Published in final edited form as:
Biostatistics (2021), **22** (3): 482–503
DOI: 10.1093/biostatistics/kxz048

The illness-death model for family studies

JOOYOUNG LEE

Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada
E-mail: j463lee@uwaterloo.ca

RICHARD J. COOK

Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada

Summary

Family studies involve the selection of affected individuals from a disease registry who provide right-truncated ages of disease onset. Coarsened disease histories are then obtained from consenting family members, either through examining medical records, retrospective reporting, or clinical examination. Methods for dealing with such biased sampling schemes are available for continuous, binary, and failure time responses, but methods for more complex life history processes are less developed. We consider a simple joint model for clustered illness-death processes which we formulate to study covariate effects on the marginal intensity for disease onset and to study the within-family dependence in disease onset times. We construct likelihoods and composite likelihoods for family data obtained from biased sampling schemes. In settings where the disease is rare and data are insufficient to fit the model of interest, we show how auxiliary data can augment the composite likelihood to facilitate estimation. We apply the proposed methods to analyze data from a family study of psoriatic arthritis carried out at the University of Toronto Psoriatic Arthritis Registry.

Keywords: Augmented likelihood, biased sample, dependence measures, family study, illness-death model

1 Introduction

Studies are often conducted to assess the nature and extent of familial aggregation of disease and to study the effect of genetic risk factors for disease onset. When present, familial aggregation suggests a shared genetic or environmental basis of disease (Li et al., 1998; Liang and Beaty, 2000). For valid inference in such settings, however, it is important to address the sampling scheme by which families are recruited. This is typically done by identifying an individual with the disease, called the proband, from a disease registry, recruiting them to the study and recording a detailed disease history including the age of onset. The age of disease onset for the proband is right-truncated since they were selected from a disease registry, and their survival time is left-truncated. Family members of the proband, called non-probands, are then contacted and upon granting consent are selected for the family study and their disease histories are recorded. In some settings, the proband may report the disease histories of their family members, but it may alternatively be acquired through clinical examination conducted by a physician; the latter is preferable when diseases are difficult to diagnose and was done in the motivating study.

A variety of frameworks for the analysis of multivariate failure time data have been developed (Hougaard, 2012). The marginal approach for the analysis of clustered failure time data has been developed in general by Lee et al. (1992) and by Liang et al. (1993) which can be used for family studies if biased sampling schemes are not employed. Clayton (1978) suggested use of the cross-ratio as a dependence measure, and Oakes (1989) showed the connection between frailty models and the cross-ratio hazard function. Frailty models have been widely used in the analysis of case-control family studies (Hsu et al., 2004; Hsu and Gorfine, 2005) where a frailty variance is interpreted as a measure of dependence in the age of onset within family members. Copula models can alternatively be used, in which case the multivariate joint distribution is formulated in terms of the marginal distributions and a copula function (Joe, 1997; Shih and Louis, 1995). Li et al. (1998), Shih and Chatterjee (2002), and Chatterjee et al. (2006) developed the copula models for case-control family studies considering the ascertainment of case-control probands. Zhong and Cook (2016) used copula functions and composite likelihood for the analysis of a combination of right-censored and current status family data while addressing complex sampling schemes; Zhong and Cook (2017) developed related methods based on estimating functions.

The aforementioned methods focus on modeling familial aggregation in disease onset times in the simple framework of failure time models. More recent work has dealt with clustered failure time data in the semi-competing risks setting, where disease onset and disease-free death are considered as competing events. Bandeen-Roche and Liang (2002) suggested a modified conditional hazard ratio to account for the cause of failure based on a frailty model and applied it to a population cohort study of dementia. Shih and Albert (2010) extended the work of Bandeen-Roche and Liang (2002) and considered two types of dependence measures with one to model the dependence in terms of the failure time of paired members and a second to model the association between the failure types given the time; they suggested use of a time-varying piecewise constant dependence measure. To examine sibship association in disease onset, Cheng et al. (2009) developed nonparametric association analysis using the bivariate cumulative incidence function defined by the cause-specific hazard function to account for the exchangeable clustered competing risks setting. Zhou et al. (2012) proposed a marginal proportional subdistribution hazard model in the clustered competing risks setting. Scheike et al. (2010) and Scheike and Sun (2012) studied a semiparametric additive model and explored a cross-odds ratio-type measure on the probability scale as the association parameters for the Danish twin data; Scheike et al. (2014) extended the model to accommodate delayed entry and to model genetic and environmental effects.

Multistate models offer another framework for dependence modeling. Aalen et al. (1980) applied the Schweder (1970) concept of local dependence to understand the interaction between two life-history events by comparing the transition intensities. Hougaard et al. (1992) and Hougaard (1999) considered dependence modeling in the lifetimes of twins via multistate models under the Markov or semi-Markov assumption.

There has been little work on the use of illness-death models in the setting of family or twin studies. The illness-death model offers a useful framework for the joint study of disease onset and mortality to better understand the nature of the disease process over an individual lifetime (Andersen, 1988). Dependence modeling for correlated illness-death processes is necessary when processes are clustered as they are in family studies. Jiang and Haneuse (2017) proposed an illness-death model with the non-parametric frailty distribution where the non-terminal event times and terminal event times are correlated. Cederkvist et al. (2018) considered the cause-specific cumulative incidence function as a basis for dependence modeling in the multivariate competing risks settings; these authors used random effects to accommodate within-cluster dependence in both risk and timing.

In this article, we develop an illness-death model using the latent variable formulation of the competing risk model for the first event (disease onset or disease-free death). A copula model is used to accommodate clustering within families in the (possibly latent) ages of disease onset. Methods are described which account for incomplete data under two types of biased sampling schemes. The use of auxiliary data is highlighted to address identifiability problems and to increase efficiency. Finally, we show how to account for incomplete genetic data when auxiliary data do not contain the desired genotype information.

The remainder of this article is organized as follows. In Section 2, we define notation and present the joint model. Two biased sampling schemes are then described and the associated likelihood is presented; composite likelihood is proposed for settings where some family sizes are large. The use of auxiliary data is discussed in Section 3 to facilitate estimation of transition intensities to the death state, and simulation studies are reported in Section 4. In Section 5, we extend the proposed methods to incorporate genotype information and present the results of further simulation studies. An application to a family study on the onset of psoriatic arthritis (PsA) from the University of Toronto is given in Section 6, and concluding remarks are given in Section 7.

2 Model Formulation

2.1 Notation and Model Formulation

We consider a four-state representation of the illness-death model to describe the joint distribution of disease onset and death (Datta et al., 2000; Xu et al., 2010). We let state 0 represent a healthy state, state 1 represent a diseased state, state 2 represent death post-disease, and state 3 represent disease-free death; see Figure 1.

Our initial interest lies in modeling the association in the age of disease onset between family members. To simplify the presentation of the joint model, we first consider dependence modeling for two individuals labeled j and k in family i, and define variables for individual j without loss of generality. We let X_{ij1} denote the age of disease onset, X_{ij2} the age at death following disease, X_{ij3} the age at disease-free death. This is a latent variable formulation of the competing risks problem for transition out of state 0 in that X_{ij1} may not be observed (or realized) if $X_{ij3} < X_{ij1}$. While unconventional and not without limitations vis-à-vis observable features, we adopt this formulation since the association in the age of disease onset is most naturally modeled in terms of $0 \to 1$ transition times. Finally, we let B_{ij} be the calendar time

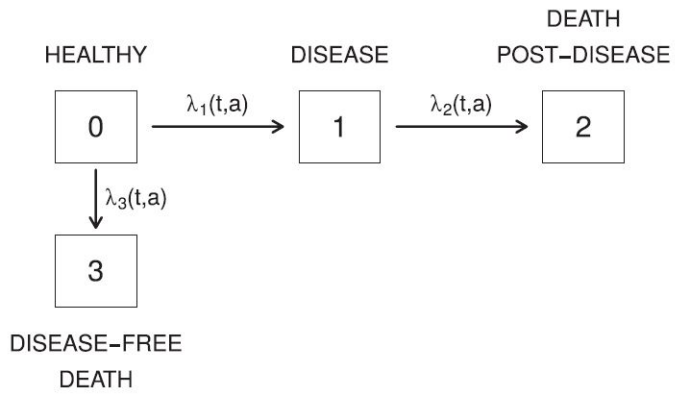

Figure 1: A four-state representation of an illness-death model.

of birth for individual j in family i, $j = 1, 2$, and let $\boldsymbol{B}_i = (B_{ij}, B_{ik})'$ be the vector of calendar times of births for individuals j and k in family i.

It will be convenient to use multistate models notation and to this end let $Z_{ij}(a)$ denote the state occupied for individual j in family i at age a and calendar time $B_{ij} + a$. We let V_{ij} denote covariates for individual j in family i and $\boldsymbol{V}_i = (V_{ij}, V_{ik})'$ be the covariate information for the (j, k) pair. Let $H_{ij}(a) = \{Z_{ij}(s), 0 \leq s < a, B_{ij}, V_{ij}\}$ denote the history for individual j in family i over age $[0, a)$ whose calendar time of birth is B_{ij}. The age- and calendar time-specific marginal intensity function for a transition from state h into l is

$$\lim_{\Delta a \downarrow 0} \frac{P(Z_{ij}(a + \Delta a^-) = l | Z_{ij}(a^-) = h, H_{ij}(a))}{\Delta a} = \lambda_{ijl}(t, a | H_{ij}(a))$$

with $t = B_{ij} + a$, where $(h, l) \in \{(0, 1), (0, 3), (1, 2)\}$. If, given the date of birth and the covariates, the disease process for each family member is Markov, we can write

$$\lambda_{ij}(t, a | H_{ij}(a)) = \lambda_l(t, a | b_{ij}, v_{ij}), \qquad l = 1, 2, 3.$$

If we assume that $\lambda_3(t, a|b_{ij}, v_{ij}) = \lambda_2(t, a|b_{ij}, v_{ij})$, the disease is incidental as it does not change the risk of death, but if $\lambda_3(t, a|b_{ij}, v_{ij}) \neq \lambda_2(t, a|b_{ij}, v_{ij})$ then survival is locally dependent of the disease process (Aalen, 2012); in this case, typically, $\lambda_2(t, a|b_{ij}, v_{ij}) > \lambda_3(t, a|b_{ij}, v_{ij})$. Andersen et al. (1985) use a Cox model to accommodate proportional mortality among diseased and disease-free individuals. In the present context, this has the form

$$\lambda_2(t, a|b_{ij}, v_{ij}) = \lambda_3(t, a|b_{ij}) \nu_0(a) \exp(v'_{ij}\beta_2), \tag{1}$$

where $\lambda_3(t, a|b_{ij})$ is the age- and calendar-time specific baseline population mortality rate. The term $\nu_0(a)$ reflects a proportional change in mortality in terms of age. We adopt the model (1) and let $\lambda_3(t, a|b_{ij}, v_{ij}) = \lambda_3(t, a|b_{ij})$ so that the subject-specific disease-free mortality rate does not depend on the covariates and can be therefore easily estimated based on the population rates. If the disease does not alter the mortality trend in age in the population, we may further assume $\nu_0(a) = \nu$.

We assume that the transition times X_{ij1} and X_{ij3} are independent, which is an unverifiable assumption as we only observe $\min(X_{ij1}, X_{ij3})$ (Kalbfleisch and Prentice, 2002). Moreover we assume that the intensity for transitions from the healthy to disease state does not depend on calendar time, in which case we may write $\lambda_1(t, a|b_{ij}, v_{ij}) = \lambda_1(a) \exp(v'_{ij}\beta_1)$. We denote the survival functions for the possibly latent disease onset time as $\mathcal{F}(a|V_{ij}; \phi_1) = 1 - \exp(-\int_0^a \lambda_1(s|V_{ij})ds)$, and let $f(a|V_{ij}; \phi_1) = -\partial\mathcal{F}(a|V_{ij}; \phi_1)/\partial a$, where ϕ_1 indexes the

marginal intensity for disease onset. To model within family association in the age of disease onset, we use a copula function to construct a joint model for X_{ij1} and X_{ik1} (Joe, 1997) in which

$$P(X_{ij1} > a_j, X_{ik1} > a_k; \boldsymbol{V}_i, \varphi) = \mathcal{C}(\mathcal{F}(a_j|V_{ij}; \phi_1), \mathcal{F}(a_k|V_{ik}; \phi_1); \rho), \tag{2}$$

with ρ indexing the copula function and $\varphi = (\phi_1', \rho)'$. We define $\phi = (\phi_1', \phi_2')'$ where ϕ_2 indexes the transition intensity from the diseased to death state and $\psi = (\phi', \rho)$. The joint density function can be written as

$$P(X_{ij1} = a_j, X_{ik1} = a_k; \boldsymbol{V}_i, \varphi) = c(\mathcal{F}(a_j|V_{ij}; \phi_1), \mathcal{F}(a_k|V_{ik}; \phi_1); \rho) f(a_j|V_{ij}; \phi_1) f(a_k|V_{ik}; \phi_1),$$

where $c(\cdot, \cdot; \rho)$ is the density of the copula. We use the Clayton copula which has the form

