# Impact of Competing Risks on the Risk Estimation of Multiple Cancers in Family Studies

Yun-Hee Choi* Laurent Briollais†

**Abstract**

In cancer family studies, individuals are often at risk of multiple cancers. Estimating the risk of successive and competing cancer events can help disentangling the familial risk and thus provide relevant clinical insights. The goal of our study is to investigate the impact of competing cancer events on the risk estimation of successive cancers in a family setting. We propose a statistical framework based on a progressive multistate model where competing events are taken into account. We also deal with the complex ascertainment of the families. We illustrate the interest of our approach through an analysis of Lynch Syndrome families identified by the Colon Cancer Family Registries, a NIH-funded initiative. These families harbour a mutation in the mismatched repair genes, whose members are at high risk of developing multiple cancers: colorectal, endometrial, ovarian, stomach etc. We estimated both relative and absolute risks of developing a first and second colorectal cancer with or without considering competing risk events. Our results showed that cancer risks can be overestimated if competing risks are ignored.

**Key Words:** Progressive multistate model, family data, successive event times, competing risks, ascertainment-corrected likelihood.

## 1. Introduction

In cancer family studies, family members are at high risk of developing successive and competing cancer events. For example, Lynch syndrome (LS) is an inherited condition that predisposes family members to elevated risks of colorectal cancer and other related cancers such as endometrial, ovarian, stomach, brain, skin etc. Especially, mutations in mismatched repair genes such as MLH1, MSH2, or MSH6 are known to increase the risk of developing various LS-related cancers within families and within individuals. The purpose of this work is to estimate the risks (absolute and relative) of developing sequential cancers associated with mutated genes in LS families while taking into account competing cancer events or not.

## 2. Methods

Progressive multistate models are a useful approach for modelling successive events experienced by an individual. Recently, Choi et al. (2014) developed a three-state progressive model for estimating successive cancer risks in family studies. A Markov model was proposed for modelling the risk of second cancers conditioning on the time to first cancer, which was used as a time-dependent covariate. However, this model assumed no competing events which could bias the cancer risk estimation (Katki et al., 2008).

In this paper, we incorporate competing cancer events into the progressive multistate model to provide accurate cancer-specific risk estimates associated with mutated genes in LS families. We first develop a joint model for successive events based on copulas (Nelsen, 2006) and then incorporate competing risks (Beyersmann et al., 2012) into this model. A

*Department of Epidemiology and Biostatistics, Western University, London, ON, Canada

†Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital; Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

retrospective likelihood approach (Carayol and Bonaïti-Pellié, 2004; Kraft and Thomas, 2000) is employed to deal with the non-random ascertainment of the families.

### 2.1 A copula model for successive event times

Suppose that the sequential event times arise from the following three-state progressive model. We let $t_1$ and $t_2$ denote the times spent in the 'Healthy' and 'Event 1' states,


**Figure 1**: Three-state progressive model

respectively, where $t_1$ represents the time to Event 1 and $t_2$ represents the time to Event 2 after Event 1—the gap time between the two events.

The proportional hazards models are assumed for $t_j$, $j = 1, 2$,

$$\lambda_j(t_j|X_j) = \lambda_{0j}(t_j)\exp\{\beta_j^\top X_j\},$$

where $\lambda_{0j}(t_j), j = 1, 2$, are the baseline hazard functions at $t_j$, and $X_j$ are the vectors of the risk factors associated with $t_j$. Denoting $\Lambda_{0j}(t_j)$, $j = 1, 2$, as the cumulative baseline hazard functions for $t_j$, the marginal survival functions for $t_j$ are given by

$$S_j(t_j|X_j) = e^{-\Lambda_{0j}(t_j|X_j)\exp\{\beta_j^\top X_j\}}.$$

To model the dependence between the two event times, we derive the joint survivor function for $t_1$ and $t_2$ using survivor copula $C$ to link the two marginal survivor functions,

$$S(t_1, t_2) = C\{S_1(t_1), S_2(t_2)\}.$$

As an important class of coupling functions, Archimedean copula (Nelsen, 2006) is considered, which takes the form

$$C(x, y) = \phi^{-1}\{\phi(x) + \phi(y)\},$$

where $\phi$ is a copula generating function from $[0, 1]$ to $[0, \infty]$ such that $\phi$ is a convex decreasing function with $\phi(1) = 0$. In particular, $\phi(t) = t^\theta - 1$ corresponds to the Clayton copula $C(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}, \theta > 0$.

