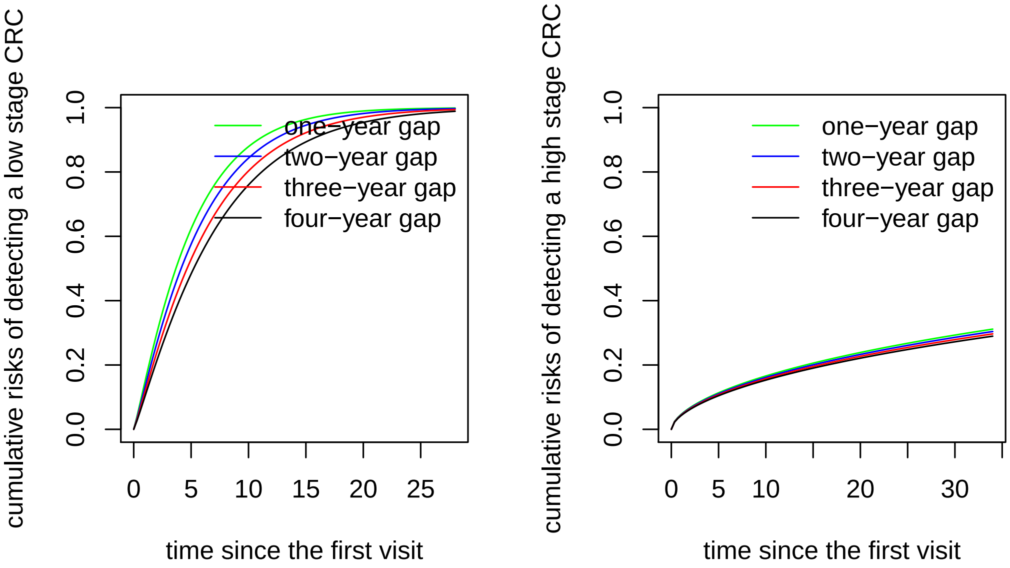

Figure C.10: MSH2: cumulative hazard for low stage cancer (left) and high stage cancer (right), specific to screening gap times. The x-axis is the time in years since the first visit. The y-axis is the cumulative hazard of developing cancers in two stages.

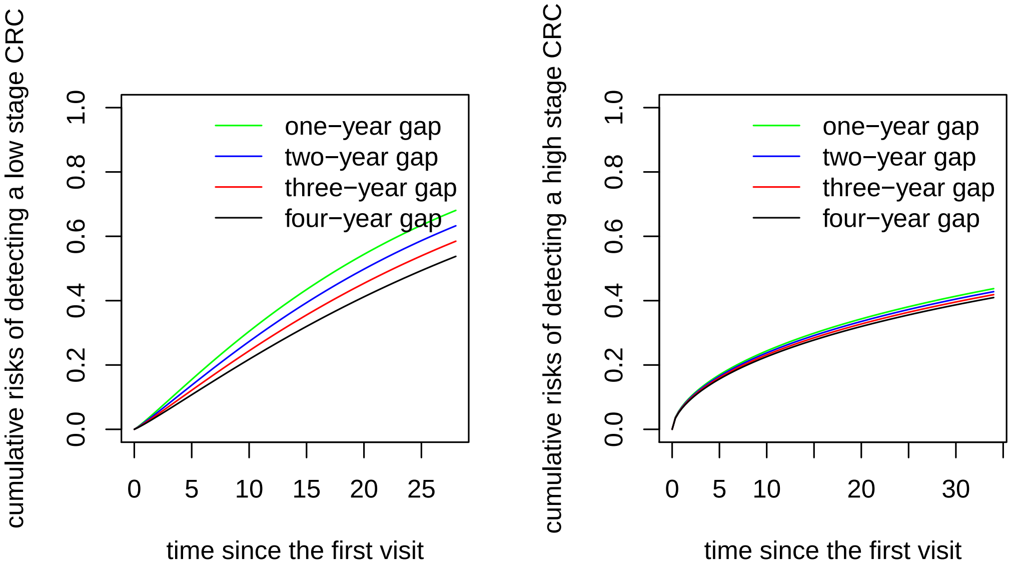

Figure C.11: MSH6: cumulative hazard for low stage cancer (left) and high stage cancer (right), specific to screening gap times. The x-axis is the time in years since the first visit. The y-axis is the cumulative hazard of developing cancers in two stages.

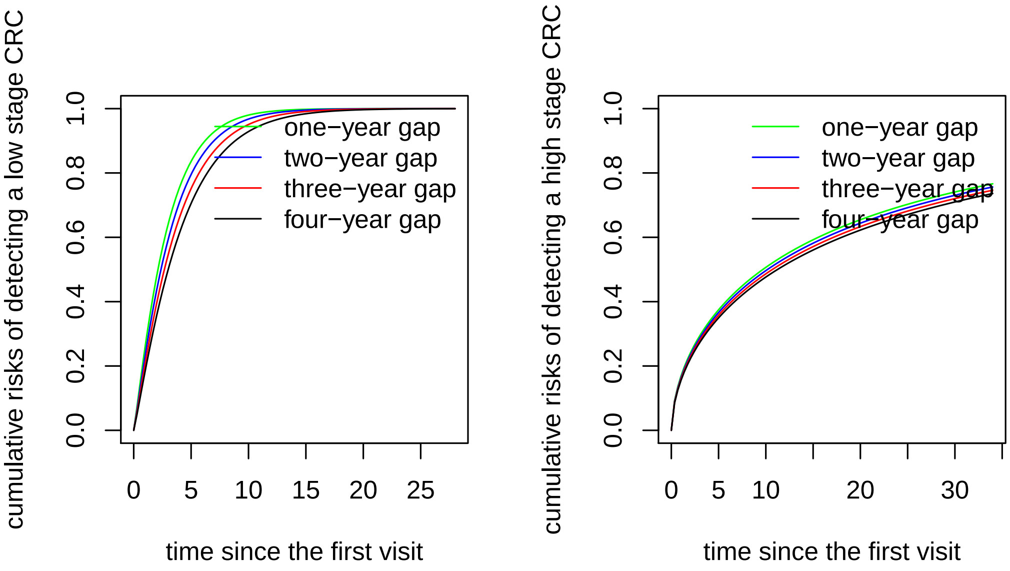

Figure C.12: PMS2: cumulative hazard for low stage cancer (left) and high stage cancer (right), specific to screening gap times. The x-axis is the time in years since the first visit. The y-axis is the cumulative hazard of developing cancers in two stages.

Curriculum Vitae

Name:	Bing(Angie) Yu
Post-Secondary Education and Degrees:	Southwestern Univerisity of Finance and Economics China 2011 - 2015 B.Sc. – Economics
	University of Western Ontario London, ON 2015 - 2017 M.Sc.
Honours and Awards:	Schulich Graduate Scholarship 2015 - 2017
Related Work Experience:	Research Assistant, University of Western Ontario, 2015 - 2017 Summer Research Student, Lunenfeld-Tanenbaum Research Institute, 2016 Teaching Assistant for Sampling Methods, University of Western Ontario, 2016

Poster Presentation:

Bing Yu (2016). Comparison of Risks of Second Colorectal Cancers Between Lynch Syndrome Families and Familial Colorectal Cancer Type X Families. London Health Research Day.