Noname manuscript No.
(will be inserted by the editor)

Second-order estimating equations for clustered current status data from family studies using response-dependent sampling

Yujie Zhong · Richard J. Cook

Received: date / Accepted: date

Abstract Studies about the genetic basis for disease are routinely conducted through family studies under response-dependent sampling in which affected individuals called *probands* are sampled from a disease registry, and their respective family members (*non-probands*) are recruited for study. The extent to which the dependence in some feature of the disease process (e.g. presence, age of onset, severity) varies according to the kinship of individuals, reflects the evidence of a genetic cause for disease. When the probands are selected from a disease registry it is common for them to provide quite detailed information regarding their disease history, but non-probands often simply provide their disease status at the time of contact. We develop conditional second-order estimating equations for studying the nature and extent of within-family dependence which recognizes the biased sampling scheme employed in family studies and the current status data provided by the non-probands. Simulation studies are carried out to evaluate the finite sample performance of different estimating functions and to quantify the empirical relative efficiency of the various methods. Sensitivity to model misspecification is also explored. An application to a motivating psoriatic arthritis family study is given for illustration.

Keywords current status data · family study · Gaussian copula · relative efficiency · response-dependent sampling · robustness · second-order estimating equations

Y. Zhong
MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge CB2 0SR, UK
E-mail: yujie.zhong@mrc-bsu.cam.ac.uk

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

1 Introduction

1.1 Introduction

The heritable nature of disease can be inferred by the structure of the within-family dependence in disease manifestation [19]. For rare diseases, population-based cohort studies are inefficient and impractical, so response-dependent biased family designs are routinely employed to obtain enriched samples with higher representation of diseased individuals and more variation in genetic markers than would be seen in the unselected population. Much work has been carried out for the analysis of such data when the disease status is modeled as a binary trait. Conditional likelihood based on generalized linear mixed models can be used for dealing with the dependence of binary phenotypes within families [4,3], or estimating equations can be formed by specifying marginal mean and dependence structures for the analysis of binary phenotypes from case-control family studies [32].

The age of onset for many chronic diseases is highly variable, however, and simply using the binary trait of the disease status does not account for the variable times individuals have been at risk for disease in family studies. MacLean et al. [21] and Shih and Chatterjee [27] point out that the estimators of the covariate effects on the disease process may be less efficient and the degree of familial aggregation may be underestimated if information on the age of onset and the effect of censoring are not addressed. Models which consider the disease onset time distribution and measure dependence in terms of these times offer a preferable framework for analysis.

When interest in genetic effect or gene-environment treatment effect, case-control or case-only family study are commonly used. Li et al. [20], Hsu et al. [15] and Shih and Chatterjee [27] proposed likelihood methods based on disease onset time for case-control family study and Chatterjee et al. [6] proposed methods to estimate the relative risk, cumulative risk, and residual familial aggregation for case-control family data and modified method for case-only family data. For their methods, modeling and estimation of the residual familial aggregation is key to adjustment for ascertainment bias but this is done using an exchangeable dependence structure in which the association is the same for different pairs of relatives. Gorfine et al. [14] use the frailty models to account for heterogeneity in familial risk, but pointed out that frailty-based methods may be affected by the uncertainty on the frailty parameter estimate.

In this article, we consider a simple family study, where an affected individual called a *proband* is selected from a registry of patients. Consenting family members (*non-probands*) of each proband are then recruited and examined to collect information on their disease status [5]. Probands are given a special designation because their disease status led to the selection of their family. Under such sampling scheme, one obtains a right-truncated onset time for the proband, and current status (type I interval-censored) data for non-probands [28]. While work has been done on the analysis of multivariate current status data [16,8], little has been done to our knowledge in the context of biased sampling schemes.

Insight into the genetic basis for disease can be gained by comparing the strength of the association in disease status between pairs of family members with different kinships [7,18]. More elaborate dependence modeling also plays a central role when studying the "parent of descent" hypothesis, where the primary goal is to estimate and compare the strength of father-child and mother-child associations in phenotype to elucidate the role of the sex chromosomes in disease transmission [30]. With this in mind, we consider copula models [22] as a basis for modeling the joint risk of disease among family members. The dependence parameters can be interpreted as reflecting "residual familial aggregation" that is not explained by covariates in the marginal models. Copula models have several advantages over frailty models. First, the marginal models still retain simple interpretation when using copula models, which is not the case under the frailty model. Second, copula models yield dependence measures which are functionally independent of the parameters in the marginal onset time distribution, so the marginal distribution can be specified in any desirable way. Third, the dependence measure is directly specified under the copula model which has clear meaning and it also provides a natural basis for regression of genetic effects, but the frailty model do not provide simple measures of within-family dependence and which is difficult to interpret the meaning of the dependence.

Analyses must address the biased sampling scheme employed in these studies. Likelihood contributions from each family which are proportional to joint probability functions for the phenotypes of non-probands *conditional* on the disease status of the proband will admit valid inference [29] under correct model specification, but enumeration of all possible sample outcomes can be computationally demanding with large families. We develop a class of conditional second-order estimating equations in the spirit of Prentice [25]. We use the term *conditional* to reflect the fact that moments in the second-order estimating equation are all conditional on the disease onset time of the proband. A supplementary estimating equation is incorporated to extract the limited information about the marginal onset time distribution from the proband.

1.2 The University of Toronto Psoriatic Arthritis Family Study

The incidence of psoriatic arthritis (PsA) is reported to be between 0.3 and 1.0% [9] and hereditary factors are thought to be important, as some studies have suggested that close blood relatives of individuals affected by psoriatic arthritis are at higher risk of developing the disease compared to the general population. Characterizing the within-family association nature and identifying important genetic risk factors are important to understand the disease etiology. Particular interest lies in assessing whether there is a higher rate of paternal, rather than maternal, transmission of the disease, which is also called "parent of origin" effect [2]. A family study of psoriatic arthritis is conducted in the Centre for Prognosis Studies in the Rheumatic Disease at the University of Toronto. Probands were selected from the members of the University of Toronto Psoriatic Arthritis Registry, and their family members were recruited into the family study with their consent. A total of 169 two-generation families ranging in size from 2 to 7 individuals were recruited; 54 families were comprised of only one non-proband and 115 have more than one non-proband. The disease onset times were only available for probands, but for other family members only the disease status is available when they are examined, yielding current status data. In total 538 individuals are in the family study and only 194 (169 probands and 25 nonprobands) were diagnosed with PsA. Except for the demographic data, information of some HLA markers is also available for individuals in the PsA family study. We focus on identifying the significant HLA markers for the psoriatic arthritis and characterizing the within-family association structure, also testing whether there is "parent of origin" effect for the psoriatic arthritis.

The remainder of this paper is organized as follows. In Sect. 2 we define notation and formulate the conditional second-order estimating equation for family data under response-dependent sampling, which are a combination of right-truncated onset time from probands and current status data from non-probands. We consider an illustrative example in which the dependence structure is governed by a Gaussian copula and work with this model in subsequent calculations and simulations where we examine specific estimating equations involving different derivative matrices and working independence assumptions. In Sect. 3 we explore the asymptotic relative efficiencies and finite sample properties of estimators from several variants of the estimating equations introduced in Sect. 2; these results also permit sample size calculations for planning studies aiming to detect effects of genetic markers. The impact of misspecification of the dependence structure on properties of estimators and power of genetic tests is investigated in Sect. 4. An application to the motivating psoriatic arthritis family study is given in Sect. 5 in which we assess the genetic basis for the disease. Concluding remarks are given in Sect. 6.

2 Conditional Estimating Equations Under Biased Sampling

2.1 Notation, Sampling and Observation Scheme for Family Studies

We consider the setting in which a registry of M individuals is created by selecting a random sample from a population, screening each individual for disease, and recruiting those found to have the condition of interest [10]. If C_{i0} denotes the age of individual 0 in family i at the time of sampling and screening, and T_{i0} denotes their age of disease onset, then this individual is recruited to the registry if $Y_{i0} = I(T_{i0} \leq C_{i0}) = 1$; we assume T_{i0} is verifiable by a review of medical records for individuals recruited to the registry. When a family study is carried out, we assume that probands are selected from the disease registry by simple random sampling and without loss of generality we label the families of selected probands $i = 1, \ldots, m$.

We let T_{ij} and X_{ij} denote the event time and a $p \times 1$ covariate vector of individual j in family i, where $j = 1, \ldots, n_i$ are the labels for the non-probands. Then if $T_i = (T_{i0}, T_{i1}, \ldots, T_{in_i})'$ and $X_i = (X'_{i0}, \ldots, X'_{in_i})'$, we write the joint cumulative distribution function (j.c.d.f) for family i as $F_i(t) = P(T_{i0} \leq t_0, \ldots, T_{in_i} \leq t_{n_i}|X_i)$. We assume $T_{ij} \perp X_i^{(-j)}|X_{ij}$ where $X_i^{(-j)} = \{X_{ij'}; j' \neq j, 0 \leq j' \leq n_i\}$, and write $F_{ij}(t;\theta) = P(T_{ij} \leq t|X_{ij};\theta)$. The marginal hazard function for the disease onset time of individual j, $j = 0, 1, \ldots, n_i$, in family i is

$$\lambda_{ij}(t|X_{ij};\theta) = \lim_{\varDelta t \downarrow 0} \frac{P(t \leq T_{ij} < t + \varDelta t | t \leq T_{ij}, X_{ij}; \theta)}{\varDelta t} \ ,$$

where we write $\lambda_{ij}(t|X_{ij};\theta) = \lambda_0(t;\alpha)\exp(X'_{ij}\beta)$ under a proportional hazards formulation. This gives $F_{ij}(t_{ij};\theta) = 1 - \exp(-\varLambda_0(t_{ij};\alpha)\exp(X'_{ij}\beta))$ where $\varLambda_0(t_{ij};\alpha) = \int_0^{t_{ij}} \lambda_0(s;\alpha)ds$, α is a $q \times 1$ vector, β is a $p \times 1$ vector of regression coefficients, $\theta = (\alpha', \beta')'$. We let γ parameterize the within-family dependence and $\psi = (\theta', \gamma')'$.