Thus, the bivariate survivor function given covariates $X_1$ and $X_2$ can be expressed using Clayton copula (Clayton, 1978):

$$\begin{aligned} S(t_1, t_2|X_1, X_2) &= \phi^{-1}\{\phi(S_1(t_1|X_1)) + \phi(S_2(t_2|X_2))\} \\ &= \{S_1(t_1|X_1)^{-\theta} + S_2(t_2|X_2)^{-\theta} - 1\}^{-1/\theta} \\ &= \{e^{\theta\Lambda_{01}(t_1|X_1)} + e^{\theta\Lambda_{02}(t_2|X_2)} - 1\}^{-1/\theta}. \end{aligned}$$

The three possible transitions for a subject are: (1) staying in Healthy (experiencing no events), (2) experiencing Event 1 but no Event 2, and (3) experiencing both Event 1 and Event 2. The likelihood contributions for the three cases are respectively expressed as the marginal survival function for $t_1$ and the first and second derivatives for the joint survivor

function:

$$\begin{aligned} S_1(t_1|X_1) &= S(t_1,0|X_1,X_2) \\ -\frac{\partial}{\partial t_1}S(t_1,t_2|X_1,X_2) &= \{S_1(t_1|X_1)^{-\theta}+S_2(t_2|X_2)^{-\theta}-1\}^{-(\theta+1)/\theta}S_1(t_1|X_1)^{-\theta}\lambda_1(t_1) \\ \frac{\partial^2}{\partial t_1\partial t_2}S(t_1,t_2|X_1,X_2) &= (\theta+1)\{S_1(t_1|X_1)^{-\theta}+S_2(t_2|X_2)^{-\theta}-1\}^{-(2\theta+1)/\theta}\times \\ & \quad S_1(t_1|X_1)^{-\theta}\lambda_1(t_1|X_1)S_2(t_2|X_2)^{-\theta}\lambda_2(t_2|X_2). \end{aligned}$$

Using Weibull baseline hazard functions, $\lambda_{0j}(t_j)=\nu_j\rho_j t_j^{\rho_j}$, for $j=1,2$, and the corresponding marginal survivor functions $S_j(t_j|X_j)=\exp\{-\nu_j t_j^{\rho_j}e^{\beta_j^\top X_j}\}$, the likelihood function contribution for each individual with censoring indicators $\delta_1$ and $\delta_2$ for $t_1$ and $t_2$ is

$$\begin{aligned} L(\theta) &= (\theta+1)^{\delta_2}\{S_1(t_1|X_1)^{-\theta}+S_2(t_2|X_2)^{-\theta}-1\}^{-(\delta_1+\delta_2+\delta_1/\theta)}\times \\ & \quad S_1(t_1|X_1)^{-(\delta_1\theta+\delta_1-1)}\lambda_1(t_1|X_1)^{\delta_1}S_2(t_2|X_2)^{-\delta_2\theta}\lambda_2(t_2|X_2)^{\delta_2} \\ &= (\theta+1)^{\delta_2}\{e^{\theta\nu_1 t_1^{\rho_1}e^{\beta_1^\top X_1}}+e^{\theta\nu_2 t_2^{\rho_2}e^{\beta_2^\top X_2}}-1\}^{-(\delta_1+\delta_2+\delta_1/\theta)}\times \\ & \quad e^{(\delta_1\theta+\delta_1-1)\nu_1 t_1^{\rho_1}e^{\beta_1^\top X_1}}(\nu_1\rho_1 t_1^{\rho_1}e^{\beta_1^\top X_1})^{\delta_1}e^{\delta_2\theta\nu_2 t_2^{\rho_2}e^{\beta_2^\top X_2}}(\nu_2\rho_2 t_2^{\rho_2}e^{\beta_2^\top X_2})^{\delta_2}. \end{aligned}$$

