Published in final edited form as:

Stat Sin. 2019 ; 29(4): 1851–1871. doi:10.5705/ss.202017.0326.

SEMIPARAMETRIC TRANSFORMATION MODELS WITH MULTILEVEL RANDOM EFFECTS FOR CORRELATED DISEASE ONSET IN FAMILIES

Baosheng Liang¹, Yuanjia Wang², Donglin Zeng³

Baosheng Liang: bliang09@hku.hk; Yuanjia Wang: yw2016@columbia.edu; Donglin Zeng: dzeng@bios.unc.edu

¹Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

²Department of Biostatistics, Columbia University, New York, USA

³Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina, USA

Abstract

Large cohort studies are commonly launched to study risk of genetic variants or other risk factors on age at onset (AAO) of a chronic disorder. In these studies, family history data including AAO of disease in family members are collected to provide additional information and can be used to improve efficiency. Statistical analysis of these data is challenging due to missing genotypes in family members and the heterogeneous dependence attributed to both shared genetic back-ground and shared environmental factors (e.g., life style). In this paper, we propose a class of semiparametric transformation models with multilevel random effects to tackle these challenges. The proposed models include both proportional hazards model and proportional odds model as special cases. The multilevel random effects contain individual-specific random effects including kinship correlation structure dependent on the family pedigree, and a shared random effect to account for unobserved environment exposure. We use nonparametric maximum likelihood approach for inference and propose an expectation-maximization algorithm for computation in the presence of missing genotypes among family members. The obtained estimators are shown to be consistent, asymptotically normal, and semiparametrically efficient. Simulation studies demonstrate that the proposed method performs well with finite sample sizes. Finally, the proposed method is applied to study genetic risks in an Alzheimer's disease study.

Key words and phrases

Family data; Multilevel random effects; Nonparametric maximum likelihood estimation; Semiparametric efficiency; Alzheimer's disease

1. Introduction

Long term cohort studies are launched to study effect of risk factors, including genetic mutations, on age at onset (AAO) of a disease (e.g., Alzheimer's disease (Tang et al., 2001)). In these studies, family history data including AAO of disease in family members are collected to provide additional information and can be used to improve efficiency of estimating associations. However, although advanced technologies have decreased cost of genotyping, ascertaining many family members and collecting their blood samples is still

resource-expensive. Therefore, family history data, which can be cheaply collected from individuals participating in cohort studies (probands), remain a useful source to augment the proband data and improve the accuracy of genetic risk assessment (Whittemore, 1995; Parmigiani et al., 1998; Whittemore and Halpern, 2003).

There are several challenges when analyzing the combined family history and proband data to study genetic risks on disease AAO (Gorfine et al., 2013). First, when the disease is polygenetic, the disease onset events of the members in the same family are correlated due to the shared genetic background; second, since the family members are likely to be exposed to the same environment such as dietary and other non-genetic risk factors, their dependence can also be attributed to these unobserved environmental factors; third, since genotype information of the family members is usually not collected due to resource constraint, the exact genotypes are not observed in relatives, but their distribution can be inferred.

A number of methods have been developed to address these challenges in the analysis of the proband-family data. For example, Gorfine et al. (2009) analyzed a case-control family study for correlated failure times using a shared frailty proportional hazards model. Moreover, a class of frailty models for familial risk prediction with unknown genetic mutation status were studied in Chen et al. (2009), Graber et al. (2011), and Gorfine et al. (2013), where they assumed that a single-level of random effects were shared by the family under proportional hazards models. Garcia et al. (2017) estimated genetic risk functions from multivariate failure time data under a proportional odds model. While these works adjust for shared environmental factors using random effects, they do not account for the dependence due to the polygenetic effects. In an earlier work, Antoniou et al. (2008) applied a polygenic model by assuming that each individual had a random effect and their correlation followed the kinship correlation structure. However, this work does not consider possible dependence due to the shared environment by all family members or missing genotypes in family members.