Classification of non-probands with respect to their disease status is made at the time of recruitment and clinical examination, yielding current status data. Let C_{ij} denote the age of non-proband j in family i at the time of assessment and let $Y_{ij} = I(T_{ij} \leq C_{ij})$; we let $\bar{C}_i = (C_{i1}, \ldots, C_{in_i})'$, $\bar{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ and $\bar{X}_i = (X'_{i1}, \ldots, X'_{in_i})'$. If $Y_i = (Y_{i0}, \bar{Y}'_i)'$, $C_i = (C_{i0}, \bar{C}'_i)'$ and $X_i = (X'_{i0}, \bar{X}'_i)'$, the family data therefore consist of $\{T_{i0}, Y_i, C_i, X_i\}$ subject to $Y_{i0} = 1$.

2.2 Second-order Estimating Functions

The association parameter γ is of central importance here so we next formulate conditional second-order generalized estimating equations in the spirit of Prentice [25] and Zhao and Prentice [33].

Let $\bar{Z}_i = (Y_{i1}Y_{i2}, Y_{i1}Y_{i3}, \ldots, Y_{i1}Y_{in_i}, Y_{i2}Y_{i3}, \ldots, Y_{i,n_i-1}Y_{in_i})'$ be an $r_i \times 1$ vector of pairwise products of the elements in $\bar{Y}_i$, where $r_i = n_i(n_i-1)/2$; we let Z_{ijk} denote the element of $\bar{Z}_i$ corresponding to the pair (j,k) in family i. To account for response-biased sampling we define conditional moments and let $\mu_i = E[\bar{Y}_i|T_{i0};\psi]$ and $\eta_i = E[\bar{Z}_i|T_{i0};\psi]$ be the contributions from the non-probands and let $\mu_{i0} = E[T_{i0}|Y_{i0} = 1;\theta]$ for the proband where we suppress the dependence on X_i and C_i. The conditional second-order estimating equations (CGEE2) denoted by $U(\psi) = \sum_{i=1}^m U_i(\psi) = 0$ have the form

$$U_i(\psi) = G'_i W_i^{-1} R_i + D'_i V_i^{-1}(T_{i0} - \mu_{i0}) \ , \tag{1}$$

with

$$G_i = \begin{pmatrix} G_{i11} & G_{i12} \\ G_{i21} & G_{i22} \end{pmatrix} \ , \quad W_i = \begin{pmatrix} W_{i11} & W_{i12} \\ W'_{i12} & W_{i22} \end{pmatrix} \ , \quad \text{and} \quad R_i = \begin{pmatrix} \bar{Y}_i - \mu_i \\ \bar{Z}_i - \eta_i \end{pmatrix} \ ,$$

where $G_{i11} = \partial\mu_i/\partial\theta'$, $G_{i12} = \partial\mu_i/\partial\gamma'$, $G_{i21} = \partial\eta_i/\partial\theta'$, and $G_{i22} = \partial\eta_i/\partial\gamma'$, $W_{i11} = \mathrm{Cov}(\bar{Y}_i, \bar{Y}_i'|T_{i0})$, $W_{i22} = \mathrm{Cov}(\bar{Z}_i, \bar{Z}_i'|T_{i0})$, and $W_{i12} = \mathrm{Cov}(\bar{Y}_i, \bar{Z}_i'|T_{i0})$; note that unlike standard GEE2, $G_{i12} \neq \mathbf{0}$ since $\mu_i = E[\bar{Y}_i|T_{i0};\psi]$ is functionally dependent on γ. The covariance matrices can be parameterized by the marginal and association parameters where the latter may be specified in terms of Kendall's τ; an example is given in Sect. 2.3. Consistent estimation of ψ is possible based on the first term in (1), but the second term $D_i'V_i^{-1}(T_{i0} - \mu_{i0})$ where $D_i = \partial\mu_{i0}/\partial\psi'$ and $V_i = \mathrm{Var}(T_{i0}|Y_{i0} = 1)$, improves efficiency by exploiting the data on the onset time from the proband.

Subject to correct specification of the conditional moments, (1) is an unbiased estimating function, so the estimator $\widehat{\psi}$ solving $U(\psi) = 0$ is consistent with an asymptotic normal distribution

$$\sqrt{m}(\widehat{\psi} - \psi) \xrightarrow{d} N\left(0, \mathcal{A}^{-1}(\psi)\mathcal{B}(\psi)\left[\mathcal{A}^{-1}(\psi))\right]'\right) \,, \tag{2}$$

where

$$\mathcal{A}(\psi) = E[-\partial U_i(\psi)/\partial\psi'] \quad \text{and} \quad \mathcal{B}(\psi) = E[U_i(\psi)U_i'(\psi)] \,.$$

Natural empirical estimates of these matrices are

$$A(\widehat{\psi}) = \frac{1}{m}\sum_{i=1}^{m}\left\{\widehat{G}_i'\,\widehat{W}_i^{-1}\widehat{G}_i + \widehat{D}_i'\,\widehat{V}_i^{-1}\widehat{D}_i\right\} \,, \tag{3}$$

and

$$B(\widehat{\psi}) = \frac{1}{m}\sum_{i=1}^{m}\left\{\widehat{G}_i'\,\widehat{W}_i^{-1}\widehat{R}_i\widehat{R}_i'\widehat{W}_i^{-1}\widehat{G}_i + \widehat{D}_i'\,\widehat{V}_i^{-1}(T_{i0} - \widehat{\mu}_{i0})^2\widehat{V}_i^{-1}\widehat{D}_i\right\} \,,$$

which yield $\widehat{\mathrm{asvar}}(\sqrt{m}(\widehat{\psi} - \psi)) = A^{-1}(\widehat{\psi})B(\widehat{\psi})[A^{-1}(\widehat{\psi})]'$.

Simplified forms of G_i can be obtained by setting $G_{i21} = \partial\eta_i/\partial\theta' = \mathbf{0}$ (denoted by G^{I}) or by letting both $G_{i12} = \partial\mu_i/\partial\gamma' = \mathbf{0}$ and $G_{i21} = \partial\eta_i/\partial\theta' = \mathbf{0}$ (denoted by G^{II}). It is also common to simplify W_i and adopt a form in which $W_{i12} = W_{i21}' = \mathbf{0}$ and $W_{i22} = \mathrm{diag}\{\eta_i(1 - \eta_i)\}$ while retaining the full structure of W_{i11}; we refer to this as a *working partial independence* (WPI) matrix. Combining these simplifications we consider four different estimating functions based on (1):

A. Full G_i and Full covariance matrix W_i denoted G-W
B. Full G_i and WPI W_i denoted G-WPI
C. G^{I} and WPI W_i denoted G^{I}-WPI, and
D. G^{II} and WPI W_i denoted G^{II}-WPI.

2.3 An Illustrative Dependence Structure Based on a Gaussian Copula

The specific form of the moments for $T_i|T_{i0}, X_i$ can be motivated by a copula model. Consider an $(n_i + 1) \times 1$ vector of uniform $[0, 1]$ variables $K_i = (K_{i0}, K_{i1}, \ldots, K_{in_i})'$, in which $K_{ij} = F_{ij}(t_{ij}; \theta)$, $j = 0, \ldots, n_i$. The j.c.d.f. for K_i, denoted by $H_{n_i+1}(k; \gamma) = P(K_{i0} \le k_{i0}, K_{i1} \le k_{i1}, \ldots, K_{in_i} \le k_{in_i}; \gamma)$, is a copula function in $n_i + 1$ dimensions indexed by an $r \times 1$ parameter γ which characterizes the dependence [17, 22]. The Gaussian copula is a member of elliptical family of the form

$$H_{n_i+1}(k_{i0}, \ldots, k_{in_i}; \gamma) = \Phi_{n_i+1}\big(\Phi^{-1}(k_{i0}), \ldots, \Phi^{-1}(k_{in_i}); \gamma\big) \ ,$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of a standard normal random variable (r.v.), $\Phi_{n_i+1}(\cdot\,; \gamma)$ is the j.c.d.f. of an $(n_i + 1) \times 1$ multivariate normal r.v. with mean zero and $(n_i + 1) \times (n_i + 1)$ correlation matrix Σ_i; Σ_i is indexed by γ and we denote the off-diagonal entries by σ_{ijk}, $j \ne k = 0, \ldots, n_i$. Specification of the Gaussian copula for K_i induces a joint distribution for $T_i|X_i$ given by

$$P(T_{i0} \le t_{i0}, \ldots, T_{in_i} \le t_{in_i}|X_i; \psi) = \int_{-\infty}^{q_{i0}} \cdots \int_{-\infty}^{q_{in_i}} \frac{\exp(-s_i' \Sigma_i^{-1} s_i/2)}{\sqrt{(2\pi)^{n_i+1}|\Sigma_i|}} ds_{i0} \cdots ds_{in_i} \ , \tag{4}$$