The corresponding log-likelihood function can be also expressed as

$$\begin{aligned} \ell(\theta) &= \delta_2\log(\theta+1)-(\delta_1+\delta_2+\delta_1/\theta)\log\{e^{\theta\nu_1 t_1^{\rho_1}e^{\beta_1^\top X_1}}+e^{\theta\nu_2 t_2^{\rho_2}e^{\beta_2^\top X_2}}-1\}+ \\ & \quad (\delta_1\theta+\delta_1-1)\nu_1 t_1^{\rho_1}e^{\beta_1^\top X_1}+\delta_1\log(\nu_1\rho_1)+\delta_1\rho_1\log(t_1)+\delta_1\beta_1^\top X_1+ \\ & \quad \delta_2\theta\nu_2 t_2^{\rho_2}e^{\beta_2^\top X_2}+\delta_2\log(\nu_2\rho_2)+\delta_2\rho_2\log(t_2)+\delta_2\beta_2^\top X_2, \end{aligned}$$

where we Weibull baseline hazard functions, $\lambda_{0j}(t_j)=\nu_j\rho_j t_j^{\rho_j}$, for $j=1,2$, and the

### 2.2 Competing risks in the progressive multistate model

We consider the progressive multistate model in the presence of competing risks. Figure 2 illustrates a situation where competing events are present for both Event 1 and Event 2; Event 3 is a competing event of Event 1 and Event 4 is a competing event of Event 2. We denote the times to Event 1, ..., Event 4 by $t_1,\ldots,t_4$, respectively.

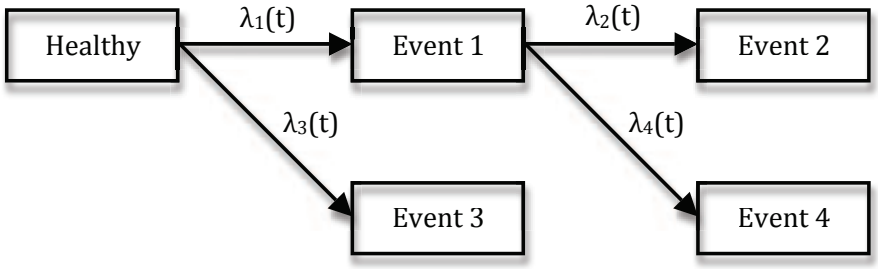


**Figure 2**: Progressive multistate model with competing risks and their cause-specific hazards $\lambda_k(t)$, $k=1,2,3,4$.

The cause-specific hazard functions for $t_k, k=1,2,3,4$, are defined by

$$\lambda_k(t_k|X_k)=\lambda_{0k}(t_k)\exp\{\beta_k^\top X_k\},$$

where $\lambda_{0k}(t_k)$ represents the baseline hazard function for transition $k = 1, 2, 3, 4$; we assume Weibull baseline hazard function $\lambda_{0k}(t_k) = \nu_k \rho_k t_k^{\rho_k}$.

Further, we let $t_A$ be the time to first event, $t_B$ be the time to the second event after the first event of interest; $t_A$ takes value $t_1$ or $t_3$ that occurs first, i.e., $t_A = \min\{t_1, t_3, age\}$, where $age$ represents age at examination as a censoring variable; $t_B$ defined as gap time takes value $t_2$ or $t_4$ that occurs first, i.e., $t_B = \min\{t_2, t_4, age - t_1\}$. We define $\delta_A = 1$ if $\delta_1 = 1$ or $\delta_3 = 1$, 0 otherwise; $\delta_B = 1$ if $\delta_2 = 1$ or $\delta_4 = 1$, 0 otherwise. Let $K_A$ represent the type of the first event, $K_A \in \{1, 3\}$, and $K_B$ the type of the second event, $K_B \in \{2, 4\}$.

The survivor functions for $t_A$ and $t_B$ can be expressed as

$$\begin{aligned} S_A(t_A) &= e^{-\Lambda_1(t_A)-\Lambda_3(t_A)} \\ S_B(t_B) &= e^{-\Lambda_2(t_B)-\Lambda_4(t_B)}, \end{aligned}$$

where $\Lambda_k(t_k)$ are the cumulative hazard functions for $t_k$, $k = 1, 2, 3, 4$.