$$\mathcal{C}(u_1, u_2; \rho) = (u_1^{-\rho} + u_2^{-\rho} - 1)^{-1/\rho}, \quad 0 \le u_j \le 1, \quad j = 1, 2,$$

with Kendall's $\tau = \rho/(\rho + 2)$. As a measure of dependence of the age of disease onset between two individuals, we consider the cross-ratio for (X_{ij1}, X_{ik1}) (Oakes, 1989) which takes the form

$$\begin{aligned} \theta(a_j, a_k) &= \frac{\lambda_1(a_k|X_{ij1} = a_j; \boldsymbol{V}_i, \varphi)}{\lambda_1(a_k|X_{ij1} > a_j; \boldsymbol{V}_i, \varphi)} \\ &= \frac{P(X_{ij1} = a_j, X_{ik1} = a_k; \boldsymbol{V}_i, \varphi) P(X_{ij1} > a_j, X_{ik1} > a_k; \boldsymbol{V}_i, \varphi)}{P(X_{ij1} = a_j, X_{ik1} > a_k; \boldsymbol{V}_i, \varphi) P(X_{ij1} > a_j, X_{ik1} = a_k; \boldsymbol{V}_i, \varphi)}, \end{aligned} \tag{3}$$

under the Clayton copula $\theta(a_j, a_k) = 1 + \rho$. We assume that the (possibly latent) age at disease-free death for an individual is independent from the life history of other family members. This assumption may not be valid in settings where the occurrence of death might be affected by shared environmental factors in a family. While we adopt this assumption we note that a within-family dependence in the marginal time of death ($\min(X_{ij2}, X_{ij3})$) accommodated in this joint model through the dependence in the disease onset time. Under the assumption of (i) conditionally independent competing risks, $X_{ij1} \perp X_{ij3}|V_{ij}$, and (ii) $X_{ij3} \perp \{Z_{ik}(s), 0 < s\}|\boldsymbol{B}_i, \boldsymbol{V}_i$ for $j \neq k$, the cause-specific cross-ratio $\theta_{11}(a_j, a_k) = \lambda_{11}(a_k|X_{ij1} = a_j, X_{ij3} > a_j; \boldsymbol{B}_i, \boldsymbol{V}_i, \varphi)/\lambda_{11}(a_k|X_{ij1} > a_j, X_{ij3} > a_j; \boldsymbol{B}_i, \boldsymbol{V}_i, \varphi)$ for the age of disease onset between two individuals is the same as the cross-odds ratio $\theta(a_j, a_k)$ in (3). For the Clayton copula $\theta(a_j, a_k) = \theta_{11}(a_j, a_k) = \theta = 1 + \rho$.

Scheike et al. (2010) introduced a cross-odds ratio as a measure of dependence in the competing risks setting given here by

$$\pi(a) = \frac{ODDS(X_{ik1} \le a, X_{ik1} < X_{ik3}|X_{ij1} \le a, X_{ij1} < X_{ij3}; \boldsymbol{B}_i, \boldsymbol{V}_i)}{ODDS(X_{ik1} \le a, X_{ik1} < X_{ik3}, B_{ik}, V_{ik})}, \tag{4}$$

where

$$P(X_{ik1} \le a, X_{ik1} < X_{ik3}, B_{ik}, V_{ik}) \tag{5}$$

is the marginal cumulative incidence function for disease onset. Note that $\pi(a)$ is not a simple expression in terms of our model formulation even with $\theta(a, a) = 1 + \rho$ under the Clayton copula since the cumulative incidence functions are complex functions of the cause-specific hazards $\lambda_1(\cdot), \lambda_2(\cdot)$, and $\lambda_3(\cdot)$.

2.2 Likelihood Construction for Family Studies

Here, we extend the model to deal with all members of family i, $i = 1, \ldots, n_F$, where n_F is the number of families recruited. We let $m_i + 1$ denote the number of individuals in family

i with the subscript 0 used to identify the proband and selected family members by $j = 1, \ldots, m_i, i = 1, \ldots, n_F$. Let $\boldsymbol{X}_{i1} = (X_{i01}, X_{i11}, \ldots, X_{im_i1})'$ denote the vector of possibly latent onset times within family i, $\boldsymbol{X}_{i3} = (X_{i03}, X_{i13}, \ldots, X_{im_i3})'$, $\boldsymbol{X}_{i2} = (X_{i02}, X_{i12}, \ldots, X_{im_i2})'$, $\boldsymbol{B}_i = (B_{i0}, \ldots, B_{im_i})'$, and $\boldsymbol{V}_i = (V_{i0}, \ldots, V_{im_i})'$. Then (2) extends to an $m_i + 1$ dimensional survival function as

$$P(X_{i01} > a_0, \ldots, X_{im_i1} > a_{m_i} | \boldsymbol{V}_i; \varphi) = \mathcal{C}(\mathcal{F}(a_0|V_{i0}; \phi_1), \ldots, \mathcal{F}(a_{m_i}|V_{im_i}; \phi_1); \rho).$$

We consider studies in which families are sampled by first selection of the proband from a disease registry. We let R_{i0} denote the calendar time of screening and recruitment of the proband to the registry and C_{i0} the age of the proband at calendar time R_{i0}. To enter the registry at R_{i0}, the proband must be alive with disease at age C_{i0}. Let R_i be the calendar time that the proband is sampled from the registry for inclusion in the family study, and A_{i0}, and $\boldsymbol{A}_i = (A_{i0}, A_{i1}, \ldots, A_{im_i})'$ denote the age at calendar time R_i for the proband (A_{i0}) and all family members (A_{ij}, $j = 1, \ldots, m_i$), respectively; let $\boldsymbol{A}_i^- = (A_{i1}, \ldots, A_{im_i})'$ denote the elements of $\boldsymbol{A}_i$ excluding the proband. More generally a superscript "-" denotes a vector with the entry for the proband excluded.

For each recruited proband we obtain data from their consenting family members (non-probands). If a non-proband died before R_i it is often possible to obtain disease history data retrospectively from medical records or via the proband. Anderson (1961) compared the accuracy of reports about disease histories of family members with physician diagnosis and found that physician assessments were necessary to ensure accurate reporting of disease related information for non-probands. We therefore also consider designs in which physicians must interview non-probands at calendar time R_i to carry out medical examinations. In this second design, non-probands must alive at calendar time R_i if they are in family i, $i = 1, \ldots, n_F$.

The Lexis diagram plays a central role in describing the incidence, path, and sampling of disease processes in a population using a calendar time $\times$ age co-ordinate system (Keiding, 1990, 2006). Figure 2 shows possible scenarios for family data on illness-death processes under the biased sampling scheme described here. In this figure, the dashed lines represent periods of calendar time and ages at which the healthy state is occupied, and the solid lines represent periods in which the diseased state is occupied. The proband, depicted in red, provides their retrospectively recorded age of disease onset, and like other individuals in the registry may be followed until death or censoring. Non-probands may give a variety of types of data: some may report retrospectively their age of disease onset, some may be disease-free at the time of examination, and for some we may simply know their date of death if they did not live long enough to be recruited and examined at the calendar time of the family study.

Here, we construct the likelihood function for two particular study designs under the biased sampling schemes, depending on whether we collect the complete history of non-probands at R_i (design I) or only examine non-probands who are alive at R_i (design II). If $\boldsymbol{a}_i = (a_{i0}, a_{i1}, \ldots, a_{im_i})'$ denotes a vector of ages of individuals in family i, we let $\boldsymbol{Z}_i(\boldsymbol{a}_i) = (Z_{i0}(a_{i0}), Z_{i1}(a_{i1}), \ldots, Z_{im_i}(a_{im_i}))'$. In both designs, the likelihood contribution of the proband is

$$L_{i0}(\phi) = P(\bar{Z}_{i0}(A_{i0}) | Z_{i0}(C_{i0}) = 1, C_{i0}, B_{i0}, V_{i0}; \phi), \tag{6}$$

where $\bar{Z}_{i0}(A_{i0}) = \{Z_{i0}(u), 0 < u \leq A_{i0}\}$. In the first design, we suppose the disease history and covariates for all non-probands are available at calendar time R_i at which the family study is conducted. The likelihood is then given as

$$L_i^I(\psi) \propto L_{i0}(\phi) P(\bar{\boldsymbol{Z}}_i^-(\boldsymbol{A}_i^-) | \bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{A}_i, \boldsymbol{B}_i, \boldsymbol{V}_i; \psi)\,, \tag{7}$$

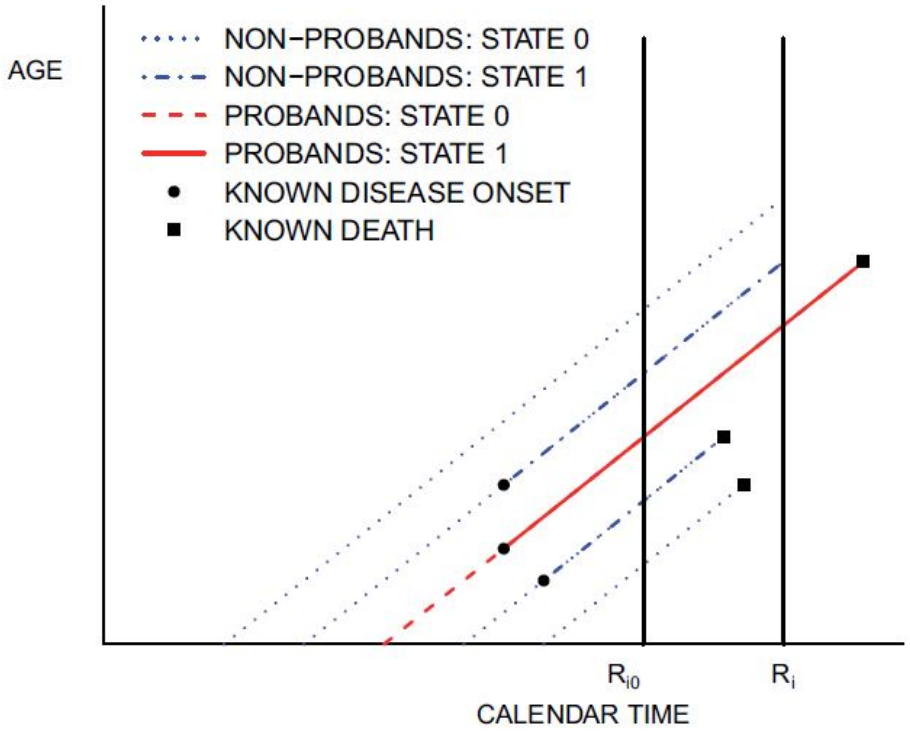

Figure 2: A Lexis diagram for family data obtained under a biased sample scheme; R_{i0} denotes the calendar time of recruitment of a proband to a registry, and R_i is the date of the family study.

where $\bar{\boldsymbol{Z}}_i(a_i) = \{Z_{ij}(u), 0 < u \leq a_{ij}, j = 0, \ldots, m_i\}$. In design II, we require non-probands to be alive at calendar time R_i in order that they be examined by a physician. This gives

$$L_i^{II}(\psi) \propto L_{i0}(\phi) P(\bar{\boldsymbol{Z}}_i^-(\boldsymbol{A}_i^-) | \bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{Z}_i^-(\boldsymbol{A}_i^-) \in \{0,1\}^{m_i}, \boldsymbol{A}_i, \boldsymbol{B}_i, \boldsymbol{V}_i; \psi). \quad (8)$$

In what follows we omit the superscript I and II indicating the design and take it as understood that L_i represents either L_i^I or L_i^{II} in a particular setting. The score vector is $S(\psi) = \sum_{i=1}^{n_F} S_i(\psi)$ where $S_i(\psi) = \partial \log L_i / \partial \psi$, and the information matrix is $I(\psi) = \sum_{i=1}^{n_F} I_i(\psi)$ where $I_i(\psi) = -\partial^2 \log L_i(\psi) / \partial \psi \partial \psi'$, respectively. We obtain the maximum likelihood estimator $\hat{\psi}$ by solving $S(\psi) = 0$ and note that asymptotically $\sqrt{n_F}(\hat{\psi} - \psi) \sim N(0, \mathcal{I}^{-1}(\psi))$ where $\mathcal{I}(\psi) = E[I_i(\psi)]$.

