Published in final edited form as:
Statistics in Medicine (2021), **40** (2): 254–270
DOI: 10.1002/sim.8772

Selection models for efficient two-phase design of family studies

YUJIE ZHONG

School of Statistics and Management,

Shanghai University of Finance and Economics, Shanghai, P.R. China

E-mail: zhong.yujie@mail.shufe.edu.cn

RICHARD J. COOK

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

Summary

Family studies routinely employ biased sampling schemes in which individuals are randomly chosen from a disease registry and genetic and phenotypic data are obtained from their consenting relatives. We view this as a two-phase study and propose the use of an efficient selection model for the recruitment of families to form a phase II sample subject to budgetary constraints. Simple randomsampling, balanced sampling and use of an approximately optimal selection model are considered where the latter is chosen to minimize the variance of parameters of interest. We consider the setting where family members provide current status data with respect to the disease and use copula models to address within-family dependence. The efficiency gains fromthe use of an optimal selection model over simple random sampling and balanced sampling schemes are investigated as is the robustness of optimal sampling to model misspecification. An application to a family study on psoriatic arthritis is given for illustration.

Keywords: Age of onset, biased sampling, clustered data, copula model, efficiency, selection model

1 Introduction

Family studies offer a powerful framework for investigating the genetic basis for disease. In twin studies, it is possible to follow individuals from birth (Leslie et al., 1993), but in family studies, families are typically identified through the selection of an affected individual, called the proband, from a disease registry (Macklin, 1954). Consenting family members of the proband are then contacted for recruitment whereupon they provide detailed information on their medical history, undergo clinical or radiological examination, and provide samples for genetic analysis. The strength of the family study design lies in the efficiency gained from the high proportion of individuals with genetic risk factors due to the shared ancestry of family members. The challenge in such designs, however, is to deal with the biased sampling scheme employed. The construction of appropriate likelihoods has been considered for many years for binary data (Whittemore and Halpern, 1997; Liang and Beaty, 1991) and most recently for failure time data (Li and Thompson, 1997). Zhong and Cook (2016) consider the use of composite likelihood and Zhong and Cook (2018) develop a class of second-order estimating functions for the study of the dependence structure within families. Lakhal-Chaieb et al. (2016) uses copula functions for the development of score tests for the effects of rare variants in family studies; see also Lakhal-Chaieb et al. (2020). In these latter papers, the dependence between onset times within families was modeled using copula functions rather than the more common approach based on frailty models (Li and Thompson, 1997).

We consider the setting in which there is a large registry of affected individuals from which probands may be selected for further study. Individuals in the disease registry have provided information on demographic features, the age of disease onset, and blood samples which we assume have been assayed for genetic testing. They may also provide summary information on their family history such as the number of diseased individuals in their family. We consider the use of this information for the efficient recruitment of families for genetic testing and confirmation of disease status. We cast this problem into the framework of a two-phase design in which the individuals in the disease registry represent a phase I sample, and the recruited probands and their respective families represent the phase II sample.

Two-phase designs have been widely used to improve statistical efficiency subject to budget constraints (Reilly and Pepe, 1995; Chatterjee et al., 2003; Zhao et al., 2009) and specifically with applications to genetic epidemiology (Whittemore and Halpern, 1997; Chen et al., 2012). Lawless et al. (1999) provides a thorough review of issues involving incomplete covariate data and two-phase designs with a focus on semiparametric methods and an emphasis on cross-sectional and retrospective settings. For family data, the family-based case-control design (Shih and Chatterjee, 2002) and kin-cohort design (Wacholder et al., 1998) are commonly used approaches that employ biased sampling schemes. McIsaac and Cook (2014) proposed response-dependent two-phase design to study the effect of biomarker on a binary response and McIsaac and Cook (2013) deal with clustered data; see also Rivera-Rodriguez et al. (2019) who consider inverse weighted marginal methods for two-phase designs with clustered data where the weights are selected by calibration. When interest lies in optimizing the selection model a challenge is that key parameters of interest are unknown; McIsaac and Cook (2015) develop adaptive two-phase designs to alleviate the need to specify values for unknown parameters at the design stage.

The remainder of the article is organized as follows. In Section 2.1, we define notation and formulate the joint model for the onset times within families, and in Section 2.2, we give the details of the two-phase design. The selection model and the optimality criteria we use are introduced in Section 2.3; we aim to minimize the variance of the effect of a genetic marker on the disease onset time distribution. Identifiability and estimability issues arise in such designs

since the onset time for the proband is right-truncated and the disease incidence is often low among family members. We discuss the incorporation of auxiliary data into the likelihood used to derive the optimal selection model for the phase II sample in Section 2.4. The results of empirical studies are given in Section 3.1 for a variety of settings where the optimal design is compared to simple random sampling and a type of balanced sampling we define for family studies. Choice of the selection model depends on many assumptions so we investigate the consequences of misspecification of design parameters in Section 3.2. An illustrative application is given in Section 4 and some extensions of the methods for general dependence structure, large family size and non-response for selected family members are discussed in Section 5. Concluding remarks and topics for further research are given in Section 6.

2 Model Formulation and Efficient Design

2.1 Notation and Formulation of the Response Model

We consider the setting in which individuals are screened for disease from cross-sectional sampling of a population and those found to have the condition of interest are recruited to a registry. This leads to a disease registry comprised of N individuals. The family study is carried out by selecting members of the registry, called probands, along with their respective family members. Let C_{i0} denote the age of proband in family i at the time of sampling and screening, and T_{i0} denote their age of disease onset; the probands need to satisfy the selection condition $T_{i0} \leq C_{i0}$ and we assume T_{i0} is verifiable by a review of medical records for individuals recruited to the registry, $i = 1, \ldots, N$. We let X_{i0} and G_{i0} denote a $p \times 1$ demographic covariate vector and genotype variable, respectively, for the proband in family i; $Z_{i0} = (X_{i0}', G_{i0})'$.

We let T_{ij} and C_{ij} denote the ages at disease onset and assessment for the proband family member (*non-probands*) j in family i which is comprised of m_i individuals, $j = 1, \ldots, m_i$. We let $T_i = (T_{i1}, \ldots, T_{im_i})'$ and $C_i = (C_{i1}, \ldots, C_{im_i})'$. Then $Y_{ij} = \mathrm{I}(T_{ij} \leq C_{ij})$ is the disease status for individual j in family i; $Y_i = (Y_{i1}, \ldots, Y_{im_i})'$ and $Y_{i0} = 1$. We let X_{ij} and G_{ij} denote the covariate vector and genotype variable for individual j in family i, $X_i = (X_{i1}', \ldots, X_{im_i}')'$, $G_i = (G_{i1}, \ldots, G_{im_i})'$ and $Z_i = (X_i', G_i')'$. An overbar is used to denote data for all individuals in a family so $\bar{T}_i = (T_{i0}, T_i')'$, $\bar{C}_i = (C_{i0}, C_i')'$, $\bar{Y}_i = (Y_{i0}, Y_i')'$, $\bar{X}_i = (X_{i0}', X_i')'$, $\bar{G}_i = (G_{i0}, G_i')'$ and $\bar{Z}_i = (Z_{i0}', Z_i')'$. We assume $T_{ij} \perp (\bar{G}_i^{(-j)}, \bar{X}_i^{(-j)}) | G_{ij}, X_{ij}$, where $\bar{G}_i^{(-j)} = \{G_{ik};\, 0 \leq k \leq m_i\,, k \neq j\}$ and $\bar{X}_i^{(-j)} = \{X_{ik};\, 0 \leq k \leq m_i\,, k \neq j\}$. The marginal cumulative distribution function for the disease onset time for individual j in family i is $F(t_{ij}|Z_{ij}\,;\theta)$, where θ indexes the marginal distribution.

A joint model for the event times in family i can be constructed by specifying an $m_i + 1$ dimensional copula function (Nelsen, 2006) that is a multivariate cumulative distribution function with uniform $[0, 1]$ margins. Specifically if $U_{ij} \sim \mathrm{unif}(0, 1)$, $j = 0, 1, \ldots, m_i$, the joint cumulative distribution function $\mathcal{C}(u_{i0}, \ldots, u_{im_i}; \phi) = P(U_{i0} \leq u_{i0}, \ldots, U_{im_i} \leq u_{im_i}; \phi)$ defines a copula function that is indexed by a $q \times 1$ parameter vector ϕ which characterizes the dependence. The Archimedean family of copulas (Genest and MacKay, 1986) can be written as

$$\mathcal{C}(u_{i0}, \ldots, u_{im_i}; \phi) = \mathcal{J}^{-1}\left(\mathcal{J}(u_{i0}; \phi) + \cdots + \mathcal{J}(u_{im_i}; \phi); \phi\right)\,,$$

where $\mathcal{J}\colon\ [0, 1] \to [0, \infty)$ is a continuous, strictly decreasing and convex generator function satisfying $\mathcal{J}(1; \phi) = 0$. Kendall's τ, a widely used measure of association with event time data, can be written as

$$\tau = 1 + 4 \int_0^1 \frac{\mathcal{J}(u; \phi)}{\mathcal{J}'(u; \phi)} du$$

for Archimedean copulas (Nelsen, 2006). We construct the joint cumulative distribution function (j.c.d.f) for the disease onset times of family i by letting $U_{ij} = F(T_{ij}|Z_{ij}\,;\theta)$ and linking the marginal distribution functions through the copula function such that

$$P(T_{i0} \le t_{i0}, \ldots, T_{im_i} \le t_{im_i}|\bar{Z}_i;\zeta) = \mathcal{C}(F(t_{i0}|Z_{i0};\theta), \ldots, F(t_{im_i}|Z_{im_i};\theta);\phi)\,, \tag{1}$$

where $\zeta = (\theta', \phi')'$ is the vector of parameters indexing the joint distribution of $\bar{T}_i|\bar{Z}_i$.

The Clayton copula is widely used in survival analysis and has generator function $\mathcal{J}(u;\phi) = \phi^{-1}(u^{-\phi} - 1)$, and then gives the joint cumulative distribution function(Joe, 1997) for $\bar{T}_i|\bar{Z}_i$ as

$$P(T_{i0} \le t_{i0}, \ldots, T_{im_i} \le t_{im_i}|\bar{Z}_i;\zeta) = (F^{-\phi}(t_{i0}|Z_{i0};\theta) + \cdots + F^{-\phi}(t_{im_i}|Z_{im_i};\theta) - m_i)^{-1/\phi}\,.$$

Under the Clayton copula, Kendall's τ characterizing the association between T_{ij} and T_{ik} given (Z_{ij}, Z_{ik}), is given(Nelsen, 2006) by $\tau = \phi/(\phi+2)$ for $0 \le j < k \le m_i$, $i = 1, \ldots, N$. Other members of the Archimedean family include the Frank copula (Nelsen, 2006) with generator $\mathcal{J}(u;\phi) = -\log((\exp(-\phi u) - 1)/(\exp(-\phi) - 1))$, and the Gumbel copula (Nelsen, 2006) with generator $\mathcal{J}(u;\phi) = (-\log u)^{\phi}$. We explore these in the application and in simulation studies reported in Supplementary Material.

2.2 A Two-Phase Framework for Family Studies

We consider this problem in the framework of a two-phase study where at phase I there is detailed information available on the individuals in the disease registry (i.e., T_{i0}, C_{i0}, Z_{i0}) along with summary information on the non-probands (i.e., Y_i, C_i, X_i) obtained from the family history and interviews with members of the registry. Therefore, at phase I, the information we have for family i is $\mathcal{H}_{i1} = \{Y_{i0} = 1, T_{i0}, G_{i0}, Y_i, \bar{C}_i, \bar{X}_i\}$, $i = 1, \ldots, N$. Then based on such information, families would be selected at phase II for the family study, where detailed genotype information for their family members would be collected. Let $R_i = \mathrm{I}(\text{family } i \text{ selected for genetic testing})$ and a phase II selection model could be written as

$$\pi_i(\alpha) = P(R_i = 1|\mathcal{H}_{i1};\alpha)\,, \tag{2}$$

which means that the selection probability for each family could depend on the available information at phase I. We consider α, a vector of design parameters, that determine the selection probabilities for the phase II sampling. Let $\mathcal{S} = \{i : R_i = 1\}$, then $n = \sum_{i=1}^{N} R_i$ families provide complete data while $(N-n)$ families do not provide information on G_i. The information available at phase II ultimately consists of $\mathcal{H}_2 = \{(\bar{Y}_i, \bar{C}_i, \bar{Z}_i, T_{i0}\,; i \in \mathcal{S})\,, (\bar{Y}_i, \bar{C}_i, \bar{X}_i, G_{i0}, T_{i0}\,; i \notin \mathcal{S})\}$.

