A Multistate Frailty Model for Sequential Events from a Family-based Study

Yun-Hee Choi[1] & Balakumar Swaminathan[2]

[1] *Department of Epidemiology and Biostatistics, Western University, London, Ontario, Canada* [2] *Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada*

Résumé. En épidémiologie génétique, les familles qui ont des maladies génétiques sont souvent prédisposées à la survenue de cancers successifs au cours de leur vie. Par exemple, les familles ayant le syndrome de Lynch portent une mutation dans un des gènes liés à la réparation des erreurs de réplication de l'ADN (MMR) et sont à risque élevé de développer des cancers multiples au cours de leur vie incluant les cancers colorectaux, de l'endomêtre, de l'ovaire, de l'estomac, etc. Notre intérêt principal est de fournir des estimés fiables des risques relatifs et des risques cumulés dépendants du temps (pénérrance) pour des cancers successifs associés au gène muté. On développe un cadre statistique pour la modélisation de deux temps de survenue séquentiels émanant d'un échantillonnage de familles issues de la population générale, en prenant en compte la dépendance de temps de survie en utilisant un modèle de fragilité partagée. En plus, on incorpore la correction nécessaire pour l'échantillonnage non-aléatoire des familles à partir d'une approche de vraisemblance rétrospective. On illustre aussi notre approche en utilisant un échantillon de 194 familles Lynch provenant des Registres Familiaux de Cancers Colorectaux et donnons les risques génétiques relatifs et les estimations de pénétrance correspondant au développement d'un premier et second cancer colorectal.

Mots-clés. Fragilité partagée, modèle multi-états, données familiales, temps d'évènements successifs, vraissemblance corrigée pour l'échantillonnage.

Abstract. In genetic epidemiology, families with genetic disorders are often predisposed to successive cancers in their lifetime. For example, Lynch syndrome families harbouring a mutation in mismatched repair (MMR) genes are at high risk of developing multiple cancers over their lifetime including colorectal, endometrial, ovarian, stomach cancers etc. Our primary interest is to provide reliable estimates of relative risk and age-dependent cumulative risks (penetrance) for successive cancers associated with the mutated gene. We develop a statistical framework for modelling two sequential survival times arising from a population-based family design by accounting for the dependence between the event times using a shared frailty model. In addition, we incorporate the necessary ascertainment correction for the non-random sampling of the families using the retrospective likelihood approach. Our simulation studies demonstrate that the proposed

method provides unbiased and reliable estimates of the disease risks associated with a mutated gene for the successive events in family-based designs. We also illustrate our approach using a population-based sample of 97 Lynch syndrome families from the Colon Cancer Family Registry and provide genetic relative risk and penetrance estimates for developing a first and second colorectal cancer.

Keywords. Shared frailty, multistate model, family data, successive event times, ascertainment-corrected likelihood.

1 A Shared frailty model for successive event times

Progressive multistate models have been a useful tool for modelling successive events experienced by an individual. Recently, Choi et al. (2014) developed a three-state progressive model for estimating successive cancer risks from family data where a Markov extension model was adopted for modelling the time to the second cancer using the time to the first cancer as a covariate. However, this model assumed the time to the first cancer to be independent of other covariates in the model, which is often not true. Instead, those successive event times should be considered as bivariate outcome variables.

In this paper, we incorporate the shared frailties to the multistate model for effectively modelling bivariate outcomes of sequential event times arising from family-based studies and for better estimating the lifetime disease risks of developing successive events associated with mutated genes. We first derive a bivariate survival distribution using a shared frailty model to account for the dependence between sequential event times, then employ the retrospective likelihood approach (Carayol and Bonaïti-Pellié, 2004; Kraft and Thomas, 2000) to correct for the complex ascertainment procedure involved in obtaining the family data. We consider a population-based family design where families are sampled through single affected individuals (also called a probands).

Suppose that the sequential event times arise from the following three-state progressive model. We let T_1 denote a non-negative continuous random variable that measures the

Figure 1: Three-state progressive model

time spent in the 'Healthy' state prior to experiencing 'Event 1' and T_2 denote another non-negative continuous random variable that measures the time spent in the state 'Event 1' prior to experiencing 'Event 2', such that T_2 represents the gap time between the two events.

We also let Z be a random frailty variable that measures the amount of dependence between two events experienced by each individual, such that conditional on the frailty,

his/her event times are independent. Therefore, the proportional hazard models for T_1 and T_2 conditional on the frailty Z shared between them can be written as

$$\begin{aligned}\lambda_1(t_1|Z,X_1) &= \lambda_{01}(t_1)Z\exp\{\beta_1^\top X_1\}\\ \lambda_2(t_2|Z,X_2) &= \lambda_{02}(t_2)Z\exp\{\beta_2^\top X_2\},\end{aligned}$$

where $\lambda_{01}(t_1)$ and $\lambda_{02}(t_2)$ are the baseline hazard functions for T_1 and T_2, respectively, and X_1 and X_2 are the vectors of the risk factors associated with T_1 and T_2, respectively.

Then, the conditional bivariate survival distribution can be derived as follows

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

where $\Lambda_{01}(t_1)$ and $\Lambda_{02}(t_2)$ are the cumulative baseline hazard functions for T_1 and T_2, respectively. By integrating out the unobserved frailty over the frailty distribution, we obtain the bivariate survival function of the form,

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

where $\mathcal{L}(.)$ is the Laplace transform of the frailty distribution.