In this paper, we consider a general class of semiparametric transformation models in Cheng et al. (1995) and Zeng and Lin (2007), which include both proportional hazards model and proportional odds model as special cases. We include multilevel random effects in the regression model and they consist of both individual-specific random effects and family-shared random effect. The former accounts for the dependence due to the shared genetic background with a kinship-structured covariance Khoury et al. (1993); while the latter represents the dependence due to the share environment effects or life style among the family members. Nonparametric maximum likelihood approach is used to handle missing genotypes for parameter estimation and inference. A fast numeric integration is used to perform multi-dimensional integration. We prove semiparametric efficiency of the proposed estimators, including establishing the invertibility of the information operator in a proper metric space. We also derive asymptotic variance formulae of the parameters, which avoids computationally intensive bootstrap procedure to compute variances. In the simulations and real data analysis, we show substantial reduction in bias, efficiency gain, and improvement in power of testing association, by properly accounting for hierarchical correlation structure among event times. We also demonstrate an improved efficiency for estimating cumulative

risk of dementia by including both probands data and family history data on their relatives in a long-term, real world study of aging and dementia.

The rest of the paper is organized as follows. We present the proposed model and the detail of an expectation-maximization algorithm for estimation in Section 2. In Section 3, the derived estimators are shown to be asymptotically efficient. Extensive simulation studies are conducted in Section 4. Finally, we apply the proposed method to study association between a causal mutation and AAO of dementia from an ongoing, large cohort study with structured family history data. Theoretical proofs are given in Supplementary Materials S.1.

2. Method

2.1. Model and likelihood function

Assume there are n i.i.d families. For the j th member in the i th family of size n_i , let T_{ij} denote the event time of age at onset, C_{ij} be the potential right censoring time, and X_{ij} be the baseline covariates including genetic polymorphisms and other risk factors. Furthermore, we denote $G_i = (G_{i1}, \dots, G_{in_i})$ as the genotype vector for all family members including the proband and let Σ_i be the kinship matrix defined in accordance with family i 's pedigree structure.

To study the genetic risk on the age at onset, we assume the following transformation model: given a family-specific random effect b_i and individual-specific random effects r_{ij} , the cumulative hazard function for T_{ij} follows

$$\Lambda_{ij}(t|G_{ij}, X_{ij}, b_i, r_{ij}) = H\{\Lambda(t) \exp(\beta G_{ij} + \gamma^T X_{ij} + b_i + r_{ij})\}, \quad (2.1)$$

where $H(\cdot)$ is a given transformation function, and both $\Lambda(t)$ and (β, γ) are unknown. Here β is the genetic association parameter of interest, and γ is a vector of parameters for other baseline covariates including confounders. Moreover, b_i is the family-specific random effect representing the family-shared environment and is assumed to follow $N(0, \sigma_b^2)$, and $(r_{i1}, \dots, r_{in_i})^T$ denotes the individual random effects for the shared polygenic effects (genetic background) and follows a multivariate normal distribution $MVN(0, \sigma_r^2 \Sigma_i)$. In model (2.1), transformation function $H(\cdot)$ is chosen from the class

$$\begin{cases} \alpha^{-1} \log(1 + \alpha x), & \alpha > 0, \\ x, & \alpha = 0, \end{cases}$$

so $\alpha = 0$ yields the proportional hazards model and $\alpha = 1$ yields the proportional odds model. Additionally, we assume that Mendelian transmission holds for G_{ij} .

In cohort studies involving family history data, G_i are often not available except for the probands who are the original participants. Instead, we observe a set $\mathcal{G}_i$ containing all possible genotypes (including the proband's genotype) that are consistent with the i th family's pedigree structure and proband's genotype. Thus, the observed data from n families

consist of $O_i = \{(Y_{ij}, X_{ij}, \mathcal{G}_i), j = 1, \dots, n_i\}, i = 1, \dots, n$, where $\delta_{ij} = I(T_{ij} \leq C_{ij})$ is the censoring indicator, and $Y_{ij} = \min(T_{ij}, C_{ij})$ is the observed event.

Assume that the censoring times $(C_{i1}, \dots, C_{in_i})^T$ are independent of $(T_{i1}, \dots, T_{in_i})^T$ and $(b_i, r_{i1}, \dots, r_{in_i})^T$ conditional on $(X_{i1}, \dots, X_{in_i})^T$. Then the observed likelihood function of the parameters in (2.1) takes form