where $S_i \sim \mathrm{MVN}_{n_i+1}(0, \Sigma_i)$, s_i is a realization, and $q_{ij} = \Phi^{-1}(F_{ij}(t_{ij}; \theta))$, $j = 0, \ldots, n_i$. Copula functions such as this are attractive for dependence modeling since pairwise associations are parameterized to be functionally independent of the marginal parameters and different pairwise associations is permitted. The Kendall's τ characterizing the association between T_{ij} and T_{ik} given X_i, for example, is given by $\tau_{ijk} = 2\mathrm{asin}(\sigma_{ijk})/\pi$, $0 \le j < k \le n_i$. Regression modeling of the within-family dependence can be achieved by specifying a second-order model of the form $g(\tau_{ijk}) = v_{ijk}'\gamma$ where $g(\cdot)$ is a 1-1 differentiable link function mapping Kendall's τ onto the real line, v_{ijk} is an $r \times 1$ covariate vector characterizing individuals j and k in family i and their relationship, and γ is the corresponding $r \times 1$ vector of coefficients. This second-order regression model can be helpful when investigating the effect of risk factors on the pairwise association as v_{ijk} could represent family-level or individual-level features, or information on the kinship of individuals j and k in family i; inference on their effects can be easily carried out based on γ. For example, in the PsA family study with two generations, when the "parent-of-origin" hypothesis is of interest, we can formulate the second-order model as

$$\begin{aligned} g(\tau_{ijk}) = \gamma_0 &+ \gamma_1 \,\mathrm{I}((j,k) \text{ pair are siblings}) + \gamma_2 \,\mathrm{I}((j,k) \text{ pair is father-child}) \\ &+ \gamma_3 \,\mathrm{I}((j,k) \text{ pair is mother-child}) \end{aligned}$$

then comparing γ_2 and γ_3 (or testing $H_0 : \gamma_2 = \gamma_3$) can inform us whether there is "parent-of-origin" effect in the onset time of PsA. More elaborate models which incorporate genetic covariates into the dependence model can also be specified.

Returning to the estimating function in (1), based on the Gaussian copula we have $\mu_{ij} = E[Y_{ij}|T_{i0}] = P(T_{ij} \le C_{ij}|T_{i0}) = \Phi((q_{ij} - \sigma_{i0j}q_{i0})/(1-\sigma_{i0j}^2)^{1/2})$ and $\eta_{ijk} = E[Y_{ij}Y_{ik}|T_{i0}] = \Phi_2\big((q_{ij} - \sigma_{i0j}q_{i0}), (q_{ik} - \sigma_{i0k}q_{i0}); \Sigma_{jk|0}\big)$ where $q_{ij} = \Phi^{-1}(F_{ij}(C_{ij}))$, $j = 1, \dots, n_i$, and $q_{i0} = \Phi^{-1}(F_{i0}(t_{i0}))$. The function $\Phi_2(\cdot, \cdot\, ; \Sigma_{jk|0})$ is the j.c.d.f of a bivariate normal r.v. with mean zero and covariance matrix $\Sigma_{jk|0}$, where

$$\Sigma_{jk|0} = \begin{pmatrix} 1-\sigma_{i0j}^2 & \sigma_{ijk} - \sigma_{i0j}\sigma_{i0k} \\ \sigma_{ijk} - \sigma_{i0j}\sigma_{i0k} & 1-\sigma_{i0k}^2 \end{pmatrix} .$$

The entries of W_i can also be derived based on the Gaussian copula where, for example, $\text{cov}(Y_{il}, Z_{ijk}|T_{i0}) = E[Y_{il}Y_{ij}Y_{ik}|T_{i0}] - \mu_{il}\eta_{ijk}$ for $k \neq l \neq j$ with

$$E[Y_{il}Y_{ij}Y_{ik}|T_{i0}] = \phi^{-1}(q_{i0}) \cdot \int_{-\infty}^{q_{il}} \int_{-\infty}^{q_{ij}} \int_{-\infty}^{q_{ik}} \phi_4\big(q_{i0}, s_{il}, s_{ij}, s_{ik}\, ;\ \Sigma_i(0,l,j,k)\big) ds_{ik} ds_{ij} ds_{il} .$$

Note that this is a j.c.d.f for a multivariate normal r.v. with mean $\mu^\dagger$ and covariance matrix $\Gamma^\dagger$ denoted by $\Phi_3(q_{il}, q_{ij}, q_{ik}\, ;\ \mu^\dagger, \Gamma^\dagger)$ where $\mu^\dagger = (\sigma_{i0l}q_{i0}, \sigma_{i0j}q_{i0}, \sigma_{i0k}q_{i0})'$ and

$$\Gamma^\dagger = \begin{pmatrix} 1-\sigma_{i0l}^2 & \sigma_{ilj} - \sigma_{i0l}\sigma_{i0j} & \sigma_{ilk} - \sigma_{i0l}\sigma_{i0k} \\ \sigma_{ilj} - \sigma_{i0l}\sigma_{i0j} & 1-\sigma_{i0j}^2 & \sigma_{ijk} - \sigma_{i0j}\sigma_{i0k} \\ \sigma_{ilk} - \sigma_{i0l}\sigma_{i0k} & \sigma_{ijk} - \sigma_{i0j}\sigma_{i0k} & 1-\sigma_{i0k}^2 \end{pmatrix} .$$

These conditional moments are easily derived under a Gaussian copula.

3 Relative Efficiency Under Particular Estimating Equations

3.1 A Study of Asymptotic Relative Efficiency

Here we examine the asymptotic relative efficiency of four different conditional estimating equations as a function of the strength of the within-family association through the functions

$$\text{ARE}_\text{B}(\widehat{\psi}) = \frac{\text{asvar}_\text{A}(\widehat{\psi})}{\text{asvar}_\text{B}(\widehat{\psi})} \ ,\ \text{ARE}_\text{C}(\widehat{\psi}) = \frac{\text{asvar}_\text{A}(\widehat{\psi})}{\text{asvar}_\text{C}(\widehat{\psi})} \ ,\ \text{and } \text{ARE}_\text{D}(\widehat{\psi}) = \frac{\text{asvar}_\text{A}(\widehat{\psi})}{\text{asvar}_\text{D}(\widehat{\psi})} \ ,$$

where asvar() denotes an asymptotic variance and its subscript indexes the adopted conditional estimating equations proposed in Sect. 2.2 . All three simplified conditional estimating equations are compared with the conditional estimating equations with full G_i and full covariance matrix W_i.

Consider two-generation families comprised of two parents and two children; $n_i = 3$. The proband is randomly selected from the four family members, and is indexed by $j = 0$. A Weibull distribution is adopted for the onset time for all family members; $\mathcal{F}(t_{ij}|X_{ij}; \theta) = \exp\big(-(\lambda t_{ij})^\kappa \exp(X_{ij}\beta)\big)$ where X_{ij} is a binary variable with $P(X_{ij} = 1) = 0.5$, $j = 0, 1, 2, 3$, and we assume that $X_{ij} \perp X_{ik}$, $j \neq k$; $\theta = (\lambda, \kappa, \beta)'$. Let $\kappa = 1.2$, $\beta = \log 1.2$, and choose λ to give a median age of 45 years for disease onset for group with $X_{ij} = 0$. The clinic entry time for the proband C_{i0} is normally distributed with mean 50 and variance 20, and families are recruited into the study only if their probands satisfy the selection condition $T_{i0} \leq C_{i0}$. For non-proband j in the selected family i, let C_{ij} be the random age of contact, following $N(\mu = 60, \sigma^2 = 10)$ for individuals in the first generation and $N(\mu = 30, \sigma^2 = 10)$ for the individuals in the second generation, $j = 1, 2, 3$; the age at contact for individuals in both generation are truncated at 90 years. We consider a Gaussian copula to induce an exchangeable within-family association for simplicity here, and let Kendall's τ vary from 0 to 0.5 to reflect independence to strong within-family association. The second-order model with a Fisher transformation link function is simply $\log\big((1+\tau_{ijk})/(1-\tau_{ijk})\big) = \gamma_0$, $0 \leq j < k \leq 3$. The asymptotic variances of estimators based on conditional estimating equations in (2) are approximated by Monte Carlo simulation based on 20,000 samples.

Fig. 1 shows the trends of asymptotic relative efficiencies of estimators under different conditional estimating equations as a function of the within-family association. It is apparent that the conditional estimating equations with full G_i and full W_i (G-W) leads to the most efficient estimators, and the efficiency gain is most appreciable for the association parameter. With the WPI matrix W_i, adopting G^{I} yields more efficient estimators than that using G^{II}, especially when the within-family association is strong. This makes sense as the former utilizes additional information about γ from the conditional mean μ_i. The conditional estimating equations with the full G_i and WPI matrix W_i (G-WPI) perform worse than other approaches when the association is less than 0.45; see Fig. 1. This indicates that with a working covariance matrix, using the full derivative matrix increases the complexity, but does not improve efficiency; on the contrary it leads to less efficient estimators. This is similar to the findings reported by Balemi and Lee [1] where they compare the performance of GEE1 and GEE2 estimators for clustered binary data.

