MODELLING SUCCESSIVE TIME-TO-EVENT OUTCOMES IN PRESENCE OF COMPETING RISK EVENTS USING COPULAS

Laurent Briollais[1,2] & Yun-Hee Choi[3] & Lajmi Lakhal-Chaieb[4]

laurent@lunenfeld.ca

[1] *Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON, Canada.*
[2] *Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada.*
[3] *Western University, Department of Epidemiology and Biostatistics, London, Canada.*
[4]*Department of Mathematics and Statistics, Laval University, Quebec, Canada.*

Résumé. Nous proposons ici un modèle d'association pour estimer la pénétrance (risque) de cancers successifs en présence d'évènements compétitifs. L'association entre les deux évènements successifs est spécifiée à partir d'une fonction Copule et un modèle de hasards proportionnels est utilisé pour chaque évènement compétitif. Ce travail est motivé par l'analyse de cancers successifs chez des individus ayant le syndrome de Lynch. La procédure d'inférence statistique est adaptée à la prise en compte de covariables génétiques manquantes ainsi que le biais de sélection induit par le recrutement de familles ayant plusieurs individus atteints d'un premier cancer colorectal. Les performances de la procédure d'estimation sont évaluées par simulations et son utilisation est illustrée par l'analyse de données provenant de registres familiaux du cancer colorectal.

Mots-clés. Évènements successifs; Risques compétitifs; Données familiales; Biais de sélection; Cancer colorectal; Syndrome de Lynch.

Abstract. We propose here an association model to estimate the penetrance (risk) of successive cancers in the presence of competing risks. The association between the successive events is modeled via a copula and a proportional hazards model is specified for each competing event. This work is motivated by the analysis of successive cancers for people with Lynch Syndrome in the presence of competing risks. The proposed inference procedure is adapted to handle missing genetic covariates and selection bias, induced by the data collection of families with multiple individuals affected with a first colorectal cancer. The performance of the proposed estimation procedure is evaluated by simulations and its use is illustrated with data from the Colon Cancer Family Registry.

Keywords. Successive events; Competing risks; Familial data; Selection bias; Colorectal cancer; Lynch syndrome.

1 Introduction

An important issue when estimating the risk associated with a single or multiple cancer events is the presence of competing events. Competing risks concern the situation where more than one

cause of failure is possible (Putter et al., 2007). A classical example relates to several causes of death (e.g. from cancer) where the occurrence of any cause of death prevents the event of interest from occurring. Treating the events of the competing causes as censored observations will lead to biased estimates of the penetrance function of the event of interest when we are in the presence of correlated competing risks (Putter et al., 2007). In this paper, we propose a general methodology to estimate the risks of observing a first cancer event and a second cancer event given the age at onset of the first cancer in people with Lynch Syndrome (LS) while accounting for the presence of competing risk events.

2 Model

Consider the following progressive multistate model with competing risks. The model includes 5 states, healthy and events 1 to 4, where events 1 and 2 are successive events of interest and events 3 and 4 represent competing events for events 1 and 2, respectively.

2.1 Marginal distributions

Let T_1 and T_3 be the times from the healthy state to events 1 and 3, respectively and $Y_1 = \min\{T_1, T_3\}$. Define ϵ_1 by $\epsilon_1 = 1$ if $T_1 < T_3$ and $\epsilon_1 = 3$, otherwise. Note that events 1 and 3 are competing risks so it is of interest to define the following cause-specific hazard functions

$$\lambda_k(y|G,X) = \lim_{dy\to 0} \frac{1}{dy} P(y < Y_1 \leq y + dy, \epsilon_1 = k \mid G, X, Y_1 > y), \quad k = 1, 3,$$

where G is the individual genotype information corresponding to the mutation carrier status (carrier=1, non-carrier=0) and X a set of measured covariates. By standard theory of competing risks,

$$h_1(y|G,X) = \lambda_1(y|G,X) + \lambda_3(y|G,X) \quad \text{and} \quad S_1(y|G,X) = \exp\{-\int_0^y h_1(u|G,X)du\}$$

are the hazard and survival functions associated to Y_1, respectively and

$$F_{11}(y|G,X) = P(Y_1 \leq y, \epsilon_1 = 1|G,X) = \int_0^y S_1(u|G,X)\lambda_1(u|G,X)du$$

is the cause-specific cumulative incidence function of event 1.
The people satisfying $\epsilon_1 = 1$ are afterwards at risk of observing either event 2 or event 4. Let T_2 and T_4 be times from event 1 to events 2 and 4, respectively and $Y_2 = \min(T_2, T_4)$. Define ϵ_2 by $\epsilon_2 = 2$ if $T_2 < T_4$ and $\epsilon_2 = 4$, otherwise. Similarly, define the conditional cause-specific hazard functions given $\epsilon_1 = 1$ by

$$\lambda_k(y|G,X) = \lim_{dy\to 0} \frac{1}{dy} P(y < Y_2 \leq y + dy, \epsilon_2 = k \mid G, X, Y_2 > y, \epsilon_1 = 1),\ k = 2, 4.$$