$$\prod_{i=1}^n \int \sum_{g_i \in \mathcal{G}_i} \left\{ p_i(g_i) \prod_{j=1}^{n_i} [\lambda(Y_{ij}) \exp(\beta g_{ij} + \gamma^T X_{ij} + b_i + r_{ij}) \times H' \{ \lambda(Y_{ij}) \exp(\beta g_{ij} + \gamma^T X_{ij} + b_i + r_{ij}) \}]^{\Delta_{ij}} \right. \\ \times \exp [- H \{ \lambda(Y_{ij}) \exp(\beta g_{ij} + \gamma^T X_{ij} + b_i + r_{ij}) \}] \times (2\pi\sigma_b^2)^{-1/2} \exp(-\frac{b_i^2}{2\sigma_b^2}) (2\pi\sigma_r^2)^{-n_i/2} |\sum_i|^{-1/2} \exp(-\frac{r_i^T \sum_i^{-1} r_i}{2\sigma_r^2}) db_i dr_i, \\ \left. - \frac{r_i^T \sum_i^{-1} r_i}{2\sigma_r^2} \right\}$$

where $\lambda(t) = d\Lambda(t)/dt$ and $H'(\cdot)$ denotes the derivative of $H(\cdot)$. Here, $p_i(g_i)$ is the conditional probability of $G_i = g_i$ given the proband's genotype. Note that $p_i(g_i), i = 1, \dots, n$, can be computed under the Mendelian transmission and the family pedigree, so they are assumed to be known in the rest of this work.

2.2. Nonparametric maximum likelihood estimation

We use the nonparametric maximum-likelihood estimation (NPMLE) approach for parameter estimation of $\Lambda(t)$. Specifically, we estimate $\Lambda(t)$ as a step function with jumps only at the observed failure times and replace $\lambda(t)$ by the corresponding jump size of Λ at t in the likelihood function. We then maximize the following function over $\theta = (\beta, \gamma^T, \sigma_b^2, \sigma_r^2)^T$ and all the jump sizes of $\Lambda(\cdot)$:

$$\prod_{i=1}^n \int \sum_{g_i \in \mathcal{G}_i} \left\{ p_i(g_i) \prod_{j=1}^{n_i} [\Lambda\{Y_{ij}\} \exp(\beta g_{ij} + \gamma^T X_{ij} + b_i + r_{ij}) \times H' \{ \Lambda(Y_{ij}) \exp(\beta g_{ij} + \gamma^T X_{ij} + b_i + r_{ij}) \}]^{\Delta_{ij}} \right. \\ \times \exp [- H \{ \Lambda(Y_{ij}) \exp(\beta g_{ij} + \gamma^T X_{ij} + b_i + r_{ij}) \}] \left. \right\} \\ \times (2\pi\sigma_b^2)^{-1/2} \exp(-\frac{b_i^2}{2\sigma_b^2}) (2\pi\sigma_r^2)^{-n_i/2} |\sum_i|^{-1/2} \exp(-\frac{r_i^T \sum_i^{-1} r_i}{2\sigma_r^2}) db_i dr_i, \quad (2.2)$$

where $\Lambda\{Y_{ij}\}$ is the jump size of Λ at Y_{ij} .

Computationally, the above maximization can be carried out using an expectation-maximization (EM) algorithm. First, we notice that the transformation $H(x)$ satisfies

$$\exp\{-H(x)\} = \int_0^\infty \exp(-tx)\psi(t)dt,$$

where $\psi(t)$ is a gamma density with shape $1/\alpha$ and scale α . Therefore, if we introduce another random variable $\xi_{ij} \sim \text{Gamma}(1/\alpha)$, then the proposed model (2.1) is equivalent to assuming the conditional hazard rate of T_{ij} given ξ_{ij} , b_i , r_i , X_{ij} and G_{ij} to be

$$\xi_{ij}\lambda(t) \exp(\beta G_{ij} + \gamma^T X_{ij} + b_i + r_{ij}). \quad (2.3)$$