Let ν index the marginal distribution of the genetic marker (e.g. the marker frequency) and $\psi = (\zeta', \nu)'$. Under the assumption that $\bar{T}_i \perp \bar{C}_i|\bar{Z}_i$ and $\bar{C}_i$ is non-informative, the likelihood contribution from family i based on this two-phase design can be written as

$$\begin{aligned} L_i(\psi) &\propto \left[P(Y_i, G_i|\bar{C}_i, \bar{X}_i, G_{i0}, Y_{i0} = 1\,;\psi)\right]^{R_i} \left[P(Y_i|\bar{C}_i, \bar{X}_i, G_{i0}, Y_{i0} = 1\,;\psi)\right]^{1-R_i} \\ &= \left[P(Y_i|\bar{Z}_i, \bar{C}_i, Y_{i0} = 1\,;\zeta)P(G_i|G_{i0}, \bar{C}_i, \bar{X}_i, Y_{i0} = 1\,;\nu)\right]^{R_i} \\ &\quad \times \left[E_{G_i|\bar{C}_i,\bar{X}_i,G_{i0},Y_{i0}=1}(P(Y_i|\bar{C}_i, \bar{Z}_i, Y_{i0} = 1\,;\psi))\right]^{1-R_i}\,, \end{aligned} \tag{3}$$

where $P(Y_i|\bar{Z}_i, \bar{C}_i, Y_{i0} = 1\,;\psi)$ can be expressed in terms of (1). Note that $P(G_i|G_{i0}, \bar{X}_i, \bar{C}_i, Y_{i0} = 1)$ can be written as

$$\frac{P(Y_{i0} = 1|G_i, G_{i0}, \bar{X}_i, \bar{C}_i)P(G_i|G_{i0}, \bar{X}_i, \bar{C}_i)}{P(Y_{i0} = 1|G_{i0}, \bar{X}_i, \bar{C}_i)} = \frac{P(Y_{i0} = 1|G_{i0}, X_{i0}, C_{i0})P(G_i|G_{i0})}{P(Y_{i0} = 1|G_{i0}, C_{i0}, X_{i0})} = P(G_i|G_{i0})\,, \tag{4}$$

which ensures that the ascertainment condition and genetic data for the non-probands are independent conditional on the proband's genetic data. The conditional expectation in (3) can therefore be rewritten as

$$E_{G_i|\bar{C}_i,\bar{X}_i,G_{i0},Y_{i0}=1}\left[P(Y_i|\bar{C}_i,\bar{Z}_i,Y_{i0}=1\,;\psi)\right] \;=\; \sum_{g_i} P(Y_i|G_i=g_i,G_{i0},\bar{X}_i,\bar{C}_i,Y_{i0}=1)P(G_i=g_i|G_{i0})\;.$$

Thus the observed log-likelihood is

$$\begin{aligned} l(\psi) \;=\; \sum_{i=1}^{N}\Bigg\{ & R_i\left[\log P(Y_i|\bar{Z}_i,\bar{C}_i,Y_{i0}=1)+\log P(G_i|G_{i0})\right] \\ & +(1-R_i)\left[\log\sum_{g_i} P(Y_i|G_i=g_i,G_{i0},\bar{X}_i,\bar{C}_i,Y_{i0}=1)P(G_i=g_i|G_{i0})\right]\Bigg\}\;. \end{aligned} \tag{5}$$

Note that we can write the observed data score vector in terms of the complete data score vector (Louis, 1982) as

$$S(\psi)=\sum_{i=1}^{N} S_i(\psi)=\sum_{i=1}^{N}\left[R_i\left\{S_{i1}(\psi)+S_{iG}(\psi)\right\}+(1-R_i)S_{i2}(\psi)\right]\;,$$

where $S_{i1}(\psi)=\partial\log P(Y_i|\bar{Z}_i,\bar{C}_i,Y_{i0}=1)/\partial\psi$, $S_{iG}(\psi)=\partial\log P(G_i|G_{i0})/\partial\psi$ are the corresponding complete data score vectors, and

$$S_{i2}(\psi)=E_{G_i|Y_i,G_{i0},\bar{X}_i,\bar{C}_i,Y_{i0}=1}\left[S_{i1}(\psi)+S_{iG}(\psi)\right]\;.$$

The maximum likelihood estimator $\hat{\psi}$ solves $S(\psi)=0$ and $\sqrt{N}(\hat{\psi}-\psi)$ is asymptotically normally distributed with mean zero and variance $\mathcal{I}^{-1}(\eta)$ where

$$\mathcal{I}(\eta)=E[S_i(\psi)S_i'(\psi)]=E[-\partial S_i(\psi)/\partial\psi']$$

and $\eta=(\psi',\alpha')'$, where α is the vector of parameters indexing the selection model used at phase II. In Section 2.3, we consider how the selection model may be chosen to yield efficient estimation.

2.3 Selection Models for Efficient Phase II Sampling of Families

The particular phase I data used in the selection model (2) and the explicit form can be chosen based on the precise nature of the information available and the scientific context. For example, when aiming to recruit families for the effect of the genetic marker, a selection model could be specified of the form

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=\alpha_0+\alpha_1 G_{i0}+\alpha_2 A_i+\alpha_3 G_{i0}A_i\;, \tag{6}$$

where $Y_{i\cdot}=\sum_{j=1}^{m_i}Y_{ij}$ and $A_i=I(Y_{i\cdot}\geq 1)$ indicates that there is at least one other affected family member. With a selection model of this form, families having affected members and families with members having a higher chance of the genetic marker may be selected with a higher probability for the phase II sample if $\alpha_k>0$, $k=1,2,3$. This kind of enrichment sampling is particularly appealing when the disease or the genetic marker is rare.

By specification of the form of the selection model (2), we know that the missing data (i.e. the genetic information for the non-probands of unselected families) is missing at random

(Little and Rubin, 2002) because $P(R_i = 1|\mathcal{H}_{i1}, G_i) = P(R_i = 1|\mathcal{H}_{i1})$. However the precise selection probabilities are as yet undetermined as (6) simply specifies the functional form of the selection model. The optimal two-phase design involves the phase II selection of families in order to minimize the asymptotic variance of the estimator of interest. When calculating the Fisher information matrix the expectation is taken with respect to $(R_i, Y_i, \bar{X}_i, \bar{G}_i, \bar{C}_i)$ given $Y_{i0} = 1$ so suppressing the argument ψ on the right-hand side we have

$$\begin{aligned}\mathcal{I}(\eta) &= E[R_i S_{i1} S_{i1}' + R_i S_{iG} S_{iG}' + R_i S_{i1} S_{iG}' + R_i S_{iG} S_{i1}' + (1 - R_i) S_{i2} S_{i2}'] \\ &= E[\pi_i S_{i1} S_{i1}' + \pi_i S_{iG} S_{iG}' + \pi_i S_{i1} S_{iG}' + \pi_i S_{iG} S_{i1}' + (1 - \pi_i) S_{i2} S_{i2}'] \; . \end{aligned} \tag{7}$$

Evaluation of the Fisher information matrix, and hence computation of the asymptotic variance, is challenging since it involves several multi-dimensional integrals; the computational burden is greater when the family size is large. In a trio family study, for example, where the proband is the child and recruited family members are their parents, $m_i = 2$ for all $i = 1, \ldots, N$; we let $j = 1$ index father and $j = 2$ index mother. Under the Clayton copula and selection model (6) we obtain

$$\begin{aligned}\log P(Y_i|\bar{C}_i, \bar{Z}_i, Y_{i0} = 1) &= Y_{i1} Y_{i2} \log P(T_{i1} \le C_{i1}, T_{i2} \le C_{i2}|T_{i0} \le C_{i0}, \bar{C}_i, \bar{Z}_i) \\ &\quad + Y_{i1}(1 - Y_{i2}) \log P(T_{i1} \le C_{i1}, T_{i2} > C_{i2}|T_{i0} \le C_{i0}, \bar{C}_i, \bar{Z}_i) \\ &\quad + (1 - Y_{i1}) Y_{i2} \log P(T_{i1} > C_{i1}, T_{i2} \le C_{i2}|T_{i0} \le C_{i0}, \bar{C}_i, \bar{Z}_i) \\ &\quad + (1 - Y_{i1})(1 - Y_{i2}) \log P(T_{i1} > C_{i1}, T_{i2} > C_{i2}|T_{i0} \le C_{i0}, \bar{C}_i, \bar{Z}_i) \; . \end{aligned}$$

If we define $F_{12}(C_i|T_{i0} \le C_{i0}, \bar{C}_i, \bar{Z}_i) = P(T_{i1} \le C_{i1}, T_{i2} \le C_{i2}|T_{i0} \le C_{i0}, \bar{C}_i, \bar{Z}_i)$, then the first term in $E[\pi_i S_{i1} S_{i1}']$ is

$$\begin{aligned}& E\left[\pi_i Y_{i1} Y_{i2} \cdot \frac{\partial \log F_{12}(C_i|T_{i0} \le C_{i0}, \bar{C}_i, \bar{Z}_i)}{\partial \psi} \cdot \frac{\partial \log F_{12}(C_i|T_{i0} \le C_{i0}, \bar{C}_i, \bar{Z}_i)}{\partial \psi'}\right] \\ = \; & E\Big[F_{12}(C_i|T_{i0} \le C_{i0}, \bar{C}_i, \bar{Z}_i) \cdot \frac{\exp(\alpha_0 + \alpha_1 G_{i0} + \alpha_2 + \alpha_3 G_{i0})}{1 + \exp(\alpha_0 + \alpha_1 G_{i0} + \alpha_2 + \alpha_3 G_{i0})} \\ & \times \frac{\partial \log F_{12}(C_i|T_{i0} \le C_{i0}, \bar{C}_i, \bar{Z}_i)}{\partial \psi} \cdot \frac{\partial \log F_{12}(C_i|T_{i0} \le C_{i0}, \bar{C}_i, \bar{Z}_i)}{\partial \psi'}\Big] \; , \end{aligned}$$

where the expectation is taken with respect to $(\bar{C}_i, \bar{G}_i, \bar{X}_i)$ given $Y_{i0} = 1$. This expectation depends on the specification of the distribution for $\bar{C}_i$, $\bar{X}_i$ and $\bar{G}_i$, and is computationally demanding. The calculation of the term $E[(1 - \pi_i) S_{i2} S_{i2}']$ in (7) poses greater computational challenges involving high dimensional integration. We therefore propose to approximate the required expectations based on the empirical distributions estimated from the available phase I data; see Appendix for details.

For illustration, we consider the selection model (6) in the context of a family study of trios at phase II. In this case, the phase I data can be partitioned into four strata defined by the values of (G_{i0}, A_i), and the stratum-specific selection probability is π_{jk} for $G_{i0} = j$ and $A_i = k$, where $j, k = 0, 1$. Then

$$\pi_i = P(R_i = 1|\mathcal{H}_{i1}, Y_{i0} = 1) = \sum_{j,k=0}^{1} \pi_{jk} \mathrm{I}(G_{i0} = j) \mathrm{I}(A_i = k) \; .$$

Budgetary constraints which limit the number of families that can be sampled at phase II are reflected by specifying some $0 < P_R \le 1$ so that

$$P(R_i = 1|Y_{i0} = 1) = \sum_{j,k=0}^{1} \pi_{jk} \frac{N_{jk}}{N} = P_R \; , \tag{8}$$

where N_{jk}, the number of probands in stratum $(G_0 = j, A = k)$, is known at the design stage. Here we adopt the similar budgetary constraint as in McIsaac and Cook (2014), which is based on the observed phase II stratum sizes rather than on the expected stratum size (Reilly and Pepe, 1995; Whittemore and Halpern, 1997).

Optimal designs under likelihood based inference can be obtained by choosing α in (6) to minimize the asymptotic variance of the parameter of interest subject to the budgetary constraints in (8). We assume that the marginal distribution of T_{ij} satisfies the proportional hazard assumption with $h(t_{ij}\,;\theta) = h_0(t_{ij}\,;\omega)\exp(X'_{ij}\delta + \beta G_{ij})$, where $h_0(t_{ij}\,;\omega)$ is the baseline hazard function indexing by a vector of parameters ω, δ is a vector of parameters associated with the demographic covariates X_{ij}, and β is the parameter of interest associated with the genetic marker; we let $\theta = (\omega', \delta', \beta)'$. We focus on finding the optimal sampling probabilities (e.g. the optimal value of α) for phase II that minimizes $\text{asvar}(\sqrt{N}(\hat{\beta} - \beta))$ subject to $P(R_i = 1|Y_{i0} = 1) = P_R$. This is done by finding the stationary point of

$$\text{asvar}(\sqrt{N}(\hat{\beta} - \beta)) - \gamma \left(\sum_{j,k=0}^{1} \pi_{jk} N_{jk}/N - P_R \right) ,$$

where γ is the Lagrange multiplier and $\text{asvar}(\sqrt{N}(\hat{\beta} - \beta))$ is the entry of $\mathcal{I}^{-1}(\eta)$ of primary interest. We denote the optimal stratum-specific sampling probabilities as π_{jk}^{opt}. To avoid under-desirable degenerate designs with near-zero selection probabilities in some strata, we constrain the stratum-specific selection probabilities to be $0.05 \leq \pi_{jk} \leq 1$ for $j, k = 0, 1$ (Reilly and Pepe, 1995; Breslow and Cain, 1988). This sampling design can be found using numerical minimization procedures and will be optimally efficient for maximum likelihood estimation whenever the models and parameter values are correctly specified at the design stage.

2.4 Efficient Sampling Incorporating Augmentation with Auxiliary Samples

Zhong and Cook (2016) argue that it was most appealing to formulate models for genetic effects in terms of the marginal onset time distribution. However, in general, the family data we obtained at phase I provide limited information about the marginal onset time distribution since the onset times of probands are right-truncated and the prevalence of disease among nonprobands is typically low. If auxiliary data are available, it could be exploited to reduce the bias and/or improve efficiency (Pitkäniemi et al., 2009). Auxiliary data in the present setting involves current status data on the presence of disease from a national cross-sectional survey (Gelfand et al., 2005). We explore the use of this data for efficient sampling under the assumption that the auxiliary processes share parameters with the processes governing the family data in this section.