When m_i is large the computational burden of evaluating the joint probability of the life histories of family members may be considerable, so we consider use of "pairwise" conditional composite likelihood (Varin et al., 2011) in which pairs are comprised of two non-probands and the contribution to the pairwise likelihood condition on the proband data for the respective family. In particular, for design II where non-probands are only selected if they are alive at R_i, the contribution from a pair to (8) is much simpler than what would be required to compute $P(\bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{Z}_i^-(\boldsymbol{A}_i^-) \in \{0,1\}^{m_i}, \boldsymbol{A}_i, \boldsymbol{B}_i, \boldsymbol{V}_i)$ under a full likelihood. The contribution to the conditional composite likelihood of family i for design k is then

$$CL_i^k(\psi) \propto L_{i0}(\phi) \prod_{1 \leq j < l \leq m_i} \{L_{ijl}^k(\psi)\}^{\frac{1}{m_i - 1}}, \qquad k = I, II, \quad (9)$$

where L_{i0} is given by (6); the weight $1/(m_i - 1)$ ensures the net contribution to the composite likelihood for the marginal function for non-probands is appropriate. Specifically in design I,

$$L_{ijl}^I(\psi) = P(\bar{\boldsymbol{Z}}_{ijl}^-(\boldsymbol{A}_{ijl}^-) | \bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)$$

and in design II,

$$L_{ijl}^{II}(\psi) = P(\bar{\boldsymbol{Z}}_{ijl}^{-}(\boldsymbol{A}_{ijl}^{-})|\bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0})=1, \boldsymbol{Z}_{ijl}^{-}(\boldsymbol{A}_{ijl}^{-}) \in \{0,1\}^2, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)$$

$$= \frac{P(\bar{\boldsymbol{Z}}_{ijl}^{-}(\boldsymbol{A}_{ijl}^{-})|\bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0})=1, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)}{P(\boldsymbol{Z}_{ijl}^{-}(\boldsymbol{A}_{ijl}^{-}) \in \{0,1\}^2|\bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0})=1, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)}$$

with $\boldsymbol{A}_{ijl} = (A_{i0}, A_{ij}, A_{il})'$, $\boldsymbol{B}_{ijl} = (B_{i0}, B_{ij}, B_{il})'$, $\boldsymbol{V}_{ijl} = (V_{i0}', V_{ij}', V_{il}')'$, $\bar{\boldsymbol{Z}}_{ijl}(s_{ijl}) = \{Z_{ih}(u), 0 < u \leq s_{ih}, h = 0, j, l; \boldsymbol{B}_{ijl}\}$.

Mesfioui and Quessy (2008) showed that the conditional Clayton copula has a useful invariance property which we exploit here. In the present context, if $X_{i1} = (X_{ij1}, X_{ik1}, X_{il1})$ follows a joint distribution governed by the Clayton copula, then the distribution for $X_{ij1}, X_{ik1}|X_{il1} = x_{il1}$ also follows a Clayton copula with parameter $\rho/(1+\rho)$, so,

$$\begin{aligned} &P(X_{ij1} > a_j, X_{ik1} > a_k|X_{il1} = a_l, V_{ij}, V_{ik}, V_{il}) \\ &\quad = C(\mathcal{F}(a_j|X_{il1} = a_l, V_{ij}, V_{il}; \phi_1, \rho), \mathcal{F}(a_k|X_{il1} = a_l, V_{ik}, V_{il}; \phi_1, \rho); \rho^*), \end{aligned} \tag{10}$$

where $\rho^* = \rho/(1+\rho)$ and

$$\mathcal{F}(a_j|X_{il1} = a_l, V_{ij}, V_{il}; \phi_1, \rho) = \left.\frac{\partial C(\mathcal{F}(a_j|V_{ij}; \phi_1), u; \rho)}{\partial u}\right|_{u=\mathcal{F}(a_l|V_{il};\phi_1)}.$$

We therefore calculate $L_{ijl}^{k}(\psi)$ using the conditional Clayton copula function based on (10). Again we suppress the superscript I or II when discussing a generic setting and we write $CL_i(\psi)$ to represent (9) in either case. The score vector for the composite likelihood is then $U(\psi) = \sum_{i=1}^{n_F} U_i(\psi)$ where $U_i(\psi) = \partial \log CL_i(\psi)/\partial\psi$ and the maximum composite likelihood estimator $\tilde{\psi}$ is obtained by solving $U(\psi) = 0$. The estimated variance of $\tilde{\psi}$ is given as $n_F^{-1}A^{-1}(\tilde{\psi})B(\tilde{\psi})A^{-1}(\tilde{\psi})$ where $A(\psi) = -n_F^{-1}\left[\partial U(\psi)/\partial\psi'\right]$ and $B(\psi) = n_F^{-1}\sum_{i=1}^{n_F} U_i(\psi)U_i'(\psi)$.

More details on how to construct the composite likelihood are given in Appendix A of *Supplementary Material* available at *Biostatistics* online using the motivating example.

3 Augmented Composite likelihood

For the motivating family study, the low incidence of disease among non-probands and bias sampling scheme employed result in limited information about the disease process. To overcome this difficulty, we exploit auxiliary data to ensure all components of the model identifiable and strengthen the analysis. The combination of data from different sources in family studies has been suggested (Pfeiffer et al., 2008; Zheng et al., 2010; Balliu et al., 2012) in which data from case-control studies or the twin-based studies are integrated with data from family studies. In the current study, the University of Toronto Psoriatic Arthritis Registry (UTPAR) provides data with right-truncated disease onset times and the left-truncated and right-censored times to death (Wong et al., 1997). The research team running the UTPAR also conducts tracing studies, which aim to yield further data on survival times for PsA patients. Another source of auxiliary data is a national cross-sectional survey conducted by the National Psoriasis Foundation in the United States; it yields current status data on the disease state of individuals (Gelfand et al., 2005b). Although this national survey only provides marginal information, efficiency of our analysis can in principle be enhanced through augmenting the composite likelihood. Since we have no data available on the time to disease-free death, we use national mortality statistics to estimate the disease-free mortality rate; the data are population-level data, and so we treat $\lambda_3(\cdot,\cdot)$ as known and define them to be the population mortality rates. We thus consider (i)

registry data with follow-up (ii) a cross-sectional survey yielding current status data on disease state, and (iii) national statistics for specification of the mortality rate.

Let $\mathcal{A}_1$ be the set of n_1 individuals in the registry but not selected as probands and $\mathcal{A}_2$ the set of n_2 individuals from the cross-sectional survey. We multiply $CL^k(\psi)$ in (9) by the corresponding marginal likelihoods which are augmentation terms and so denoted AL_1 and AL_2 based on the auxiliary data from sources (i) and (ii), respectively. For individuals from the registry in $\mathcal{A}_1$, we let X_{i1} denote the age at onset, C_i the age at recruitment, X_{i2} the age at death following disease (if available), $A_i^* = \min(C_i^*, X_{i2})$ with C_i^* the last assessment time, B_i the calendar time of birth, and V_i a vector of covariates for an individual i. Then, $AL_1 = \prod_{i\in\mathcal{A}_1} AL_{1i}$ where

$$AL_{1i} \propto P(\bar{Z}_i(A_i^*)|Z_i(C_i) = 1, C_i, B_i, V_i; \phi)\,.$$

For individuals from the national survey in $\mathcal{A}_2$, we let C_i denote the age at contact. Then, $AL_2 = \prod_{i\in\mathcal{A}_2} AL_{2i}$ where

$$AL_{2i} \propto \prod_{h\in\{0,1\}} P(Z_i(C_i) = h|Z_i(C_i) \in \{0,1\}, B_i, V_i; \phi)^{I(Z_i(C_i)=h)}\,.$$

For the disease-free death intensity, we obtain $\lambda_3(t, a)$ based on published population mortality data which are given by calendar times and age-specific intervals (Robert, 2017). Figure 3 shows the age-specific population mortality rates across calendar periods between 1921 to 2011. A decreasing trend in the age-specific mortality rates over the last 90 years is apparent, so it is important to accommodate this if the registry includes individuals born over a wide range of calendar time and vary in age a great deal.

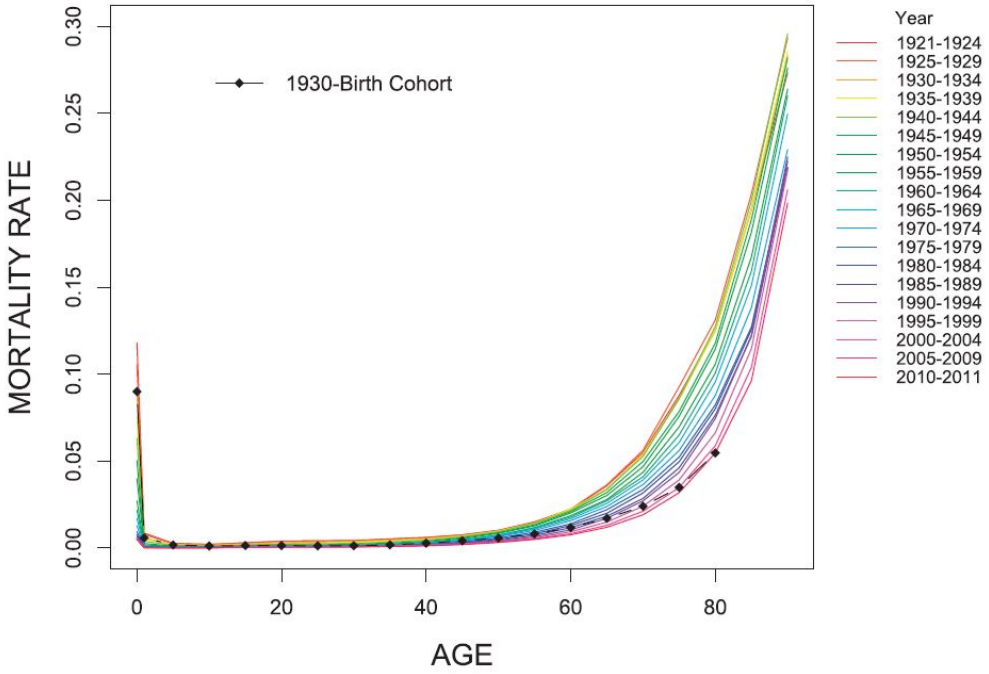

Figure 3: Age-specific population mortality rates by calendar period in Canada from 1921 to 2011.

To examine the asymptotic distribution of the estimator $\tilde{\psi}$, we construct the augmented

composite likelihood

$$ACL(\psi) \propto \prod_{i\in\mathcal{S}_F} CL_i(\psi) \prod_{i\in\mathcal{A}_1} AL_{1i} \prod_{i\in\mathcal{A}_2} AL_{2i}\,, \tag{11}$$

where $\mathcal{S}_F$ is the set of indices for probands (and hence their families) selected for the family study. We may then write

$$U_i(\psi) = \frac{\partial \log CL_i(\psi)}{\partial\psi},$$

$$U_{1i}(\phi) = \frac{\partial \log AL_{1i}(\phi)}{\partial\psi},$$

and

$$U_{2i}(\phi) = \frac{\partial \log AL_{2i}(\phi)}{\partial\psi}.$$

The score vector for the augmented composite likelihood is

$$\bar{U}(\psi) = \sum_{i\in\mathcal{S}_F} U_i(\psi) + \sum_{i\in\mathcal{A}_1} U_{1i}(\phi) + \sum_{i\in\mathcal{A}_2} U_{2i}(\phi)$$

and the maximum augmented pairwise likelihood estimator $\tilde{\psi}$ is obtained by solving $\bar{U}(\psi) = 0$. The estimated variance of $\tilde{\psi}$ is given as $n^{-1}A^{-1}(\tilde{\psi})B(\tilde{\psi})A^{-1}(\tilde{\psi})'$ where

$$A(\psi) = -\frac{1}{n}\left(\sum_{i\in\mathcal{S}_F} \frac{\partial^2 \log CL_i(\psi)}{\partial\psi\partial\psi'} + \sum_{i\in\mathcal{A}_1} \frac{\partial^2 \log AL_{1i}(\phi)}{\partial\psi\partial\psi'} + \sum_{i\in\mathcal{A}_2} \frac{\partial^2 \log AL_{2i}(\phi)}{\partial\psi\partial\psi'}\right),$$

and

$$B(\psi) = \frac{1}{n}\left(\sum_{i\in\mathcal{S}_F} U_i(\psi)U_i'(\psi) + \sum_{i\in\mathcal{A}_1} U_{1i}(\phi)U_{1i}'(\phi) + \sum_{i\in\mathcal{A}_2} U_{2i}(\phi)U_{2i}'(\phi)\right)$$

with $n = n_F + n_1 + n_2$.

4 Simulation Studies

Here, we assess the performance of the methods introduced in Sections 2 and 3 through simulation studies. To mimic more closely the PsA study, we consider the age- and calendar time-specific mortality rates based on the population mortality rates $\lambda_3(t,a)$ and assume $\lambda_2(t,a) = \nu\lambda_3(t,a)$. We set the rate of occurrence of disease $\lambda_1 = 0.01$ as a constant value. We consider the Clayton copula with Kendall's $\tau = 0.2$ and 0.4. We generate the time to disease-free death from the age and time-specific population mortality rates with $\nu = 1.1$. We generate the family size with 4 or 6 members having two parents and 2 or 4 children in family where $P(m_i + 1 = 4) = 2/3$ and $P(m_i + 1 = 6) = 1/3$. Then, we randomly choose an individual from the family members and generate the date of birth from the uniform distribution $(1920, 1950)$ if the individual is a parent or $(1950, 1980)$ otherwise. Then, we generate an individual path from the marginal distribution. We generate the individual sampling date from the uniform distribution (1980, 2010) and select those who are alive and diseased at the sampling date. We set the family sample size $n_F = 1000$, the size of registry $n_1 = 2000$. Among $n_1 + n_F$ individuals who are alive at the family sampling date on July 1, 2010, we randomly select probands and generate the data for non-probands given the proband data with the family size n_F. If the proband is a parent, the birth dates of a spouse or child are obtained by adding the uniform distribution $(0, 10)$ or $(20, 30)$ to the birth date of proband, respectively, and conduct similarly

when the proband is a child. In design I, we include all non-probands data in analysis, whereas we only include alive non-probands data in design II. In this simulation, we consider both types of auxiliary data: the registry data with follow-up and the current status survey data. The registry follow-up data including probands are assumed to be collected until July 1, 2010 with the record of death post disease. For the current status survey data with the survey size $n_2 = 1000$, we generate the date of birth from the uniform distribution $(1930, 1980)$ and set the sampling date as July 1, 2000. Here, the augmented pairwise estimations are carried out and the results are reported in Table 1. We also compare the performance of the estimators from the proposed model to those based on a clustered failure time model using a Clayton copula in which the risk of death is not considered.