Treating ξ_{ij} , b_i , r_i , G_i as missing data, the observed likelihood function in (2.2) is equivalent to the likelihood arising from the complete data consisting of $\xi_i = (\xi_{i1}, \dots, \xi_{in})^T$ and $O_i^* = (G_i, b_i, r_i, O_i)$ for $i = 1, \dots, n$, whose joint density is

$$\begin{aligned} f(\xi_i, O_i^*) &= \prod_{j=1}^{n_i} \left\{ \xi_{ij} \Lambda\{Y_{ij}\} \exp(\beta G_{ij} + \gamma^T X_{ij} + b_i + r_{ij}) \right\}^{\Delta_{ij}} \\ &\times \exp[-\xi_{ij} \Lambda(Y_{ij}) \exp(\beta G_{ij} + \gamma^T X_{ij} + b_i + r_{ij})] \frac{\alpha^{-1/\alpha}}{\Gamma(1/\alpha)} \xi_{ij}^{1/\alpha - 1} \exp(-\xi_{ij}/\alpha) \\ &\times |\Sigma_i|^{-1/2} (2\pi\sigma_b^2)^{-1/2} (2\pi\sigma_r^2)^{-n_i/2} \exp\left(-\frac{b_i^2}{2\sigma_b^2} - \frac{r_i^T \Sigma_i^{-1} r_i}{2\sigma_r^2}\right) p_i(G_i). \end{aligned}$$

Each iteration in the EM algorithm consists of the following E- and M-steps. In the E-step, we calculate the conditional expectation of some integral function $Q(\xi_i, O_i^*)$ given O_i denoted by $\hat{E}\{Q(\xi_i, O_i^*) | O_i\}$, using the following equation

$$\hat{E}\{Q(\xi_i, O_i^*) | O_i\} = \frac{\sum_{g_i \in \mathcal{G}_i} \int_{\xi_i, b_i, r_i} Q(\xi_i, O_i^*) f(\xi_i, O_i^*) I(G_i = g_i) d\xi_i db_i dr_i}{\sum_{g_i \in \mathcal{G}_i} \int_{\xi_i, b_i, r_i} f(\xi_i, O_i^*) I(G_i = g_i) d\xi_i db_i dr_i}.$$

The multi-dimensional integration to compute the above conditional expectation can be challenging due to multiple level of random effects (both family level and individual level), especially when the family size is large. We propose to use numeric approximation with a multi-dimension adaptive Gaussian quadrature approach (Abramowitz and Stegun, 1972; Liu and Pierce, 1994; Evans and Swartz, 2000) to evaluate the above integral, where we apply a sparse grid technique (Heiss and Winschel, 2008) to generate nodes for the quadratures with dimension larger than four.

In the M-step, we maximize the conditional expectation of the complete data log-likelihood function, which is given by

$$\begin{aligned}
& \sum_{i=1}^n \sum_{j=1}^{n_i} \{ \Delta_{ij} (\log \Lambda(Y_{ij}) + \log \xi_{ij} + \beta G_{ij} + \gamma^T X_{ij} + b_i + r_{ij}) \\
& - \xi_{ij} \Lambda(Y_{ij}) \exp(\beta G_{ij} + \gamma^T X_{ij} + b_i + r_{ij}) \} + \sum_{i=1}^n \log \{ p_i(G_i) \} \\
& - \frac{1}{2} \sum_{i=1}^n \left[\log(2\pi\sigma_b^2) + \frac{b_i^2}{\sigma_b^2} + n_i \log(2\pi\sigma_r^2) + \log \left| \sum_i \right| + \frac{1}{\sigma_r^2} r_i^T \sum_i^{-1} r_i \right],
\end{aligned}$$

given the observed data and the current parameter values. By simple algebra, we update σ_b^2 and σ_r^2 by

$$\sigma_b^2 = \frac{1}{n} \sum_{i=1}^n \hat{E}(b_i^2 | O_i), \quad \sigma_r^2 = \frac{1}{\sum_{i=1}^n n_i} \sum_{i=1}^n \hat{E}(r_i^T \sum_i^{-1} r_i | O_i).$$

We then update β and γ by solving

$$\begin{aligned}
0 = & \sum_{i=1}^n \sum_{j=1}^{n_i} \Delta_{ij} \left\{ \left(\hat{E}(G_{ij} | O_i), X_{ij}^T \right)^T - \right. \\
& \left. \frac{\sum_{k=1}^n \sum_{l=1}^{n_i} I(Y_{kl} \geq Y_{ij}) \hat{E}\{ (G_{kl}, X_{kl}^T)^T \xi_{kl} \exp(\beta G_{kl} + \gamma^T X_{kl} + b_k + r_{kl}) | O_i \}}{\sum_{k=1}^n \sum_{l=1}^{n_i} I(Y_{kl} \geq Y_{ij}) \hat{E}\{ \xi_{kl} \exp(\beta G_{kl} + \gamma^T X_{kl} + b_k + r_{kl}) | O_i \}} \right\}
\end{aligned}$$

using one-step Newton-Raphson method. Based on the updated values for β and γ , we calculate the jump size of Λ at Y_{kl} as

$$\frac{\sum_{i=1}^n \sum_{j=1}^{n_i} \Delta_{ij} I(Y_{ij} = Y_{kl})}{\sum_{i=1}^n \sum_{j=1}^{n_i} I(Y_{ij} \geq Y_{kl}) \hat{E}\{ \xi_{ij} \exp(\beta G_{ij} + \gamma^T X_{ij} + b_i + r_{ij}) | O_i \}}. \quad (2.4)$$