Then, the joint survivor function can be constructed using survivor copula $C$ to link the two marginal survivor functions as

$$S(t_A, t_B) = C\{S_A(t_A), S_B(t_B)\} = \phi^{-1}\{\phi(S_A(t_A)) + \phi(S_B(t_B))\}.$$

Using Clayton copula, we obtain the joint survivor function and its first and second derivatives:

$$S(t_A, t_B) = \{S_A(t_A)^{-\theta} + S_B(t_B)^{-\theta} - 1\}^{-1/\theta}$$

$$\begin{aligned} -\frac{\partial}{\partial t_A} S(t_A, t_B) &= \{S_A(t_A)^{-\theta} + S_B(t_B)^{-\theta} - 1\}^{-(\theta+1)/\theta} S_A(t_A)^{-\theta} \lambda_1(t_A) \\ \frac{\partial^2}{\partial t_A \partial t_B} S(t_A, t_B) &= (\theta + 1)\{S_A(t_A)^{-\theta} + S_B(t_B)^{-\theta} - 1\}^{-(2\theta+1)/\theta} S_A(t_A)^{-\theta} \lambda_1(t_A) S_B(t_B)^{-\theta} \lambda_2(t_B). \end{aligned}$$

Subjects can take one of the following five possible transitions and their contribution to the likelihood are respectively:

Case 1: $\delta_A = 0$ and $\delta_B = 0$

$$L = S_A(t_A)$$

Case 2: $\delta_A = 1, K_A = 1$, and $\delta_B = 0$

$$L = -\frac{\partial}{\partial t_1} S(t_A, t_B)$$

Case 3: $\delta_A = 1, K_A = 3$

$$L = -\frac{\partial}{\partial t_3} S_A(t_3) = S_A(t_3) h_3(t_3)$$

Case 4: $\delta_A = 1, K_A = 1$ and $\delta_B = 1, K_B = 2$

$$L = \frac{\partial^2}{\partial t_1 \partial t_2} S(t_A, t_B).$$

Case 5: $\delta_A = 1, K_A = 1$ and $\delta_B = 1, K_B = 4$

$$L = \frac{\partial^2}{\partial t_1 \partial t_4} S(t_A, t_B).$$

Thus, combining these components together the likelihood function can be written as

$$\begin{aligned} L(\theta) &= \left\{\frac{\partial^2}{\partial t_1 \partial t_2} S(t_1,t_2)\right\}^{\delta_1\delta_2} \left\{\frac{\partial^2}{\partial t_1 \partial t_4} S(t_1,t_4)\right\}^{\delta_1\delta_4} \times \\ &\quad \left\{-\frac{\partial}{\partial t_1} S(t_1,t_B)\right\}^{\delta_1(1-\delta_B)} \left\{-\frac{\partial}{\partial t_3} S_A(t_3)\right\}^{\delta_3} S_A(t_A)^{(1-\delta_A)(1-\delta_B)}. \\ &= (\theta+1)^{\delta_B}\{S_A(t_A)^{-\theta} + S_B(t_B)^{-\theta} - 1\}^{-\{(\theta+1)\delta_1/\theta+\delta_B\}} \times \\ &\quad S_A(t_A)^{-(\theta\delta_1+\delta_1-1)} S_B(t_B)^{-\theta\delta_B} \lambda_1(t_A)^{\delta_1} \lambda_3(t_A)^{\delta_3} \lambda_2(t_B)^{\delta_2} \lambda_4(t_B)^{\delta_4}. \quad (1) \end{aligned}$$