2 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 study. For individual i, $i=1,\ldots,n_f$, in family f, we observe the vector of the phenotype, $Y_{fi}=(t_{fi1},\delta_{fi1},t_{fi2},\delta_{fi2})$, containing the event times, t_{fi1},t_{fi2}, and censoring indicators $\delta_{fi1},\delta_{fi2}$ for the first and second events and the genotype G_{fi} which is coded as 1 for mutation carriers and 0 for mutation non-carriers. For simplicity, we express the bivariate model adjusting only for the genetic effect G. However, it can easily accommodate other risk factors. 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:

$$L_f^c = P(G_f|Y_f, Asc_f) = \frac{P(Asc_f|Y_f, G_f)P(Y_f|G_f)}{P(Y_f, Asc_f|G_f)}, \tag{1}$$

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

In the numerator, $P(Y_f|G_f)$ represents the prospective likelihood based on modelling the time-to-event data given the family members' genotypes, assuming the family members are independent given their genotypes but their event times are correlated, which has the form:

$$\begin{aligned} P(Y_f|G_f) = L_f(\theta) &= \prod_{i=1}^{n_f} \left\{ \frac{\partial^2}{\partial t_{fi1} \partial t_{fi2}} S(t_{fi1}, t_{fi2}|G_{fi}) \right\}^{\delta_{fi1}\delta_{fi2}} \times \\ &\quad \left\{ -\frac{\partial}{\partial t_{fi1}} S(t_{fi1}, t_{fi2}|G_{fi}) \right\}^{\delta_{fi1}(1-\delta_{fi2})} S(t_{fi1}, t_{fi2}|G_{fi})^{(1-\delta_{fi1})(1-\delta_{fi2})}. \end{aligned}$$

Here, the likelihood contribution of each individual is one of three cases based on the bivariate distribution: 1) occurring both events at t_{fi1} and t_{fi2}, respectively, 2) occurring only first event at t_{fi1} but no second event by one's current age a_{fi}, i.e., $t_{fi2} = a_{fi} - t_{fi1}$, or 3) occurring no events by a_{fi} i.e., $t_{fi1} = a_{fi}$, $t_{fi2} = 0$.

The denominator of equation (1) 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 to arise from a population-based design where the ascertainment of families is only based on the probands, who are randomly sampled from a diseased population. Therefore, the ascertainment probability for family f can be obtained simply by calculating the probability that the proband is affected by the first event prior to his/her age at examination, which can be written as

$$P(Y_f, Asc_f|G_f) = P(T_1 < a_{fp}|G_{fp}) = 1 - S_1(a_{fp}|G_{fp}),$$

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

We also accommodate the robust variance estimators (White, 1982) for the disease risks derived from the model to handle possible model mis-specifications.

3 Application to Lynch Syndrome Families

A total of 97 population-based Lynch syndrome families were identified through the Colon Cancer Family Registries, a NIH-funded initiative. These families harbour a mutation in

mismatched repair genes such as MLH1 or MSH2 and members of those families are at high risk of developing multiple cancers over their lifetime including colorectal, endometrial, ovarian, stomach etc. Based on our proposed frailty multistate model with ascertainment correction, we estimated the disease risks of developing the first and second cancers for these family members arising from a population-based family design. In this study, we are particularly interested in the first colorectal cancer as our first event and any Lynch syndrome related cancers including colorectal, endometrial, stomach, ovarian, brain, liver, small bowel etc. followed by the first colorectal cancer as our second event. Our data include 156 first cancer patients and of them 39 experienced a second Lynch syndrome cancer.

The following table summarizes various disease risks of developing the first and second cancers depending on gender and carrier status. The first penetrance represents the lifetime risk of developing the first cancer by age 70 and the second penetrance represents the risk of developing the second cancer within 5 years after the first cancer given the age of the first cancer at t_1.

		First Cancer Penetrance[†]	5-year Second Cancer Penetrance[‡]		
		By age 70	$t_1 = 30$	$t_1 = 40$	$t_1 = 50$
Carriers	Male	34.76%	13.13%	9.80%	6.12%
	Female	22.06%	19.50%	17.46%	13.98%
Noncarriers	Male	26.99%	7.58%	6.33%	4.55%
	Female	15.25%	11.38%	10.65%	9.23%

[†] the risk of developing the first cancer by age 70

[‡] the risk of developing the second cancer within 5 years after the first cancer given the age of the first cancer, t_1.

For relative risks, we found that genetic risks were similar for both first and second cancers. The estimated hazards ratios between carriers and noncarriers were 1.88 (se=0.42) for the first cancer and 1.89 (se=0.74) for the second cancer when the same value of frailty is given. However, the gender effect was higher on the first cancer (hazards ratio between male and female = 2.83, se=1.10) than on the second cancer (hazards ratio = 0.66, se=0.27). In addition, the frailty parameter estimate was 0.22 (se=0.09) indicating relatively strong correlation between the first and second cancers.

4 Discussion

We developed the share frailty model for the correlated times to successive cancers arising from cancer families by taking complex ascertainment criteria of families into account. The proposed model can provide disease risks in genetic counselling for those susceptible individuals from high risk cancer families. We analyzed Lynch syndrome families sampled from Colon Cancer Family Registries and provided both relative and absolute disease risks

for first and second cancers. In our analysis, it is noteworthy that their age-specific risks of the second cancer varied depending on their age at the first cancer; females have higher risk of developing a second cancer within 5 years after the first cancer than males. We also observed that the earlier the first cancer occurred the higher the risk of developing a second cancer was. On the other hand, males had higher hazards than females for both first and second cancers given the same value of frailty.

Bibliographie

[1] 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.
[2] 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.
[3] 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.
[4] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1-25.