Web-based Supplementary Materials for "Modelling of Successive Cancer Risks in Lynch Syndrome Families in the presence of competing risks using Copulas"

By Yun-Hee Choi, Laurent Briollais, Aung K Win, John Hopper, Dan Buchanan, Mark Jenkins, and Lajmi Lakhal-Chaieb

Web Appendix A: Testing partial independence given by (2)

(i) Background. Consider a random intercept logistic regression

$$P(W = 1|b, \nu, Z) = L(b + \phi^{-1}Z^\top \nu),$$

where W is a binary response variable, $L(u) = e^u/(1 + e^u)$ is the inverse logit function, b is a random intercept with mean 0, $\phi \in [0, 1]$ is a known scale parameter, Z is a set of covariates including an intercept and ν is a vector of corresponding regression coefficients. According to Parzen et al. (2011), the marginal probability of success is

$$P(W = 1|\nu, Z) = \int_{-\infty}^{\infty} L(b + \phi^{-1}Z^\top \nu)\psi_\phi(b)db = L(Z^\top \nu),$$

when b follows a bridge distribution whose density function

$$\psi_\phi(b) = \frac{1}{2\pi}\frac{\sin(\phi\pi)}{\cosh(\phi b) + \cos(\phi\pi)}, \ \ (-\infty < b < \infty),$$

is indexed by ϕ. An interesting feature of this distribution is that b converges to 0 when $\phi \to 1$, which corresponds to $P(W = 1|b, \nu, Z) = P(W = 1|\nu, Z)$.

In our application, the conditional and marginal probabilities of success are respectively

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

and

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

where $A(y_2|G,X) = \log\{\lambda_2(y_2|G,X)/\lambda_4(y_2|G,X)\}$ and these probabilities satisfy

$$P(\epsilon_2 = 2|Y_2 = 2, \epsilon_1 = 1, G, X) =$$

$$\int_0^\infty P(\epsilon_2 = 2|Y_1 = y_1, Y_2 = y_2, \epsilon_1 = 1, G, X) f_1(y_1|\epsilon_1 = 1, Y_2 = y_2, G, X) dy_1. \quad \text{(A.2)}$$

(ii) Modelling partial dependence. When equation (2) does not hold, equations (A.1) and (A.2) prompts us to assume

$$P(\epsilon_2 = 2|Y_1 = y_1, Y_2 = y_2, \epsilon_1 = 1, G, X) = L\{b(y_1, y_2|G,X) + \phi^{-1}A(y_2|G,X)\},$$

where

$$\begin{aligned} b(y_1, y_2|G,X) &= \Psi_\phi^{-1}\left[P(Y_1 \le y_1|\epsilon_1 = 1, Y_2 = y_2, G, X)\right] \\ &= \Psi_\phi^{-1}\left[\mathcal{C}_\gamma^{01}\left\{F_{11}(y_1|G,X)/p(G,X), S_2(y_2|G,X)\right\}\right], \end{aligned}$$

and

$$\Psi_\phi^{-1}(u) = \frac{1}{\phi}\log\left[\frac{\sin(\phi\pi u)}{\sin\{\phi\pi(1-u)\}}\right]$$

is the inverse of the cumulative distribution function of a bridge distribution. The estimation procedure for the model with the added parameter ϕ is obtained by replacing equation (6) by

$$\begin{aligned} & l_2(\theta_2, \theta_4, \gamma|\hat\theta_1, \hat\theta_3, \tilde Y_1, \tilde Y_2, \tilde\epsilon_2, G, X) = I(\tilde\epsilon_2 = 0)\log\left[\mathcal{C}_\gamma^{10}\{\hat F_{11}(\tilde Y_1|G,X)/\hat p(G,X), S_2(\tilde Y_2|G,X)\}\right] \\ & + \sum_{k\in\{2,4\}} I(\tilde\epsilon_2 = k)\log\left[\mathcal{C}_\gamma^{11}\{\hat F_{11}(\tilde Y_1|G,X)/\hat p(G,X), S_2(\tilde Y_2|G,X)\}\right] \\ & + \sum_{k\in\{2,4\}} I(\tilde\epsilon_2 = k)\log\left[S_2(\tilde Y_2|G,X)h_2(\tilde Y_2|G,X)L\{b(y_1, u|G,X) + \phi^{-1}A(u|G,X)\}\right]. \end{aligned}$$

(iii) Testing partial independence. Testing the partial independence under the model assumed above consists of testing $H_0 : \phi = 1$ versus $H_1 : \phi < 1$. Our test statistics is $\hat\phi$ and its p-value is computed using a parametric bootstrap procedure that works following these steps:

Step 1: Compute the maximum likelihood estimators $\{\hat\theta_1, \hat\theta_2, \hat\theta_3, \hat\theta_4, \hat\gamma, \hat\phi\}$ from the original data.