The corresponding log-likelihood function has the form

$$\begin{aligned} \ell(\theta) &= \delta_B \log(\theta+1) - \{(\theta+1)\delta_1/\theta + \delta_B\}\{S_A(t_A)^{-\theta} + S_B(t_B)^{-\theta} - 1\} \\ &\quad -(\theta\delta_1+\delta_1-1)\log S_A(t_A) - \theta\delta_B \log S_B(t_B) + \\ &\quad \delta_1 \log\lambda_1(t_A) + \delta_3 \log\lambda_3(t_A) + \delta_2 \log\lambda_2(t_B) + \delta_4 \log\lambda_4(t_B) \\ &= \delta_B \log(\theta+1) - \{\delta_1 + \delta_1/\theta + \delta_B\}\{e^{\theta\Lambda_1(t_A)+\theta\Lambda_3(t_A)} + e^{\theta\Lambda_2(t_B)+\theta\Lambda_4(t_B)} - 1\} + \\ &\quad (\theta\delta_1+\delta_1-1)\{\Lambda_1(t_A)+\Lambda_3(t_A)\} + \theta\delta_B\{\Lambda_2(t_B)+\Lambda_4(t_B)\} + \\ &\quad \delta_1 \log\lambda_1(t_A) + \delta_3 \log\lambda_3(t_A) + \delta_2 \log\lambda_2(t_B) + \delta_4 \log\lambda_4(t_B). \end{aligned}$$

### 2.3 Ascertainment-corrected retrospective likelihood for family data

A general form of the ascertainment corrected likelihood for $n$ families (Le Bihan *et al.*, 1995) can be expressed as

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

where $L_f^c$ is the ascertainment corrected likelihood function for family $f$, $f = 1,\ldots,n$. The numerator $N_f$ is the likelihood contribution for the members of family $f$ and the denominator $A_f$ is the probability of family $f$ being ascertained into the study. We employ the ascertainment corrected retrospective likelihood approach (Carayol and Bonaïti-Pellié, 2004; Kraft and Thomas, 2000) to account for complex ascertainment criteria of families into study.

The ascertainment corrected retrospective likelihood for family $f$ is obtained by conditioning on all the phenotypes $Y_f$ in the family and ascertainment scheme $Asc_f$, which can be expressed as the conditional distribution of $Y_f$ given $G_f$ divided by the ascertainment correction probability,

$$L_f^c = P(G_f|Y_f, Asc_f) = \frac{P(Asc_f|Y_f,G_f)P(Y_f|G_f)}{P(Y_f, Asc_f|G_f)}, \quad (2)$$

where $P(Asc_f|Y_f,G_f)$ is equal to 1 if a family satisfies the ascertainment scheme, and 0 otherwise.

The numerator $P(Y_f|G_f)$ can be obtained by the likelihood function expressed in equation (2). The denominator represents the ascertainment probability of observing the phenotypes of the members through whom the family is ascertained into the study.

In this paper, we consider the families arise from a population-based design that the ascertainment of a family is only based on a single proband affected with a first colorectal cancer, randomly sampled from a diseased population. Therefore, the ascertainment probability for family $f$ can be calculated as the probability that the proband is affected by the

first event of interest prior to his/her age at examination,

$$\begin{aligned} P(Y_f, Asc_f|G_f) &= P(T_A < a_{fp}, K_A = 1|G_{fp}) = \int_0^{a_{fp}} P(T_A > u|G_{fp})\lambda_1(u|G_{fp})du \\ &= P(T_1 < a_{fp}, T_3 > a_{fp}|G_{fp}) = \{1 - S_1(a_{fp}|G_{fp})\}S_3(a_{fp}|G_{fp}), \end{aligned}$$

where $a_{fp}$ is the age at examination of the proband, $G_{fp}$ is the genotype of the proband, and $S_k(t_k|G_{fp})$ are the survivor functions for $t_k$, $k = 1, 3$; this probability represents that the proband's first event of interest has occurred by age $a_{fp}$ but the competing event has not.

### 2.4 Penetrance functions

We define two penetrance functions relating to the first and second events of interest to describe absolute disease risks. The penetrance function for the first event of interest ($k = 1$) represents the probability that the subject survives all causes up to time $t_1$, then develops cancer of type 1 by time $t_1$, which can be obtained as

$$\begin{aligned} P(T_A < t_1, K_A = 1|X) &= \int_0^{t_1} S_A(u|X)\lambda_k(u|X)du \\ &= \int_0^{t_1} e^{-\Lambda_1(u|X)-\Lambda_3(u|X)}\lambda_1(u|X)du \\ &= \{1 - S_1(t_1|X)\}S_3(t_1|X), \end{aligned}$$

where $X$ represents the covariates associated with the first event.