The conditional hazard and survival functions associated to Y_2 given $\epsilon_1 = 1$ are then, respectively,

$$h_2(y|G,X) = \lambda_2(y|G,X) + \lambda_4(y|G,X) \quad \text{and} \quad S_2(y|G,X) = \exp\{-\int_0^y h_2(u|G,X)du\}.$$

We assume that the cause-specific hazard for the event k, $k = 1,2,3,4$, follows a proportional hazards regression model

$$\lambda_k(y|G,X) = \lambda_{k0}(y)e^{\beta_k^\top X + \beta_{g_k} G},$$

where λ_{k0} is the baseline hazard function and β_k and β_{g_k} the regression coefficients related to event k. Two approaches are considered in this paper: (*i*) a parametric approach where a parametric distribution is specified for each λ_{k0} and (*ii*) a piecewise constant hazard approach where λ_{k0} is assumed to be constant within each interval of a partition of $[0,\infty)$. In both cases, we denote by θ_k the set of baseline distribution parameters and regression coefficients related to event k.

2.2 Association model

For the people satisfying $\epsilon_1 = 1$, we model the dependence in the pair (Y_1, Y_2) through a semi-survival copula, $\mathcal{C}_\gamma$, (Lakhal-Chaieb et al. 2006) defined as follows:

$$\begin{aligned} P(Y_1 \leq y_1, Y_2 > y_2|\epsilon_1 = 1, G, X) &= \mathcal{C}_\gamma \left\{P(Y_1 \leq y_1|\epsilon_1 = 1, G, X), P(Y_2 > y_2|\epsilon_1 = 1, G, X)\right\} \\ &= \mathcal{C}_\gamma \left\{F_{11}(y_1|G,X)/p(G,X), S_2(y_2|G,X)\right\}, \end{aligned}$$

where the parameter γ measures the conditional dependency in the pair (Y_1, Y_2) given $\epsilon_1 = 1$ and $p(G,X) = P(\epsilon_1 = 1|G,X) = \lim_{t\to\infty} F_{11}(t|G,X)$.
The model is completed by specifying $P(\epsilon_2 = 2|G,X,Y_1 = y_1, Y_2 = y_2, \epsilon_1 = 1)$. This probability has to satisfy

$$\begin{aligned} P(\epsilon_2 = 2|G,X,Y_2 = y_2, \epsilon_1 = 1) &= E_{Y_1}\left\{P(\epsilon_2 = 2|G,X,Y_1,Y_2 = y_2,\epsilon_1 = 1)\right\} \\ &= \frac{\lambda_2(y_2|G,X)}{\lambda_2(y_2|G,X) + \lambda_4(y_2|G,X)}, \end{aligned} \tag{1}$$

where the expectation is taken with respect to Y_1. A natural and mathematically convenient strategy to ensure that (1) holds is to assume

$$P(\epsilon_2 = 2|G,X,Y_1 = y_1, Y_2 = y_2, \epsilon_1 = 1) = P(\epsilon_2 = 2|G,X,Y_2 = y_2, \epsilon_1 = 1). \tag{2}$$

When this condition is not fulfilled, we are in the presence of an additional aspect of the dependency between the successive competing risks. In Web Appendix A, we present a procedure to test equation (2). Applying this test to the LS families cancer data suggests that it is plausible to assume (2) in our case. Therefore, the developments presented throughout the rest of this paper are made under this assumption.

2.3 Penetrance functions

The penetrance functions are defined as cause-specific cumulative incidence functions. The penetrance for event 1 is $\mathcal{P}_1(y_1; G, X) = F_{11}(y_1|G, X)$, which is the cumulative risk of developing event 1 by age y_1 in the presence of the competing event 3. The penetrance function for event 2 is the cause-specific cumulative incidence function conditional on the age at onset of event 1. When the assumption (2) is satisfied, we show in Appendix A that this penetrance function equals

$$\begin{aligned}\mathcal{P}_2(y_2; y_1, G, X) &= P(Y_2 \leq y_2, \epsilon_2 = 2|Y_1 = y_1, \epsilon_1 = 1, G, X) \\ &= \int_0^{y_2} \mathcal{C}_\gamma^{11}\{F_{11}(y_1|G, X)/p(G, X), S_2(u|G, X)\} S_2(u|G, X)\lambda_2(u|G, X)du, \end{aligned} \tag{3}$$

where $\mathcal{C}_\gamma^{ij}(u, v) = \partial^{i+j}\mathcal{C}_\gamma(u, v)/\partial^i u \partial^j v$. It is the probability of developing event 2 within y_2 since event 1 which has occurred at y_1. One is often interested in a 5-year or 10-year penetrance for second event.