3.2 Finite Sample Study of the Conditional Estimating Equations

Here we conduct a simulation study to assess the validity and finite sample performance of these four conditional estimating equations for family data from response-dependent sampling. The parameter settings are the same as in Sect. 3.1 and we let Kendall's $\tau = 0.0, 0.2$ and 0.4 for an exchangeable Gaussian copula. We also consider a more general case where the within-family association is induced by a Gaussian copula with structured correlation matrix. For the two-generation families comprised of two parents and two children, let Kendall's $\tau_{pp} = 0.1$ for parents, Kendall's $\tau_{pc} = 0.2$ for parent-child, and Kendall's $\tau_{ss} = 0.4$ for siblings to reflect the existence of both environmental and genetic effect on the age of onset. One thousand datasets of $m = 200$ and 1000 ascertained families are generated, the four proposed conditional estimating equations are used for analysis and the empirical properties of estimates of β and γ_0 are summarized in Table 1 for the exchangeable Gaussian copula and β and $\gamma = (\gamma_0, \gamma_1, \gamma_2)'$ for the structured Gaussian copula in Table 2. The

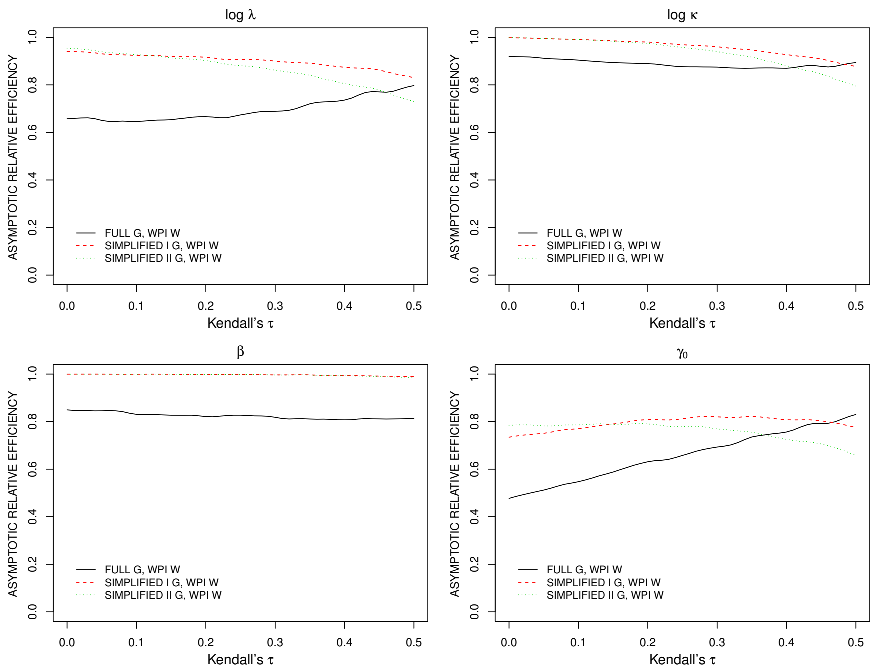

Fig. 1 Asymptotic relative efficiencies of estimators under conditional estimating equations with full G_i and WPI W_i (ARE_B), simplified G^I and WPI W_i (ARE_C), simplified G^II and WPI W_i (ARE_D) compared with that under conditional estimating equation with full G_i and full covariance matrix W_i; within-family dependence of disease onset times is induced by a Gaussian copula with exchangeable structure with Kendall's τ varying from 0 to 0.5; $(\log\lambda, \log\kappa, \beta) = (-4.11, \log 1.2, \log 1.2)$, $n_i = 3$, $m = 20,000$.

performance of the estimators for the parameters of the baseline hazard was excellent under the (correct) Weibull specification in all settings and so we do not tabulate these results.

The results under the exchangeable Gaussian copula in Table 1 show that when the Weibull model is specified for the onset time distribution the empirical biases are negligible for all conditional estimating equations; there is very slight finite sample bias for the association parameter when $m = 200$ and the within-family association is strong (Kendall's $\tau = 0.4$). The empirical standard errors (ESE) agree with the average standard errors (ASE) based on the robust variance form, and the empirical coverage probabilities (ECP) of nominal 95% confidence intervals are in general within the acceptable range. Consistent with the theoretical results of Sect. 3.1, the greatest efficiency came from the conditional estimating equations with the full derivative matrix and full covariance matrix (G-W), followed by those with G^I and WPI matrix W_i (G^I-WPI). The empirical performance of the conditional estimating equations with the full G_i and WPI matrix W_i is worse than others, again in alignment with the conclusion based on Fig. 1.

Table 1 Empirical properties of estimates under conditional estimating equations for family data from response-dependent sampling, where within-family association is induced by Gaussian copula with exchangeable structure; parametric margins and piecewise constant baseline hazards (3 pieces) are considered†; $n_i = 3$, $nsim = 1000$.

		Weibull Margin								Piecewise Constant Margin)†								
		β				γ_0				β				γ_0				
τ	EE‡	EBIAS	ESE	ASE	ECP%	EBIAS	ESE	ASE	ECP%	EBIAS	ESE	ASE	ECP%	EBIAS	ESE	ASE	ECP%	NC %
									$m = 200$									
0.0	G-W	0.002	0.113	0.115	95.5	-0.002	0.067	0.065	94.1	0.002	0.113	0.115	95.7	-0.000	0.070	0.067	93.4	0.2
	G-WPI	0.001	0.124	0.125	95.7	0.000	0.100	0.096	93.9	0.002	0.125	0.125	95.7	0.004	0.111	0.102	94.0	1.3
	G$^{\text{I}}$-WPI	0.001	0.113	0.115	95.7	-0.001	0.080	0.077	93.2	0.002	0.114	0.115	95.6	0.000	0.084	0.080	93.7	0.0
	G$^{\text{II}}$-WPI	0.001	0.113	0.115	95.7	-0.001	0.078	0.074	93.2	0.001	0.113	0.115	95.7	0.001	0.081	0.076	93.2	0.2
0.2	G-W	-0.006	0.109	0.109	94.5	0.002	0.106	0.105	94.6	-0.006	0.109	0.108	94.3	0.021	0.115	0.112	95.2	0.1
	G-WPI	-0.010	0.120	0.119	95.1	0.009	0.136	0.133	95.0	-0.011	0.120	0.118	95.1	0.050	0.163	0.151	93.8	2.2
	G$^{\text{I}}$-WPI	-0.006	0.109	0.109	94.2	0.005	0.118	0.117	94.8	-0.007	0.109	0.108	94.2	0.032	0.132	0.126	95.0	0.0
	G$^{\text{II}}$-WPI	-0.007	0.109	0.109	94.3	0.006	0.119	0.119	95.4	-0.007	0.108	0.108	94.1	0.031	0.133	0.128	95.8	0.1
0.4	G-W	0.001	0.092	0.094	96.0	0.008	0.152	0.154	95.6	-0.003	0.090	0.092	96.3	0.065	0.174	0.164	94.3	0.1
	G-WPI	0.001	0.103	0.104	96.4	0.015	0.173	0.175	95.7	-0.005	0.101	0.101	95.8	0.100	0.208	0.197	93.3	1.6
	G$^{\text{I}}$-WPI	0.001	0.093	0.094	96.0	0.013	0.170	0.173	96.6	-0.004	0.091	0.092	95.6	0.083	0.195	0.189	96.6	0.0
	G$^{\text{II}}$-WPI	0.000	0.093	0.094	96.1	0.016	0.183	0.186	96.0	-0.004	0.091	0.092	95.7	0.091	0.217	0.211	98.4	2.0
									$m = 1000$									
0.0	G-W	-0.001	0.051	0.052	95.6	0.001	0.030	0.029	94.7	-0.002	0.049	0.052	95.9	0.001	0.030	0.030	95.3	0.0
	G-WPI	-0.001	0.055	0.056	95.4	-0.000	0.042	0.043	95.1	-0.002	0.054	0.056	95.6	0.001	0.046	0.046	94.9	0.0
	G$^{\text{I}}$-WPI	-0.001	0.051	0.052	95.8	-0.000	0.035	0.035	94.0	-0.002	0.049	0.052	96.0	0.001	0.036	0.036	94.9	0.0
	G$^{\text{II}}$-WPI	-0.001	0.051	0.052	95.8	0.000	0.034	0.034	93.9	-0.002	0.049	0.052	96.1	0.002	0.035	0.034	95.1	0.0
0.2	G-W	-0.004	0.050	0.049	93.2	0.002	0.048	0.047	95.2	-0.005	0.050	0.048	93.4	0.021	0.051	0.050	93.5	0.0
	G-WPI	-0.004	0.054	0.054	93.6	0.004	0.059	0.059	95.6	-0.005	0.054	0.053	93.5	0.039	0.068	0.067	92.7	0.0
	G$^{\text{I}}$-WPI	-0.004	0.050	0.049	93.4	0.003	0.053	0.052	95.0	-0.005	0.050	0.048	93.4	0.031	0.057	0.056	93.2	0.0
	G$^{\text{II}}$-WPI	-0.004	0.050	0.049	93.4	0.004	0.054	0.053	94.7	-0.005	0.050	0.048	93.4	0.030	0.058	0.057	92.8	0.0
0.4	G-W	0.000	0.042	0.042	95.0	0.003	0.068	0.068	95.0	-0.004	0.041	0.041	94.7	0.058	0.076	0.073	88.9	0.0
	G-WPI	-0.000	0.045	0.047	95.5	0.006	0.078	0.077	95.1	-0.007	0.044	0.045	95.3	0.088	0.092	0.088	83.7	0.0
	G$^{\text{I}}$-WPI	0.000	0.042	0.042	95.0	0.005	0.075	0.075	95.9	-0.005	0.041	0.041	94.7	0.076	0.085	0.083	86.7	0.0
	G$^{\text{II}}$-WPI	-0.000	0.042	0.042	94.9	0.005	0.080	0.080	96.0	-0.005	0.041	0.041	94.8	0.079	0.094	0.090	88.6	0.0

† Empirical properties are summarised based on replicates leading to convergence for the piecewise constant model. The percentages of replicates failing to converge (out of $nsim = 1000$) are indicated in the last column (all replicates converged for the approach based on the correct Weibull marginal model); ‡ G$^{\text{I}}$ corresponds to G_i with $G_{i21} = \partial \eta_i / \partial \theta' = 0$; G$^{\text{II}}$ corresponds to G_i with $G_{i12} = 0$ also; WPI represents the working partial independence assumption with $W_{i22} = \text{diag}\{\eta_i(1-\eta_i)\}$, $W_{i12} = W'_{i21} = 0$.