We iterate between the E- and M-steps till convergence, and denote the final estimators for θ and Λ as $\hat{\theta}$ and $\hat{\Lambda}$, respectively.

3. Asymptotic Properties

To establish the asymptotic distribution of the nonparametric maximum likelihood estimator $(\hat{\theta}, \hat{\Lambda})$, we assume that the family size n_i is a bounded random variable and that conditional on n_i the pedigree structure of the family, denoted by $\mathcal{K}_i$, has a discrete distribution with finite choices so Σ_i , the kinship matrix, is also random. Denote the true values for θ and $\Lambda(t)$ as θ_0 and $\Lambda_0(t)$, respectively. We need the following assumptions.

- A.1** The true value θ_0 lies in a known compact set Θ in the interior of the domain for θ . Moreover, the true function Λ_0 is continuously differentiable with $\Lambda'_0(t) > 0$ in $[0, \tau]$, where τ is the study duration assumed to be finite.
- A.2** With probability one, X is bounded and there exists a positive constant δ such that $P\{\sum_{j=1}^{n_i} I(Y_{ij} \geq \tau) \geq 1 | X_{ij}, j = 1, \dots, n_i, n_i\} \geq \delta$.
- A.3** There exists a constant n_0 such that the family size n_i satisfies $P(1 \leq n_i \leq n_0) = 1$ and $P(n_i \geq 2) > 0$.
- A.4** Conditional on n_i , let $e_{ij} = (1, 0, \dots, 0, 1, 0, \dots, 0)^T$ be a $(n_i+1) \times 1$ vector with only the first and the j th elements equal to 1 for $j = 1, \dots, n_i$. If $(\beta, \gamma, c_1, c_2)$ is a constant vector and $v(t)$ is a deterministic function satisfying

$$v(t) + \beta g_{ij} + \gamma^T X_{ij} + e_{ij}^T \begin{pmatrix} c_1 & 0 \\ 0 & c_2 \sum_i \end{pmatrix} e_{ij} = 0, \quad e_{ij}^T \begin{pmatrix} c_1 & 0 \\ 0 & c_2 \sum_i \end{pmatrix} e_{ij'} = 0$$

for $j \neq j', j, j' = 1, \dots, n_i$, g_{ij} and $t \in [0, \tau]$ with probability 1 (note that both X_{ij} and Σ_j corresponding to the pedigree $\mathcal{K}_i$ are random), then $\beta = 0$, $\gamma = 0$, $v(t) = 0$ for $t \in [0, \tau]$, and $c_1 = c_2 = 0$.

- A.5** Conditional on n_i , if $\sum (g_1, \dots, g_{n_i}) q(g_1, \dots, g_{n_i}) p(g_1, \dots, g_{n_i}) I((g_1, \dots, g_{n_i}) \sim \mathcal{K}_i) = 0$ with probability one for some function $q(\cdot)$, where the summation is over all genotype vectors $(g_1, \dots, g_{n_i})$ and $(g_1, \dots, g_{n_i}) \sim \mathcal{K}_i$ means that $(g_1, \dots, g_{n_i})$ is compatible with the pedigree $\mathcal{K}_i$, then $q(g_1, \dots, g_{n_i}) = 0$.

Conditions (A.1)–(A.2) are standard assumptions for clustered survival data. Condition (A.3) assumes that the family size is bounded and that there exist at least some families with at least two members. Both conditions (A.4) and (A.5) are used to obtain parameter identifiability in the presence of multilevel random effects. In particular, (A.4) holds if $(1, X_{ij}, g_{ij})$ has a full rank with positive probability and Σ_j has non-zero off-diagonal elements with a positive probability. In (A.5), we require that for a fixed family size, the matrix from the joint probabilities of $\{p(g_i)\}$ across all possible pedigree structures is a full rank matrix. Under the above assumptions, we obtain the following asymptotic results.