For the proposed methods in two designs, the biases are negligible, the empirical standard errors (ESE) are in a good agreement with the average standard errors (ASE), and the empirical coverage probability (ECP) of nominal 95% confidence intervals are all within an acceptable range. The estimators under the full likelihood have smaller ASE compared to the pairwise likelihood with the registry data; however, the current status auxiliary data improve efficiency so that the estimators obtained by the pairwise likelihood are as efficient as those by the full likelihood. Since the current status auxiliary data have no time to death data, the efficiency of ν is not improved. Comparing design I and II, the estimators have better efficiency under design I. Also, the estimators λ_1 and τ under design II are as nearly as efficient as those under design I with current status data. We find that the bias of λ_1 is present if we do not adjust the condition of being alive in biased sampling. Since we assume the independent competing risks, the dependence parameter τ shows small bias compared to λ_1 which may be induced by the biased estimates of λ_1. The standard errors of all estimators are very close between two models. The relative mean square error (RMSE) is defined as the ratio of the MSE for the estimator from the clustered failure time model to that of the proposed model. This RMSE is greater than 1 for all estimators under all parameter settings indicating that when the true disease process is actually a clustered illness-death process there is a price to pay in terms of bias and MSE when a clustered failure time model is used which does not account for the risk of death.

5 Assessment of Genetic Risk Factors

If familial aggregation is identified by the proposed model in Sections 2 and 3, interest may lie in the effect of genetic factors on disease onset to explain familial aggregation. However, if some individuals in the study are not genotyped, incomplete genetic data must be dealt with. For example, in design I, we may obtain the disease history for non-probands who died but cannot sample their DNA. Also, the national current status survey data do not provide the genetic information. Chatterjee et al. (2006) proposed an analysis for a kin-cohort case-control and case-only family data with genotype and phenotype. Gong et al. (2010) categorized two family designs: the population and the clinic designs and present the simulation studies to examine the performance of phenotype-/genotype-based methods. Zhang et al. (2010) suggested statistical methods in estimating age-dependent penetrance under a case-family design.

In this section, we accommodate genetic data in our model but deal with missing genetic information. We let G_{ij} denote the genotype (gene carrier indicator), which is tentatively related to disease with $P(G_{ij} = 0) = q^2, P(G_{ij} = 1) = p^2 + 2pq$ with the allele frequency p and $q = 1 - p$ for individual j in family i and $\boldsymbol{G}_i = (G_{i0}, \ldots, G_{im_i})'$. We denote $W_{ij} = (V'_{ij}, G_{ij})'$ a vector of covariates and genotypes and $\boldsymbol{W}_i = (\boldsymbol{V}'_i, \boldsymbol{G}'_i)'$. The transition intensities, then, are written as $\lambda_l(t, a|b_{ij}, w_{ij})$ for $l = 1, 2, 3$ where v_{ij} is replaced with w_{ij}. The joint probability of disease onset also needs to replace $\boldsymbol{V}_i$ with $\boldsymbol{W}_i$ but the cross-ratio or cause-specific hazard

Table 1: Frequency properties of estimators based on the augmented pairwise likelihood for family data given $\lambda_3(\cdot,\cdot)$ under biased sampling scheme for the proband and disease history of non-probands available (design I) with two auxiliary data: the registry follow-up data and the current status survey data; Clayton copula with Kendall's τ=0.2, 0.4; $n_F = 1000$, $n_1 = 2000$, $n_2 = 1000$, and $nsim = 1000$

			Illness-death Model								Failure Time Model under Clayton Copula									
			Registry Data				Registry + Current Status Data				Registry Data					Registry + Current Status Data				
Design	τ	PARAMETER	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	RMSE	EBIAS	ESE	ASE	ECP	RMSE
			Full Likelihood								*Full Likelihood*									
I	0.2	$\log\lambda_1$	-0.003	0.056	0.058	0.968	-0.002	0.038	0.040	0.956	-0.007	0.056	0.059	0.958	1.027	-0.007	0.038	0.040	0.959	1.031
		$\log\nu$	0.000	0.039	0.038	0.945	0.000	0.039	0.038	0.946	-	-	-	-	-	-	-	-	-	-
		τ	0.001	0.028	0.028	0.951	0.001	0.022	0.022	0.952	0.003	0.028	0.029	0.949	1.023	0.003	0.022	0.022	0.948	1.020
	0.4	$\log\lambda_1$	-0.003	0.078	0.079	0.949	-0.001	0.044	0.045	0.950	-0.011	0.079	0.080	0.949	1.037	-0.007	0.044	0.045	0.951	1.023
		$\log\nu$	-0.001	0.036	0.037	0.954	-0.001	0.036	0.037	0.953	-	-	-	-	-	-	-	-	-	-
		τ	0.000	0.033	0.033	0.955	-0.000	0.021	0.022	0.960	0.003	0.033	0.034	0.958	1.023	0.002	0.021	0.022	0.959	1.005
			Pairwise Likelihood								*Pairwise Likelihood*									
	0.2	$\log\lambda_1$	-0.003	0.059	0.061	0.952	-0.002	0.039	0.041	0.952	-0.009	0.059	0.062	0.957	1.039	-0.007	0.039	0.041	0.956	1.037
		$\log\nu$	0.000	0.039	0.038	0.947	0.000	0.039	0.038	0.946	-	-	-	-	-	-	-	-	-	-
		τ	0.001	0.030	0.030	0.947	0.001	0.023	0.0723	0.955	0.004	0.030	0.031	0.955	1.033	0.003	0.023	0.023	0.952	1.024
	0.4	$\log\lambda_1$	-0.002	0.081	0.082	0.951	-0.001	0.045	0.045	0.953	-0.011	0.083	0.082	0.954	1.039	-0.007	0.045	0.045	0.955	1.022
		$\log\nu$	-0.001	0.037	0.036	0.954	-0.001	0.036	0.037	0.954	-	-	-	-	-	-	-	-	-	-
		τ	-0.000	0.035	0.035	0.953	-0.001	0.022	0.022	0.962	0.004	0.035	0.035	0.956	1.023	0.002	0.022	0.022	0.960	1.003
			Pairwise Likelihood								*Pairwise Likelihood*									
II	0.2	$\log\lambda_1$	-0.001	0.064	0.065	0.955	-0.001	0.040	0.042	0.957	-0.017	0.065	0.067	0.960	1.093	-0.010	0.040	0.042	0.957	1.057
		$\log\nu$	-0.002	0.049	0.047	0.935	-0.002	0.049	0.047	0.936	-	-	-	-	-	-	-	-	-	-
		τ	0.001	0.033	0.033	0.951	0.001	0.025	0.025	0.952	0.004	0.034	0.033	0.955	1.025	0.001	0.025	0.025	0.952	1.001
	0.4	$\log\lambda_1$	-0.003	0.086	0.086	0.953	-0.001	0.045	0.046	0.955	-0.023	0.089	0.087	0.950	1.151	-0.010	0.045	0.045	0.950	1.040
		$\log\nu$	-0.002	0.049	0.047	0.937	-0.002	0.049	0.047	0.938	-	-	-	-	-	-	-	-	-	-
		τ	0.001	0.037	0.037	0.952	0.001	0.023	0.023	0.963	0.006	0.038	0.037	0.950	1.069	0.001	0.023	0.023	0.964	1.003

ratio under the Clayton copula remains the same as θ. We make the following additional assumptions: (i) The process is in Hardy-Weinberg equilibrium and the Mendelian law holds, (ii) $G_{ij} \perp V_{ij}$, (iii) $\bar{Z}_{ij}(s)|G_{ij} \perp G_{ik}$ $\forall s$ for $j \neq k$, (iv) $\lambda_1(t, a|b_{ij}, w_{ij}) = \lambda_1(a)\exp(g_{ij}\alpha + v'_{ij}\beta_1)$, and (v) $\lambda_2(t, a|b_{ij}, w_{ij}) = \lambda_2(t, a|b_{ij}, v_{ij})$ and $\lambda_3(t, a|b_{ij}, w_{ij}) = \lambda_3(t, a|b_{ij})$.

5.1 Composite Likelihood with Incomplete Genetic Data

Here, we focus on the augmented pairwise conditional likelihood in Section 3. First, we consider design II with two sources of auxiliary data: (i) the family study data and the registry data and (ii) the family study data, the registry data, and current status data from the survey. In the former case, all individuals are genotyped in the family study and the registry since they are all examined, so we can assume that the genotypes are given and the pairwise composite likelihood does not change the form of likelihood which has the genotype variable as a covariate. However, the genotype data are not available in the survey, so in the latter setting, we need to model G_{ij}. The contribution of the proband to the likelihood is then

$$\begin{aligned} L_{i0}(\phi) &= P(\bar{Z}_{i0}(A_{i0}), G_{i0}|Z_{i0}(C_{i0}) = 1, C_{i0}, B_{i0}, V_{i0}; \phi) \\ &= \frac{P(\bar{Z}_{i0}(A_{i0})|G_{i0}, C_{i0}, B_{i0}, V_{i0}; \phi)P(G_{i0})}{\sum_{g\in(0,1)} P(Z_{i0}(C_{i0}) = 1|C_{i0}, B_{i0}, G_{i0} = g, V_{i0}; \phi)P(G_{i0} = g)}, \end{aligned} \tag{12}$$

where we select the proband based only on phenotype (disease status) at R_{i0}. Then the contribution from the non-probands $L_i^{II}(\psi)$ is

$$L_{ijl}^{II}(\psi) = P(\bar{\boldsymbol{Z}}_{ijl}^-(\boldsymbol{A}_{ijl}^-), \boldsymbol{G}_{ijl}^-|\bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, G_{i0}, \boldsymbol{Z}_{ijl}^-(\boldsymbol{A}_{ijl}^-) \in \{0,1\}^2, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)$$

$$= \frac{P(\bar{\boldsymbol{Z}}_{ijl}^-(\boldsymbol{A}_{ijl}^-)|\bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{G}_{ijl}, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)P(\boldsymbol{G}_{ijl}^-|G_{i0})}{\sum_{g\in\{0,1\}^2} P(\boldsymbol{Z}_{ijl}^-(\boldsymbol{A}_{ijl}^-) \in \{0,1\}^2|\bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, G_{i0}, \boldsymbol{G}_{ijl}^- = g, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)P(\boldsymbol{G}_{ijl}^- = g|G_{i0})},$$

where $\boldsymbol{G}_{ijl} = (G_{i0}, G_{ij}, G_{il})'$ and $P(\boldsymbol{G}_{ijl}^-|G_{i0})$ can be calculated using the allele frequency f and family structure; see Appendix B of *Supplementary Material* available at *Biostatistics* online. For the auxiliary data, we let G_i denote the genotype of individual i in $\mathcal{A}_1$ or $\mathcal{A}_2$ in Section 3. The likelihood terms based on the auxiliary data AL_1 and AL_2 are then given as

$$AL_1 \propto \prod_{i\in\mathcal{A}_1} P(\bar{Z}_i(A_i^*), G_i|Z_i(C_i) = 1, C_i, B_i, V_i),$$

and

$$\begin{aligned} AL_2 \propto \prod_{i\in\mathcal{A}_2}\prod_{h\in\{0,1\}} \Big\{ &\sum_g P(Z_i(C_i) = h|Z_i(C_i) \in \{0,1\}, B_i, G_i = g, V_i) \\ &\times P(G_i = g|Z_i(C_i) \in \{0,1\}, B_i, V_i)\Big\}^{I(Z_i(C_i)=h)} \end{aligned}$$

where

$$P(G_i = g|Z_i(C_i), B_i, V_i) = \frac{P(Z_i(C_i)|G_i = g, B_i, V_i)P(G_i = g)}{\sum_{g\in\{0,1\}} P(Z_i(C_i)|G_i = g, B_i, V_i)P(G_i = g)}.$$

Secondly, we only observe the genotype of non-probands who are alive in design I, and non-probands who did not survive to R_i are not genotyped. In this case

$$\begin{aligned} L_{ijl}^I(\psi) &= P(\bar{\boldsymbol{Z}}_{ijl}^-(\boldsymbol{A}_{ijl}^-), \boldsymbol{G}_{ijl,o}^-|\bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, G_{i0}, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi) \\ &= \sum_{\boldsymbol{G}_{ijl,m}} P(\bar{\boldsymbol{Z}}_{ijl}^-(\boldsymbol{A}_{ijl}^-)|\bar{Z}_{i0}(A_{i0}), Z_{i0}(A_{i0}) = 1, \boldsymbol{G}_{ijl}, \boldsymbol{A}_{ijl}, \boldsymbol{B}_{ijl}, \boldsymbol{V}_{ijl}; \psi)P(\boldsymbol{G}_{ijl}^-|G_{i0}) \end{aligned}$$

where $\boldsymbol{G}_{ijl,o}$ is a vector of observed genotype in family i for the pair of family member j and l with the proband genotype on the first component, $\boldsymbol{G}_{ijl,m}$ is a vector of missing genotypes for the family member j and l in family i, and $\boldsymbol{G}_{ijl} = (\boldsymbol{G}'_{ijl,o}, \boldsymbol{G}'_{ijl,m})'$.