Let $\mathcal{A}$ denote the set of indexes for individuals in an auxiliary sample of size M, and suppose information available from auxiliary data consists of $\{Y_j, C_j, X_j; j \in \mathcal{A}\}$, where C_j and Y_j are current age and disease status, respectively. Demographic data are available, but data on the genetic marker are missing for the individuals in the survey. We augment the log-likelihood (5) to obtain

$$\log L^{\text{aug}}(\psi) \;=\; \sum_{i=1}^{N} \log L_i(\psi) + \sum_{j \in \mathcal{A}} \log P(Y_j|C_j, X_j\,;\psi) \;, \tag{9}$$

and the augmented score function is

$$S^{\text{aug}}(\psi) \;=\; \sum_{i=1}^{N} [R_i S_{i1} + R_i S_{iG} + (1 - R_i) S_{i2}] + \sum_{j \in \mathcal{A}} S_{jA} \;,$$

where $S_{jA} = \partial \log P(Y_j|C_j, X_j\,;\psi)/\partial\psi$ and are independent with S_{i1}, S_{iG} and S_{i2}. It is easy to show that asymptotically

$$\frac{1}{M}\sum_{j\in\mathcal{A}} S_{jA} \sim N(0, \mathcal{I}_A(\psi))$$

as $M \to \infty$, where $\mathcal{I}_A(\psi) = E[S_{jA}S'_{jA}]$. Therefore, as both N and M go to infinity, if $N/(N+M) \to \rho \neq 0$, the maximum likelihood estimator $\hat{\psi}^{\text{aug}}$ based on the augmented likelihood has the following asymptotic distribution:

$$\sqrt{N+M}\,(\hat{\psi}^{\text{aug}} - \psi) \sim N(0, (\rho\,\mathcal{I}(\eta) + (1-\rho)\,\mathcal{I}_A(\psi))^{-1})\,.$$

When there is no auxiliary data $\rho = 1$ and $\hat{\psi}^{\text{aug}} = \hat{\psi}$, but when there is auxiliary data $0 < \rho < 1$ and it is self-evident that the maximum augmented likelihood is more efficient. The optimal sampling probabilities for the two-phase family study on genetic association when utilizing the auxiliary data are denoted by π_{jk}^{aug} and can be obtained by finding the stationary point of

$$\text{asvar}(\sqrt{N+M}\,(\hat{\beta}^{\text{aug}} - \beta)) - \gamma\left(\sum_{j,k=0}^{1}\pi_{jk}N_{jk}/N - P_R\right)\,, \tag{10}$$

where $\text{asvar}(\sqrt{N+M}\,(\hat{\beta}^{\text{aug}} - \beta))$ is the corresponding entry of $(\rho\,\mathcal{I}(\eta) + (1-\rho)\,\mathcal{I}_A(\psi))^{-1}$.

In the next section, we investigate the finite sample performance of the estimators from the two-stage design with a focus on the efficiency gains over simple random sampling of probands from the registry and a balanced sampling scheme (Breslow and Cain, 1988), in which the phase II sampling probabilities are inversely proportional to the size of the strata, that is, $\pi_{jk} = (NP_R/4)\cdot N_{jk}^{-1}$.

3 Empirical Studies

3.1 Efficient Phase II Selection Models for Assessing Genetic Effects

Here we report on simulation studies designed to assess the efficiency of the optimal two-phase trio family study on the genetic association; then $m_i = 2$. In this context, we consider a phase I sample of $N = 2000$ probands is recruited in a disease registry whose onset time are right-truncated by their clinic entry time. Detailed information on the proband's age at onset, age at screening, demographic variables and genetic markers are available. For their parents, the disease status, age at contact and some demographic information are also recorded through interviewing the probands. We assign proband, proband's father and proband's mother label 0, 1, and 2, respectively, and assume that all family members have a common marginal onset time distribution with $F(t_{ij}|G_{ij};\theta) = 1 - \exp(-(\lambda t_{ij})^{\kappa}e^{\beta G_{ij}})$, $j = 0, 1, 2$, $i = 1, \ldots, N$, where G_{ij} is the genetic marker of interest with marker frequency ν and consider $\nu = 0.25$; $\theta = (\log\lambda, \log\kappa, \beta)'$. The parameters λ and κ are chosen so that the non-carrier penetrance at age 45 and 70 are $F(45|G=0) = p_1$ and $F(70|G=0) = p_2$, respectively; let $p_1 = 0.15$ and $p_2 = 0.30$. Assume $\beta = \log 2$ to represent the scenario that having the mutation increases the risk of developing the disease. The proband's clinic entry time C_{i0} is normally distributed with mean $\mu = 45$ and variance $\sigma^2 = 20$, and conditional on this right-truncation time we generate $T_{i0}|T_{i0} < C_{i0}$. The latent onset times for the non-probands are then generated as $T_{i1}, T_{i2}|T_{i0}$ using a copula function, here we consider an exchangeable association structure based on the Clayton copula with Kendall's $\tau = 0.25$. Then the observed family data are created following

the generation of the assessment times. Specifically, for the non-probands (i.e., parents in the trio family study), the age at contact follows $N(\mu = 70, \sigma^2 = 20)$ and they are truncated at 90 years for all individuals. Therefore the phase I data consist of $\{T_{i0}, G_{i0}, \bar{Y}_i, \bar{C}_i\,;\, i = 1, \ldots, N\}$. We consider the selection model (6) with budgetary constraint (8) at phase II, then the phase I data could be stratified into 4 strata based on G_{i0} and $A_i = I(Y_{i\cdot} \geq 1)$; each stratum has its own sampling probability π_{jk}, $j, k = 0, 1$ and let $P_R = 0.4$. Simple random sampling, balanced sampling and optimal sampling based on the likelihood method are applied to sample families at phase II and the genetic information will be collected for the non-probands in the selected families. Each of the 1000 simulated incomplete datasets were analysed using maximum likelihood method. We know that the balanced sampling and optimal sampling are based on the phase I data, so these designs depend on each simulated dataset. Therefore the stratum-specific selection probabilities $\pi = [\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}]$ employed by simple random sampling is $\pi^{\text{srs}} = [0.40, 0.40, 0.40, 0.40]$, but the average selection probabilities for the balanced sampling and optimal design based on maximum likelihood are quite different at $\pi^{\text{bal}} = [1.00, 1.00, 0.36, 0.25]$ and $\pi^{\text{opt}} = [0.88, 0.05, 0.05, 0.65]$, respectively. Table 1 summarizes the empirical bias, empirical standard error (ESE), average robust standard error (ASE), ASE evaluated based on phase I data only, and the empirical coverage probability of nominal 95% confidence intervals for the different designs.

When there is auxiliary current status data (absent data on the genetic marker), the augmented likelihood (9) can be used and we obtain different optimal sampling probabilities via (10). Under the same parameter setting, we generate such current status data with the underlying onset time distribution for individuals in the auxiliary sample the same as in the family study; we set $M = 2000$. The assessment times of the current status auxiliary sample are generated from the same distribution as that for the clinic entry time for the probands. The auxiliary data then consist of $\{C_j, Y_j; j = 1, \ldots, M\}$. One thousand replicates were generated with simple random sampling, balanced sampling, and optimal sampling based on the augmented likelihood. Simple random sampling and balanced sampling designs do not depend on the availability of auxiliary data so the average selection probabilities are the same as before. For the optimal design the average selection probabilities become $\pi^{\text{aug}} = [0.76, 0.05, 0.05, 0.67]$ which are slightly different from those without the augmentation data. The empirical properties of estimates based on the augmented likelihood while employing simple random sampling, balanced sampling and optimal sampling are also summarised in Table 1. We find that all biases are negligible and that the ESEs are close to the ASEs. The ASEs evaluated based on the phase I data only ($\text{ASE}^{\dagger}$) also agree well with the ASE, which supports the use of our proposed approach to approximate the Fisher information based on the phase I data (see Appendix). The empirical coverage probability of nominal 95% confidence intervals are all within an acceptable range. We also note that the ASEs are bigger for all parameters under the likelihood analysis compared to those based on the augmented likelihood, in alignment with expectations based on Section 2.4. Furthermore we find that the ASEs of the estimate of genetic effects are smallest under the optimal sampling design illustrating the gain from attempts to select the most informative phase II sample. When there is no auxiliary data, the asymptotic relative efficiency of the estimates of genetic effect under the simple random sampling is 0.80 compared to under optimal sampling, and this relative efficiency is 0.78 when there is auxiliary current status data. These findings illustrate the improved efficiency in estimation of genetic effect under optimal design, particularly when the auxiliary data are available. Note that there was not much evidence of improved efficiency under the balanced sampling scheme compared to simple random sampling.

We consider another scenario where penetrance and marker frequency are much smaller; so we let $p_1 = 0.05$, $p_2 = 0.10$ and $\nu = 0.10$. Other parameters are set to the same values as before.

Table 1: Empirical properties of estimates from analysing 1000 simulated datasets consisting of $N = 2000$ individuals without and with auxiliary current status data of size $M = 2000$ at phase I while employing simple random sampling (srs), balanced sampling (bal) or optimal sampling (opt) for phase II design with an expected phase II sample size of $P(R_i = 1|Y_{i0} = 1) = 0.4$; Parameter Setting (i): $p_1 = 0.15\,, p_2 = 0.30\,, \nu = 0.25$.

		NO AUGMENTATION					AUGMENTATION				
		BIAS	ESE	ASE	ASE†	ECP	BIAS	ESE	ASE	ASE†	ECP
β	srs	0.007	0.098	0.098	0.097	0.950	0.000	0.088	0.092	0.092	0.964
	bal	0.007	0.103	0.103	0.102	0.939	-0.003	0.097	0.098	0.098	0.947
	opt	0.004	0.088	0.087	0.087	0.955	0.004	0.082	0.081	0.081	0.952
$\log\lambda$	srs	-0.021	0.125	0.120	0.120	0.958	-0.010	0.111	0.106	0.105	0.949
	bal	-0.021	0.125	0.121	0.121	0.954	-0.009	0.114	0.107	0.107	0.944
	opt	-0.020	0.122	0.118	0.117	0.950	0.011	0.109	0.104	0.104	0.947
$\log\kappa$	srs	-0.017	0.232	0.225	0.224	0.962	-0.007	0.104	0.102	0.102	0.950
	bal	-0.017	0.231	0.225	0.225	0.964	-0.007	0.105	0.102	0.102	0.942
	opt	-0.016	0.229	0.225	0.224	0.964	-0.007	0.104	0.102	0.102	0.946
$\log\frac{1+\tau}{1-\tau}$	srs	-0.002	0.046	0.044	0.044	0.938	-0.000	0.043	0.042	0.042	0.940
	bal	-0.002	0.046	0.044	0.044	0.937	-0.001	0.044	0.042	0.042	0.937
	opt	-0.002	0.045	0.044	0.044	0.946	0.000	0.043	0.042	0.042	0.937
$\log\frac{\nu}{1-\nu}$	srs	-0.002	0.062	0.062	0.062	0.954	-0.002	0.061	0.062	0.062	0.954
	bal	-0.003	0.064	0.065	0.065	0.943	-0.007	0.062	0.065	0.065	0.972
	opt	-0.005	0.064	0.066	0.066	0.947	-0.005	0.066	0.065	0.065	0.949

† The average of the square root of the asymptotic variance evaluated based on phase I data only

We still consider the simple random sampling, balanced sampling and optimal sampling to sample families at phase II and likelihood or augmented likelihood methods are used for estimation. Among the 1000 generated samples, analysis of 21, 21 and 20 samples featured convergence issues for the respective sampling methods with the likelihood method, but no convergence issues arose when auxiliary current status data ($M = 2000$) were incorporated. The summaries for this parameter setting are based on the converged replicates only. The average selection probabilities for the balanced sampling and optimal design are $\pi^{\text{bal}} = [0.61, 1.00, 0.20, 0.46]$ and $\pi^{\text{opt}} = [0.77, 0.05, 0.05, 1.00]$ without augmentation, and the probabilities for optimal design using the auxiliary data become $\pi^{\text{aug}} = [0.74, 0.05, 0.09, 1.00]$. The empirical properties of estimates based on the likelihood and augmented likelihood under this parameter setting are reported in Table 2. Here we obtain similar results in this scenario, with the relative efficiency of the estimates of the genetic effect under the simple random sampling of 0.65 compared to under the optimal sampling design when there is no augmentation; this becomes 0.62 when auxiliary current status data are available. Therefore, when the disease is rare or the genetic mutation rate is low, employing optimal sampling can be more beneficial, particularly in the presence of auxiliary data.

Table 2: Empirical properties of estimates from analysing 1000 simulated datasets consisting of $N = 2000$ individuals without and with auxiliary current status data of size $M = 2000$ at phase I while employing simple random sampling (srs), balanced sampling (bal) or optimal sampling (opt) for phase II design with an expected phase II sample size of $P(R_i = 1|Y_{i0} = 1) = 0.4$; Parameter Setting (ii): $p_1 = 0.05\,, p_2 = 0.10\,, \nu = 0.10$.