The penetrance function for the second cancer we define as the probability of developing any second cancer in $t_2$ years after the first cancer of interest occurred at $t_1$ given covariates $X$, i.e.,

$$\begin{aligned} P(T_B < t_2|T_1 = t_1, X) &= 1 - \frac{-\frac{\partial}{\partial t_1}S(t_1, t_2|X)}{-\frac{\partial}{\partial t_1}S_A(t_1|X)} \\ &= 1 - \frac{\{S_A(t_1|X)^{-\theta} + S_B(t_2|X)^{-\theta} - 1\}^{-(\theta+1)/\theta}S_A(t_1|X)^{-\theta}\lambda_1(t_1|X)}{S_A(t_1|X)\lambda_1(t_1|X)} \\ &= 1 - \{S_A(t_1|X)^{-\theta} + S_B(t_2|X)^{-\theta} - 1\}^{-(\theta+1)/\theta}S_A(t_1|X_1)^{-(\theta+1)}. \end{aligned}$$

## 3. Application to Lynch Syndrome Families

A total of 97 population-based Lynch Syndrome families were identified through the Colon Cancer Family Registries. These families harbour a mutation in one of the mismatched repair genes such as MLH1 or MSH2. Members of those families are at high risk of developing multiple cancers over their lifetime. In this study, our goal is to estimate the risks of a first colorectal cancer (CRC1) and a second colorectal cancer (CRC2) when endometrial cancers (EC) are competing cancer events. Figure 3 displays the data with the number of cancers observed. In the 97 LS families, there are 835 individuals; 157 of them developed a first CRC and 9 experienced EC. Of 157 first CRCs, 22 had the second CRC and 8 had EC.

Based on our proposed progressive multistate model with competing risks, we estimated both relative and absolute risks of developing the first and second CRC for these family members. Ascertainment correction was used because these families were sampled via affected mutation carrier probands. Each transition assumes Weibull baseline hazard and includes gender and mutation status as covariates.

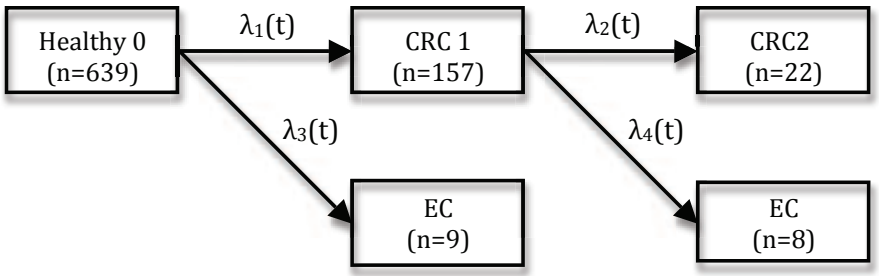


**Figure 3**: Successive cancers and competing cancer events observed in 97 population-based Lynch syndrome families identified from the Colon Cancer Family Registries

**Table 1**: Relative risk estimates in terms of hazard ratios (HR) of developing the first and second colorectal cancers (CRC) for mutation status and gender, with and without adjusting for competing risks.

With Competing Risks

| | 1st CRC | 2nd CRC |
|---|---|---|
| HR (carriers vs. non-carriers) | 1.58 (se=0.42) | 2.08 (se=1.54) |
| HR (males vs. females) | 2.32 (se=0.42) | 1.25 (se=0.52) |

Without Competing Risks

| | 1st CRC | 2nd CRC |
|---|---|---|
| HR (carriers vs. non-carriers) | 2.01 (se=0.53) | 1.92 (se=1.46) |
| HR (males vs. females) | 2.18 (se=0.58) | 1.16 (se=0.54) |