3 Observed data and inference procedures

3.1 Maximum likelihood estimation

In this section, we describe the observed data and derive an estimation procedure for the parameters $\{\theta_1, \theta_2, \theta_3, \theta_4, \gamma\}$. In the LS families cancer data, Y_1 is right-censored by the age of last follow-up a. The observed data related to the events 1 and 3 is then $\{a, \tilde{Y}_1, \tilde{\epsilon}_1\}$, where $\tilde{Y}_1 = \min(Y_1, a)$ and $\tilde{\epsilon}_1 = \epsilon_1 \times I(Y_1 < a) \in \{0, 1, 3\}$. For those satisfying $\tilde{\epsilon}_1 = 1$, we also observe $\tilde{Y}_2 = \min(Y_2, a - Y_1)$ and $\tilde{\epsilon}_2 = \epsilon_2 \times I(Y_2 < a - Y_1) \in \{0, 2, 4\}$.
The observations are clustered into I families. The data is then

$$\Delta = \{(a_{ij}, \tilde{Y}_{1ij}, \tilde{\epsilon}_{1ij}, \tilde{Y}_{2ij}, \tilde{\epsilon}_{2ij}, G_{ij}, X_{ij}), i = 1, \cdots, I, j = 1, \cdots, n_i\},$$

where n_i is the size of the i^{th} family.
A family is included into the study if and only if the first examined person or proband has observed either event 1 or event 3 by age a. We assume a unique proband per family, which we index by the subscript $j = 1$. Close relatives of this proband for whom some genotype and cancer history information is available from the corresponding family unit. As this data collection protocol induces a selection bias, an ascertainment correction is required. To this end, we employ a conditional likelihood approach where the contribution of each family is corrected for its probability of being ascertained. For parameter estimation, we consider a two-stage estimation procedure. In the first stage, we estimate the parameters related to events 1 and 3 by maximizing the conditional log-likelihood function

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i} l_1(\theta_1, \theta_3|\tilde{Y}_{1ij}, \tilde{\epsilon}_{1ij}, G_{ij}, X_{ij}) - \sum_{i=1}^{I} l_c(\theta_1, \theta_3|a_{i1}, G_{i1}, X_{i1}), \tag{4}$$

where

$$l_1(\theta_1, \theta_3 | \tilde{Y}_1, \tilde{\epsilon}_1, G, X) = \sum_{k \in \{1,3\}} I(\tilde{\epsilon}_1 = k) \times \log\{\lambda_k(\tilde{Y}_1 | G, X)\} - \int_0^{\tilde{Y}_1} h_1(u | G, X) du$$

is the standard contribution of an individual to the log-likelihood function and

$$l_c(\theta_1, \theta_3 | a, G, X) = \log\{P(Y_1 < a | G, X)\} = \log\{1 - S_1(a | G, X)\} \quad (5)$$

is the familial ascertainment correction term. This log-likelihood function is derived under the assumption of conditional independence of ages at onset of cancer of family members given their mutation carrier statuses. This assumption is plausible in our case given the strong association between the genotype and the risk of developing cancer.
At the second stage, we estimate the parameters related to events 2 and 4 as well as the copula parameter γ by maximizing the log-likelihood function

$$\sum_{i=1}^{I} \sum_{j=1}^{n_i} I(\tilde{\epsilon}_{1ij} = 1) l_2(\theta_2, \theta_4, \gamma | \hat{\theta}_1, \hat{\theta}_3, \tilde{Y}_{1ij}, \tilde{Y}_{2ij}, \tilde{\epsilon}_{2ij}, G_{ij}, X_{ij}),$$

where

$$l_2(\theta_2, \theta_4, \gamma | \hat{\theta}_1, \hat{\theta}_3, \tilde{Y}_1, \tilde{Y}_2, \tilde{\epsilon}_2, G, X) = I(\tilde{\epsilon}_2 = 0) \log\left[\mathcal{C}_\gamma^{10}\{\hat{F}_{11}(\tilde{Y}_1 | G, X)/\hat{p}(G, X), S_2(\tilde{Y}_2 | G, X)\}\right]$$
$$+ \sum_{k \in \{2,4\}} I(\tilde{\epsilon}_2 = k) \log\left[\mathcal{C}_\gamma^{11}\{\hat{F}_{11}(\tilde{Y}_1 | G, X)/\hat{p}(G, X), S_2(\tilde{Y}_2 | G, X)\} S_2(\tilde{Y}_2 | G, X) \lambda_k(\tilde{Y}_2 | G, X)\right] \quad (6)$$

and $\hat{\theta}_1$ and $\hat{\theta}_3$ are obtained from the first stage.