		NO AUGMENTATION					AUGMENTATION				
		BIAS	ESE	ASE	ASE†	ECP	BIAS	ESE	ASE	ASE†	ECP
β	srs	0.002	0.122	0.125	0.124	0.963	-0.001	0.119	0.123	0.122	0.959
	bal	-0.001	0.122	0.124	0.123	0.953	-0.001	0.113	0.121	0.121	0.964
	opt	0.001	0.099	0.101	0.101	0.952	-0.002	0.094	0.096	0.096	0.956
$\log\lambda$	srs	-0.137	0.454	0.493	0.492	0.939	-0.070	0.358	0.335	0.335	0.937
	bal	-0.137	0.451	0.504	0.502	0.941	-0.071	0.358	0.335	0.335	0.941
	opt	-0.133	0.450	0.497	0.496	0.934	-0.070	0.362	0.334	0.334	0.940
$\log\kappa$	srs	-0.032	0.263	0.256	0.255	0.964	-0.021	0.184	0.175	0.175	0.953
	bal	-0.031	0.264	0.256	0.255	0.965	-0.022	0.183	0.175	0.175	0.957
	opt	-0.032	0.263	0.255	0.254	0.964	-0.021	0.184	0.175	0.175	0.949
$\log\frac{1+\tau}{1-\tau}$	srs	-0.000	0.038	0.037	0.037	0.947	0.002	0.036	0.035	0.035	0.948
	bal	-0.000	0.038	0.037	0.037	0.951	0.002	0.036	0.035	0.035	0.945
	opt	-0.001	0.037	0.037	0.037	0.949	0.002	0.036	0.035	0.035	0.946
$\log\frac{\nu}{1-\nu}$	srs	-0.007	0.084	0.087	0.087	0.962	-0.008	0.089	0.087	0.087	0.942
	bal	-0.007	0.094	0.094	0.093	0.949	-0.005	0.098	0.093	0.093	0.946
	opt	-0.006	0.092	0.093	0.093	0.949	-0.006	0.093	0.093	0.093	0.957

† The average of the square root of the asymptotic variance evaluated based on phase I data only

3.2 Empirical Studies on the Effect of Parameter Misspecification

The optimal selection model was derived in Section 3.1 based on the true parameter values, which are unknown in practice. Here we explore the sensitivity of optimal design to misspecification of the parameter values at the design stage. We consider the following eight scenarios: (a) we overestimate the non-carrier penetrance, (b) we underestimate the non-carrier penetrance, (c) we overestimate the genetic effect, (d) we underestimate the genetic effect, (e) we overestimate the within-family association, (f) we underestimate the within-family association, (g) we overestimate the marker frequency, and (h) we underestimate the marker frequency. The parameter values for each misspecified scenario under the two parameter settings are summarized in Table 3. The other parameter settings are the same as in Section 3.1.

A total of 2000 potential families are generated based on the true parameter values, simple random sampling, balanced sampling and optimal sampling where the sampling probabilities are based on the misspecified parameter values are applied. Then each of the 1000 simulated incomplete datasets are analysed using maximum likelihood method. Since simple random sampling and balanced sampling do not depend on the specified parameter values at design stage, the sampling probabilities are the same in both scenarios. However, since the optimal sampling probabilities depend on the specified parameter values, the average selection probabilities for these eight misspecified scenarios are summarized in Table 3 where there is apparent variation in the approximate optimal selection models. There are estimability problems in the second

parameter setting when auxiliary data are not available and so we do not report on results for the two-phase design for this setting.

Table 3: Parameter values and the corresponding average stratum-specific optimal selection probabilities under maximum likelihood method for the eight misspecification scenarios under two parameter settings without and with auxiliary current status data of size $M = 2000$.

	Parameter value†					No Augmentation				Augmentation			
Scenarios	p_1	p_2	e^β	τ	ν	π_{00}	π_{10}	π_{01}	π_{11}	π_{00}	π_{10}	π_{01}	π_{11}
Parameter Setting (i): $p_1 = 0.15$, $p_2 = 0.30$ and $\nu = 0.25$													
True	0.15	0.30	2.0	0.25	0.25	0.88	0.05	0.05	0.65	0.76	0.05	0.05	0.67
(a)	0.30	0.50	×	×	×	0.45	0.06	0.05	0.70	0.26	0.07	0.05	0.72
(b)	0.05	0.10	×	×	×	0.95	0.05	0.05	0.64	0.91	0.05	0.05	0.65
(c)	×	×	3.0	×	×	0.79	0.05	0.05	0.66	0.37	0.06	0.05	0.71
(d)	×	×	1.2	×	×	0.88	0.05	0.05	0.65	0.89	0.05	0.05	0.65
(e)	×	×	×	0.40	×	0.95	0.05	0.05	0.64	0.91	0.05	0.05	0.65
(f)	×	×	×	0.10	×	0.67	0.05	0.05	0.68	0.63	0.05	0.05	0.68
(g)	×	×	×	×	0.40	1.00	0.05	0.07	0.62	1.00	0.05	0.07	0.63
(h)	×	×	×	×	0.10	0.05	0.07	0.05	0.74	0.05	0.06	0.05	0.75
Parameter Setting (ii): $p_1 = 0.05$, $p_2 = 0.10$ and $\nu = 0.10$													
True	0.05	0.10	2.0	0.25	0.10	0.77	0.05	0.05	1.00	0.74	0.05	0.09	1.00
(a)	0.02	0.08	×	×	×	-	-	-	-	0.69	0.05	0.11	1.00
(b)	0.02	0.08	×	×	×	-	-	-	-	0.76	0.05	0.08	1.00
(c)	×	×	3.0	×	×	-	-	-	-	0.84	0.05	0.06	1.00
(d)	×	×	1.2	×	×	-	-	-	-	0.50	0.05	0.17	1.00
(e)	×	×	×	0.4	×	-	-	-	-	0.83	0.05	0.06	1.00
(f)	×	×	×	0.1	×	-	-	-	-	0.60	0.05	0.13	1.00
(g)	×	×	×	×	0.25	-	-	-	-	0.68	0.05	0.11	1.00
(h)	×	×	×	×	0.05	-	-	-	-	0.75	0.05	0.09	1.00

† × means the value of this parameter is correctly specified

Table 4 displays the finite sample properties of estimators from likelihood and augmented likelihood analyses under quasi-optimal selection at phase II based on misspecified parameter values; here we report only the results for estimation of β, the parameter of primary interest. As expected we find negligible empirical bias since estimates should remain consistent in this setting. Moreover, there appears to be a good degree of robustness to the types of misspecification we considered since the ESEs of the regression coefficient appears remarkably stable across all scenarios.

Table 4: The empirical properties of estimates of β from analysing 1000 simulated datasets consisting of $N = 2000$ individuals without and with auxiliary current status data of size $M = 2000$ at phase I while employing optimal sampling using misspecified parameter values for phase II design with an expected phase II sample size of $P(R_i = 1|Y_{i0} = 1) = 0.4$.

	NO AUGMENTATION					AUGMENTATION				
Scenarios	BIAS	ESE	ASE	ASE†	ECP	BIAS	ESE	ASE	ASE†	ECP
	Parameter Setting (i): $p_1 = 0.15$, $p_2 = 0.30$ and $\nu = 0.25$									
True	0.004	0.088	0.087	0.087	0.955	0.004	0.082	0.081	0.081	0.952
(a)	0.005	0.088	0.088	0.087	0.947	0.004	0.082	0.082	0.082	0.945
(b)	0.004	0.088	0.087	0.087	0.953	0.004	0.082	0.081	0.081	0.950
(c)	0.004	0.088	0.087	0.087	0.953	0.003	0.081	0.082	0.082	0.945
(d)	0.004	0.088	0.087	0.087	0.953	0.004	0.082	0.081	0.081	0.946
(e)	0.005	0.088	0.087	0.087	0.951	0.003	0.082	0.081	0.081	0.948
(f)	0.004	0.087	0.087	0.087	0.954	0.003	0.081	0.081	0.081	0.948
(g)	0.005	0.088	0.088	0.088	0.948	0.003	0.082	0.082	0.082	0.949
(h)	0.004	0.089	0.088	0.088	0.946	0.004	0.082	0.082	0.082	0.945
	Parameter Setting (ii): $p_1 = 0.05$, $p_2 = 0.10$ and $\nu = 0.10$									
True	-	-	-	-	-	-0.002	0.094	0.096	0.096	0.956
(a)	-	-	-	-	-	-0.002	0.094	0.096	0.096	0.958
(b)	-	-	-	-	-	-0.002	0.094	0.096	0.096	0.955
(c)	-	-	-	-	-	-0.003	0.095	0.096	0.096	0.955
(d)	-	-	-	-	-	-0.003	0.093	0.097	0.097	0.953
(e)	-	-	-	-	-	-0.003	0.095	0.096	0.096	0.957
(f)	-	-	-	-	-	-0.002	0.094	0.096	0.096	0.956
(g)	-	-	-	-	-	-0.003	0.094	0.096	0.096	0.953
(h)	-	-	-	-	-	-0.002	0.094	0.096	0.096	0.957

† The average of the square root of the asymptotic variance evaluated based on phase I data only

4 An Illustrative Family Study Involving Psoriatic Arthritis

There is known to be a genetic basis for the development of psoriatic arthritis and interest lies in characterizing the effect of a human leukocyte antigen marker HLA-B27 on the risk of developing psoriatic arthritis while addressing the within-family association in disease process. Here we consider the implications of different selection models applied to a phase I sample comprising members of the University of Toronto Psoriatic Arthritis Registry (UTPSA) (Gladman and Chandran, 2011). Specifically we consider the use of an optimal sampling approach with a focus on testing the effect of HLA-B27 on the risk of developing psoriatic arthritis. To illustrate the design however, we need to create complete data for all families with a member in the UTPSA. We do this using data from the family study of Pollock et al. (2015) as described in Section 4.1.

4.1 Exploiting Available Family Data for Creation of a Complete Data Set

To complete the family data for all members of the UTPSA, we use data from the selected families of Pollock et al. (2015). The size and composition of families for individuals not selected in Pollock et al. (2015) is determined by resampling the family composition of those in Pollock et al. (2015). To do this we stratify the families from Pollock et al. (2015) into five groups based on the age of onset of the proband: $[0, 20)$, $[20, 30)$, $[30, 40)$, $[40, 50)$, and $[50, \infty)$ years of age. We then match an unselected potential proband from the registry with a selected proband from the same stratum and assign the unselected potential probands family members to be the same age at examination as those of their selected matched pair. The remaining data are generated by simulation using models and estimates from earlier analyses as follows.

We set the marginal marker frequency to $\nu = 0.06$ and use the kinship of family members to complete the genetic data based on (4). Given the underlying HLA-B27 indicator, we generate the age at onset for all non-probands based on the conditional j.c.d.f $P(T_1 \leq t_1, T_2 \leq t_2|T_0)$ where we assume that the marginal distribution is Weibull with one covariate (HLA-B27); the parameter values are set to be the same as those obtained in Zhong and Cook (2018). We assume the within-family structure is characterized by a Clayton copula with Kendall's $\tau = 0.2$; this value of τ is obtained by pooling the estimates of Kendall's τ for a father-child pair and a mother-child pair reported in Zhong and Cook (2018). We next consider the available phase I data and the implementation of the three two-phase designs.

4.2 Implementation of the Phase II Sampling Schemes

For illustration, we focus on families with at least two non-probands and target the two oldest family members for recruitment so that they will have been at risk the longest time for the development of psoriatic arthritis. Based on these phase I data created in Section 4.1 we apply the simple random sampling, balanced sampling and the proposed optimal sampling with selection model (6) under the constraint $P_R = 0.15$ to reflect the fact that about 15% families could be selected. The total number of probands in each stratum are [341, 73, 348, 81] and the corresponding sampling probabilities are [0.09, 0.43, 0.09, 0.39] under balanced sampling, and [0.06, 0.05, 0.06, 1.00] under the optimal sampling scheme. When we incorporate the auxiliary survey data providing current status data with sample size of 15,307 (Gelfand et al., 2005), the optimal selection probabilities become [0.07, 0.05, 0.05, 1.00]. The numbers of selected probands in each stratum using these different sampling approaches are summarized in Table 5.

Table 5: The total number of individuals and the selected number of probands in each stratum $(G_0 = j, A_i = k)$, $j, k = 0, 1$ under simple random sampling, balanced sampling and optimal sampling, for the psoriatic arthritis study.

	$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$	Total
N	341	73	348	81	843
n^{srs}	49	11	56	12	128
n^{bal}	29	27	34	27	117
n^{opt}	22	3	22	81	128
n^{aug}	22	5	11	81	119

We next compute the likelihood (5) and the augmented likelihood (9) for the resulting data. The resulting estimates of the effect of HLA-B27 on the age of onset for PsA are given in Table 6.

Note that the point estimates vary somewhat across the different selection models but these are within a reasonable range given the standard errors. We find that without auxiliary data, the standard error is the largest under simple random sampling, followed by that from balanced sampling. The optimal sampling method provides the estimate of the genetic effect with the smallest standard error. When incorporating the auxiliary data, the efficiency improves for all analyses but the optimal sampling scheme still leads to the estimate with the smallest standard error.

Table 6: Log hazard ratio estimates, standard errors and its 95% CI of the effect of HLA-B27 on the onset time distribution for PsA based on the selected families using different sampling schemes without (no augmentation) and with (augmentation) auxiliary current status survey data.

	No Augmentation			Augmentation		
	srs	bal	opt	srs	bal	opt
EST	0.859	0.815	1.054	0.824	0.744	0.945
SE	0.531	0.419	0.306	0.499	0.396	0.287
95% CI	(-0.182, 1.900)	(-0.006, 1.636)	(0.454, 1.654)	(-0.154, 1.802)	(-0.032, 1.520)	(0.382, 1.508)

5 Some Extensions

5.1 Dependence Misspecification and More General Dependence Modeling

In the current formulation, the dependence structure is modeled via a copula function; see equation (1). With response-dependent sampling schemes of the sort we discuss, consistent estimation requires correct specification of the joint model and hence the copula function must be correct at the analysis stage. To explore the sensitivity of the parameter estimates to misspecification of the copula function, we conducted additional simulation studies. We first consider the same family study setting as described in Section 3.1 where we use the Clayton copula to model the within-family association at the analysis stage, but consider the setting where the correct joint model involves the Frank copula; see Section 1.1 of the Supplementary Material.