When considering a more flexible marginal model with a piecewise constant (3 pieces) baseline hazard function, we set the cut-points at $t = 20$ and 40. For the large sample size $m = 1000$, performance was excellent for inference regarding β and very good for the dependence parameter γ_0 when the association was small; properties of the estimator of γ_0 became worse with stronger within-family dependence, possibly as a result of the crude approximation of the piecewise constant hazard. While one might expect superior performance if more pieces were accommodated, convergence problems arose even with just three pieces under the smaller sample sizes for some replicates (typically less than 2.5%); the percentages of replicates failing to converge are reported in the last column and where necessary the properties of estimators from converged replicates are given. The G^{I}-WPI estimating equation always resulted in convergence. The convergence issues likely arose due to the right-truncated nature of the proband onset time and the severe censoring from a current status observation scheme of non-probands; these combine to yield little information to estimate the hazard function in small samples.

Under more general association structure, results under G^{II}-WPI estimating equation are not summarized because of high non-convergence percentage for such more general association structure. For other three conditional estimating equations, their performance was again excellent under the correct Weibull model and again 100% of the replicates lead to convergence for $m = 200$ and $m = 1000$; see Table 2. Empirical biases were general small, there was good agreement between the empirical and average robust standard errors, and the empirical coverage probability was generally within the acceptable range. Under the piecewise constant model, convergence rate was 100% when $m = 1000$ and the empirical properties of the estimators for β and γ were good in such settings. When $m = 200$ performance remained good but with small finite sample bias and good empirical coverage probability.

4 Impact of Misspecifying the Dependence Structure

4.1 Limiting Bias under Misspecified Conditional Estimating Equations

While standard GEE1 only requires correct specification of the marginal mean for consistent estimation of the marginal parameters, the conditional estimating equations requires correct specification of the marginal distribution and the dependence structure for consistent estimation, even for the simplified conditional estimating equations G^{I}-WPI and G^{II}-WPI. As is often the case, the efficiency gains coming from the use of higher-order moments in the conditional estimating equations such as G-W, comes at the cost of poorer robustness. We explore the limiting behaviour of estimators from misspecified models here based on large sample theory [31]. Specifically if $U(\psi)$ is an estimating function for ψ based on a misspecified model then the solution $\widehat{\psi}$ for $U(\psi) = 0$ asymptotically follows,

$$\sqrt{m}(\widehat{\psi} - \psi^*) \sim N\left(0, \bar{\mathcal{A}}^{-1}(\psi^*)\bar{\mathcal{B}}(\psi^*)[\bar{\mathcal{A}}^{-1}(\psi^*)]'\right) \qquad (5)$$

Table 2 Empirical properties of estimates under conditional estimating equations for family data from response-dependent sampling, where within-family association is induced by Gaussian copula with structure, $\tau_{pp} = 0.1, \tau_{ss} = 0.4$ and $\tau_{pc} = 0.2$; parametric margins and piecewise constant baseline hazards (3 pieces) are considered[†]; $n_i = 3$, $nsim = 1000$.

		β				γ_0				γ_1				γ_2				
Margins	EE[‡]	BIAS	ESE	ASE	ECP%	BIAS	ESE	ASE	ECP%	BIAS	ESE	ASE	ECP%	BIAS	ESE	ASE	ECP%	NC %
								$m = 200$										
Weibull	G-W	0.001	0.103	0.104	96.3	-0.002	0.172	0.169	94.4	0.015	0.194	0.196	95.2	0.004	0.143	0.138	94.7	0.0
	G-WPI	0.001	0.115	0.114	95.3	0.006	0.239	0.233	94.2	0.015	0.267	0.262	94.3	0.001	0.196	0.191	94.2	0.0
	G^{I}-WPI	0.000	0.103	0.104	96.3	-0.001	0.216	0.210	93.3	0.018	0.262	0.256	94.8	0.004	0.193	0.187	94.6	0.0
PWC-3[†]	G-W	-0.000	0.102	0.103	96.2	0.011	0.179	0.175	94.4	0.031	0.207	0.208	95.0	0.011	0.147	0.142	94.3	0.2
	G-WPI	-0.001	0.114	0.112	95.1	0.032	0.255	0.246	93.7	0.058	0.317	0.303	94.0	0.016	0.211	0.205	94.3	3.9
	G^{I}-WPI	-0.001	0.102	0.103	96.3	0.013	0.227	0.219	93.3	0.051	0.298	0.286	94.8	0.016	0.205	0.197	94.7	0.4
								$m = 1000$										
Weibull	G-W	-0.002	0.045	0.047	95.9	-0.000	0.073	0.075	97.2	0.006	0.084	0.088	96.2	0.001	0.060	0.062	95.2	0.0
	G-WPI	-0.001	0.050	0.051	95.6	0.005	0.102	0.104	95.0	0.005	0.114	0.117	95.9	-0.001	0.084	0.085	95.8	0.0
	G^{I}-WPI	-0.002	0.045	0.047	96.0	0.002	0.091	0.094	95.6	0.007	0.111	0.115	96.0	-0.000	0.082	0.084	96.4	0.0
PWC-3[†]	G-W	-0.003	0.045	0.046	95.6	0.014	0.075	0.078	96.5	0.021	0.092	0.093	95.6	0.007	0.063	0.063	95.0	0.0
	G-WPI	-0.003	0.050	0.051	95.3	0.030	0.108	0.110	94.4	0.040	0.135	0.135	94.2	0.011	0.090	0.092	95.9	0.0
	G^{I}-WPI	-0.004	0.045	0.046	95.7	0.019	0.095	0.098	95.0	0.036	0.126	0.128	94.9	0.010	0.087	0.088	95.5	0.0

[†] Empirical properties are summarised based on replicates leading to convergence for the piecewise constant model. The percentages of replicates failing to converge (out of $nsim = 1000$) are indicated in the last column (all replicates converged for the approach based on the correct Weibull marginal model); [‡] G^{I} corresponds to G_i with $G_{i21} = \partial\eta_i/\partial\theta' = 0$; G^{II} corresponds to G_i with $G_{i12} = 0$ also; WPI represents the working partial independence assumption with $W_{i22} = \mathrm{diag}\{\eta_i(1-\eta_i)\}$, $W_{i12} = W_{i21}' = 0$.

as $m \to \infty$, where $\bar{\mathcal{A}}(\psi) = E[-\partial U_i(\psi)/\partial \psi' \, ; \zeta]$, $\bar{\mathcal{B}}(\psi) = E[U_i(\psi)U_i'(\psi) \, ; \zeta]$ and ψ^* is the solution to $E[U(\psi) \, ; \zeta] = 0$, where $E[\,\cdot\, ; \zeta]$ denotes an expectation taken with respect to the true distribution indexed by ζ. Note that $E[U(\psi); \zeta]$ can be written as

$$\sum_{i=1}^{m} E\{U_i(\psi) \, ; \zeta\} = \sum_{i=1}^{m} E\left\{G_i' \, W_i^{-1} \begin{pmatrix} \mu_i^*(\zeta) - \mu_i \\ \eta_i^*(\zeta) - \eta_i \end{pmatrix} + D_i' V_i^{-1}(T_{i0} - \mu_{i0})\right\} \quad (6)$$

where $\mu_i^*(\zeta) = E[\bar{Y}_i | T_{i0}, X_i, C_i]$ and $\eta_i^*(\zeta) = E[\bar{Z}_i | T_{i0}, X_i, C_i]$ are the conditional expectation of $\bar{Y}_i$ and $\bar{Z}_i$ given $\{T_{i0}, X_i, C_i\}$ under the true model. The expectation on the right hand side of (6) is taken with respect to the remaining random variables $\{T_{i0}, X_i, C_i\}$. Of course when the model is correctly specified, then $\psi^* = \zeta$ but this is not the case more generally; we investigate the limiting bias of estimators under the misspecified model by examining $\psi^* - \zeta$.

Here we consider two-generation families comprised of two parents and two children, and the proband is randomly selected from the four family members. The probands are recruited into the registry only if $T_{i0} \leq C_{i0}$. We adopt the same parameter settings as in Sect. 3.1 but assume here that the true within-family association structure is induced by the Clayton copula

$$H(k_0, k_1, \ldots, k_{n_i} \, ; \phi) = (k_0^{-\phi} + k_1^{-\phi} + \cdots + k_{n_i}^{-\phi} - n_i)^{-1/\phi} \, , \quad (7)$$

where Kendall's $\tau = \phi/(\phi + 2)$. The adopted estimating functions are misspecified in that the dependence structure is modeled based on a Gaussian copula with an exchangeable association structure. We consider values of Kendall's τ ranging from 0 to 0.5 to reflect independence to strong within-family dependence. We evaluate the limiting relative biases of estimators by using Monte Carlo methods to take the expectation in (6) and solving the resulting equation.