Theorem 1

Under (A.1)–(A.5), we have $|\hat{\theta}_n - \theta_0| \rightarrow 0$ and $\|\hat{\Lambda}_n - \Lambda_0\|_\infty \rightarrow 0$ almost surely, where $|\cdot|$ is the Euclidean norm and $\|\cdot\|_\infty$ is the supremum norm in the interval $[0, \tau]$.

The proof of Theorem 1 is given in the Supplementary Materials S.1. The key idea of the proof is to first show that $\hat{\Lambda}$ is bounded so is weakly compact, and then show any convergence sequence of $(\hat{\theta}, \hat{\Lambda})$ should converge to the true parameters. The latter makes use of the Glivenko–Cantelli theorem for empirical processes and the identifiability results under conditions (A.4) and (A.5) to be established in the proof.

To establish the asymptotic distribution, let $\mathcal{L} = \{h(t) : h(t) \in BV[0, \tau], \|h\|_{BV} = 1\}$, where $\|h\|_{BV}$ denotes the total variation of h , then by defining $\hat{\Lambda}_n(h) = \int_0^\tau h(s) d\hat{\Lambda}_n(s)$ for $h \in \mathcal{L}$, $\hat{\Lambda}_n(t)$ is considered as a bounded linear function in $\mathcal{F}^\infty(\mathcal{L})$. Hence, $(\hat{\theta}_n - \theta_0, \hat{\Lambda}_n - \Lambda_0)$ is viewed as a random element in the metric space $\mathbb{R}^d \times \mathcal{F}^\infty(\mathcal{L})$, where d is the dimension of θ_0 . Then the asymptotic distribution of the estimators is given as follows.

Theorem 2

Under (A.1)–(A.5), it holds that $\sqrt{n}(\hat{\theta}_n - \theta_0, \hat{\Lambda}_n - \Lambda_0)$ converges weakly to a mean zero Gaussian process in the metric space $\mathbb{R}^d \times \mathcal{F}^\infty(\mathcal{L})$. In addition, the asymptotic covariance matrix of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ attains the semiparametric efficiency bound.

Theorem 2 concludes that $\hat{\theta}_n$ is an efficient estimator for θ_0 . The proof of Theorem 2 mostly follows the standard arguments for the semiparametric transformation models in Zeng and Lin (2007). One key challenge in the proof is to establish the invertibility of the information operator under the proposed multilevel random effect models. Furthermore, following the result of Theorem 4 in Zeng and Lin (2007), the asymptotic covariance for $\hat{\theta}$ can be estimated as the inverse of the observed information matrix. The latter can be calculated using the Louis formula method (Louis, 1982). Thus, using the estimated covariance matrix, we can construct the asymptotic confidence intervals for β_0 and conduct score test for the null hypothesis $\beta_0 = 0$.

4. Simulation Studies

Extensive simulation studies were conducted to assess the performance of the proposed method in finite samples. In the simulations, we considered transformation $H(x) = x$ and $\log(1 + x)$ corresponding to the proportional hazards model and proportional odds model, respectively. We generated one binary covariate X_{ij} from Bernoulli(0.5) and random effect $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b = 0.5$. To generate both random effects r_i and G_i , we considered two types of family pedigree structures as given in Figure 1, namely Case I and Case II. In particular, Case I is a two-generation pedigree including four members: Father, Mother, Sibling and Proband, and Case II is a three-generation pedigree with five members: Father (F), Mother (M), Sibling (S), Child (C) and Proband (P). Corresponding to each case, the resulting kinship matrices are respectively given as

$$\Sigma_1 = \begin{pmatrix} \text{F} & \text{M} & \text{S} & \text{P} \\ 1.00 & 0.00 & 0.50 & 0.50 \\ 0.00 & 1.00 & 0.50 & 0.50 \\ 0.50 & 0.50 & 1.00 & 0.50 \\ 0.50 & 0.50 & 0.50 & 1.00 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} \text{F} & \text{M} & \text{S} & \text{C} & \text{P} \\ 1.00 & 0.00 & 0.50 & 0.25 & 0.50 \\ 0.00 & 1.00 & 0.50 & 0.25 & 0.50 \\ 0.50 & 0.50 & 1.00 & 0.50 & 0.50 \\ 0.25 & 0.25 & 0.50 & 1.00 & 0.50 \\ 0.50 & 0.50 & 0.50 & 0.50 & 1.00 \end{pmatrix}.$$