The simulation results show that when the copula function is misspecified, all estimators have non-negligible empirical bias; Table S.2 of the Supplementary Material reveals that the empirical bias is larger with stronger within-family association. If auxiliary data are available for synthesis with the family study data the biases can be larger under misspecification of the copula as shown in Table S.3 of the Supplementary Material. When the penetrance among individuals free of the genetic marker is lower and the frequency of the marker is smaller the biases can be appreciable as shown in Table S.4 of the Supplementary Material. We note, however, that the estimator of the effect of the genetic marker tends to have a smaller empirical bias; this bias is particularly small when the strength of the dependence is mild, but it can be appreciable when the dependence is strong.

When the within-family association is driven primarily by genetic factors, a more general dependence structure may be appealing in which separate dependence parameters accommodate an association that depends on the kinship of pairs of family members. Zhong and Cook (2016)

used a three-parameter Gaussian copula to accommodates this kind of association, while Lakhal-Chaieb et al. (2020) used a Gaussian copula with a single dependence parameter which was scaled according to the kinship of different pairs of family members. The Gaussian copula with a general correlation matrix can be written as

$$\mathcal{C}(u_{i0}, u_{i1}, \ldots, u_{im_i}, \phi) \;=\; \Phi_{m_i+1}\left(\Phi^{-1}(u_{i0}), \Phi^{-1}(u_{i1}), \ldots, \Phi^{-1}(u_{im_i}) ; \phi\right) ,$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of a standard normal random variable (r.v.), and $\Phi_{m_i+1}(\cdot\,;\phi)$ is the cumulative distribution function of an $(m_i+1)\times 1$ multivariate normal r.v. with mean zero and $(m_i+1)\times(m_i+1)$ covariance matrix $\Sigma_i(\phi)=\Sigma_i$ with off-diagonal entries σ_{ijk}. From (1), this gives

$$\begin{aligned} & P(T_{i0}\le t_{i0}, T_{i1}\le t_{i1}, \ldots, T_{im_i}\le t_{im_i}|\bar{Z}_i, \zeta) \\ = & \int_{-\infty}^{r_{i0}}\cdots\int_{-\infty}^{r_{im_i}} \frac{\exp\left(-s_i'\Sigma_i^{-1}s_i/2\right)}{\sqrt{(2\pi)^{m_i+1}|\Sigma_i|}} ds_{i0}\ldots ds_{im_i} , \end{aligned} \tag{11}$$

where if $S_i\sim \mathrm{MVN}_{m_i+1}(0,\Sigma_i)$, s_i is a realization, and $r_{ij}=\Phi^{-1}(F(t_{ij}|Z_{ij}\,;\theta))$, $j=0,1,\ldots,m_i$. Results from simulation studies examining the finite sample bias of estimators obtained when the true copula is Gaussian but the Clayton copula is used at the design and analysis stages are reported in Section 1.2 of the Supplementary Material. Again we see that when the strength of the within-family association is mild the biases for the regression coefficient can be mild but it can become appreciable when there is stronger within-family dependence; see Tables S.6 and S.7 of the Supplementary Material.

5.2 Pairwise Composite Likelihood with Large Varying Family Size

The likelihood and efficient sampling approach are both developed allowing different family size m_i in Section 2, while the empirical studies and the illustrative family study we considered are for trio-family study, i.e. $m_i=2$. When family size m_i varies and especially when it is large it can be challenging to compute and maximize the full likelihood. However, composite likelihood (Lindsay, 1988; Cox and Reid, 2004) comprised of contributions based on lower dimensional subsets of individuals in each family can be considered. Zhong and Cook (2016) proposed two conditional composite likelihoods based on all pairs or all triplets of family members including the proband for the family study under biased sampling. They showed that composite likelihood can simplify the analytical expression and computation when the family size is large and varying, and the efficiency loss incurred can be modest when either family sizes are small or the within-family associations are modest and the loss can be offset by exploitation of auxiliary data when it is available. Our approach can be extended to deal with this case, following which one would derive the optimal sampling scheme based on composite likelihood for two-phase family studies. The contribution from family i of the phase I sample following phase II selection can be written as

$$\begin{aligned} CL_i(\psi) \;\propto\; & \left[\prod_{1\le j<k\le m_i} P(Y_i^{(j,k)}, G_i^{(j,k)}|C_i^{(j,k)}, X_i^{(j,k)}, G_{i0}, Y_{i0}=1\,;\psi)\right]^{R_i} \\ & \times\left[\prod_{1\le j<k\le m_i} P(Y_i^{(j,k)}|C_i^{(j,k)}, X_i^{(j,k)}, G_{i0}, Y_{i0}=1\,;\psi)\right]^{1-R_i} ; \end{aligned}$$

which is the conditional pairwise composite likelihood contribution where $Y_i^{(j,k)}=(Y_{ij},Y_{ik})$, $G_i^{(j,k)}=(G_{ij},G_{ik})$, $X_i^{(j,k)}=(X_{ij}',X_{ik}')$ and $C_i^{(j,k)}=(C_{ij},C_{ik})$ are the disease status, genetic marker, demographic covariates, and ages at assessment for pair (j,k) in family i, $i=$

$1, 2, \ldots, N$, respectively. The estimates $\tilde{\psi}$ can be obtained by maximizing the composite likelihood for all families $CL(\psi) = \prod_{i=1}^{N} CL_i(\psi)$. Zhong and Cook (2016) give the limiting distribution of estimators based on composite likelihood, which are consistent and asymptotically normally distributed with a covariance matrix based on the robust sandwich variance formula. The optimal phase II sampling probabilities can be found in a similar way to the approach used for (2) based on the full likelihood; that is the selection probabilities can be obtained by minimizing the asymptotic robust variance of the parameter of interest based on the composite likelihood subject to the budgetary constraints.

5.3 Non-response for Some Members of Families Selected in Phase II

In some settings it may arise that not all members of selected families provide samples for genetic testing. To accommodate this we let $\Delta_{ij} = I(G_{ij} \text{ is observed})$, $j = 1, \ldots, m_i$, $\Delta_i = (\Delta_{i1}, \ldots, \Delta_{im_i})'$, and we informally let $G_i = (G_i^{\circ}, G_i^{m})$ denote G_i partitioned according to individuals with observed and missing data, respectively, $i = 1, \ldots, N$. Since Δ_i is random the likelihood analogous to (3), but accommodating incomplete observation of G_i, can be written as

$$\begin{aligned} L_i &\propto \left[P(Y_i, G_i^{\circ}, \Delta_i | \bar{C}_i, \bar{X}_i, G_{i0}, Y_{i0} = 1)\right]^{R_i} \left[P(Y_i | \bar{C}_i, \bar{X}_i, G_{i0}, Y_{i0} = 1)\right]^{1-R_i} \qquad (12) \\ &= \left[\textstyle\sum_{G_i^m} P(Y_i, G_i | \Delta_i, \bar{C}_i, \bar{X}_i, G_{i0}, Y_{i0} = 1) P(\Delta_i | \bar{C}_i, \bar{X}_i, G_{i0}, Y_{i0} = 1)\right]^{R_i} \\ &\quad \times \left[E_{G_i | \bar{C}_i, \bar{X}_i, G_{i0}, Y_{i0}=1}(P(Y_i | \bar{C}_i, \bar{Z}_i, Y_{i0} = 1\,; \psi))\right]^{1-R_i} \end{aligned}$$

If, missingness is clustered within-families and joint modeling of Δ_i is required in (12), then models for multivariate binary data (Prentice, 1988) may be useful for the evaluation of (12). If however, $\Delta_i \perp G_i | Y_i, \bar{C}_i, \bar{X}_i, G_{i0}, Y_{i0} = 1$ and incomplete participation of family members is non-informative for families with $R_i = 1$, then we may ignore the term $P(\Delta_i | \bar{C}_i, \bar{X}_i, G_{i0}, Y_{i0} = 1)$ and focus on the partial likelihood contribution for family i that can be written as

$$\begin{aligned} &\textstyle\prod_{\delta_i} \left[E_{G_i^m | \Delta_i = \delta_i, G_i^{\circ}, \bar{C}_i, \bar{X}_i, G_{i0}, Y_{i0}=1}\{P(Y_i | \bar{C}_i, \bar{Z}_i, Y_{i0} = 1\,; \psi)\}\right]^{R_i I(\Delta_i = \delta_i)} \\ &\quad \times \left[E_{G_i | \bar{C}_i, \bar{X}_i, G_{i0}, Y_{i0}=1}\{P(Y_i | \bar{C}_i, \bar{Z}_i, Y_{i0} = 1\,; \psi)\}\right]^{1-R_i} \qquad (13) \end{aligned}$$

where $\prod_{\delta_i}$ represents a product taken over all 2^{m_i} possible realizations of Δ_i. The independence assumption $\Delta_i \perp G_i | Y_i, \bar{C}_i, \bar{X}_i, G_{i0}, Y_{i0} = 1$ will often be reasonable since individuals will not typically be aware of their status with respect to the marker of interest. Of course while this approach can lead to consistent estimation, some loss of efficiency will result from incomplete collection of specimens for genetic testing. If refusal rates of families as a whole, or individuals within selected families, are high, then adaptive two-phase design may be needed to mitigate loss of power from incomplete data.

6 Discussion

We consider the setting where the goal is to estimate the effect of genetic markers on disease onset time. Casting this problem into a failure time model is necessary since family members can vary a great deal in their age and therefore will have been at risk for varying amounts of time. In many settings, such as in metabolic diseases or other complex disorders, the disease onset time may be difficult to ascertain precisely and it may be preferable simply to use current status data on the onset time for family members; the onset time is typically available for the

proband as they often have been referred to the tertiary care clinic maintaining the registry by a family physician. Use of retrospectively reported onset times is possible but given these are right-truncated by the age at examination in the family study, they will typically convey relatively little additional information. It would be interesting, however, to investigate how use of such data might alter the optimal selection model for the phase II sample.

The empirical studies conducted here have shown that the properties of estimators are sensitive to misspecification of the copula function. The use of more highly parameterized copula functions enlarges the range of within-family dependence structures but these are still typically going to be specified within a parametric family. In some settings, inverse probability of selection weights can be used to deal with truncated samples but in this clustered data setting such weights may still need to be based on a model for the within-family dependence and so it is unclear whether this approach could lead to greater robustness. Fortunately within-family dependencies are often smaller than the larger values we explored in the empirical studies, so if interest lies in the effects of the genetic markers then biases are typically modest. Nevertheless, the development of more robust approaches for dealing with response-biased samples in family studies is a challenging area worthy of research.

Patient records will provide some information on family history but we have assumed here that it is detailed enough that the disease status is reported on each member. Often it will only be reported in aggregated form in that it will only be known how many family members are affected, or even if at least one family member is affected. The fact that the family members' disease status will change over time is what motivates the use of a failure time model as the basis for the study of genetic effects. In the family studies conducted at the Centre for Prognosis Studies in Rheumatic Disease, the report on the family members disease status by the members of the registries can be incorrect, so an interesting extension would be to model the misclassification probability of the proband's report on the disease status of their family members. Given auxiliary data on this misclassification process alternative optimal selection probabilities may be obtained.

Lee and Cook (2019) recently considered the use of illness-death models (Fix and Neyman, 1951) for the analysis of family studies. The rationale is that the onset time distribution is in fact improper since not all individuals, even with a high risk configuration of genetic markers, will develop the disease in their lifetime. Moreover, when diseases alter risk of death and attention is restricted to non-probands who are alive to attend a clinic for examination, there is a more subtle aspect to the biased sampling scheme based on right-truncation of the survival distribution which characterizes the absorption time to the death state in the illness-death process. Generalization of the proposed design framework for these more complex processes is of interest and the topic of ongoing work. An alternative would be to adopt a mixture model in which individuals have a latent susceptibility indicator and these, along with the relevant onset times, are correlated within families in the spirit of Chatterjee and Shih (2001).

The general framework we have outlined here has been used primarily for the estimation of the effect of particular genetic markers of interest, but interest could lie in recruiting families with a view to better understanding parent of origin effects (Burden et al., 1998; Pollock et al., 2015). To do this, the design would be best recast using a Gaussian copula that admits a more general dependence structure within families (Zhong and Cook, 2016; Lakhal-Chaieb et al., 2016). In this case, interest may lie in estimating a function of the father-child and mother-child association parameters; see Zhong and Cook (2016).

Acknowledgements

This research was supported by grants from National Natural Science Foundation of China (NSFC-11901376), Shanghai Pujiang Program (2019PJC051) and SUFE Innovation Funding (2019110051) to Y Zhong, a Discovery Grant and Supplement Award from the Natural Science and Engineering Research Council of Canada to RJ Cook (RGPIN 155849 and RGPIN 04207), and a grant from the Canadian Institutes for Health Research (FRN 13887). RJ Cook is a Faculty of Mathematics Research Chair, University of Waterloo. The authors thank Drs. Dafna Gladman, Vinod Chandran and Remy Pollock for collaboration on family studies and permission to use the data, and Dr. Michael McIsaac for helpful discussion.

Conflict of Interest

The authors declare no potential conflict of interests.

Data Availability Statement

The data for the family study are confidential and held by the University of Toronto Psoriatic Arthritis Clinic.

References

Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75(1):11–20.

Burden, A. D., Javed, S., Bailey, M., Hodgins, M., Connor, M., and Tillman, D. (1998). Genetics of psoriasis: paternal inheritance and a locus on chromosome 6p. *Journal of Investigative Dermatology*, 110:958–960.

Chatterjee, N., Chen, Y. H., and Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98(461):158–168.

Chatterjee, N. and Shih, J. H. (2001). A bivariate cure-mixture approach for modeling familial association in diseases. *Biometrics*, 57:779–786.

Chen, Z., Craiu, R. V., and Bull, S. B. (2012). Two-phase stratified sampling designs for regional sequencing. *Genetic Epidemiology*, 36(4):320–332.