5.2 Simulation Studies with Genetic Data are Incomplete

We conducted further simulation studies to assess performance of the proposed model with genetic risk factors. We considered a binary indicator G_{ij} with the allele frequency $p = 0.06$ with a hazard ratio of $\exp(\alpha) = 1.5$; we do not consider additional covariates for simplicity and otherwise adopt the same simulation settings as in Section 4.

We first generate the genotype for family members based on the family structure under the Mendelian law and given the genotype we generate family members' lifetime paths based on the proposed model. The selection criteria remains the same as in Section 4. The empirical properties of the estimators for the parameters based on design I and II are reported in Table 2 and 3, respectively.

Table 2: Frequency properties of estimators based on the augmented pairwise likelihood for family data with genotype information given $\lambda_3(\cdot,\cdot)$ under biased sampling scheme for the proband and disease history of non-probands available (design I) with two auxiliary data: the registry follow-up data and the current status survey data; Clayton copula with Kendall's τ=0.2, 0.4; $n_F = 1000$, $n_1 = 2000$, $n_2 = 1000$, and $nsim = 1000$

		Registry Data				Registry + Current Status Data			
τ	PARAMETER	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
0.2	$\log\lambda_1$	0.002	0.057	0.059	0.952	-0.000	0.041	0.040	0.954
	α	-0.003	0.064	0.064	0.948	-0.002	0.064	0.064	0.947
	$\log\nu$	-0.001	0.037	0.037	0.956	-0.001	0.037	0.037	0.957
	$\log p$	0.002	0.055	0.056	0.947	0.002	0.055	0.055	0.947
	τ	-0.000	0.028	0.029	0.963	0.001	0.023	0.023	0.953
0.4	$\log\lambda_1$	0.002	0.076	0.078	0.950	0.000	0.045	0.044	0.952
	α	-0.002	0.057	0.058	0.948	-0.002	0.058	0.058	0.945
	$\log\nu$	-0.001	0.035	0.036	0.952	-0.001	0.035	0.036	0.946
	$\log p$	0.002	0.053	0.054	0.946	-0.003	0.053	0.053	0.947
	τ	-0.001	0.032	0.033	0.953	-0.000	0.022	0.022	0.949

Here we can observe the same findings pointed out in Section 5. The current status survey data do not affect the efficiency α and p because the genetic marker is not available in the survey, however, they increase the efficiency of λ_1 and Kendall's τ in design I. This highlights the value of the current status data when disease onset times are right-truncated even for the dependence parameter. In design II, the current status data improve efficiency of each estimator except the one for ν. It is therefore advantageous for score tests, in particular, when interest lies in testing genetic effects on disease onset as it may increase the power of such tests.

6 Application to the Psoriatic Arthritis Family Study

Psoriasis is an inflammatory skin disease occurring about 2-3% of the general population and PsA(Psoriatic Arthrists) is an inflammatory arthritis disease affecting about 30% of patients

Table 3: Frequency properties of estimators based on the augmented pairwise likelihood for family data with genotype information given $\lambda_3(\cdot,\cdot)$ under biased sampling scheme for the proband and alive non-probands data available (design II) with two auxiliary data: the registry follow-up data and the current status survey data; Clayton copula with Kendall's τ=0.2, 0.4; $n_F = 1000$, $n_1 = 2000$, $n_2 = 1000$, and $nsim = 1000$

		Registry Data				Registry + Current Status Data			
τ	PARAMETER	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
0.2	$\log\lambda_1$	0.001	0.061	0.064	0.951	-0.000	0.042	0.042	0.941
	α	-0.003	0.070	0.071	0.953	-0.003	0.065	0.065	0.949
	$\log\nu$	-0.001	0.047	0.046	0.951	-0.001	0.046	0.046	0.949
	$\log p$	-	-	-	-	0.002	0.055	0.055	0.949
	τ	0.000	0.030	0.032	0.956	0.001	0.024	0.024	0.952
0.4	$\log\lambda_1$	0.004	0.081	0.082	0.955	0.001	0.046	0.045	0.951
	α	-0.001	0.063	0.062	0.949	-0.002	0.059	0.058	0.948
	$\log\nu$	-0.002	0.047	0.046	0.948	-0.002	0.046	0.046	0.950
	$\log p$	-	-	-	-	0.002	0.053	0.053	0.942
	τ	-0.001	0.035	0.035	0.949	0.001	0.023	0.023	0.941

with psoriasis (Gladman, 1991; Langley et al., 2005; Eder et al., 2012). Patients with PsA are at higher risk for death compared to the general population of Ontario with a standardized mortality ratio of 1.36 (Gladman, 2008). Many studies showed that psoriasis is a heritable disease; Pedersen et al. (2008) reported an increased concordance measure in monozygotic relative to dizygotic twins and Chandran et al. (2009) confirmed a high familial recurrence risk of PsA based on family studies as shown in Moll and Wright (1973). To obtain a better sense of heredity, Gladman and Farewell (1995); Pedersen et al. (2008); Chandran and Raychaudhuri (2010); Eder et al. (2012) identified genes related to psoriasis and PsA and explored environmental factors which increase the risk of PsA. We consider the Human Leucocyte Antigens (HLA)-B27, and HLA-C06 by the findings of the genetic etiology of psoriasis and psoriatic arthritis in the literature.

We consider data from the Centre for Prognosis Studies in Rheumatic Disease at the University of Toronto which recruited University of Toronto Psoriatic Arthritis Registry and among 1436 individuals from the registry, 150 were selected for family studies as probands. In this family studies, family members were recruited to conduct a thorough examination including genotype information, therefore, this study design belongs to the biased sampling scheme design II. To simplify the analysis, we generate a number of 167 pseudo-families from the original 150 families where two-generation families are considered with the non-missing date of birth and genotype information and we use this pseudo-family data. In the pseudo-family data, the family sizes range from 2 to 7 individuals; 55 families have 2 family members (1 proband and 1 non-proband), and 112 families have at least three members. One hundred and ninety-two individuals were diagnosed with PsA among a total of 530 individuals. One hundred and forty-four families have one member with PsA (i.e. proband), 21 families having two members with PsA, and 2 families with three PsA patients in their family.

As a source of auxiliary data, we use the survey of US population in which Gelfand et al. (2005a) reported the prevalence of psoriatic arthritis in 2001. In this survey, subjects 18 years of age or older were randomly selected and provided the status of psoriasis and psoriatic arthritis; 328 have psoriatic arthritis among 15,307 respondents.

We begin with the model not using the genotype information. We fit a marginal model for the age at PsA onset with piecewise constant hazards with cut-points 28, 38 and 48 corresponding to 25%, 50% and 75% quantiles of the onset age of PsA in the registry data and assume $\lambda_2(t,a) = \lambda_3(t,a)\nu$. In the registry data, individuals with missing genotype are dealt with similarly in the survey data. Table 4 summarizes the estimates of fitted model without genetic variable in the first column followed by two univariate models with genotype HLA-B27, and HLA-C06 variables including the allele frequency p for each genetic markers.

Table 4: Estimates of parameters based on the augmented pairwise likelihood; auxiliary data include the UTPAR and the survey from Gelfand et al. (2005a) without/with genotype variable under the piecewise constant marginal model for age at PsA onset with cut-points 28, 38 and 48

MARKER	α_{marker}	ν	τ	p_{marker}
-	-	1.201 (0.081)	0.329 (0.094)	-
B27	0.336 (0.054)	1.199 (0.081)	0.326 (0.095)	0.065 (0.013)
C06	-0.214 (0.033)	1.199 (0.081)	0.321 (0.094)	0.169 (0.023)

First, based on the model without genetic markers, we find that $\hat{\nu} = 1.201$ indicating that the ratio of the hazard of death post PsA to PsA-free death is 1.201, which is lower than the reported value in Gladman (2008). As expected, PsA is not lethal while it increases the risk of death. The estimate of dependence parameter is $\hat{\tau} = 0.329$ (95% CI: 0.145, 0.513; p< 0.001) which indicates significant association between family members. We find that HLA-B27 positive and HLA-C06 positive have insignificant effect on the risk of PsA. The allele frequency of HLA-B27 is 0.065, which is compatible with the value of 0.061 from the national USA prevalence of HLA-B27 (Reveille et al., 2012). HLA-C06 has the allele frequency 0.169 which is more prevalent that HLA-B27. After adjusted significant genetic marker HLA-B27, and HLA-C06, Kendall's $\hat{\tau}$ decreases to 0.326 (95% CI: 0.140, 0.512; p=0.001), 0.321 (95% CI: 0.137, 0.505; p=0.001), respectively since HLA-B27 and HLA-C06 partially explain the residual familial aggregation.

Figure 4 shows the cross-odds ratio defined in (4) for a sibling given other sibling born in the same year 1930, 1940, 1950, 1960 (the left panel) and a child born in 1930, 1940, 1950, 1960 given a parent born in 1905, 1915, 1925, 1935 (the right panel), respectively. For the sibling pairs, two siblings are governed by the same mortality rates belonging the same birth cohort. The cross-odds ratio before 30 almost plateaus but showed a decreasing trend as they age because the mortality rate increases. There is a drastic decrease in the cross-odds ratio as age increases for the child-parent pairs compared to the sibling pairs. This difference arises due to the higher mortality rates for parents; see Figure 3. Similar patterns of the cross-odds ratio for different birth cohorts are observed, but the variation exists.

Figure 5 shows the marginal probability of death (states 2 and 3) and the cumulative incidence function for the age of PsA defined in (5) for different birth cohorts at 1930, 1940, 1950, and 1960. We find that PsA itself is a rare disease with the low cumulative incidence function.

7 Discussion

In this article, we have proposed an illness-death model for family studies which incorporates within-family dependence in the age of disease onset via a copula model. The illness-death

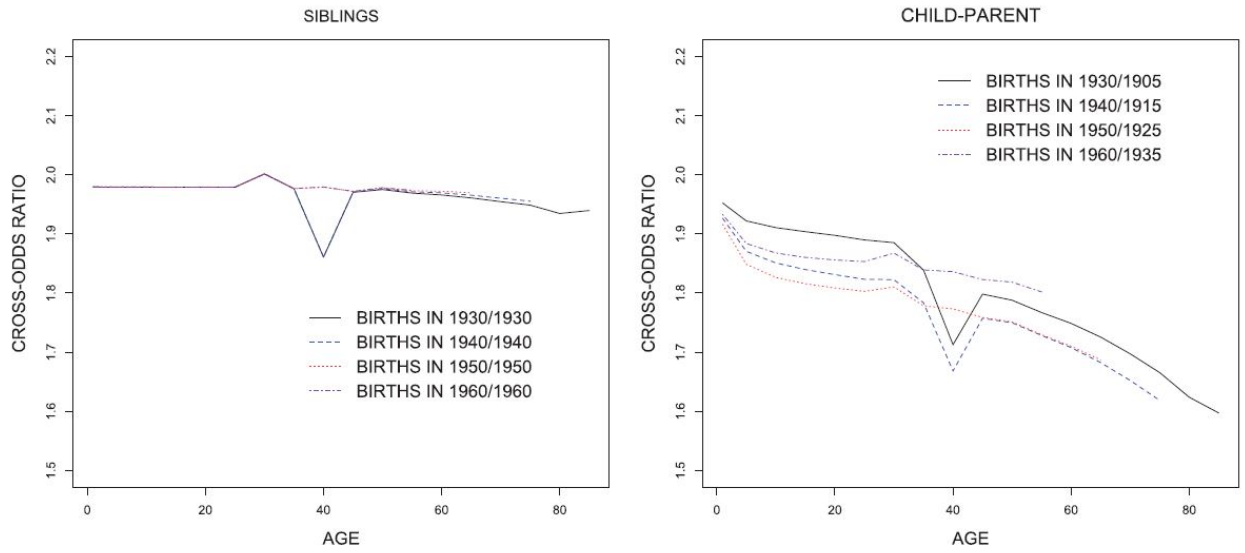

Figure 4: The cross-odds ratio for two siblings born in the same year 1930, 1940, 1950, 1960 (left panel) and a child born in 1930, 1940, 1950, or 1960 with a parent born in 1905, 1915, 1925, or 1935 (right panel) based on the fitted model with no effect of a genetic marker.

model offers a natural framework to consider survival bias and the Clayton copula models retain simple interpretations of cross ratio/cause-specific cross ratio and marginal interpretations of estimates of covariance. We explore two study designs for family studies with biased sampling schemes and developed statistical methods for analysis. Pairwise composite likelihood is utilized to ease the computational burden. We exploit auxiliary data to address identifiability and estimability issues. Age- and calendar time-specific population mortality rates adequately address the trend of mortality rates in family studies where more than two generations are considered. We extend our model to study the effect of genetic markers on risk of disease in which the availability of genotype data depends on the study design.