Step 2: Compute the maximum likelihood estimators $\{\hat{\theta}_1^{(0)}, \hat{\theta}_2^{(0)}, \hat{\theta}_3^{(0)}, \hat{\theta}_4^{(0)}, \hat{\gamma}^{(0)}\}$ from the original data under the null hypothesis $H_0 : \phi = 1$.

Step 3: For $m = 1, \cdots, M$, generate a sample with same features as the original dataset using the parameters $\{\theta_1 = \hat{\theta}_1^{(0)}, \theta_2 = \hat{\theta}_2^{(0)}, \theta_3 = \hat{\theta}_3^{(0)}, \theta_4 = \hat{\theta}_4^{(0)}, \phi = 1\}$ and compute $\hat{\phi}_m$, the maximum likelihood estimator of ϕ from the m^{th} generated sample.

Step 4: The p-value is then equal to $\sum_{m=1}^{M} I(\hat{\phi}_m < \hat{\phi})/M$.

The procedure that generates samples under the null hypothesis $H_0 : \phi = 1$ in Step 3 is detailed below.

(iv) Algorithm to generate data under the partial independence assumption $\phi = 1$. For the i^{th} family, given $\{(a_{ij}, X_{ij}), j = 1, \cdots, n_i\}$, the current age and sex of each of its members, we generate $\{(G_{ij}, \tilde{Y}_{1ij}, \tilde{\epsilon}_{1ij}, \tilde{Y}_{2ij}, \tilde{\epsilon}_{2ij}), j = 1, \cdots, n_i\}$, the genotype and the observed event times following these steps:

Step 1 - Proband:

1-a) Generate G_{i1}, the genotype of the proband, from a Bernoulli distribution with a probability of success equal to $P(G_{i1} = 1|\tilde{Y}_{1i1} < a_{i1}, \epsilon_{1i1} = 1, X_{i1})$. This probability is computed using Bayes rule.

1-b) Generate Y_{1i1} from the conditional distribution function $P(Y_{1i1} \leq y|Y_{1i1} < a_{i1}, \epsilon_{1i1} = 1, G_{i1}, X_{i1})$. This probability is also computed using Bayes rule.

1-c) Generate Y_{2i1} from the conditional survival function $P(Y_{2i1} > y|Y_{1i1} = y_{1i1}, \epsilon_{1i1} = 1, G_{i1}, X_{i1})$. Generate ϵ_{2i1} from a Bernoulli distribution, which equals to 2 with probability $P(\epsilon_{2i1} = 2|Y_{2i1} = y_{2i1}, \epsilon_{1i1} = 1, G_{i1}, X_{i1})$ and to 4, otherwise.

1-d) The observed data for the proband is

$$\{G_{i1}, \tilde{Y}_{1i1} = y_{1i1}, \tilde{\epsilon}_{1i1} = 1, \tilde{Y}_{2i1} = \min(y_{2i1}, a_{i1} - y_{1i1}), \tilde{\epsilon}_{2i1} = I(y_{2i1} < a_{i1} - y_{1i1}) \times \epsilon_{2i1}\}.$$

Step 2 - Other members of the family: For $j = 2, \cdots, n_i$,

2-a) Generate G_{ij} from a Bernoulli distribution with a probability of success equal to $P(G_{ij=1}|G_{i1})$. This probability depends only on the relationship between the proband and the j^{th} member of the family and is computed using the Mendelian inheritance rule.

2-b) Generate Y_{1ij} from the survival function $P(Y_{1ij} > y|G_{ij}, X_{ij})$.

2-c) Generate ϵ_{1ij} from a Bernoulli distribution, which equals to 1 with probability $P(\epsilon_{1ij} = 1|Y_{1ij} = y, G_{ij}, X_{ij})$ and to 3, otherwise.