Cox, D. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91:729–737.

Fix, E. and Neyman, J. (1951). A simple stochastic model of recovery, relapse, death and loss of patients. *Human Biology*, 23(3):205–241.

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*, 53:573–586.

Genest, C. and MacKay, J. (1986). The joy of copulas: bivariate distributions with uniform marginals. *The American Statistician*, 40(4):280–283.

Gladman, D. D. and Chandran, V. (2011). Observational cohort studies: lessons learnt from the University of Toronto Psoriatic Arthritis Program. *Rheumatology*, 50(1):25–31.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall, London.

Lakhal-Chaieb, L., Cook, R. J., and Zhong, Y. (2020). Testing the heritability and parent-of-origin hypothesis for ages at onset of psoriatic arthritis under biased sampling. *Biometrics*, 76(1):293–303.

Lakhal-Chaieb, L., Oualkacha, K., Richards, B. J., and Greenwood, C. M. T. (2016). A rare variant association test in family-based designs and non-normal quantitative traits. *Statistics in Medicine*, 35(6):905–921.

Lawless, J. F., Kalbfleisch, J. D., and Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):413–438.

Lee, J. Y. and Cook, R. J. (2019). The illness-death model for family studies (submitted manuscript). *Biostatistics*.

Leslie, R. D. G., Pyke, D. A., Braun, M. M., Caporaso, N., Duffy, D. L., and Macdonald, A. M. (1993). Twin studies in medical research. *The Lancet*, 341(8857):1418–1419.

Li, H. and Thompson, E. A. (1997). Semiparametric estimation of major gene and family-specific random effects for age of onset. *Biometrics*, 53:282–293.

Liang, K. Y. and Beaty, T. H. (1991). Measuring familial aggregation by using odds-ratio regression models. *Genetic Epidemiology*, 8:361–370.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80:221–239.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, 2nd Edition.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):226–233.

Macklin, M. T. (1954). Methods of selection of probands and controls. *American Journal of Human Genetics*, 6(1):86.

McIsaac, M. A. and Cook, R. J. (2013). Response-dependent sampling with clustered and longitudinal data. In *ISS-2012 proceedings volume on longitudinal data analysis subject to measurement errors, missing values, and/or outliers*, pages 157–181. Springer.

McIsaac, M. A. and Cook, R. J. (2014). Response-dependent two-phase sampling designs for biomarker studies. *Canadian Journal of Statistics*, 42(2):268–284.

McIsaac, M. A. and Cook, R. J. (2015). Adaptive sampling in two-phase designs: a biomarker study for progression in arthritis. *Statistics in Medicine*, 34(21):2899–2912.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer, New York.

Pitkäniemi, J., Varvio, S. L., and Corander, J. (2009). Full likelihood analysis of genetic risk with variable age at onset disease - combining population-based registry data and demographic information. *PLoS ONE*, 4:e6836.

Pollock, R. A., Thavaneswaran, A., Pellett, F., Chandran, V., Petronis, A., Rahman, P., and Gladman, D. D. (2015). Further evidence supporting a parent-of-origin effect in psoriatic disease. *Arthritis Care and Research*, page doi: 10.1002/acr.22625.

Prentice, R. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, pages 1033–1048.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82(2):299–314.

Rivera-Rodriguez, C., Spiegelman, D., and Haneuse, S. (2019). On the analysis of two-phase designs in cluster-correlated data settings. *Statistics in Medicine*, 38(23):4611–4624.

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

Wacholder, S., Hartge, P., Struewing, J. P., Pee, D., McAdams, M., Brody, L., and Tucker, M. (1998). The kin-cohort study for estimating penetrance. *American Journal of Epidemiology*, 148(7):623–630.

Whittemore, A. S. and Halpern, J. (1997). Multi-stage sampling in genetic epidemiology. *Statistics in Medicine*, 16(2):153–167.

Zhao, Y., Lawless, J. F., and McLeish, D. L. (2009). Likelihood methods for regression models with expensive variables missing by design. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(1):123–136.

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. (2018). Second-order estimating equations for clustered current status data from family studies using response-dependent sampling. *Statistics in Biosciences*, 10(1):160–183.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of this article.

APPENDIX

A Approximation of Asymptotic Variance Using Phase I Data

As discussed in Section 2.3, evaluating the Fisher information matrix requires to take expectation of $S_i(\psi)S_i'(\psi)$ with respect to $(R_i, Y_i, \bar{X}_i, \bar{G}_i, \bar{C}_i)$ given the ascertainment condition $Y_{i0} = 1$, where $S_i(\psi) = R_i\{S_{i1}(\psi) + S_{iG}(\psi)\} + (1 - R_i)S_{i2}(\psi)$, $S_{i1}(\psi) = \partial \log P(Y_i|\bar{Z}_i, \bar{C}_i, Y_{i0} = 1)/\partial\psi$, $S_{iG}(\psi) = \partial \log P(G_i|G_{i0})/\partial\psi$, and $S_{i2}(\psi) = E_{G_i|Y_i,G_{i0},\bar{X}_i,\bar{C}_i,Y_{i0}=1}[S_{i1}(\psi) + S_{iG}(\psi)]$. This is challenging since it involves several multi-dimensional integrals, particularly when the family size is large. Therefore we propose to approximate the required expectations based on the empirical distributions estimated from the available phase I data. Note that the information available at phase I consists of $\mathcal{H}_1 = \{(\bar{C}_i, \bar{X}_i, G_{i0}, Y_i, Y_{i0} = 1)\,; i = 1, 2\ldots, N\}$. G_i is the possible missing information at phase II if ith family is not selected for family study, $i = 1, \ldots, N$. Note that $S_{i1}(\psi)$ is a function of $(Y_i, \bar{G}_i, \bar{X}_i, \bar{C}_i)$ so we let $Q(Y_i, \bar{G}_i, \bar{X}_i, \bar{C}_i\,; \psi) = S_{i1}(Y_i, \bar{G}_i, \bar{X}_i, \bar{C}_i\,; \psi)S_{i1}'(Y_i, \bar{G}_i, \bar{X}_i, \bar{C}_i\,; \psi)$. Therefore under the selection model (6), the first term of $\mathcal{I}(\eta)$ in (7) can be rewritten as

$$\begin{aligned} & E[R_iS_{i1}(Y_i, \bar{G}_i, \bar{X}_i, \bar{C}_i\,; \psi)S_{i1}'(Y_i, \bar{G}_i, \bar{X}_i, \bar{C}_i\,; \psi)] \qquad \text{(A1)} \\ = \;& E\{\pi_i(Y_i, G_{i0}\,; \alpha)Q(Y_i, \bar{G}_i, \bar{X}_i, \bar{C}_i\,; \psi)\} \\ = \;& E_{\bar{C}_i,\bar{X}_i,Y_i,G_{i0}|Y_{i0}=1}\Big\{\pi_i(Y_i, G_{i0}\,; \alpha)E_{G_i|G_{i0},\bar{C}_i,\bar{X}_i,Y_i,Y_{i0}=1}\big[Q(Y_i, \bar{G}_i, \bar{X}_i, \bar{C}_i\,; \psi)\big]\Big\}\,. \end{aligned}$$

Since the ascertainment condition and genetic mutation for the non-probands are independent conditional on the proband's genetic mutation (see equation (3) of the main body of the manuscript) the inner conditional expectation in (A1) can be calculated as

$$\begin{aligned} & E_{G_i|G_{i0},\bar{C}_i,\bar{X}_i,Y_i,Y_{i0}=1}\big[Q(Y_i, \bar{G}_i, \bar{X}_i, \bar{C}_i\,; \psi)\big] \\ = \;& \sum_{g_i} P(G_i = g_i|G_{i0}, \bar{C}_i, \bar{X}_i, Y_i, Y_{i0} = 1)Q_i(Y_i, g_i, G_{i0}, \bar{X}_i, \bar{C}_i\,; \psi) \\ = \;& \sum_{g_i}\left[\frac{P(Y_i|Y_{i0} = 1, G_i = g_i, G_{i0}, \bar{X}_i, \bar{C}_i)P(G_i = g_i|G_{i0})}{\sum_{l_i} P(Y_i|Y_{i0} = 1, G_i = l_i, G_{i0}, \bar{X}_i, \bar{C}_i)P(G_i = l_i|G_{i0})}Q_i(Y_i, g_i, G_{i0}, \bar{X}_i, \bar{C}_i\,; \psi)\right] \\ \stackrel{def}{=} \;& M(Y_i, G_{i0}, \bar{C}_i, \bar{X}_i\,; \psi)\,, \qquad \text{(A2)} \end{aligned}$$

where g_i is one of the possible realizations of G_i. Phase I data is collected through the ascertained probands (i.e. $Y_{i0} = 1$), therefore we could use such data to approximate the outer conditional expectation in (A1), that is,

$$\begin{aligned} & E[R_iS_{i1}(Y_i, \bar{G}_i, \bar{X}_i, \bar{C}_i\,; \psi)S_{i1}'(Y_i, \bar{G}_i, \bar{X}_i, \bar{C}_i\,; \psi)] \\ = \;& E_{\bar{C}_i,\bar{X}_i,Y_i,G_{i0}|Y_{i0}=1}\Big\{\pi_i(Y_i, G_{i0}\,; \alpha)M(Y_i, G_{i0}, \bar{C}_i, \bar{X}_i\,; \psi)\Big\}\,. \end{aligned}$$

We approximate this by

$$\frac{1}{N}\sum_{i=1}^{N}\pi_i(y_i, g_{i0}\,; \alpha)M_i(y_i, g_{i0}, \bar{c}_i, \bar{x}_i\,; \psi)\,, \qquad \text{(A3)}$$

which is equivalent to

$$
\begin{aligned}
& E[R_i S_{i1}(Y_i, \bar{G}_i, \bar{X}_i, \bar{C}_i\,; \psi) S'_{i1}(Y_i, \bar{G}_i, \bar{X}_i, \bar{C}_i\,; \psi)] \qquad \text{(A4)} \\
\approx\ & \frac{1}{N}\sum_{i=1}^{N} \pi_i(y_i, g_{i0}) \frac{\sum_{g_i} P(Y_i = y_i | Y_{i0} = 1, g_i, g_{i0}, \bar{c}_i, \bar{x}_i) P(G_i = g_i | G_{i0} = g_{i0}) Q_i(y_i, g_i, g_{i0}, \bar{c}_i, \bar{x}_i)}{\sum_{l_i} P(Y_i = y_i | Y_{i0} = 1, l_i, g_{i0}, \bar{c}_i, \bar{x}_i) P(G_i = l_i | G_{i0} = g_{i0})} \\
=\ & \frac{1}{N}\sum_{i=1}^{N} \left[\sum_{l_i} P(Y_i = y_i | Y_{i0} = 1, l_i, g_{i0}, \bar{c}_i, \bar{x}_i) P(G_i = l_i | G_{i0} = g_{i0}) \right]^{-1} \times \left[\sum_{g_i} \pi_i(y_i, g_{i0}) \right. \\
& \left. \times P(Y_i = y_i | Y_{i0} = 1, g_i, g_{i0}, \bar{c}_i, \bar{x}_i) P(G_i = g_i | G_{i0} = g_{i0}) S_{i1}(y_i, g_i, g_{i0}, \bar{c}_i, \bar{x}_i) S'_{i1}(y_i, g_i, g_{i0}, \bar{c}_i, \bar{x}_i) \right] .
\end{aligned}
$$

Using (A4) to approximate $E[R_i S_{i1}(Y_i, \bar{G}_i, \bar{X}_i, \bar{C}_i\,; \psi) S'_{i1}(Y_i, \bar{G}_i, \bar{X}_i, \bar{C}_i\,; \psi)]$ can save the effort to calculate several integrals, also eliminate the effects of misspecification of the distribution for age at assessment ($\bar{C}_i$) and the distribution for other covariates ($\bar{X}_i$). Similarly, we could use the same strategy to approximate other expectation terms in $\mathcal{I}(\eta)$ based on phase I data.

Published in final edited form as:
Statistics in Medicine (2021), **40** (2): 254–270
DOI: 10.1002/sim.8772

Supplementary Material for *Selection models for efficient two-phase design of family studies*

YUJIE ZHONG

School of Statistics and Management,
Shanghai University of Finance and Economics, Shanghai, P.R. China
E-mail: zhong.yujie@mail.shufe.edu.cn

RICHARD J. COOK

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

Summary

In this supplementary material, we report additional simulation results for the effect of misspecification of copula function discussed in the paper "Selection models for efficient two-phase design of family studies".

1 Simulation to investigate the effect of copula misspecification

Consider a family study of trios, and assume that all family members have a common marginal onset time distribution with survivor function $\mathcal{F}(t_{ij}|G_{ij}) = \exp(-(\lambda t_{ij})^{\kappa} e^{\beta G_{ij}})$, where G_{ij} is the genetic marker with frequency ν, and $j = 0, 1, 2$ index proband (child), father and mother in the ith family; $i = 1, \ldots, N$. The parameters λ and κ are chosen to satisfy $F(45|G = 0) = p_1$ and $F(70|G = 0) = p_2$ and we let $\beta = \log 2$ to represent the scenario that being positive for the genetic marker increases the risk of developing the disease. The clinic entry time for the proband C_{i0} is taken to be normally distributed with mean 45 and variance 20, and the age of contact for parents follows normal distribution with mean 70 and variance 20; the age at contact for all individuals are truncated at 90 years. As in the main body of the manuscript we consider two parameter settings, (i) $p_1 = 0.15, p_2 = 0.3, q = 0.25$; (ii) the penetrances for the non-carrier and marker frequency are smaller and set at $p_1 = 0.05$, $p_2 = 0.1$, and $q = 0.1$.