We restrict our attention to the case-only probands family studies. If case-control probands are available, it would be useful to compare the robustness to misspecification of model assumption (Chatterjee et al., 2006) and compare the efficiency with the case-only probands family studies. It is natural to extend our model to allow for different dependence structures in families using a more flexible Gaussian copula (Zhong and Cook, 2016; Lakhal-Chaieb et al., 2018). As we have shown in the simulation studies in Sections 4 and 5.2, the use of auxiliary data improves efficiency in estimating the marginal parameters related to disease onset and the dependence parameter, so further exploration of the relative value of different types of auxiliary data would be of interest as this would have bearing on the power of the design.

The assumption of independent competing failure times (i.e. $X_{ij1} \perp X_{ij3}$) is not checkable directly and is a limitation in any competing risks analysis based on models for cause-specific hazards. A correlation between these potentially latent times may arise from omitted shared covariates, for example, so enriching the covariate vector to include factors that might, if omitted, induce such a dependence may be advisable. Note, however, that this will lead to a marker effect with a different interpretation and will change the meaning of the measure of within-family association in disease onset time. We examine sensitivity of parameter estimates to violations of the independent competing risks assumption in a brief simulation study in Appendix C.1 of *Supplementary Material* available at *Biostatistics* online. Here, we introduce a shared multiplicative gamma-distributed random effect which acts on the $0 \to 1$ and $0 \to 3$ intensities; we

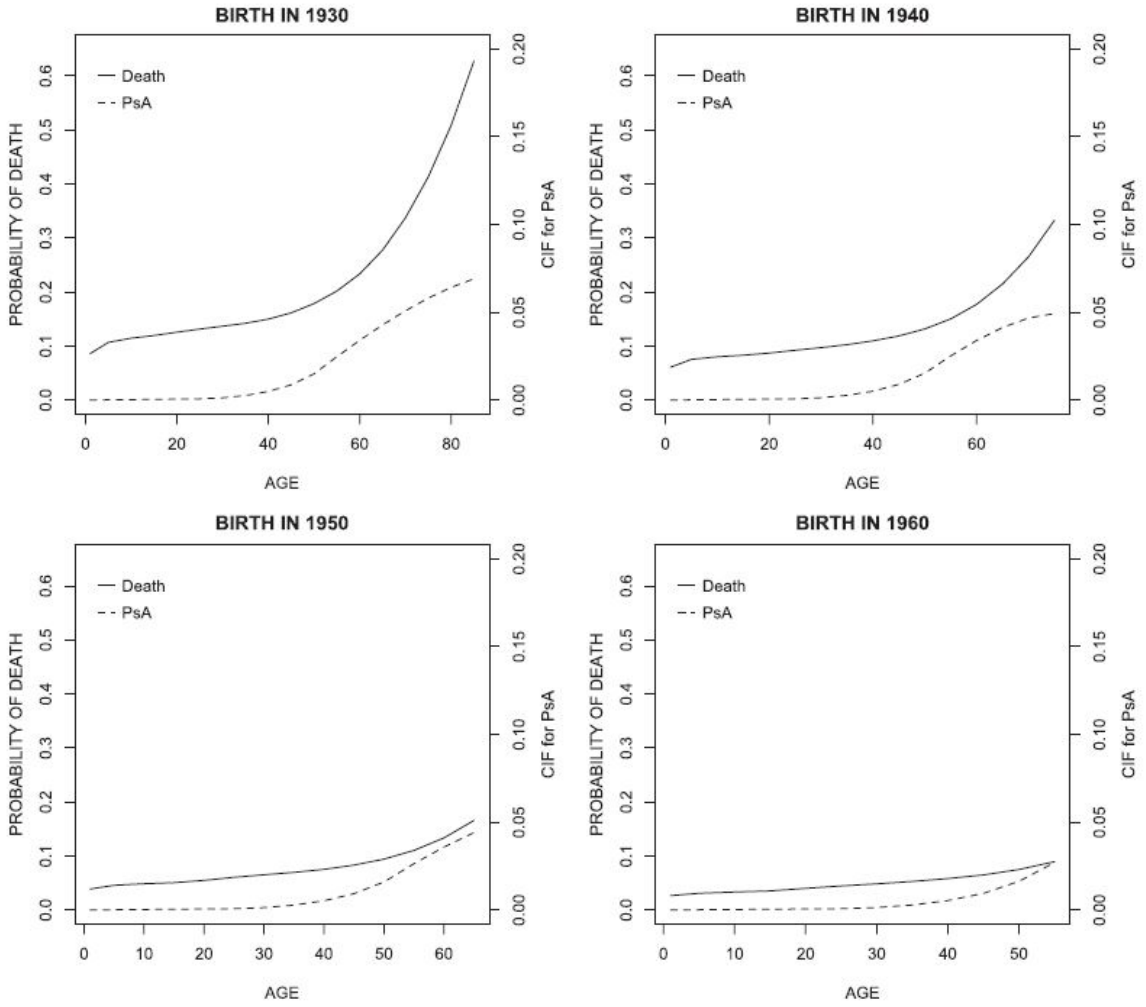

Figure 5: The marginal probability of death and the cumulative incidence of PsA by the year of birth of 1930, 1940, 1950, or 1960 based on the fitted model with no effect of a genetic marker.

took this to have mean 1 and variance 0.2. The bias of estimator for τ is appreciable in all settings considered, which is not unexpected since since omission of this random effect induces dependence between X_{ij1} and X_{ij3}, introduces an extra source of between-individual variation in the intensities, and leads to model violations of the proportional mortality assumption for $1 \to 2$ and $0 \to 3$ transitions. We expect to report on a more thorough study of the impact of dependent competing risks in this setting using a more flexible copula model in which separate dependence parameters can be specified and sensitivity analyses may be carried out.We note however, that there is an identifiability problem which prohibits modeling this dependence in such a way that model assumptions can be checked. Finally in Appendix C.2 *Supplementary Material* available at *Biostatistics* online, we also examined the sensitivity of our conclusions in the psoriatic arthritis family study to the specification of the copula function and to the proportional mortality assumption among diseased vs. disease-free individuals. There we show results based on the Frank copula function as well as the Clayton copula function, and note that apart from the estimate of Kendall's τ the parameters appeared similar for the two copula models. To be more flexible, we also allow $\nu(a)$ to be piecewise functions in our application and found that the effects of the HLA markers were again quite comparable in the fitted models.

We assume that the subject-specific disease-free mortality rate does not depend on covari-

ates, which is the same as age-, time-specific population mortality. This is our limitation since we do not have available data at hand to explore the effect of any covariates on the disease-free mortality from the University of Toronto Arthritis Registry due to the sampling scheme for this cohort. It may be useful to adopt age-, time-, and gender-specific population mortality rates.

In our motivating example, we formed the pseudo-families comprised of at most two generations since calculation of the joint distribution of alleles for multi-generational families is computationally complex. More formal treatment of multi-generational family studies may help to disentangle genetic effects from the effect of shared family environment in family studies; extensions of this sort warrant development.

PsA occurs in 10-20% of patients with psoriasis and the genetic marker HLA-C06 mostly contributes to develop psoriasis (Queiro et al., 2015). As an extension of our application, to distinguish the genetic risk factors for psoriasis with those for PsA, we may introduce the state of psoriasis in our analysis. Another extension would be to use multiple allele in our analysis. This leads to computational burden due to the summation of all possible combination of genetic markers for missing genotypes.We may use other sources of population studies to calculate the allele frequency, and we may exploit this value to assume that the allele frequency p is known in our proposed model. This will reduce the number of parameters to be estimated.

Code used for the data generation and analysis is available for download at https://github.com/joolee0918/clusteridm.

Supplementary Material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgements

Conflict of Interest: None declared.

Funding

The Natural Sciences and Engineering Research Council of Canada (RGPIN 155849 and RGPIN 04207); and the Canadian Institutes for Health Research (FRN 13887). Richard Cook is a Tier I Canada Research Chair in Statistical Methods for Health Research.

References

Aalen, O. O. (2012). Armitage lecture 2010: understanding treatment effects: the value of integrating longitudinal data and survival analysis. *Statistics in Medicine*, 31(18):1903–1917.

Aalen, O. O., Borgan, Ø., Keiding, N., and Thormann, J. (1980). Interaction between life history events. nonparametric analysis for prospective and retrospective data in the presence of censoring. *Scandinavian Journal of Statistics*, pages 161–171.

Andersen, P. K. (1988). Multistate models in survival analysis: a study of nephropathy and mortality in diabetes. *Statistics in Medicine*, 7(6):661–670.

Andersen, P. K., Borch-Johnsen, K., Deckert, T., Green, A., Hougaard, P., Keiding, N., and Kreiner, S. (1985). A Cox regression model for the relative mortality and its application to diabetes mellitus survival data. *Biometrics*, pages 921–932.

Anderson, V. E. (1961). Statistical studies of probands and their relatives. *Annals of the New York Academy of Sciences*, 91(1):781–796.

Balliu, B., Tsonaka, R., van der Woude, D., Boehringer, S., and Houwing-Duistermaat, J. J. (2012). Combining family and twin data in association studies to estimate the noninherited maternal antigens effect. *Genetic Epidemiology*, 36(8):811–819.

Bandeen-Roche, K. and Liang, K. (2002). Modelling multivariate failure time associations in the presence of a competing risk. *Biometrika*, 89(2):299–314.

Cederkvist, L., Holst, K. K., Andersen, K. K., and Scheike, T. H. (2018). Modeling the cumulative incidence function of multivariate competing risks data allowing for within-cluster dependence of risk and timing. *Biostatistics*, 20(2):199–217.

Chandran, V. and Raychaudhuri, S. P. (2010). Geoepidemiology and environmental factors of psoriasis and psoriatic arthritis. *Journal of Autoimmunity*, 34(3):J314–J321.

Chandran, V., Schentag, C. T., Brockbank, J. E., Pellett, F. J., Shanmugarajah, S., Toloza, S. M., Rahman, P., and Gladman, D. D. (2009). Familial aggregation of psoriatic arthritis. *Annals of the Rheumatic Diseases*, 68(5):664–667.

Chatterjee, N., Kalaylioglu, Z., Shih, J. H., and Gail, M. H. (2006). Case–control and case-only designs with genotype and family history data: estimating relative risk, residual familial aggregation, and cumulative risk. *Biometrics*, 62(1):36–48.

Cheng, Y., Fine, J. P., and Kosorok, M. R. (2009). Nonparametric association analysis of exchangeable clustered competing risks data. *Biometrics*, 65(2):385–393.

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.

Datta, S., Satten, G. A., and Datta, S. (2000). Nonparametric estimation for the three-stage irreversible illness–death model. *Biometrics*, 56(3):841–847.

Eder, L., Chandran, V., Pellett, F., Shanmugarajah, S., Rosen, C. F., Bull, S. B., and Gladman, D. D. (2012). Differential human leucocyte allele association between psoriasis and psoriatic arthritis: a family-based association study. *Annals of the Rheumatic Diseases*, 71(8):1361–1365.

Gelfand, J. M., Gladman, D. D., Mease, P. J., Smith, N., Margolis, D. J., Nijsten, T., Stern, R. S., Feldman, S. R., and Rolstad, T. (2005a). Epidemiology of psoriatic arthritis in the population of the United States. *Journal of the American Academy of Dermatology*, 53(4):573–e1.

Gelfand, J. M., Weinstein, R., Porter, S. B., Neimann, A. L., Berlin, J. A., and Margolis, D. J. (2005b). Prevalence and treatment of psoriasis in the United Kingdom: a population-based study. *Archives of Dermatology*, 141(12):1537–1541.

Gladman, D. D. (1991). Psoriatic arthritis. In *Prognosis in the Rheumatic Diseases*, pages 153–166. Springer.

Gladman, D. D. (2008). Mortality in psoriatic arthritis. *Clinical & Experimental Rheumatology*, 26(5):S62.

Gladman, D. D. and Farewell, V. T. (1995). The role of HLA antigens as indicators of disease progression in psoriatic arthritis. *Arthritis and Rheumatology*, 38(6):845–850.

Gong, G., Hannon, N., and Whittemore, A. S. (2010). Estimating gene penetrance from family data. *Genetic Epidemiology*, 34(4):373–381.

Hougaard, P. (1999). Multi-state models: a review. *Lifetime Data Analysis*, 5(3):239–264.

Hougaard, P. (2012). *Analysis of Multivariate Survival Data*. Springer Science & Business Media.

Hougaard, P., Harvald, B., Holm, N. V., Flournoy, N., Islam, M. A., and Singh, K. P. (1992). Assessment of dependence in the life times of twins. In *Survival Analysis: State of the Art*, pages 77–97. Springer.

Hsu, L., Chen, L., Gorfine, M., and Malone, K. (2004). Semiparametric estimation of marginal hazard function from case–control family studies. *Biometrics*, 60(4):936–944.

Hsu, L. and Gorfine, M. (2005). Multivariate survival analysis for case–control family data. *Biostatistics*, 7(3):387–398.