2-d) If $Y_{1ij} < a_{ij}$ and $\epsilon_{1ij} = 1$, generate Y_{2ij} from the conditional survival function $P(Y_{2ij} > y|Y_{1ij} = y_{1ij}, \epsilon_{1ij} = 1, G_{ij}, X_{ij})$ and generate ϵ_{2ij} from a Bernoulli distribution, which equals to 2 with probability $P(\epsilon_{2ij} = 2|Y_{2ij} = y_{2ij}, \epsilon_{1ij} = 1, G_{ij}, X_{ij})$ and to 4, otherwise.

2-e) Generation of missing genotypes: Set $\tilde{G}_{ij} = -1$ with a probability equal to a predetermined missing rate for the genotypes. Otherwise, set $\tilde{G}_{ij} = G_{ij}$.

2-f) The observed data for the j^{th} member of the family is

$$\{\tilde{G}_{ij}, \tilde{Y}_{1ij} = \min(y_{1ij}, a_{ij}), \tilde{\epsilon}_{1ij} = \epsilon_{1ij} \times I(y_{1ij} < a_{ij}),$$
$$\tilde{Y}_{2ij} = I(\tilde{\epsilon}_{1ij} = 1) \times \min(y_{2ij}, a_{ij} - y_{1ij}), \tilde{\epsilon}_{2ij} = I(\tilde{\epsilon}_{1ij} = 1) \times I(y_{2ij} < a_{ij} - y_{1ij}) \times \epsilon_{2ij}\}.$$

Web Appendix B: Derivation of the penetrance function for the second cancer

The penetrance function for the second cancer is:

$$\begin{aligned}\mathcal{P}_2(y_2; y_1, G, X) &= P(Y_2 \leq y_2, \epsilon_2 = 2|Y_1 = y_1, \epsilon_1 = 1, G, X) \\ &= \int_0^{y_2} f_{2,\epsilon_2|1}(u, \epsilon_2 = 2|Y_1 = y_1, \epsilon_1 = 1, G, X)du \\ &= \int_0^{y_2} \frac{P(\epsilon_2 = 2|Y_1 = y_1, Y_2 = u, \epsilon_1 = 1, G, X) f_{12}(y_1, u|\epsilon_1 = 1, G, X)}{f_1(y_1|\epsilon_1 = 1, G, X)} du,\end{aligned}$$

where the generic notation f refers to conditional density functions. Assuming equation (2) and employing standard manipulations of copula models yields equation (3).

Web Appendix C: testing the proportional hazards assumption

A standard way to test the proportionality assumption is to fit a model with an additional age-dependent parameter $\lambda_k(y|G,X) = \lambda_{0k}(y)e^{\beta_k^\top X+\beta_{g_k}G+\beta_y G\times\log(y)}$ and test $H_0 : \beta_y = 0$. Under the Weibull baseline hazard model, one has $\lambda_{0k}(y) = \rho_k\nu_k y^{\rho_k-1}$ and therefore $\lambda_k(y|G,X) = \rho_k\nu_k y^{\rho_k+G\beta_y-1}e^{\beta_k^\top X+\beta_{g_k}G}$ and

$$\Lambda_k(y) = \frac{\rho_k\nu_k y^{\rho_k+G\beta_y}e^{\beta_k^\top X+\beta_{g_k}G}}{\rho_k+G\beta_y}.$$

Appendix D: No competing risks model

The "No competing risks model" treats failure times corresponding to T_3 and T_4 as independent right-censored observations. This model assumes that:

(i) The pair (T_3, T_4) is independent of (T_1, T_2) given the covariates X and G.

(ii) The marginal distributions of T_1 and T_2 follow standard proportional hazard models so that $P(T_k > t_k|G,X) = e^{-\Lambda_k^T(t|G,X)}$, where $\Lambda_k^T(t|G,X) = \int_0^t \lambda_k^T(u|G,X)du$ is the conditional cumulative hazard function of T_k, given the covariates $k = 1, 2$.

(iii) The joint distribution of (T_1, T_2) follows a semi-survival copula $\mathcal{C}_\gamma$.