We consider the cases that Clayton copula is used to introduce the residual within-family association between disease onset times at the *design and analyses stage*, whereas the true within-family association is induced by Frank copula (Section 1.1) or Gaussian copula (Section 1.2).

The potential families are being recruited to the phase I sample only if their probands satisfy $T_{i0} < C_{i0}$. At phase II, the potential recruited families are stratified into 4 strata based on G_{i0} and $I(Y_{i\cdot} \geq 1)$; each stratum has its own sampling probability π_{kl}, $k, l = 0, 1$. We impose the constraint that $P(R_i = 1|Y_{i0} = 1) = P_R = 0.4$. Simple random sampling, balanced sampling and optimal sampling proposed based on maximum likelihood method are applied to sample the families at phase II for 1000 simulated data sets with $N = 2000$.

1.1 Within-Family Association Induced by the Frank Copula

First, we assume the true within-family dependence structure is induced by Frank copula but Clayton copula is used for design and analyses. Among 1000 replicates, the average number of individuals for the four strata at phase I, $[N_{00}, N_{10}, N_{01}, N_{11}]$ are $[263, 275, 558, 904]$ for Kendall's $\tau = 0.1$ under parameter setting (i). The average stratum-specific selection probability and average number of selected families in each stratum based on simple random sampling, balanced sampling and optimal sampling are shown in Table S.1. The empirical properties of estimates from analysing 1000 simulated data sets while employing these sampling schemes at the second phase with an expected phase II sample size of $P(R_i = 1|Y_{i0} = 1) = 0.4$ are reported in Table S.2. When the dependence structure is misspecified, all estimates based on likelihood method are biased, but the bias for the estimate of genetic effect is quite small. The empirical standard errors of β under optimal sampling is 8% lower compared to simple random sampling for Kendall's $\tau = 0.1$. When within-family association becomes stronger, for example, Kendall's $\tau = 0.4$, the average number of individuals for the four strata at phase I, $[N_{00}, N_{10}, N_{01}, N_{11}]$ are $[82, 92, 738, 1088]$. The average stratum-specific sampling probability and average number of selected families in each stratum are reported in Table S.1 and the empirical properties of resulting estimates are summarized in Table S.2 as well. Similarly, the estimates for all parameters are biased and the biases become larger when the dependence increase. The empirical standard errors of β under optimal sampling is 6% lower compared to simple random sampling for Kendall's $\tau = 0.4$.

When there are auxiliary current status data with sample size $M = 2000$, the sampling probabilities did not change too much. The average stratum-specific selection probability and average number of selected families in each stratum based on these sampling schemes are shown in Table S.1 and the empirical properties of resulting estimates are summarized in Table S.3. Similar patterns could be observed when there are auxiliary data. Interestingly the biases become larger for all estimators when an auxiliary sample is available. The empirical standard errors of β under optimal sampling is 13% and 11% lower compared to simple random sampling for Kendall's $\tau = 0.1$ and 0.4, respectively.

We also consider all these scenarios with auxiliary current status data, but for parameter setting (ii), where the penetrance for non-carrier and marker frequency are smaller. When Kendall's $\tau = 0.1$, among 1000 replicates, 9, 4, and 8 replicates do not converge under the simple random sampling, balanced sampling and optimal sampling, respectively. Based on the

converged replicates (in total 979 replicates), the average number of individuals for the four strata at phase I, $[N_{00}, N_{10}, N_{01}, N_{11}]$ are $[982, 393, 390, 235]$. When within-family association becomes stronger, for example, Kendall's $\tau = 0.4$, among 1000 replicates, 8, 15 and 10 replicates do not converge under the simple random sampling, balanced sampling and optimal sampling, respectively. Based on the converged replicates (in total 968 replicates), the average number of individuals for the four strata at phase I, $[N_{00}, N_{10}, N_{01}, N_{11}]$ are $[599, 215, 773, 413]$. The average stratum-specific selection probability and average number of selected families in each stratum based on these sampling schemes are shown in Table S.1 for those scenarios. The empirical properties of corresponding estimates are summarized in Table S.4. The bias for all parameters increases when the within-family association become stronger. The empirical standard errors of β under optimal sampling is 20% and 19% lower compared to simple random sampling for Kendall's $\tau = 0.1$ and 0.4, respectively.

Table S.1: Average stratum-specific sampling probabilities and the average number of selected families in each stratum at phase II under simple random sampling, balanced sampling, and optimal designs under maximum likelihood when the true within-family association is induced by Frank copula; $P(R_i = 1|Y_{i0} = 1) = 0.4$.

	Parameter Setting I						Parameter Setting II		
	No Augmentation			With Augmentation			With Augmentation		
	srs	bal	opt_{ML}	srs	bal	opt_{ML}	srs	bal	opt_{ML}
				Kendall's $\tau = 0.1$					
π_{00}	0.40	0.76	0.05	0.4	0.76	0.05	0.40	0.20	0.16
π_{10}	0.40	0.73	0.05	0.4	0.73	0.05	0.40	0.51	0.05
π_{01}	0.40	0.36	0.05	0.4	0.36	0.05	0.40	0.51	1.00
π_{11}	0.40	0.22	0.82	0.4	0.22	0.82	0.40	0.85	1.00
n_{00}	105	200	13	105	200	13	393	199	154
n_{10}	110	200	15	110	200	15	157	200	21
n_{01}	224	200	28	224	200	28	156	200	390
n_{11}	362	200	743	362	200	743	94	200	235
				Kendall's $\tau = 0.4$					
π_{00}	0.40	1.00	1.00	0.40	1.00	1.00	0.40	0.33	0.30
π_{10}	0.40	1.00	0.05	0.40	1.00	0.05	0.40	0.93	0.05
π_{01}	0.40	0.42	0.05	0.40	0.42	0.05	0.40	0.26	0.26
π_{11}	0.40	0.29	0.62	0.40	0.29	0.62	0.40	0.49	1.00
n_{00}	33	82	81	33	81	81	239	200	177
n_{10}	37	92	5	37	92	5	86	199	11
n_{01}	296	313	42	296	314	42	309	200	199
n_{11}	435	313	672	436	314	673	165	200	413

Table S.2: Empirical properties of estimates from analysing data sets consisting of $N = 2000$ individuals when true within-family association is induced by Frank copula while employing simple random sampling, balanced sampling or optimal sampling for phase II design using Clayton copula with an expected phase II sample size of $P(R_i = 1|Y_{i0} = 1) = 0.4$; *Parameter Setting (i)*.

	Simple Random Sampling					Balanced Sampling					Optimal Sampling				
	BIAS	ESE	ASE	ASE†	ECP	BIAS	ESE	ASE	ASE†	ECP	BIAS	ESE	ASE	ASE†	ECP
							Kendall's $\tau = 0.1$								
$\log\lambda$	0.064	0.142	0.133	0.134	0.780	0.064	0.145	0.134	0.136	0.784	0.064	0.140	0.132	0.133	0.789
$\log\kappa$	-0.046	0.246	0.239	0.241	0.969	-0.047	0.247	0.239	0.241	0.968	-0.045	0.249	0.239	0.241	0.965
β	-0.006	0.090	0.089	0.088	0.938	-0.006	0.103	0.098	0.097	0.931	-0.006	0.083	0.080	0.079	0.940
$\log\frac{\nu}{1-\nu}$	0.000	0.061	0.062	0.062	0.961	-0.004	0.062	0.064	0.064	0.957	-0.003	0.067	0.065	0.065	0.941
τ	-0.073	0.020	0.020	0.020	0.024	-0.074	0.021	0.021	0.021	0.041	-0.073	0.020	0.020	0.020	0.029
							Kendall's $\tau = 0.4$								
$\log\lambda$	0.394	0.054	0.054	0.053	0.000	0.396	0.057	0.055	0.054	0.000	0.392	0.053	0.052	0.051	0.000
$\log\kappa$	0.076	0.195	0.206	0.205	0.912	0.075	0.197	0.206	0.205	0.916	0.078	0.194	0.205	0.205	0.911
β	0.036	0.088	0.090	0.089	0.946	0.032	0.093	0.092	0.091	0.928	0.040	0.083	0.082	0.081	0.925
$\log\frac{\nu}{1-\nu}$	0.001	0.063	0.062	0.062	0.949	-0.000	0.062	0.063	0.063	0.947	-0.003	0.066	0.066	0.066	0.955
τ	-0.236	0.027	0.026	0.026	0.000	-0.237	0.028	0.026	0.026	0.000	-0.236	0.027	0.026	0.026	0.000

Table S.3: Empirical properties of estimates from analysing data sets consisting of $N = 2000$ individuals with auxiliary current status data of size $M = 2000$ at phase I when true within-family association is induced by Frank copula while employing simple random sampling, balanced sampling or optimal sampling for phase II design using Clayton copula with an expected phase II sample size of $P(R_i = 1|Y_{i0} = 1) = 0.4$; *Parameter Setting (i)*.

	Simple Random Sampling					Balanced Sampling					Optimal Sampling				
	BIAS	ESE	ASE	ASE†	ECP	BIAS	ESE	ASE	ASE†	ECP	BIAS	ESE	ASE	ASE†	ECP
							Kendall's $\tau = 0.1$								
$\log\lambda$	0.127	0.077	0.074	0.075	0.542	0.125	0.081	0.078	0.078	0.568	0.130	0.073	0.071	0.071	0.511
$\log\kappa$	0.136	0.077	0.078	0.078	0.558	0.136	0.077	0.078	0.079	0.569	0.137	0.076	0.077	0.078	0.549
β	0.026	0.090	0.087	0.087	0.930	0.028	0.100	0.096	0.096	0.944	0.019	0.079	0.078	0.078	0.948
$\log\frac{\nu}{1-\nu}$	-0.007	0.063	0.062	0.062	0.945	-0.002	0.066	0.064	0.064	0.952	-0.005	0.063	0.065	0.065	0.956
τ	-0.063	0.017	0.017	0.017	0.037	-0.063	0.018	0.017	0.017	0.049	-0.064	0.017	0.017	0.017	0.035
							Kendall's $\tau = 0.4$								
$\log\lambda$	0.352	0.042	0.042	0.043	0.000	0.350	0.041	0.043	0.044	0.000	0.353	0.041	0.041	0.042	0.000
$\log\kappa$	0.456	0.059	0.058	0.058	0.000	0.455	0.059	0.058	0.058	0.000	0.455	0.059	0.059	0.058	0.000
β	0.142	0.088	0.090	0.092	0.662	0.145	0.087	0.092	0.095	0.659	0.137	0.078	0.081	0.083	0.621
$\log\frac{\nu}{1-\nu}$	-0.009	0.061	0.062	0.062	0.950	-0.005	0.062	0.063	0.063	0.951	-0.017	0.065	0.066	0.066	0.950
τ	-0.215	0.020	0.019	0.019	0.000	-0.215	0.019	0.020	0.020	0.000	-0.215	0.020	0.019	0.019	0.000

Table S.4: Empirical properties of estimates from analysing data sets consisting of $N = 2000$ individuals with auxiliary current status data of size $M = 2000$ at phase I when true within-family association is induced by Frank copula while employing simple random sampling, balanced sampling or optimal sampling for phase II design using Clayton copula with an expected phase II sample size of $P(R_i = 1|Y_{i0} = 1) = 0.4$; based on converged replicates, *Parameter Setting (ii)*.

	Simple Random Sampling					Balanced Sampling					Optimal Sampling				
	BIAS	ESE	ASE	ASE†	ECP	BIAS	ESE	ASE	ASE†	ECP	BIAS	ESE	ASE	ASE†	ECP
							Kendall's $\tau = 0.1$								
$\log\lambda$	0.329	0.262	0.247	0.254	0.570	0.334	0.258	0.238	0.246	0.568	0.341	0.254	0.234	0.241	0.544
$\log\kappa$	0.212	0.164	0.161	0.166	0.642	0.215	0.164	0.159	0.163	0.636	0.218	0.163	0.157	0.162	0.627
β	0.054	0.133	0.134	0.135	0.946	0.047	0.117	0.114	0.115	0.937	0.042	0.106	0.105	0.106	0.939
$\log\frac{\nu}{1-\nu}$	-0.005	0.083	0.087	0.087	0.952	-0.008	0.089	0.087	0.087	0.949	-0.008	0.085	0.085	0.085	0.949
τ	-0.086	0.012	0.012	0.012	0.000	-0.086	0.012	0.012	0.012	0.000	-0.086	0.012	0.012	0.012	0.000
							Kendall's $\tau = 0.4$								
$\log\lambda$	0.863	0.074	0.077	0.079	0.000	0.865	0.074	0.076	0.079	0.000	0.874	0.070	0.072	0.075	0.000
$\log\kappa$	0.644	0.080	0.082	0.085	0.000	0.645	0.080	0.082	0.084	0.000	0.650	0.079	0.081	0.083	0.000
β	0.200	0.120	0.121	0.122	0.640	0.190	0.117	0.120	0.121	0.670	0.163	0.097	0.097	0.098	0.630
$\log\frac{\nu}{1-\nu}$	-0.012	0.084	0.087	0.087	0.947	-0.016	0.090	0.091	0.091	0.955	-0.022	0.087	0.088	0.088	0.951
τ	-0.332	0.011	0.012	0.012	0.000	-0.332	0.011	0.012	0.012	0.000	-0.333	0.011	0.012	0.012	0.000

1.2 Within-Family Association is Induced by the Gaussian Copula

Furthermore, we consider a more complicated scenario where the true within-family association is induced by Gaussian copula, where τ_{fm}, τ_{fc} and τ_{mc} are the Kendall's τ for parents, father-child, mother-child pairs, respectively. With auxiliary current status data of size $M = 2000$, for parameter setting (i), when the pairwise associations are the same and Kendall's $\tau_{fm} = \tau_{fc} = \tau_{mc} = 0.1$, the average number of individuals for the four strata at phase-I, $[N_{00}, N_{10}, N_{01}, N_{11}]$ are $[271, 297, 548, 884]$ among 1000 replicates. When the dependence strength increases, for example, $\tau_{fm} = \tau_{fc} = \tau_{mc} = 0.4$, among the 1000 replicates, the average number of individuals for the four strata at phase I, $[N_{00}, N_{10}, N_{01}, N_{11}]$ are $[102, 133, 718, 1047]$. When $\tau_{fm} = 0.2, \tau_{fc} = 0.6, \tau_{mc} = 0.4$, among the 1000 replicates, 12, 8 and 7 replicates under simple random sampling, balanced sampling and optimal sampling do not converge, respectively. Based on the converged replicates (in total 974 replicates), the average number of individuals for the four strata at phase I, $[N_{00}, N_{10}, N_{01}, N_{11}]$ are $[17, 27, 804, 1152]$. The corresponding average stratum-specific selection probabilities and average number of selected families in each stratum based on those three sampling schemes are reported in Table S.5, and the empirical properties of estimates are summarized in Table S.6.