What follows are the results from estimating all the parameters in the model including baseline hazards parameters ($\nu's, \rho's$), regression parameters ($\beta's$), and copula association parameter $\theta$. To examine the impact of competing risks on disease risks estimations, we fitted two models: one accounts for competing risk events and the other ignores them. Table 1 summarizes hazard ratios (HR) for the first and second CRCs between males and females and mutation carriers and non-carriers. When taking competing risks into account, mutation carriers have 1.58 times higher hazards than noncarriers for the first CRC but about twice higher hazards for the second CRC; however, when ignoring competing risks, the genetic effect was slightly overestimated (HR=2.01, se=0.53) for the first cancer but underestimated (HR=1.92, se=1.46) for the second cancer. The gender effect was also larger for the first CRC (HR=2.32, se=0.42) than the second CRC (HR=1.25, se=0.52) when accounting for competing risks; when ignoring competing risks, the gender effects were slightly underestimated for both cancers.

For absolute disease risks associated with mutated genes for the first and second CRC, we graphically display the penetrance functions for mutation carriers, separated by males and females, also by with and without adjusting for competing risks. Figure 4 displays the age-specific penetrance functions for the first CRC, broken down by males and females; the solid lines are from the competing risks model and dotted lines from the model without them. We notice that the risks were increased with age and males had higher risks than females over all age values. When competing risks were ignored, the penetrances were overestimated for both males and females.

The penetrance function for the second cancer, defined as $P(T_B < t_2 | T_1 = t_1, X)$ in

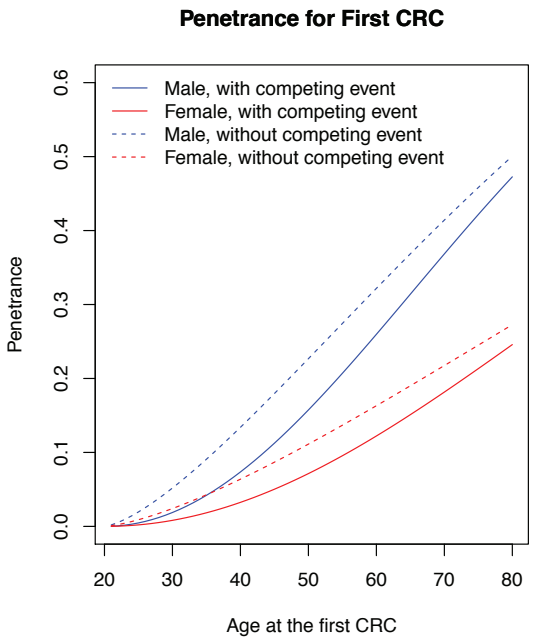


**Figure 4**: Penetrance function for the first colorectal cancer (CRC)

§2.4, can be displayed in two ways: Figure 5 (a) presents the penetrance as a function of $t_2$ (time since the first cancer) provided that the first CRC occurred at the age 30 $(t_1 = 30)$ and Figure 5 (b) presents 10-year penetrance $P(T_B < 10|T_1 = t_1, X)$ by varying $t_1$.

As shown in Figure 5 (a), when first CRC occurred at age 30, the risk of developing a second CRC was estimated much higher when taking competing risks into account compared to when ignoring them. We also observe that the risks of developing a second CRC increase over time and males have slightly higher risks of second CRC than females. Another way of viewing the penetrance for the second cancer is as a risk at a fixed duration of time of 10 years after the first CRC, which varies by age at first CRC, i.e., the probability of developing a second CRC in 10 years after the first CRC for different ages at onset for the first CRC (Figure 5(b)). We observe that the risk of developing a second cancer within a 10 years interval decreases as the first cancer occurs at older age. It is noteworthy that if the first cancer occurred before age 40, males tend to have higher risk of getting a second cancer in a 10 years interval compared to females. However, for the first cancer after age 40, females have higher risk of getting a second cancer in 10 years compared to males.

## 4. Discussion

We developed a general statistical framework for modelling successive cancers in the presence of competing cancers associated with gene mutation in Lynch Syndrome families. The proposed multistate-competing risk model provides cancer-specific transition intensities and probabilities (penetrance) that can be of clinical interest for characterizing cancer risks in LS families and designing efficient preventive strategies. Our approach is not restricted to the two cancers considered (CRC and EC) but can be generalized to the other cancers observed in LS families, provided that enough events are observed for accurate estimation.