From Fig. 2 we see that the conditional estimating equations with the full G_i and WPI matrix W_i is the most sensitive to misspecification. Although one might anticipate that the full G_i and full covariance matrix W_i (G-W) would be less robust than G^{I}-WPI or G^{II}-WPI, the asymptotic relative biases of estimators defined through G-W are in general no larger than those under G^{I}-WPI and G^{II}-WPI when Kendall's τ is less than 0.3; the sensitivity of estimators from G-W to misspecification become more apparent, compared to those based on G^{I}-WPI and G^{II}-WPI, when Kendall's τ is larger (i.e. > 0.3); G^{I}-WPI is slightly more sensitive to this form of misspecification than G^{II}-WPI. Furthermore, the asymptotic relative biases for β under the conditional estimating equations are all relatively modest when Kendall's τ is small to modest. If one is primarily interested in estimation of β, then the proposed conditional estimating equations are reasonably robust to misspecification of the copula function for modest Kendall's τ, but the asymptotic biases of the dependence parameters are appreciable under misspecification of the dependence structure. This conclusion is analogous to those made regarding misspecification of the random effect distribution with response-dependent sampling [11,13, 23]. We also conducted supplementary simulation studies demonstrating good agreement between the finite sample and asymptotic biases in studies with 200 and 1000 families (See Supplementary Material).

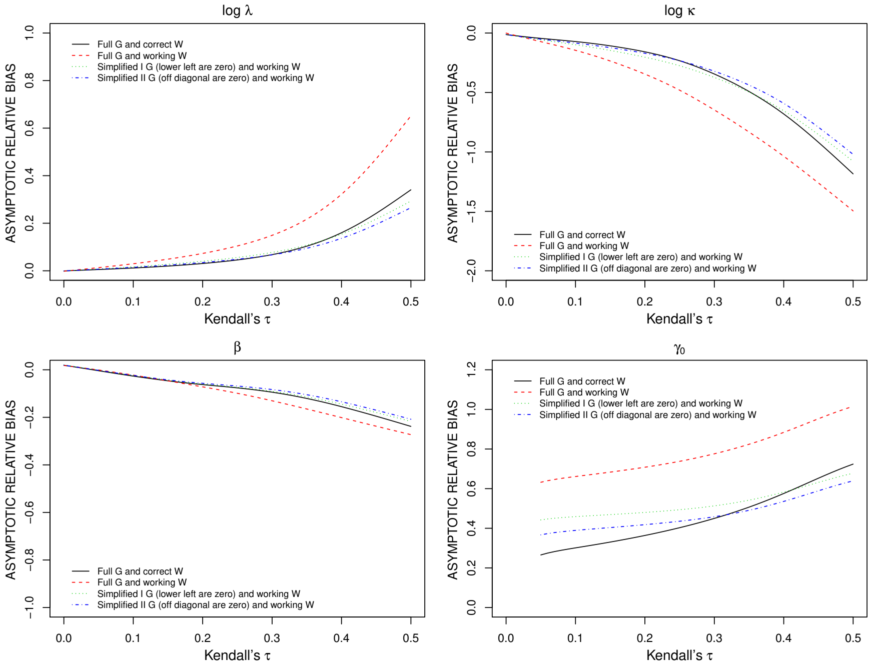

Fig. 2 Asymptotic relative biases of estimators under conditional estimating equations when a Gaussian copula with an exchangeable structure is adopted for within-family dependence modeling; the true within-family dependence structure is induced by a Clayton copula; $(\log\lambda, \log\kappa, \beta) = (-4.11, \log 1.2, \log 1.2)$.

4.2 Power Implications of Dependence Structure Misspecification

We next investigate the effect of dependence structure misspecification on the power of tests regarding covariate effects. Based on our previous findings regarding asymptotic relative efficiency and robustness, we focus attention here on the preferred estimating functions G-W and G$^{\mathrm{I}}$-WPI. We consider a test of $H_0 : \beta = \beta_0 = 0$ vs $H_A : \beta \neq 0$, and let β_A be the clinically important effect. When both the marginal and association models are correctly specified from (2) we have,

$$\sqrt{m}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2(\psi)) \ , \tag{8}$$

as $m \to \infty$, where $\sigma^2(\psi)$ is the diagonal element in the robust covariance matrix $\mathcal{A}^{-1}(\psi)\mathcal{B}(\psi)\left[\mathcal{A}^{-1}(\psi)\right]'$ corresponding to β. Under a two-sided Wald test with significance level $100\alpha_1\%$, the required number of families to ensure $100(1-\alpha_2)\%$ power to detect β_A is the smallest m satisfying

$$m \geq \left\{ \frac{z_{\alpha_1/2}\ \sigma(\psi_0) + z_{\alpha_2}\ \sigma(\psi_A)}{\beta_A} \right\}^2 , \tag{9}$$

where $\sigma(\psi_0)$ and $\sigma(\psi_A)$ are the square root of asymptotic variances of $\sqrt{m}(\widehat{\beta}-\beta)$ under the null and alternative hypotheses; $\psi_0 = (\lambda, \kappa, \beta_0, \gamma')$ and $\psi_A = (\lambda, \kappa, \beta_A, \gamma')$. z_u is the $100(1-u)\%$ percentile of standard normal distribution.

When the dependence structure is misspecified, the limiting value of estimators under the conditional estimating equations is $\psi^*(\neq \zeta)$ (Sect. 4.1). Then based on (6), we can calculate the limiting values of $\widehat{\psi}$ under the null and alternative hypotheses when the dependence structure is misspecified, and denote them as ψ_0^* and ψ_A^*, respectively. Furthermore, we can show that under the null hypothesis, the estimator based on the misspecified conditional estimating equations satisfies

$$\sqrt{m}(\widehat{\psi} - \psi_0^*) \xrightarrow{d} N(0, \mathit{\Gamma}_0^*) , \tag{10}$$

as $m \to \infty$, and under the alternative hypothesis

$$\sqrt{m}(\widehat{\psi} - \psi_A^*) \xrightarrow{d} N(0, \mathit{\Gamma}_A^*) , \tag{11}$$

where

$$\mathit{\Gamma}_0^* = \bar{\mathcal{A}}^{-1}(\psi)\bar{\mathcal{B}}(\psi)[\bar{\mathcal{A}}^{-1}(\psi)]'\Big|_{\psi=\psi_0^*} , \quad \text{and} \quad \mathit{\Gamma}_A^* = \bar{\mathcal{A}}^{-1}(\psi)\bar{\mathcal{B}}(\psi)[\bar{\mathcal{A}}^{-1}(\psi)]'\Big|_{\psi=\psi_A^*} .$$

Hence the asymptotic properties of $\widehat{\beta}$ can be determined by considering the corresponding component of $\widehat{\psi}$. When the copula model is misspecified, the actual power of such two-sided Wald test of $H_0 : \beta = \beta_0 = 0$ vs $H_A : \beta \neq 0$, at the clinically important effect β_A given sample size m and significance level α_1, is

$$\text{POWER} = \mathit{\Phi}\left(\frac{-z_{\alpha_1/2}\,\sigma_0^* - \sqrt{m}\ \beta_A^*}{\sigma_A^*} \right) + \mathit{\Phi}\left(\frac{-z_{\alpha_1/2}\,\sigma_0^* + \sqrt{m}\ \beta_A^*}{\sigma_A^*} \right) , \tag{12}$$

where σ_0^* and σ_A^* are the square roots of the diagonal elements of $\mathit{\Gamma}_0^*$ and $\mathit{\Gamma}_A^*$, respectively, corresponding to β.

Here we report on an asymptotic study to examine the effect of copula misspecification on the power. Assume each family consist of two parents and two children, and the proband is randomly selected from the four family members. As before we presume families are recruited to the study only if $T_{i0} \leq C_{i0}$. The parameter settings are the same as in Sect. 3.1 but we consider two specific scenarios: *i)* the true within-family association structure is based on a Gaussian copula with an exchangeable association structure and *ii)* the true within-family association structure is based on a Clayton copula (7); in both cases we set Kendall's $\tau = 0.4$. At the design stage we adopt a Gaussian copula with an exchangeable association structure for the within-family dependence, and let Kendall's $\tau = 0.4$. We therefore only consider the case in which the form of the dependence structure is misspecified. In this setting we calculate the required sample size to achieve 80% power to reject H_0 at $\beta_A = \log 1.2$ by (9), where $\sigma(\psi)$ is obtained from the Gaussian copula. The minimum number of families are 420 and 422 based on estimating equation G-W and G^{I}-WPI, respectively. Under these sample sizes, the actual power of such a design can be computed by (12) for values of β ranging from 0 to $\log 1.2$. The power curves are plotted in Fig. 3 from which infer that when the association model is correctly specified, tests based on the conditional estimating equations G-W and G^{I}-WPI have the desired power at the clinically important effect; as expected the power decreases when the true value of β approaches 0. When the copula is misspecified (i.e. the true dependence structure is set by a Clayton copula but a Gaussian copula is used for sample size calculation), tests based on both conditional estimating equations lead to a loss in power, with a greater loss in power under G-W compared to G^{I}-WPI. This is reasonable since the G-W estimating equations exploit information from higher-order dependencies more than G^{I}-WPI, which is less robust than the latter. In summary, based on the comprehensive investigation of these conditional estimating equations in terms of the efficiency and robustness, estimating equation G^{I}-WPI is suitable in the absence of information about the association structure, but if information is available about the structure, estimating equation G-W could be adopted to achieve higher efficiency.

5 Application to The Psoriatic Arthritis Family Study

Hereditary factors are thought to be important in psoriatic arthritis, as some studies have suggested that close blood relatives of affected individuals are at higher risk of developing the disease compared to the general population. Interest therefore lies in characterizing the effect of genetic markers on risk of disease; we consider four human leukocyte antigen (HLA) markers reported in the literature as being associated with psoriasis or psoriatic arthritis including HLA-B8, HLA-B27, HLA-C6 and HLA-C12. Characterizing the nature of the within-family association structure can also provide useful insight into the genetic basis for disease. Particular interest lies in assessing the "parent of origin" effect; preliminary evidence suggests there may be a stronger risk of paternal transmission, over maternal transmission, of risk of disease; we refer readers to Pollock et al. [24] for associated results based on binary analyses.