Under these assumptions, one has

$$P(T_1 \le t_1, T_2 > t_2|G,X) = \mathcal{C}_\gamma\left\{1-e^{-\Lambda_1^T(t_1|G,X)}, e^{-\Lambda_2^T(t_2|G,X)}\right\}$$

The penetrance functions for the first and second CRC cancers are then $P(T_1 \le t_1|G,X)$ and

$$P(T_2 \le t_2|T_1 = t_1, G, X) = 1 - \mathcal{C}_\gamma^{10}\left\{1-e^{-\Lambda_1^T(t_1|G,X)}, e^{-\Lambda_2^T(t_2|G,X)}\right\},$$

respectively. Note that unlike the penetrance functions defined in Section (2.3), which are (conditional) cause-specific cumulative incidence functions, those specified above are (conditional) survival functions. Therefore, extreme care must be taken by practitioners while estimating, comparing and interpreting these quantities. The estimation procedure for the parameters of the no-competing-risks model is obtained by replacing equations (4), (5) and (6) respectively by

$$l_1(\theta_1, \theta_3|\tilde{Y}_1, \delta_1 G, X) = I(\tilde{\epsilon}_1 = 1)\log\{\lambda_1^T(\tilde{Y}_1|G, X)\} - \Lambda_1^T(\tilde{Y}_1|G, X),$$

$$l_c(\theta_1, \theta_3|a, G, X) = \log\{1 - e^{-\Lambda_1^T(a|G,X)}\},$$

and

$$l_2(\theta_2, \theta_4, \gamma|\hat{\theta}_1, \hat{\theta}_3, \tilde{Y}_1, \tilde{Y}_2, \delta_2, G, X) = I(\tilde{\epsilon}_2 \neq 2)\log\left[\mathcal{C}_\gamma^{10}\left\{1 - e^{-\Lambda_1^T(\tilde{Y}_1|G,X)}, e^{-\Lambda_2^T(\tilde{Y}_2|G,X)}\right\}\right]$$
$$+ I(\tilde{\epsilon}_2 = 2)\log\left[\mathcal{C}_\gamma^{11}\left\{1 - e^{-\Lambda_1^T(\tilde{Y}_1|G,X)}, e^{-\Lambda_2^T(\tilde{Y}_2|G,X)}\right\}\lambda_2^T(\tilde{Y}_2|G, X)e^{-\Lambda_2^T(\tilde{Y}_2|G,X)}\right].$$

Web Appendix E: Descriptive statistics and additional analysis results for the Lynch syndrome families data

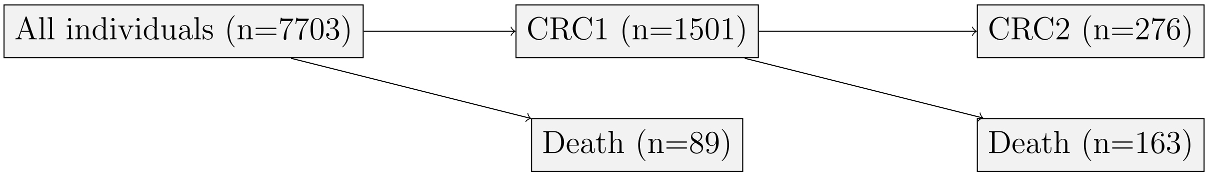

Figure 1: Successive cancers and competing events observed in 781 Lynch syndrome families identified from the Colon CFR

Table 1: Summary of data for 781 Lynch syndrome families

	Carrier	CRC1		CRC2		DEATH1[†]		DEATH2[‡]		No Events	
Gene	status	M	F	M	F	M	F	M	F	M	F
MLH1	+	180	167	46	34	1	6	6	7	66	138
	−	3	3	0	0	0	0	0	0	120	158
	NA	138	103	23	19	11	11	31	17	787	760
MSH2	+	185	199	42	52	3	4	9	10	129	207
	−	3	8	0	0	0	0	0	0	164	218
	NA	164	134	23	22	14	24	39	24	1039	936
MSH6	+	46	31	5	2	0	2	3	2	41	72
	−	2	0	0	0	0	0	2	0	40	44
	NA	31	26	1	5	1	6	7	2	343	313
PMS2	+	31	26	0	0	1	0	2	0	19	26
	−	1	2	0	0	0	0	0	1	16	28
	NA	2	6	0	0	3	2	0	0	228	173
EPCAM	+	3	2	1	0	0	0	0	70	1	3
	−	0	0	0	0	0	0	0	0	5	3
	NA	2	3	1	0	0	0	0	1	23	13
Any[*]	+	445	425	94	88	5	12	20	19	256	446
	−	9	13	0	0	0	0	2	1	345	451
	NA	337	272	48	46	29	43	77	44	2420	2195
Total		1501		276		89		163		6113	

[†] death before CRC1 due to other LS related cancers, a competing event for CRC1