We also dealt with with missing genotypes. For those individuals with missing geno-

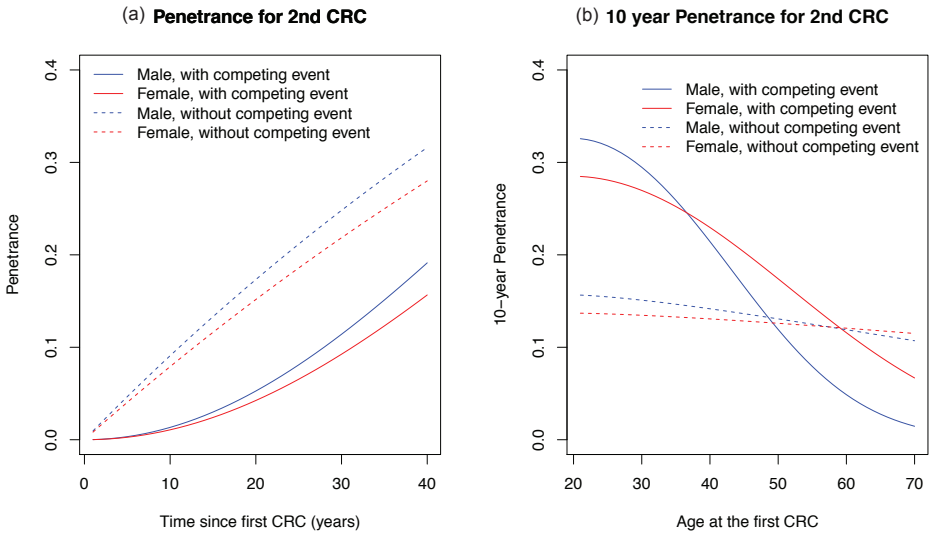


**Figure 5**: Penetrance functions for the second colorectal cancer (CRC); Left panel is the penetrance when the age at onset for the first cancer is fixed at 30 years and right panel is 10-year penetrance depending on different ages at onset of the first CRC.

types, we calculated a weighted log-likelihood where the weights are given by the mutation carrier probabilities conditional on observed genetic and phenotype information. The genetic effect associated with a second cancer was not precisely estimated due to the missing data and the small number of second cancers. In addition, we obtained a robust variance estimator (White, 1982) to account for the within-family dependence and possible model misspecification.

In terms of parameter estimation, the copula parameter $\theta$ was estimated at 6.27 (se=3.49) when accounting for competing risks, which corresponds to a large dependence between the two successive cancers (Kendall's $\tau = 0.75$). In contrary, when the competing risks were ignored, the copula parameter was estimated at ($\theta$ = 0.79 (se=2.54); Kendall's $\tau = 0.28$). This suggests that the presence of competing events can confound the estimation of association between successive cancer events and thus can be of importance when predicting the risk of cancer recurrence in these LS families.

Finally, our analysis showed that ignoring competing events can have a substantial impact on cancer risk estimation in LS families, in particular by overestimating the relative and absolute risk of first CRC.

## REFERENCES

Beyersmann, J. Schumacher, M. and Allignol, A. (2012). *Competing risks and multistate models with R.* Springer, New York.

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

Choi, Y.-H., Briolllais, L., Green, J. Parfrey, P. and Kopciuk, K. (2014). "Estimating Successive Cancer Risk in Lynch Syndrome Families using a Progressive Three-state Model." *Statistics in Medicine*, 33(4), 618–638.

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.

Katki, H. A., Blackford, A., Chen, S. and Parmigiani, G. (2008). "Multiple diseases in carrier probability estimation: Accounting for surviving all cancers other than breast and ovary in BRCAPRO." *Statistics in Medicine*, 27(22), 4532–4548.

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

Nelsen, R.B. (2006). *An introduction to copulas.* 2nd ed. Springer, New York.

White, H. (1982). "Maximum likelihood estimation of misspecified models." *Econometrica*, 50, 1-25.