Using the kinship matrix, we generated r_i from a multivariate normal distribution with mean zero and covariance matrix $\sigma_r^2 \Sigma_k$ for $k = 1, 2$ with $\sigma_r^2 = 0.5$ or 1.0 . The proband's alleles were obtained with equal probabilities. Next, we used the Monte-Carlo method to obtain

$p_i(\cdot)$ s for the i th family under its given pedigree structure: assuming the minor allele frequency to be 0.02 and using the Mendelian principles of inheritance, we used the pedigree structure to simulate all genotypes for the whole family members including the proband by starting from the top level of the pedigree. We repeated this simulation 30,000 times and retained those relatives' genotypes for which the proband's genotype matched with his or her observed genotype. Then $p_i(\cdot)$ was calculated as the proportion of the relatives' genotypes in the retained samples. Finally, the disease onset time, T_{ij} , was generated from model (2.1) with $\beta = 0.4$, $\gamma = -0.5$ and $\lambda(t) = 1/2$. The censoring time was simulated from the uniform distribution in $[0, 6]$ to yield around 30–35% censoring rates. In the simulations, we considered sample size $n = 300$ and 500 and conducted simulations with 1,000 replicates.

For each simulated data, we used the proposed EM algorithm with multi-dimension adaptive Gaussian quadrature to calculate the NPMLEs. In the EM algorithm, the initial values of β and γ were set to be 0, the initial values of σ_b^2 and σ_r^2 were set to be 1, and the initial values of $\Lambda(t)$'s jump sizes were set to be $1/m$, where m was the total count number of the disease events. We used the inverse of observed information matrix, computed by the Louis formula, to estimate the covariances, and used the Satterthwaite approximation (e.g., Satterthwaite, 1946; Burdick and Graybill, 1992) to construct the 95% confidence interval for σ_b^2 and σ_r^2 : $(v\hat{\sigma}^2/\chi_{v, .975}^2, v\hat{\sigma}^2/\chi_{v, .025}^2)$, where $v = 2\{\hat{\sigma}^2/\widehat{SE}(\hat{\sigma}^2)\}^2$, $\widehat{SE}(\hat{\sigma}^2)$ is the estimated standard error of $\hat{\sigma}^2$, $\chi_{v, q}^2$ is the q -quantile of the chi-squared distribution with v degrees of freedom, and $\hat{\sigma}^2 = \hat{\sigma}_b^2$ or $\hat{\sigma}_r^2$. Additionally, we compared analyses under the proposed model with analyses under a misspecified model similar to (2.1) except that no r_{ij} was included. Thus, in the misspecified model, only the family-shared random effect was used to account for the dependence within the family.

Table 1 and 2 display the simulation results for Case I and II with sample size $n = 300$ and 500, $\sigma_r^2 = 0.5$ and 1.0. As seen in these tables, the estimates of $(\beta_0, \gamma_0, \sigma_{b0}^2, \sigma_{r0}^2)$ and $\Lambda_0(\cdot)$ are virtually unbiased when sample sizes increase. Because the parameter σ_{r0}^2 is non-negative, the empirical distribution of $\hat{\sigma}_r^2$ is not symmetric in practice. The estimated standard errors (SE) of σ_{r0}^2 are slightly larger than the standard deviations (SD) with small sample sizes.

However, as the sample size increases, the performance can be significantly improved. Overall, the estimated standard errors accurately reflect the true variability. Furthermore, when comparing to the misspecified model which ignores the dependence due to the shared genetic background, the estimates from our multilevel method are more accurate and reliable than the estimates from the misspecified model, especially for estimating the genetic association parameter β . The coverage probabilities obtained under the misspecified model are much lower than the nominal level. Table 3 further presents the root mean of integrated squared error for estimating the cumulative hazard functions under both proposed model and misspecified model. Clearly, we observe that the misspecified model can lead to estimates with large bias and variability. The above conclusion is also reflected in Figure 2, where the relative bias of the estimated Euclidean parameter θ is defined as $|(\hat{\theta}_n - \theta_0)/\theta_0|$ and the relative efficiency improvement of the proposed model over the misspecified model in terms

of the root mean integrated squared errors (abbreviated as MSE) is defined as $|(MSE(\hat{\Lambda}_n) - MSE_{missp}(\hat{\Lambda}_n)) / MSE(\hat{\Lambda}_n)|$. For example, the relative bias under misspecified single level model can be up to 20% and the relative efficiency gain of multi-level model can be up to 30%.