Here we consider an application to the motivating family study on the genetic basis for psoriatic arthritis conducted in the Centre for Prognosis Studies in the Rheumatic Diseases at the University of Toronto. One hundred and sixty-nine families comprised of 2 to 7 members, including the proband, were recruited for study. The date of disease onset is available for probands from the clinic registry but only the disease status of other individuals is available when they are examined, yielding current status data. A Weibull model is adopted

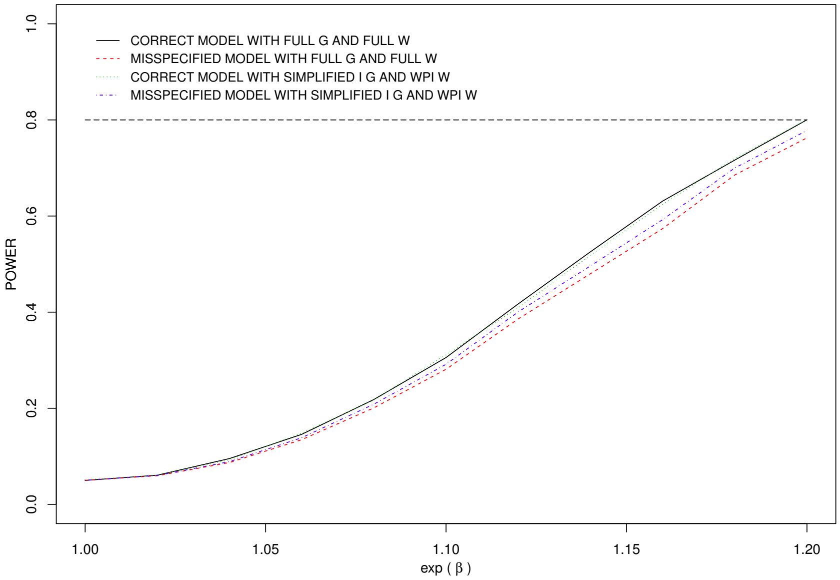

Fig. 3 Power curves of a two-sided Wald test for $H_0 : \beta = 0$ under conditional estimating equations G-W and G^{I}-WPI when the within-family dependence structure is correct specified or misspecified; true within-family dependence is induced by Gaussian copula with exchangeable structure or Clayton copula, adopted family dependence structure in the design stage is Gaussian copula with exchangeable association; Kendall's $\tau = 0.4$, $\beta_A = \log 1.2$.

for the marginal distribution of the PsA onset time with survivor function $\mathcal{F}(t|X_{ij};\theta) = \exp(-(\lambda t)^\kappa \exp(X_{ij}'\beta))$ where $\theta = (\lambda, \kappa, \beta')'$, $j = 1, \ldots, n_i$, and $i = 1, \ldots, 169$. A flexible model for the within-family dependence is formulated based on a Gaussian copula with different pairwise dependencies between parents (τ_{pp}), between siblings (τ_{ss}), between a father and his child (τ_{fc}), and between a mother and her child (τ_{mc}). This can be formulated in terms of a second-order regression model given by

$$\log\big((1+\tau_{ijk})/(1-\tau_{ijk})\big) = \gamma_0 + \gamma_1 v_{ijk1} + \gamma_2 v_{ijk2} + \gamma_3 v_{ijk3} \ , \qquad (13)$$

where $v_{ijk1} = I((j,k) \text{ pair are siblings})$, $v_{ijk2} = I((j,k) \text{ pair is father} - \text{child})$, and $v_{ijk3} = I((j,k) \text{ pair is mother} - \text{child})$. The hypotheses $H_0 : \gamma_2 - \gamma_3 = 0$ and $H_A : \gamma_2 - \gamma_3 \neq 0$ are the basis of a test regarding the parent of origin question. There are only 8 pairs of parents, which leads to insufficient data to estimate the intercept in (13) and so we constrain that parameter to be zero with the implicit assumption that there are no environmental determinants of PsA.

Table 3 summarizes the results with the top half obtained from the full derivative and covariance matrices (G-W) and the bottom half reporting the results from G^{I}-WPI. A model with no HLA covariate is given in the first column followed by four univariate models, with the last column containing results from a multivariate model including all four markers. The estimates for the association parameters are given in terms of γ and the three Kendall's τ parameters.

Based on the model with no HLA covariates, we find $\hat{\tau}_{ss} = 0.337$ (95% CI : 0.113, 0.528; p$-$value = 0.002), indicating highly significant association between siblings in the disease onset time. The father-child association is lower at $\hat{\tau}_{fc} = 0.225$ (95% CI : -0.030, 0.452) and not quite statistically significant (p$-$value = 0.072). For the mother-child association we find $\hat{\tau}_{mc} = 0.130$ (95% CI : -0.153, 0.393) which is weaker still and insignificant (p$-$value = 0.364). A test of the parent of origin hypothesis based on $H_0 : \gamma_2 - \gamma_3 = 0$ yields a Wald statistic of 1.435 (p$-$value = 0.151). As this is not statistically significant at the 5% significance level, there is insufficient evidence to claim a statistically significant "parent of origin" effect. The results are broadly comparable for the HLA regression analyses based on the other conditional estimating equation (G^{I}-WPI). For the association parameters the estimates are somewhat lower with $\hat{\tau}_{ss} = 0.220$ (95% CI : -0.003, 0.423; p$-$value = 0.046), $\hat{\tau}_{fs} = 0.104$ (95% CI : -0.128, 0.324; p$-$value = 0.378), and $\hat{\tau}_{ms} = -0.018$ (95% CI : -0.256, 0.222; p$-$value = 0.886). The Wald statistic of 1.682 (p$-$value = 0.092) does not suggest a "parent of origin" effect.

The large sample theory we develop can be used to plan a future family study and it is possible to calculate how many families would be required to ensure adequate power to test the parent of origin hypothesis in a future study. In a new study we may consider recruitment of families of members of the registry, and presume that the distribution of family members, ages at assessment and other factors are similar in the new study. We use the sample size formula similar to (9) but for $\gamma_2 - \gamma_3$ and determine that 627 families would be required to ensure 80% power to detect a significant difference between the father-child and mother-child association using estimating function G-W when the true effects correspond to those seen in the first column of Table 3. The current study therefore appears to be grossly under-powered to formally test the parent of origin hypothesis.

None of the HLA markers were shown to have a significant effect on the time to the onset of PsA. Based on the G-W estimating equations, there is a trend toward a reduction in risk with HLA-B8 and a trend toward an increased risk with the presence of each of the other HLA markers.

6 Discussion

Estimating functions have been developed to model the nature and extent of within-family dependence in disease onset times from family studies under response-dependent sampling. A novel aspect of this work is the formulation of the dependence measures on the basis of the disease onset time and the recognition that the available data on family members are handled more nat-

Table 3 Estimates of analyses of HLA markers and time to disease onset based on conditional estimating equations using response-biased psoriatic arthritis family data; associated standard errors are in the parentheses.

	Univariate models					Multivariate
	No HLA	HLA-B8	HLA-B27	HLA-C6	HLA-C12	model
Full G and Full W (G-W)						
$\log\lambda$	-5.461 (0.531)	-5.353 (0.455)	-5.573 (0.609)	-5.650 (0.590)	-5.342 (0.464)	-5.544 (0.543)
$\log\kappa$	1.347 (0.086)	1.347 (0.086)	1.348 (0.086)	1.349 (0.086)	1.346 (0.087)	1.349 (0.086)
β_{B8}	-	-0.483 (0.709)	-	-	-	-0.195 (0.751)
β_{B27}	-	-	1.161 (0.755)	-	-	1.308 (0.771)
β_{C6}	-	-	-	0.397 (0.593)	-	0.458 (0.625)
β_{C12}	-	-	-	-	0.788 (0.769)	0.982 (0.830)
γ_1	0.700 (0.242)	0.660 (0.224)	0.741 (0.261)	0.777 (0.242)	0.620 (0.234)	0.695 (0.244)
γ_2	0.458 (0.264)	0.414 (0.242)	0.470 (0.281)	0.536 (0.259)	0.378 (0.264)	0.423 (0.269)
γ_3	0.261 (0.291)	0.220 (0.268)	0.300 (0.310)	0.343 (0.285)	0.173 (0.289)	0.252 (0.287)
τ_{ss}	0.337 (0.107)	0.318 (0.101)	0.354 (0.114)	0.370 (0.104)	0.301 (0.106)	0.334 (0.108)
τ_{fc}	0.225 (0.125)	0.204 (0.116)	0.231 (0.133)	0.262 (0.121)	0.187 (0.127)	0.209 (0.129)
τ_{mc}	0.130 (0.143)	0.110 (0.132)	0.149 (0.151)	0.170 (0.138)	0.086 (0.143)	0.125 (0.141)
G^{I} and WPI W (G^{I}-WPI)						
$\log\lambda$	-5.037 (0.326)	-5.007 (0.313)	-5.105 (0.358)	-5.034 (0.325)	-5.052 (0.317)	-5.105 (0.331)
$\log\kappa$	1.337 (0.087)	1.338 (0.087)	1.339 (0.087)	1.337 (0.087)	1.338 (0.088)	1.340 (0.087)
β_{B8}	-	-0.490 (0.644)	-	-	-	-0.325 (0.669)
β_{B27}	-	-	0.849 (0.587)	-	-	0.969 (0.580)
β_{C6}	-	-	-	-0.037 (0.492)	-	0.060 (0.537)
β_{C12}	-	-	-	-	0.767 (0.644)	0.911 (0.667)
γ_1	0.448 (0.232)	0.447 (0.234)	0.481 (0.237)	0.447 (0.238)	0.437 (0.227)	0.465 (0.238)
γ_2	0.208 (0.237)	0.199 (0.234)	0.208 (0.238)	0.206 (0.241)	0.195 (0.241)	0.180 (0.244)
γ_3	-0.036 (0.249)	-0.039 (0.246)	-0.005 (0.250)	-0.037 (0.256)	-0.059 (0.251)	-0.036 (0.254)
τ_{ss}	0.220 (0.110)	0.220 (0.111)	0.236 (0.112)	0.220 (0.113)	0.215 (0.108)	0.228 (0.113)
τ_{fc}	0.104 (0.117)	0.099 (0.116)	0.104 (0.118)	0.103 (0.119)	0.097 (0.119)	0.090 (0.121)
τ_{mc}	-0.018 (0.124)	-0.019 (0.123)	-0.002 (0.125)	-0.019 (0.128)	-0.030 (0.125)	-0.018 (0.127)