4 Simulation Study

We conducted a simulation study to evaluate the performance of our proposed successive competing risks model by examining the accuracy and precision of estimates of model parameters and penetrance functions. We simulated samples of 781 families with structures and inclusion criteria similar to those of the Lynch Syndrome families from the Colon Cancer Family Registry (Colon CFR). For each family member, the times to the first and second events of interest were generated in the presence of competing events based on the proposed model assuming Weibull baseline hazard functions and a Clayton copula, with parameters estimated from the Colon CFR's data in order to mimic realistic disease risks. We considered 0% (no missing), 50% and 80% of missing genotypes among family members of the probands for studying the impact of missing genotypes. For each genotype missing rate, we generated 1000 samples and for each generated sample, we estimated the parameters of the model and deduced plug-in estimators for the penetrance functions for the first and second cancers. We fitted the simulated data assuming various forms for the baseline hazard functions: parametric Weibull, log-logistic, and gamma distributions and piecewise constant hazards where λ_{01} and λ_{03} were assumed to be constant within the intervals $(0, 5], (5, 10], \cdots, (60, \infty)$ and λ_{02} and λ_{04} within $(0, 5], \cdots, (30, \infty)$.

Table 1: Accuracy and precision of estimates of log relative risks and penetrance for mutation carriers by age 70 for the first cancer, $\mathcal{P}_1(70; X)$, given gender X, male (M) and female (F) based on 1000 simulations of sample size of 781 families. For each simulation, data were generated assuming Weibull baselines, and different baseline distributions assumptions were applied for fitting the data.

Baseline		True	No missing genotypes			50% Missing genotypes			80% Missing genotypes		
distribution		value	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
Weibull	β_{1sex}	0.3706	0.0073	0.1399	0.1401	0.0187	0.1606	0.1617	0.0574	0.1947	0.2030
	β_{1gene}	3.5206	0.0182	0.2256	0.2264	-0.0028	0.2801	0.2802	-0.1273	0.4026	0.4223
Log-logistic	β_{1sex}	0.3706	0.0100	0.1488	0.1491	0.0123	0.1587	0.1591	0.0503	0.1998	0.2061
	β_{1gene}	3.5206	0.0109	0.2394	0.2396	-0.0037	0.2891	0.2891	-0.1253	0.4411	0.4585
Penetrance for the first cancer by age 70											
Weibull	$\mathcal{P}_1(70; M)$	0.6250	0.0001	0.0177	0.0177	0.0010	0.0175	0.0176	-0.0017	0.0179	0.0179
	$\mathcal{P}_1(70; F)$	0.4922	-0.0015	0.0460	0.0460	-0.0044	0.0533	0.0535	-0.0188	0.0628	0.0655
Log-logistic	$\mathcal{P}_1(70; M)$	0.6250	0.0007	0.0239	0.0239	0.0000	0.0172	0.0172	-0.0020	0.0177	0.0178
	$\mathcal{P}_1(70; F)$	0.4922	-0.0019	0.0502	0.0502	-0.0037	0.0526	0.0527	-0.0171	0.0648	0.0670

SE is empirical standard error; RMSE is root mean square error.

5 Conclusion

Our simulation studies demonstrated the good performances of our approach in terms of bias and precision of the estimates of interest. For the first event, the estimation of covariate effects (gender, mutation status) and penetrance function was quite robust to the presence of missing genotypes, misspecification of the baseline and familial ascertainment. For the second event, although we noted larger biases of the covariate effects when the baseline hazard function was misspecified, the estimation of the penetrance function was generally unbiased even in the presence of missing genotypes. This is an important result since our main interest is in this penetrance function for the second event. Application of the method to LS families will be further discussed during the presentation.

Bibliography

[1] Putter H., Fiocco W., Geskus, R.B. (2007) Tutorial in biostatistics: Competing risks and multi-state models. Stat. Med. *26*: 2389-2430.

[2] Lakhal-Chaieb, M. L., Rivest, L.-P., Abdous, B. (2006). Estimating survival under dependent truncation. Biometrika *93*, 655-669.