When Kendall's $\tau_{fm} = \tau_{fc} = \tau_{mc} = 0.1$, the empirical standard error of estimate for genetic effect under optimal sampling is just 12% lower compared to that under simple random sampling. The average of estimated Kendall's τ under those three sampling schemes are all 0.072 with empirical standard errors 0.0165, 0.0170 and 0.0164 respectively.

When Kendall's $\tau_{fm} = \tau_{fc} = \tau_{mc} = 0.4$, the empirical standard error of estimate for genetic

effect under optimal sampling is 12% lower compared to that under simple random sampling. The average of estimated Kendall's τ under those three sampling schemes are all 0.29 with empirical standard errors 0.026.

When $\tau_{fm} = 0.2, \tau_{fc} = 0.6, \tau_{mc} = 0.4$, based on the converged replicates, the empirical standard error of estimate for genetic effect under optimal sampling is 17% lower compared to that under simple random sampling. The average of estimated Kendall's τ are all 0.21, with standard errors 0.0218, 0.0214, and 0.0215 under the simple random sampling, balanced sampling and optimal sampling schemes respectively.

For parameter setting (ii) with auxiliary current status data, when $\tau_{fm} = \tau_{fc} = \tau_{mc} = 0.1$, among the 1000 replicates, 5, 6 and 8 replicates under simple random sampling, balanced sampling and optimal sampling do not converge, respectively. Based on the converged replicates (in total 981 replicates), the average number of individuals for the four strata at phase I, $[N_{00}, N_{10}, N_{01}, N_{11}]$ are $[939, 386, 433, 242]$. When the dependence strength increase, i.e., $\tau_{fm} = \tau_{fc} = \tau_{mc} = 0.4$, among the 1000 replicates, 3, 4 and 3 replicates under simple random sampling, balanced sampling and optimal sampling do not converge, respectively. Based on the converged replicates (in total 993 replicates), the average number of individuals for the four strata at phase I, $[N_{00}, N_{10}, N_{01}, N_{11}]$ are $[457, 197, 918, 428]$. When $\tau_{fm} = 0.2, \tau_{fc} = 0.6, \tau_{mc} = 0.4$, among the 1000 replicates, 12, 20 and 6 replicates under simple random sampling, balanced sampling and optimal sampling do not converge, respectively. Based on the converged replicates (in total 962 replicates), the average number of individuals for the four strata at phase I, $[N_{00}, N_{10}, N_{01}, N_{11}]$ are $[150, 76, 1222, 552]$. We see that the number of individuals for the four strata at phase I are quite different when the within-family dependence strength changes. The corresponding average stratum-specific selection probabilities and average number of selected families in each stratum based on those three sampling schemes are reported in Table S.5 and the empirical properties of estimates are summarized in Table S.7. When the copula function is misspecified, the estimates of all parameters are biased and biases increase when the within-family dependence becomes stronger.

When $\tau_{fm} = \tau_{fc} = \tau_{mc} = 0.1$, based on the converged replicates, the average of estimated Kendall's τ are all 0.046, with standard errors 0.013, 0.013, and 0.012 under the simple random sampling, balanced sampling and optimal sampling, respectively. The average of estimated Kendall's τ are all 0.226 with standard errors 0.018 under three different sampling schemes when $\tau_{fm} = \tau_{fc} = \tau_{mc} = 0.4$. The average of estimated Kendall's τ are 0.121, 0.121, and 0.118, with standard errors 0.016, 0.015, and 0.014 under the simple random sampling, balanced sampling and optimal sampling, respectively, when $\tau_{fm} = 0.2, \tau_{fc} = 0.6, \tau_{mc} = 0.4$.

The empirical standard errors of estimate for genetic effect under optimal sampling are 25%, 23% and 37% lower compared to those under simple random sampling for Kendall's $\tau = 0.1$, 0.4, and different pairwise Kendall's τ, respectively.

Here the ESEs and ASEs for $\hat{\beta}$ are quite different in the setting when $\tau_{fm} = 0.2, \tau_{fc} = 0.6$, and $\tau_{mc} = 0.4$.

Table S.5: Average stratum-specific sampling probabilities and the average number of selected families in each stratum at phase II under simple random sampling, balanced sampling, and optimal designs under maximum likelihood when within-family dependent structure is induced by Gaussian copula with auxiliary current status data of size $M = 2000$; $P(R_i = 1|Y_{i0} = 1) = 0.4$.

	Parameter Setting I			Parameter Setting II		
	srs	bal	opt_{ML}	srs	bal	opt_{ML}
	Kendall's $\tau_{fm} = \tau_{fc} = \tau_{mc} = 0.1$					
π_{00}	0.40	0.74	0.05	0.40	0.21	0.11
π_{10}	0.40	0.68	0.05	0.40	0.52	0.05
π_{01}	0.40	0.36	0.05	0.40	0.46	1.00
π_{11}	0.40	0.23	0.84	0.40	0.83	1.00
n_{00}	108	200	14	376	200	106
n_{10}	118	200	15	154	200	19
n_{01}	219	200	28	173	201	433
n_{11}	354	200	743	97	200	242
	Kendall's $\tau_{fm} = \tau_{fc} = \tau_{mc} = 0.4$					
π_{00}	0.40	1.00	1.00	0.40	0.44	0.63
π_{10}	0.40	1.00	0.05	0.40	1.00	0.05
π_{01}	0.40	0.39	0.05	0.40	0.22	0.09
π_{11}	0.40	0.27	0.62	0.40	0.47	1.00
n_{00}	41	103	102	182	202	286
n_{10}	53	133	7	79	193	10
n_{01}	287	282	41	367	202	76
n_{11}	419	282	651	171	202	428
	Kendall's $\tau_{fm} = 0.2, \tau_{fc} = 0.6, \tau_{mc} = 0.4$					
π_{00}	0.40	1.00	1.00	0.40	1.00	1.00
π_{10}	0.40	1.00	0.05	0.40	1.00	0.05
π_{01}	0.40	0.47	0.05	0.40	0.24	0.08
π_{11}	0.40	0.32	0.64	0.40	0.52	1.00
n_{00}	7	17	17	60	150	150
n_{10}	11	28	1	30	76	5
n_{01}	322	378	43	488	287	93
n_{11}	461	376	736	221	287	552

Table S.6: Empirical properties of estimates from analysing data sets consisting of $N = 2000$ individuals with auxiliary current status data of size $M = 2000$ at phase I when true within-family association is induced by Gaussian copula while employing simple random sampling, balanced sampling or optimal sampling for phase II design using Clayton copula with an expected phase II sample size of $P(R_i = 1|Y_{i0} = 1) = 0.4$; based on converged replicates, *Parameter Setting (i).*

	Simple Random Sampling					Balanced Sampling					Optimal Sampling				
	BIAS	ESE	ASE	ASE†	ECP	BIAS	ESE	ASE	ASE†	ECP	BIAS	ESE	ASE	ASE†	ECP
							Kendall's $\tau_{fm} = \tau_{fc} = \tau_{mc} = 0.1$								
$\log \lambda$	-0.022	0.098	0.101	0.103	0.966	-0.021	0.105	0.104	0.107	0.960	-0.020	0.096	0.097	0.098	0.959
$\log \kappa$	0.004	0.089	0.092	0.092	0.951	0.006	0.090	0.093	0.093	0.952	0.003	0.089	0.092	0.092	0.954
β	0.066	0.091	0.092	0.095	0.909	0.064	0.103	0.102	0.106	0.912	0.065	0.080	0.081	0.083	0.877
$\log \frac{\nu}{1-\nu}$	-0.000	0.062	0.062	0.062	0.951	0.006	0.066	0.064	0.064	0.940	-0.012	0.064	0.065	0.065	0.951
							Kendall's $\tau_{fm} = \tau_{fc} = \tau_{mc} = 0.4$								
$\log \lambda$	0.070	0.115	0.098	0.090	0.751	0.073	0.115	0.099	0.093	0.757	0.070	0.114	0.097	0.089	0.753
$\log \kappa$	0.090	0.127	0.108	0.095	0.746	0.084	0.128	0.109	0.097	0.759	0.091	0.124	0.107	0.095	0.762
β	0.070	0.092	0.090	0.091	0.871	0.042	0.092	0.092	0.093	0.926	0.079	0.081	0.081	0.081	0.816
$\log \frac{\nu}{1-\nu}$	-0.008	0.062	0.062	0.062	0.953	-0.004	0.063	0.064	0.064	0.953	-0.022	0.063	0.066	0.066	0.948
							Kendall's $\tau_{fm} = 0.2, \tau_{fc} = 0.6, \tau_{mc} = 0.4$								
$\log \lambda$	0.437	0.039	0.035	0.034	0.000	0.427	0.040	0.036	0.034	0.000	0.447	0.036	0.033	0.033	0.000
$\log \kappa$	0.614	0.046	0.049	0.053	0.000	0.613	0.046	0.048	0.053	0.000	0.615	0.046	0.049	0.054	0.000
β	0.193	0.124	0.103	0.089	0.523	0.242	0.129	0.106	0.092	0.380	0.134	0.103	0.089	0.078	0.666
$\log \frac{\nu}{1-\nu}$	-0.015	0.064	0.062	0.062	0.945	-0.023	0.062	0.062	0.062	0.937	-0.001	0.064	0.066	0.066	0.958

Table S.7: Empirical properties of estimates from analysing data sets consisting of $N = 2000$ individuals with auxiliary current status data of size $M = 2000$ at phase I when true within-family association is induced by Gaussian copula while employing simple random sampling, balanced sampling or optimal sampling for phase II design using Clayton copula with an expected phase II sample size of $P(R_i = 1|Y_{i0} = 1) = 0.4$; based on converged replicates, *Parameter Setting (ii).*

	Simple Random Sampling					Balanced Sampling					Optimal Sampling				
	BIAS	ESE	ASE	ASE†	ECP	BIAS	ESE	ASE	ASE†	ECP	BIAS	ESE	ASE	ASE†	ECP
							Kendall's $\tau_{fm} = \tau_{fc} = \tau_{mc} = 0.1$								
$\log \lambda$	-0.098	0.476	0.435	0.439	0.922	-0.087	0.469	0.421	0.424	0.914	-0.083	0.455	0.415	0.418	0.922
$\log \kappa$	-0.016	0.225	0.215	0.216	0.939	-0.013	0.223	0.211	0.212	0.936	-0.012	0.219	0.210	0.211	0.938
β	0.084	0.150	0.151	0.153	0.929	0.076	0.124	0.127	0.129	0.930	0.076	0.112	0.115	0.116	0.920
$\log \frac{\nu}{1-\nu}$	-0.002	0.092	0.087	0.087	0.931	-0.005	0.084	0.087	0.087	0.967	-0.013	0.087	0.086	0.086	0.949
							Kendall's $\tau_{fm} = \tau_{fc} = \tau_{mc} = 0.4$								
$\log \lambda$	-0.401	0.581	0.515	0.467	0.976	-0.399	0.577	0.514	0.467	0.979	-0.383	0.572	0.504	0.454	0.982
$\log \kappa$	-0.171	0.245	0.224	0.202	0.942	-0.172	0.243	0.224	0.202	0.944	-0.160	0.241	0.221	0.199	0.948
β	0.022	0.141	0.134	0.129	0.930	0.014	0.138	0.132	0.127	0.943	0.069	0.108	0.105	0.102	0.893
$\log \frac{\nu}{1-\nu}$	-0.010	0.083	0.087	0.087	0.956	-0.023	0.091	0.092	0.093	0.952	-0.028	0.092	0.092	0.093	0.939
							Kendall's $\tau_{fm} = 0.2, \tau_{fc} = 0.6, \tau_{mc} = 0.4$								
$\log \lambda$	1.150	0.045	0.037	0.036	0.000	1.151	0.042	0.036	0.035	0.000	1.162	0.035	0.033	0.034	0.000
$\log \kappa$	0.977	0.051	0.051	0.056	0.000	0.976	0.050	0.051	0.056	0.000	0.984	0.047	0.050	0.055	0.000
β	0.054	0.184	0.135	0.108	0.845	0.049	0.157	0.123	0.101	0.868	-0.024	0.115	0.094	0.083	0.876
$\log \frac{\nu}{1-\nu}$	-0.007	0.085	0.087	0.087	0.954	-0.032	0.090	0.094	0.094	0.955	-0.014	0.094	0.096	0.095	0.955