Jiang, F. and Haneuse, S. (2017). A semi-parametric transformation frailty model for semi-competing risks survival data. *Scandinavian Journal of Statistics*, 44(1):112–129.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. CRC Press.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, volume 360. John Wiley & Sons.

Keiding, N. (1990). Statistical inference in the Lexis diagram. *Phil. Trans. R. Soc. Lond. A*, 332(1627):487–509.

Keiding, N. (2006). Event history analysis and the cross-section. *Statistics in Medicine*, 25(14):2343–2364.

Lakhal-Chaieb, L., Cook, R. J., and Zhong, Y. (2018). Testing the heritability and parent-of-origin hypotheses for ages at onset of psoriatic arthritis under biased sampling. *Biometrics (provisionally accepted)*.

Langley, R. G. B., Krueger, G. G., and Griffiths, C. E. M. (2005). Psoriasis: epidemiology, clinical features, and quality of life. *Annals of the Rheumatic Diseases*, 64(suppl 2):ii18–ii23.

Lee, E. W., Wei, L. J., Amato, D. A., and Leurgans, S. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, pages 237–247. Springer.

Li, H., Yang, P., and Schwartz, A. G. (1998). Analysis of age of onset data from case-control family studies. *Biometrics*, pages 1030–1039.

Liang, K. and Beaty, T. H. (2000). Statistical designs for familial aggregation. *Statistical Methods in Medical Research*, 9(6):543–562.

Liang, K., Self, S. G., and Chang, Y. (1993). Modelling marginal hazards in multivariate failure time data. *Journal of the Royal Statistical Society. Series B*, pages 441–453.

Mesfioui, M. and Quessy, J. (2008). Dependence structure of conditional Archimedean copulas. *Journal of Multivariate Analysis*, 99(3):372–385.

Moll, J. M. and Wright, V. (1973). Familial occurrence of psoriatic arthritis. *Annals of the Rheumatic Diseases*, 32(3):181.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84(406):487–493.

Pedersen, O. B., Svendsen, A. J., Ejstrup, L., Skytthe, A., and Junker, P. (2008). On the heritability of psoriatic arthritis. Disease concordance among monozygotic and dizygotic twins. *Annals of the Rheumatic Diseases*, 67(10):1417–1421.

Pfeiffer, R. M., Pee, D., and Landi, M. T. (2008). On combining family and case-control studies. *Genetic Epidemiology*, 32(7):638–646.

Queiro, R., Morante, I., Cabezas, I., and Acasuso, B. (2015). HLA-B27 and psoriatic disease: a modern view of an old relationship. *Rheumatology*, 55(2):221–229.

Reveille, J. D., Hirsch, R., Dillon, C. F., Carroll, M. D., and Weisman, M. H. (2012). The prevalence of HLA–B27 in the US: data from the US national health and nutrition examination survey, 2009. *Arthritis & Rheumatology*, 64(5):1407–1411.

Robert, B. (2017). Mortality data for Canada. https://www.mortality.org/cgi-bin/hmd/country.php?cntr=CAN&level=1.

Scheike, T. H., Holst, K. K., and Hjelmborg, J. B. (2014). Estimating heritability for cause specific mortality based on twin studies. *Lifetime Data Analysis*, 20(2):210–233.

Scheike, T. H. and Sun, Y. (2012). On cross-odds ratio for multivariate competing risks data. *Biostatistics*, 13(4):680–694.

Scheike, T. H., Sun, Y., Zhang, M., and Jensen, T. K. (2010). A semiparametric random effects model for multivariate competing risks data. *Biometrika*, 97(1):133–145.

Schweder, T. (1970). Composable Markov processes. *Journal of Applied Probability*, 7(2):400–410.

Shih, J. H. and Albert, P. S. (2010). Modeling familial association of ages at onset of disease in the presence of competing risk. *Biometrics*, 66(4):1012–1023.

Shih, J. H. and Chatterjee, N. (2002). Analysis of survival data from case–control family studies. *Biometrics*, 58(3):502–509.

Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, pages 1384–1399.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42.

Wong, K., Gladman, D. D., Husted, J., Long, J. A., and Farewell, V. T. (1997). Mortality studies in psoriatic arthritis. Results from a single outpatient clinic. I. Causes and risk of death. *Arthritis & Rheumatology*, 40(10):1868–1872.

Xu, J., Kalbfleisch, J. D., and Tai, B. (2010). Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics*, 66(3):716–725.

Zhang, H., Olschwang, S., and Yu, K. (2010). Statistical inference on the penetrances of rare genetic mutations based on a case–family design. *Biostatistics*, 11(3):519–532.

Zheng, Y., Heagerty, P. J., Hsu, L., and Newcomb, P. A. (2010). On combining family-based and population-based case–control data in association studies. *Biometrics*, 66(4):1024–1033.

Zhong, Y. and Cook, R. J. (2016). Augmented composite likelihood for copula modeling in family studies under biased sampling. *Biostatistics*, 17(3):437–452.

Zhong, Y. and Cook, R. J. (2017). Second-order estimating equations for clustered current status data from family studies using response-dependent sampling. *Statistics in Biosciences*, pages 1–24.

Zhou, B., Fine, J., Latouche, A., and Labopin, M. (2012). Competing risks regression for clustered data. *Biostatistics*, 13(3):371–383.

Published in final edited form as:
Biostatistics (2021), **22** (3): 482–503
DOI: 10.1093/biostatistics/kxz048

Supplementary material for *The illness-death model for family studies*

JOOYOUNG LEE
Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada
E-mail: j463lee@uwaterloo.ca

RICHARD J. COOK
Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada

APPENDIX

A An illustrative Construction of the Composite Likelihood

Here we illustrate how to construct the composite conditional likelihood given in Section 3. We consider a particular family consisting of two parents and one proband from the family study of the Centre for Prognosis Studies in Rheumatic Disease at the University of Toronto. In what follows we omit the subscript i labeling the family and suppress the dependence on covariates for simplicity. The details of this family are as follows:

- the family was recruited in 2007;
- the father was born on May 30, 1929 and was alive, aged 77.59 years, and PsA-free at the point the family was recruited;
- the mother was born on July 21, 1934 and was alive and aged 72.45 years at the point the family was recruited, and she developed PsA at age 65;
- the proband was born on June 7, 1955 and developed PsA at age 37; the date of their first clinic visit was July 30, 2001 when they were 46.08 years of age; they were 51.57 years of age when the family was recruited to the family study.

The likelihood contribution (8) of the proband can written as

$$P(\bar{Z}_0(A_0)|Z_0(C_0)=1,C_0,B_0;\phi)=\frac{P(\bar{Z}_0(A_0),B_0;\phi)}{P(Z_0(C_0)=1,C_0,B_0;\phi)}, \tag{A.1}$$

with $A_0 = 51.57, C_0 = 46.08$, and $B_0 = 1955.43$. Then, numerator of (A.1) is

$$\begin{aligned} P(\bar{Z}_0(A_0), B_0; \phi) &= P(X_{01} = 37, X_{03} > 37, X_{02} > 51.57, B_0 = 1955.43; \phi) \\ &= \lambda_1(37) \exp\left(-\int_0^{37} \lambda_1(s)ds\right) \exp\left(-\int_0^{37} \lambda_3(1955.43 + s, s)ds\right) \\ &\times \exp\left(-\int_{37}^{51.57} \lambda_2(1955.43 + s, s)ds\right), \end{aligned}$$

and the denominator of (A.1) is given as

$$\begin{aligned} P(Z_0(C_0) = 1, C_0, B_0; \phi) &= P(X_{01} < 46.08, X_{03} > X_{01}, X_{02} > 46.08, B_0 = 1955.43; \phi) \\ &= \int_0^{46.08} \lambda_1(s) \exp(-\int_0^s \lambda_1(u)du) \exp(-\int_0^s \lambda_3(1955.43 + u, u)du) \\ &\times \exp(-\int_s^{46.08} \lambda_2(1955.43 + u, u)du)ds. \end{aligned} \tag{A.2}$$

Note that the data in the motivating example were obtained by design II. We denote the father with subscript $j = 1$ and the mother with subscript $j = 2$. The contribution to the augmented composite likelihood can then be written as

$$L_{23}^{II}(\psi) = \frac{P(\bar{\boldsymbol{Z}}_{12}^-(\boldsymbol{A}_{12}^-))|\bar{Z}_0(A_0), Z_0(A_0) = 1, \boldsymbol{A}_{12}, \boldsymbol{B}_{12}; \psi)}{P(\boldsymbol{Z}_{12}^-(\boldsymbol{A}_{12}^-) \in \{0,1\}^2|\bar{Z}_0(A_0), Z_0(A_0) = 1, \boldsymbol{A}_{12}, \boldsymbol{B}_{12}; \psi)},$$

where $\boldsymbol{A}_{12} = (51.57, 77.59, 72.45)', \boldsymbol{B}_{12} = (1955.43, 1929.41, 1934.55)'$, the numerator is given as

$$P(X_{21} > 77.59, X_{23} > 77.59, X_{31} = 65, X_{32} > 72.45, X_{33} > 65|X_{01} = 37, X_{02} > 51.57, X_{03} > 37; \psi),$$

and the denominator is given as

$$\begin{aligned} &P(\boldsymbol{Z}_{12}^-(\boldsymbol{A}_{12}^-) \in \{0,0\}|\bar{Z}_0(A_0), Z_0(A_0) = 1, \boldsymbol{A}_{12}, \boldsymbol{B}_{12}; \psi) + P(\boldsymbol{Z}_{12}^-(\boldsymbol{A}_{12}^-) \in \{0,1\}|\bar{Z}_0(A_0), Z_0(A_0) = 1, \\ &\boldsymbol{A}_{12}, \boldsymbol{B}_{12}; \psi) + P(\boldsymbol{Z}_{12}^-(\boldsymbol{A}_{12}^-) \in \{1,0\}|\bar{Z}_0(A_0), Z_0(A_0) = 1, \boldsymbol{A}_{12}, \boldsymbol{B}_{12}; \psi) + P(\boldsymbol{Z}_{12}^-(\boldsymbol{A}_{12}^-) \in \{1,1\}| \\ &\bar{Z}_0(A_0), Z_0(A_0) = 1, \boldsymbol{A}_{12}, \boldsymbol{B}_{12}; \psi). \end{aligned} \tag{A.3}$$

The first term in (A.3) is

$$\begin{aligned} &P(X_{21} > 77.59, X_{23} > 77.59, X_{31} > 72.45, X_{33} > 72.45|X_{01} = 37, X_{02} > 51.57, X_{03} > 37; \psi) \\ &= C(\mathcal{F}(77.59|X_{01} = 37; \phi_1, \rho), \mathcal{F}(72.45|X_{01} = 37; \phi_1, \rho); \rho^*) \exp\left(-\int_0^{77.59} \lambda_3(1929.41 + v, v)dv\right) \\ &\times \exp\left(-\int_0^{72.45} \lambda_3(1934.55 + v, v)dv\right), \end{aligned}$$

the second term in (A.3) is

$$P(X_{21} < 77.59, X_{23} > X_{21}, X_{22} > 77.59, X_{31} > 72.45, X_{33} > 72.45 | X_{01} = 37, X_{02} > 51.57, X_{03} > 37; \psi)$$
$$= \int_0^{77.59} \left. \frac{\partial C(\mathcal{F}(u|X_{01} = 37; \phi_1, \rho), \mathcal{F}(72.45|X_{01} = 37; \phi_1, \rho); \rho^*)}{\partial u} \right|_{u=s} \exp\left(- \int_0^s \lambda_3(1929.41 + v, v) dv \right)$$
$$\times \exp\left(- \int_s^{77.59} \lambda_2(1929.41 + v, v) dv \right) ds \exp\left(- \int_0^{72.45} \lambda_3(1934.55 + v, v) dv \right),$$

and the third term in (A.3) is $P(X_{21} > 77.59, X_{23} > 77.59, X_{31} < 72.45, X_{33} > X_{31}, X_{32} > 72.45 | X_{01} = 37, X_{02} > 51.57, X_{03} > 37; \psi)$ which is obtained in a fashion similar to the second term. The last term in (A.3), $P(X_{21} < 77.59, X_{23} > X_{21}, X_{22} > 77.59, X_{31} < 72.45, X_{33} > X_{31}, X_{32} > 72.45 | X_{01} = 37, X_{02} > 51.57, X_{03} > 37; \psi)$, is calculated as

$$\int_0^{72.45} \int_0^{77.59} \left. \frac{\partial^2 C(\mathcal{F}(u|X_{01} = 37; \phi_1, \rho), \mathcal{F}(w|X_{01} = 37; \phi_1, \rho); \rho^*)}{\partial u \partial w} \right|_{u=s, w=y}$$
$$\times \exp\left(- \int_0^s \lambda_3(1929.41 + v, v) dv \right) \exp\left(- \int_s^{77.59} \lambda_2(1929.41 + v, v) dv \right)$$
$$\times \exp\left(- \int_0^y \lambda_3(1934.55 + v, v) dv \right) \exp\left(- \int_y^{72.45} \lambda_2(1934.55 + v, v) dv \right) ds dy.$$

