WEB APPENDIX A

In prospective data under the shared frailty model, the estimation for $\Lambda_0(t)$ follows the idea of Johansen (1983) and gives an estimator that can be viewed as if it were obtained through a direct maximization of the expected complete likelihood with respect to jump sizes $\Delta\Lambda_0(t)$. If we attempt to directly maximize the complete likelihood from our retrospective data with respect to $\Delta\Lambda_0(t)$, we obtain the following expression:

$$\Lambda_0(t) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^t \frac{dN_{ij}(u)}{S_0(u;\Lambda_0,\beta,\theta)},$$

where $S_0(u;\Lambda_0,\beta,\theta)$ is given by

$$\begin{aligned} S_0(u;\Lambda_0,\beta,\theta) &= \sum_{i=1}^{n}\sum_{j=1}^{n_i}\overline{\omega_i}Y_{ij}(u)\exp(\beta' Z_{ij}) + \sum_{i=1}^{n}\overline{\omega_i}Y_{i0}(u)\exp(\beta' Z_{i0}) \\ &\quad -\sum_{i=1}^{n}\frac{1+\theta\delta_{i0}}{1+\theta\Lambda_0(X_{i0})\exp(\beta' Z_{i0})}Y_{i0}(u)\exp(\beta' Z_{i0}). \end{aligned}$$

Baseline hazard function $\Lambda_0(t)$ is involved on both sides of the equation, therefore we cannot directly obtain a closed-form estimator. However, if we first evaluate the part that involves $\Lambda_0(t)$ on the right hand side at the current parameter estimates, we can use the above expression to obtain an updated $\widehat{\Lambda}_0(t)$. Note that $\{1+\widehat{\theta}\delta_{i0}\}/\{1+\widehat{\theta}\Lambda_0(X_{i0})\exp(\widehat{\beta}' Z_{i0})\}$ is the conditional expectation of the frailty given the data from the proband. Denote it by $\overline{\omega_{i0}}$, the proposed estimator for $\Lambda_0(t)$ is

$$\widehat{\Lambda}_0(t) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^t \frac{1}{S(u;\,\widehat{\beta})}dN_{ij}(u), \tag{1}$$

where $S(u;\,\widehat{\beta})$ is given by

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\overline{\omega_i}Y_{ij}(u)\exp(\widehat{\beta}' Z_{ij}) + \sum_{i=1}^{n}\{\overline{\omega_i}-\overline{\omega_{i0}}\}Y_{i0}(u)\exp(\widehat{\beta}' Z_{i0}).$$

This formula gives a closed form expression for updating $\widehat{\Lambda}_0(t)$. In this estimator, the first term in $S(u;\,\widehat{\beta})$ represents those relatives who are still at risk at time u. The second term in $S(u;\,\widehat{\beta})$ comes from those probands who are at-risk, and for each proband the function is proportional to the difference between the expected frailty conditional on variables from both the proband and the relatives and the expected frailty conditional on variables only from the proband. The two expectations are close to each other on average, therefore the contribution from the relatives in the first term dominates $S(u;\,\widehat{\beta})$.

WEB APPENDIX B

Define $f_{ij}(g_{i0}) = P(g_{ij} = 1 \,|\, g_{i0};\, q)$, where q is the allele frequency for the high risk allele. The function $f_{ij}(g_{i0})$ is the probability of the relative carrying the high risk allele conditional on the carrier-noncarrier status of the proband. This probability depends on the relationship between the two individuals and the allele frequency q. One easy way to calculate such probability is to use the transitional matrices proposed in Li and Sacks (1954).

In the ECM algorithm, at current parameter estimates, denote

$$\begin{aligned}
a_{ij} &= 1/\widehat{\theta} + \delta_{i0} + \delta_{ij}, \\
b_{ij} &= 1/\widehat{\theta} + \widehat{\Lambda}_0(X_{i0}) \exp(\widehat{\gamma}' Z_{i0} + \widehat{\beta} g_{i0}) + \widehat{\Lambda}_0(X_{ij}) \exp(\widehat{\gamma}' Z_{ij}) \exp(\widehat{\beta}), \\
c_{ij} &= 1/\widehat{\theta} + \widehat{\Lambda}_0(X_{i0}) \exp(\widehat{\gamma}' Z_{i0} + \widehat{\beta} g_{i0}) + \widehat{\Lambda}_0(X_{ij}) \exp(\widehat{\gamma}' Z_{ij}), \\
d_{ij} &= \exp\{(\widehat{\gamma}' Z_{i0} + \widehat{\beta} g_{i0})\delta_{i0} + (\widehat{\gamma}' Z_{ij} + \widehat{\beta})\delta_{ij}\}, \\
e_{ij} &= \exp\{(\widehat{\gamma}' Z_{i0} + \widehat{\beta} g_{i0})\delta_{i0} + \widehat{\gamma}' Z_{ij}\delta_{ij}\},
\end{aligned}$$

the expectations involving missing data are

$$\begin{aligned}
\tilde{E}\{g_{ij}\} &= \frac{d_{ij} b_{ij}^{-a_{ij}} f_{ij}(g_{i0})}{d_{ij} b_{ij}^{-a_{ij}} f_{ij}(g_{i0}) + e_{ij} c_{ij}^{-a_{ij}} \{1 - f_{ij}(g_{i0})\}}, \\
\tilde{E}\{\omega_i\} &= a_{ij} \frac{d_{ij} b_{ij}^{-(a_{ij}+1)} f_{ij}(g_{i0}) + e_{ij} c_{ij}^{-(a_{ij}+1)} \{1 - f_{ij}(g_{i0})\}}{d_{ij} b_{ij}^{-a_{ij}} f_{ij}(g_{i0}) + e_{ij} c_{ij}^{-a_{ij}} \{1 - f_{ij}(g_{i0})\}}, \\
\tilde{E}\{\ln \omega_i\} &= \phi(a_{ij}) - \frac{d_{ij} b_{ij}^{-a_{ij}} \ln(b_{ij}) f_{ij}(g_{i0}) + e_{ij} c_{ij}^{-a_{ij}} \ln(c_{ij}) \{1 - f_{ij}(g_{i0})\}}{d_{ij} b_{ij}^{-a_{ij}} f_{ij}(g_{i0}) + e_{ij} c_{ij}^{-a_{ij}} \, (1 - f_{ij}(g_{i0}))}, \\
\tilde{E}\{\omega_i \exp(\beta g_{ij})\} &= a_{ij} \frac{\exp(\beta) d_{ij} b_{ij}^{-(a_{ij}+1)} f_{ij}(g_{i0}) + e_{ij} c_{ij}^{-(a_{ij}+1)} \{1 - f_{ij}(g_{i0})\}}{d_{ij} b_{ij}^{-a_{ij}} f_{ij}(g_{i0}) + e_{ij} c_{ij}^{-a_{ij}} \{1 - f_{ij}(g_{i0})\}}.
\end{aligned}$$

References

Li, C. C. and Sacks, L. (1954). The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics* **10**, 347–360.