[‡] death after CRC1 due to other LS related cancers, a competing event for CRC2

[*] any mutation in MLH1, MSH2, MSH6, PMS2, EPCAM genes

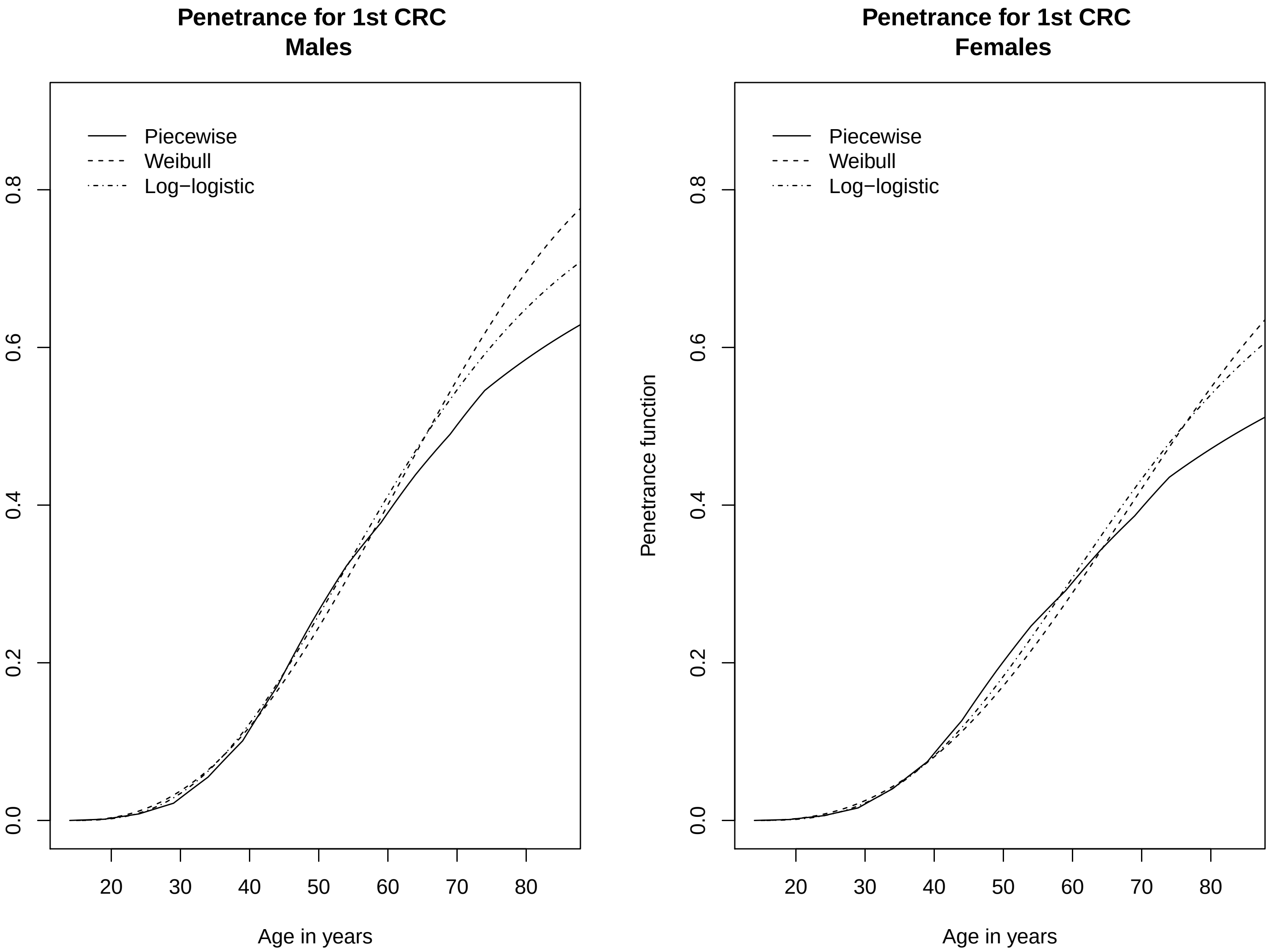

Figure 2: LS Cancer data: Penetrance estimates for first CRC among male and female mutation carriers, assuming different baseline hazard functions

Bibliography

1. Parzen, M. and Ghosh, S. and Lipsitz, S. and Sinha, D. and Fitzmaurice, G. M. and Mallick, B.K. and Ibrahim, J. G. (2011). A generalized linear mixed model for longitudinal binary data with a marginal logit link function. *The Annals of Applied Statistics* **5**, 449–467.