urally as current status data rather than binary data. This approach utilises all available data from probands and their relatives in assessing association between age of onset and covariates, and in evaluating association structure of age of onset among family members. The biased sampling scheme typically employed in family studies is addressed by the use of conditional estimating equations where the conditioning event reflects the selection criteria. Several specific estimating functions within the class proposed are assessed in terms of efficiency and robustness; these results complement the standard results of second-order estimating functions since all moments in the proposed equations are conditional. We also outline how sample size requirements for family studies can be assessed based on this framework to ensure power objectives are met. Code for solving the conditional second-order estimating equations (1) and for obtaining the variance estimates of Section 2.2 are available at Github `https://github.com/Yujie-Zhong/CGEE2`.

We have focused on the use of estimating functions for the analysis of family data in part because the likelihood can be challenging to compute when the size of the family is large. Nevertheless, some assessment of the loss of efficiency in comparison to this optimal approach would be worthwhile. The validity of the proposed conditional second-order estimating equations hinges on correct specification of the dependence structure, a requirement that is analogous to the need for correct specification of the mixing distribution in random effects models for data obtained based on a response-dependent sampling scheme [23]. Assessing model adequacy is best done by testing for the need for model expansion; this could be carried out by testing the need for more cut-points in the baseline hazard function to accommodate a more flexible hazard function, or the need to test for a more general dependence structure. In the present setting the dependence structure is most easily formulated by selecting a working copula model for the joint distribution of the onset times in the population. If this dependence structure is misspecified, inconsistent estimates are obtained, and we examine the consequences of that in Sect. 4 to make recommendations on the use of a particular derivative and working covariance matrix. The properties of estimators under model misspecification can be explored using large sample theory [31], but these will be influenced by response-dependent sampling schemes and so more general study of the effect of misspecification in this framework represents an important area for further research.

We have restricted attention to parametric models for the onset time distribution. Natural extensions would be to introduce non-parametric or semi-parametric methods for estimating the marginal distributions. In the latter case one can look at multiplicative Cox models, accelerated failure time models, Aalen's additive model, among many other methods. Joint estimation based on the most general conditional estimating equation can be challenging in this setting, but two-stage estimation procedures may be feasible; this is an area of current research. The preliminary work based on the piecewise constant baseline hazard model, however, suggests studies may need to recruit a lot of families if the incidence rate is low to estimate the marginal onset time distribution. If the disease onset times are available for all or even some of the non-probands found to have the disease, this data could help in estimation; the estimating equations we present can be modified in this case to incorporate such data. Auxiliary samples can also be useful to enhance inferences.

While there is an increasing amount of attention given to the use of disease onset time as a basis for modeling within-family dependence, there remain challenging issues that warrant further attention. The primary challenge is in quantifying dependence in the presence of the competing risk of death [26, 12]. The classical illness-death process is a natural framework for modeling the occurrence of disease in individuals who are at risk, and generalization of this set-up to model within-family dependence is an area warranting attention. This issue is not unique to analyses based on disease onset times; when current status data are treated as binary data the requirement that individuals are alive at the time of contact is ignored.

Acknowledgements The authors thank Drs. Dafna Gladman, Vinod Chandran and Lihi Eder for stimulating collaboration and helpful discussions involving the psoriatic arthritis research program. This research was financially supported by grants from the UK Medical Research Council [Unit programme no. MC_UP_1302/3], the Natural Sciences and Engineering Research Council of Canada (RGPIN 155849) 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

1. Balemi A, Lee A (2009) Comparison of GEE1 and GEE2 estimation applied to clustered logistic regression. Journal of Statistical Computation and Simulation 79(4):361–378
2. Burden AD, Javed S, Bailey M, Hodgins M, Connor M, Tillman D (1998) Genetics of psoriasis: paternal inheritance and a locus on chromosome 6p. Journal of Investigative Dermatology 110(6):958–960
3. Burton PR, Palmer LJ, Jacobs K, Keen KJ, Olson JM, Elston RC (2000) Ascertainment adjustment: where does it take us? The American Journal of Human Genetics 67(6):1505–1514
4. Burton PR, Tiller KJ, Gurrin LC, Cookson WO, Musk AW, Palmer LJ (1999) Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. Genetic Epidemiology 17(2):118–140
5. Cannings C, Thompson EA (1977) Ascertainment in the sequential sampling of pedigrees. Clinical Genetics 12(4):208–212
6. Chatterjee N, Kalaylioglu Z, Shih JH, Gail MH (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
7. Dorman JS, Trucco M, LaPorte RE, Kuller LH (1988) Family studies: the key to understanding the genetic and environmental etiology of chronic disease? Genetic Epidemiology 5(5):305–310
8. Dunson DB, Dinse GE (2002) Bayesian models for multivariate current status data with informative censoring. Biometrics 58(1):79–88
9. Gelfand JM, Gladman DD, Mease PJ, Smith N, Margolis DJ, Nijsten T, Stern RS, Feldman SR, Rolstad T (2005) Epidemiology of psoriatic arthritis in the population of the United States. Journal of the American Academy of Dermatology 53 (4):573–586
10. Gladman DD, Schentag CT, Tom B, Chandran V, Brockbank J, Rosen C, Farewell VT (2009) Development and initial validation of a screening questionnaire for psoriatic arthritis: the Toronto Psoriatic Arthritis Screen (ToPAS). Annals of the Rheumatic Diseases 68(4):497–501
11. Glidden DV, Vittinghoff E (2004) Modelling clustered survival data from multicentre clinical trials. Statistics in Medicine 23(3):369–388
12. Gorfine M, Hsu L (2011) Frailty-based competing risks model for multivariate survival data. Biometrics 67(2):415–426
13. Gorfine M, De-Picciotto R, Hsu L (2012) Conditional and marginal estimates in case-control family data–extensions and sensitivity analyses. Journal of Statistical Computation and Simulation 82(10):1449–1470
14. Gorfine M, Hsu L, Parmigiani G (2013) Frailty Models for familial risk with application to breast cancer. Journal of the American Statistical Association 108 (504):1205–1215
15. Hsu L, Prentice RL, Zhao LP (1999) On dependence estimation using correlated failure time data from case-control family studies. Biometrika 86 (4):743–753
16. Jewell NP, Van Der Laan M, Lei X (2005) Bivariate current status data with univariate monitoring times. Biometrika 92(4):847–862
17. Joe H (1997) Multivariate Models and Multivariate Dependence Concepts. Chapman and Hall, London
18. Khoury MJ, Beaty TH, Cohen BH (1993) Fundamentals of Genetic Epidemiology. Oxford University Press, New York
19. Laird NM, Lange C (2006) Family-based designs in the age of large-scale gene-association studies. Nature Reviews Genetics 7(5): 385–394
20. Li H, Yang P, Schwartz AG (1998) Analysis of age of onset data from case-control family studies. Biometrics 54 (3):1030–1039

21. MacLean CJ, Neale MC, Meyer JM, Kendler KS, Rao DC (1990) Estimating familial effects on age at onset and liability to schizophrenia. II. Adjustment for censored data. Genetic Epidemiology 7 (6):419–426
22. Nelsen RB (2006) An Introduction to Copulas. Springer, New York
23. Neuhaus JM, McCulloch CE (2011) The effect of misspecification of random effects distributions in clustered data settings with outcome-dependent sampling. Canadian Journal of Statistics 39(3):488–497
24. Pollock RA, Thavaneswaran A, Pellett F, Chandran V, Petronis A, Rahman P, Gladman DD (2015) Further evidence supporting a parent-of-origin effect in psoriatic disease. Arthritis Care & Research 67(11):2151–4658
25. Prentice RL (1988) Correlated binary regression with covariates specific to each binary observation. Biometrics 44(4):1033–1048
26. Shih JH, Albert PS (2010) Modeling familial association of ages at onset of disease in the presence of competing risk. Biometrics 66(4):1012–1023
27. Shih JH, Chatterjee N (2002) Analysis of survival data from case-control family studies. Biometrics 58 (3):502–509
28. Sun J (2006) The Statistical Analysis of Interval-Censored Failure Time Data. Springer, New York
29. Thompson EA (1993) Sampling and ascertainment in genetic epidemiology: A tutorial review. Department of Statistics, University of Washington, Seattle, Washington
30. Weinberg CR (1999) Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. The American Journal of Human Genetics 65(1):229–235
31. White H (1982) Maximum likelihood estimation of misspecified models. Econometrica 50(1):1–26
32. Zhao LP, Hsu L, Holte S, Chen Y, Quiaoit F, Prentice RL (1998) Combined association and aggregation analysis of data from case-control family studies. Biometrika 85 (2):299–315
33. Zhao LP, Prentice RL (1990) Correlated binary regression using a quadratic exponential model. Biometrika 77(3):642–648