The follow-up date is available for the proband as they are in the registry. In the above example, the likelihood contribution is therefore

$$P(\bar{Z}_0(A_0^*) | Z_0(C_0) = 1, C_0, B_0; \phi) = \frac{P(\bar{Z}_0(A_0^*), B_0; \phi)}{P(Z_0(C_0) = 1, C_0, B_0; \phi)}, \tag{A.4}$$

with $A_0^* = 61.74$. The numerator of (A.4) is then given as

$$P(\bar{Z}_0(A_0^*), B_0; \phi) = P(X_{01} = 37, X_{03} > X_{01}, X_{02} > 61.74, B_0 = 1955.43; \phi)$$
$$= \lambda_1(37) \exp\left(- \int_0^{37} \lambda_1(s) ds \right) \exp\left(- \int_0^{37} \lambda_3(1955.43 + s, s) ds \right) \exp\left(- \int_{37}^{61.74} \lambda_2(1955.43 + s, s) ds \right),$$

and the denominator of (A.4) has the same form as (A.2). From the cross-sectional survey, we consider an individual who developed the disease by the age at contact for survey, denoted C_i. The likelihood contribution is

$$P(Z_i(C_i) = 1 | Z_i(C_i) \in \{0, 1\}, B_i) = \frac{P(Z_i(C_i) = 1, B_i; \phi)}{P(Z_i(C_i) = 0; \phi) + P(Z_i(C_i) = 1; \phi)}$$

where

$$P(Z_i(C_i) = 0, B_i; \phi) = \exp\left(- \int_0^{C_i} \lambda_1(u) du \right) \exp\left(- \int_0^{C_i} \lambda_3(B_i + u, u) du \right),$$

and

$$P(Z_i(C_i)=1,B_i;\phi)=\int_0^{C_i}\lambda_1(s)\exp\left(-\int_0^s\lambda_1(u)du\right)\exp\left(-\int_0^s\lambda_3(B_i+u,u)du\right)\times\exp\left(-\int_s^{C_i}\lambda_2(B_i+u,u)du\right)ds.$$

We use numerical integration based on Gaussian-Quadrature with 20 nodes to evaluate the integrals.

B Calculation of $P(G_{ijl})$

Recall $G_{ijl}=(G_{i0},G_{ij},G_{il})'$ is a vector of genetic markers for the proband and members j and l of family i; for families with two members we let $G_{ij}=(G_{i0},G_{ij})'$. We can calculate $P(\boldsymbol{G}_{ijl})$ based on the assumption that the process is in Hardy-Weinberg equilibrium and following Mendel's law, with a risk allele frequency p (Elandt-Johnson, 1971). In Table B.1 we consider the possible relationships between two or three members of a family and use G_p and G_c to denote the genotype of a parent or child respectively. If there are two parents we use G_{p_1} and G_{p_2} to distinguish them and if there are two children we use G_{c_1} and G_{c_2} respectively. The combination of binary markers within a pair (top half of Table B.1) or triple (bottom half of Table B.1) are given in the left column while the probabilities for a given set of relationships are given in the different columns.

C Sensitivity Analyses for the PsA Family Study

C.1 Sensitivity to the Assumption of Independent Competing Risks

In Section 2, we adopted the conventional assumption of independent competing risks when modeling cause-specific hazards for the disease onset time and disease-free mortality at the individual level. Here we report on a small simulation study conducted to examine the sensitivity of the finding regarding the within-family association in disease onset times to violations of the independent competing risks assumption (i.e. to violations of the assumption that X_{ij1} and X_{ij3} are independent). We adopt the same setting as in Section 4 of the manuscript but introduce a random effect u_{ij}, $j=0,\ldots,m_i$ which acts multiplicatively on the conditional cause-specific hazards for individual j in family i via

$$\lambda_1(t,a|b_{ij},u_{ij})=u_{ij}\lambda_1(a)\quad\text{and}\quad\lambda_3(t,a|b_{ij},u_{ij})=u_{ij}\lambda_3(t,a|b_{ij})\ .$$

When generating the data we take U_{ij} to be gamma distributed with mean 1 and variance 0.2, but we omit it from the analysis which was conducted as described in Section 4. Omission of this random effect will mimic the effect of omitting a shared covariate acting on the respective cause-specific hazards, which induces a dependence between X_{ij1} and X_{ij3}. This random effect

Table B.1: The joint model for the genetic markers for two (top) or three (bottom) family members according to their relationships

Joint distribution of alleles for different types of pairs of family members

$\boldsymbol{G}$		$P(G_p, G_p)$	$P(G_p, G_c)$	$P(G_c, G_c)$
1	1	$(1-q^2)^2$	$p^2q + p$	$\frac{1}{4}p^2(1+p)^2 + pq(2p+1)$
1	0	$(1-q^2)q^2$	pq^2	$\frac{1}{4}p^2q^2 + \frac{1}{2}pq^2(1+q)$
0	1	$(1-q^2)q^2$	pq^2	$\frac{1}{4}p^2q^2 + \frac{1}{2}pq^2(1+q)$
0	0	q^4	q^3	$\frac{1}{4}q^2(1+q)^2$

Joint distribution of alleles for different types of triples of family members

$\boldsymbol{G}$			$P(G_{p_1}, G_{p_2}, G_c)$	$P(G_p, G_{c_1}, G_{c_2})$	$P(G_{c_1}, G_{c_2}, G_{c_3})$
1	1	1	$p^2(1+2q)$	$\frac{1}{4}p^2(1+p)(5-3p) + \frac{1}{2}pq(p+pq+1)$	$\frac{1}{16}p^2(1+3p)(7-3p) + \frac{1}{4}pq(6p+3pq+2)$
1	1	0	p^2q^2	$\frac{1}{4}p^2q^2 + \frac{1}{2}pq^2$	$\frac{5}{16}p^2q^2 + \frac{1}{4}pq^2(1+q)$
1	0	1	pq^2	$\frac{1}{4}p^2q^2 + \frac{1}{2}pq^2$	$\frac{5}{16}p^2q^2 + \frac{1}{4}pq^2(1+q)$
1	0	0	pq^3	$\frac{1}{4}pq^2(1+q)$	$\frac{1}{16}p^2q^2 + \frac{1}{8}pq^2(1+3q)$
0	1	1	pq^2	$\frac{1}{2}pq^2(1+p)$	$\frac{5}{16}p^2q^2 + \frac{1}{4}pq^2(1+q)$
0	1	0	pq^3	$\frac{1}{2}pq^3$	$\frac{1}{16}p^2q^2 + \frac{1}{8}pq^2(1+3q)$
0	0	1	0	$\frac{1}{2}pq^3$	$\frac{1}{16}p^2q^2 + \frac{1}{8}pq^2(1+3q)$
0	0	0	q^4	$\frac{1}{2}q^3(1+q)$	$\frac{1}{16}q^2(1+3q)^2$

also introduces another component of variation which we anticipate will create problems when estimating the within-family dependence parameter indexing the copula (i.e. τ). Finally this will lead to a violation of the assumption of proportional mortality rates (i.e. the disease-free and post-disease mortality rates). While we change the nature of the data generation, we adopt the same model as before and report the empirical properties of the resultant estimates in Table C.1.

We find that bias in the estimates of τ can be substantial and in particular that this mis-specification leads to an underestimation of the dependence among the disease onset within families; this bias becomes larger with larger values of τ. Bias in estimation of λ_{01} is also apparent. Inclusion of current status data through the use of the augmented composite likelihood accentuates the bias of λ_{01} when $\tau = 0.2$. The bias in ν is likewise greater for design II compared to design I.

In summary, as is the case with any models based on cause-specific hazard functions, the findings from the proposed analyses are sensitive to this type of violation of the independent competing risks assumption. In the current setting, we focus on estimation of the dependence parameter of the copula function which is conservatively biased in the settings considered,

implying that one could under-estimate the extent of familial aggregation. We expect to report on further research exploring the use of different dependence models for X_{ij1} and X_{ij3} in a future manuscript; these may involve use of a more highly parameterized and higher dimensional copula function accommodating different types of dependence.

Table C.1: Results of a simulation study exploring the sensitivity of the proposed model to violations of the independent competing risks assumption arising from a shared gamma distributed random effect with mean one and variance 0.20; design I and II are considered with auxiliary registry data alone and in combination with current status survey data; a Clayton copula is used with Kendall's τ=0.2, 0.4; $n_F = 1000$, $n_1 = 2000$, $n_2 = 1000$, and $nsim = 1000$

			Registry Data				Registry + Current Status Data			
Design	τ	PARAMETER	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
I	0.2	$\log \lambda_1$	-0.003	0.057	0.054	0.934	-0.038	0.041	0.041	0.841
		$\log \nu$	-0.004	0.038	0.039	0.945	-0.003	0.038	0.039	0.945
		τ	-0.050	0.027	0.028	0.517	-0.033	0.023	0.023	0.716
	0.4	$\log \lambda_1$	0.059	0.070	0.068	0.826	-0.023	0.045	0.045	0.933
		$\log \nu$	-0.001	0.038	0.037	0.943	-0.001	0.038	0.037	0.942
		τ	-0.098	0.031	0.031	0.132	-0.062	0.023	0.023	0.243
II	0.2	$\log \lambda_1$	-0.005	0.063	0.058	0.932	-0.039	0.042	0.042	0.862
		$\log \nu$	0.016	0.047	0.047	0.944	0.014	0.046	0.047	0.947
		τ	-0.054	0.029	0.029	0.525	-0.038	0.024	0.023	0.627
	0.4	$\log \lambda_1$	0.056	0.072	0.071	0.852	-0.022	0.047	0.046	0.928
		$\log \nu$	0.021	0.047	0.048	0.933	0.016	0.046	0.047	0.942
		τ	-0.105	0.033	0.033	0.135	-0.070	0.024	0.024	0.174

C.2 The proportional mortality assumption among those diseased vs. disease-free

To address the assumptions of proportional mortality and the choice of copula functions, we now perform additional sensitivity analyses in the psoriatic arthritis family study. Specifically we use a Clayton copula function and a Frank copula function and specify a piecewise proportional hazards model relating the post-disease vs. disease-free. The cut points are set at 60 and 70 years of age to ensure roughly one third of the post-disease deaths occurred in each interval. Thus in (1), we let $\nu(a) = \nu_j$ for $a_{j=1} \leq \nu_j < a_j$, $j = 1, 2, 3$, where $a_0 = 0$, $a_1 = 60$, $a_2 = 70$, and $a_3 = \infty$. The results are reported in Table C.2

We find a substantial increase in the mortality among younger individuals with disease compared to those disease-free, but the relative mortality decreases as individuals enter the older age intervals. The estimate of Kendall's τ appears relatively insensitive to the proportional mortality assumption since the estimate is quite close to the estimate reported in Section 6 for the Clayton copula.

When the Frank copula function is used, a smaller value of Kendall's *tau* is obtained compared to the Clayton copula function. This may be due to the fact that unlike the Frank

copula, the Clayton copula is asymmetric in that there is a greater dependence in the negative tail than in the positive tail. However, all parameter estimates apart from Kendall's τ are quite similar for the fitted model with the Clayton copula function. Importantly, the estimates of the effect of the HLA markers B27 and C06 appear relatively robust to the proportional mortality assumption and the copula function.

Table C.2: Sensitivity analysis of the proposed model to violations of proportional mortality assumption for post-disease death and a Clayton copula function. Parameter estimates are based on the augmented pairwise likelihood; auxiliary data include the University of Toronto Psoriatic Arthritis Registry and data from the national survey by Gelfand et al. (2005) without/with genotype data under the piecewise constant marginal model for the age at PsA onset with cut points 28, 38 and 48 years of age; a piecewise proportional hazards model was adopted for post- disease death (compared to disease-free death) with cut-points at 60 and 70 years of age; models based on Clayton and Frank copula functions fitted.

MARKER	α_{marker}	ν_1	ν_2	ν_3	τ	p_{marker}
			Clayton copula			
-	-	1.542 (0.215)	1.374 (0.176)	1.053 (0.094)	0.331 (0.094)	-
B27	0.336 (0.233)	1.537 (0.215)	1.368 (0.175)	1.054 (0.094)	0.328 (0.095)	0.065 (0.013)
C06	-0.214 (0.181)	1.537 (0.215)	1.368 (0.175)	1.054 (0.094)	0.323 (0.094)	0.169 (0.023)
			Frank copula			
-	-	1.542 (0.216)	1.374 (0.176)	1.053 (0.094)	0.178 (0.060)	-
B27	0.343 (0.236)	1.537 (0.215)	1.368 (0.175)	1.054 (0.094)	0.177 (0.061)	0.065 (0.013)
C06	-0.215 (0.181)	1.537 (0.215)	1.368 (0.175)	1.054 (0.094)	0.173 (0.060)	0.169 (0.023)

References

Elandt-Johnson, R. C. (1971). Joint genotype distributions of s children and a parent, and of s siblings: multiple alleles. *American Journal of Human Genetics*, 23(5):442.

Gelfand, J. M., Gladman, D. D., Mease, P. J., Smith, N., Margolis, D. J., Nijsten, T., Stern, R. S., Feldman, S. R., and Rolstad, T. (2005). Epidemiology of psoriatic arthritis in the population of the United States. *Journal of the American Academy of Dermatology*, 53(4):573–e1.