Finally, we compare the type-I errors of the score test under the true model and misspecified model, where $\beta = 0$ and the other parameters remain the same. Our simulation results show that, for both proportional hazard model and proportional odds model, the type-I errors of misspecified model are heavily inflated. Specifically, for the Cox proportional hazard model, when the sample size is 300, the type-I error of misspecified model is around 0.112 with $\sigma_r^2 = 1.0$, and it becomes 0.094 when the sample size is 500. The type-I error rates are 0.120 and 0.109 for the proportional odds model. The type-I error rates from our approach are mostly around the significance level of 0.05.

5. Application

The Washington Heights-Inwood Community Aging Project (WHICAP) is a prospective, community-based cohort study of aging, dementia, and Alzheimer's disease (AD) in Northern Manhattan, New York City (Tang et al., 2001). The ongoing, decades-long study collected rich information on neuropsychiatric measures, environmental exposures, and genetic risk factors in thousands of participants. Incidence of dementia in proband participants was carefully monitored. The familial aggregation of dementia and AD is well established. First-degree relatives of AD patients have approximately twice the expected life time risk of dementia compared with persons without an affected first-degree relative (Devi et al., 2000). Among African Americans, increased relative risk of AD among first-degree relatives of patients was also observed (ranges from 1.4 in a community based sample to 2.6 in a clinic based sample). These findings suggest a substantial heritable contribution to AD and dementia, as well as non-genetic pathways due to shared environmental risk factors or life styles.

A structured family history interview (Maestre et al., 1995) was administered to the probands. The family history interview collected pedigree structure, birth order, vital status, level of education, family history of cognitive impairment, dementia, and AAO among first-degree relatives, including the non-WHICAP parent and among siblings. In our analysis, 1705 probands and 1342 relatives from 692 families were included in data analysis. The censoring rates of probands' and relatives' observations were 89.6% and 91.99%, respectively. Apolipoprotein-E (APOE) genotypes were obtained for each proband. Of interest is to estimate the cumulative incidence of dementia for APOE- $\epsilon 4$ mutation carrier group and non-carrier group adjusting for two potential confounders, gender and education.

To select the best transformation, we varied α in $H(\cdot)$ from 0 to 4 and chose the optimal model which gave the largest likelihood function. The optimal value of α is 2 (see Figure 3), the variance calculated using the negative inverse of the curvature of the log profile likelihood function at the estimated α is 0.249, and the 95% confidence interval is [1.02, 2.98]. The analyses results based on the transformation model with $\alpha = 2$ are presented in Table 4. From the table, the estimated log hazard ratio for the APOE- $\epsilon 4$ is 0.930 with $p <$

0.0001, indicating a significant higher risk of the dementia occurrence in the APOE- $\epsilon 4$ carrier group. Gender is not significant in either analysis. The table also shows that the lower education has a significant effect of an increased risk of dementia.

In Table 5, we report the estimated cumulative distribution functions for carrier and non-carrier groups with other covariates fixed at the sample mean level. Furthermore, to demonstrate the usefulness of the family history data on relatives, we also report the estimates from the best transformation using the proband data only, in which no random effects were used and the best transformation was $\alpha = 0.6$. Interestingly, for the carrier group, both analyses give similar results, but in the non-carrier group, using both proband and family data gives slightly lower estimates of cumulative risk of dementia with a smaller variability (also see Figure 4). This analysis demonstrates the efficiency gain of including the family history data on relatives.

6. Discussion

In this article, we studied a class of transformation models with multilevel random effects for the correlated disease age at onset. The proposed method incorporated multilevel random effects to explain polygenic heterogeneity and shared environment within the family and accounted for missing genotype information in the relatives in a family. The NPMLE has been shown to perform well in small sample. Our application shows that using family history data can increase accuracy of disease risk estimation in non-carrier group. It can be conjectured that adding family history data will lead to improved power to detect the genetic association in genome wide association studies.

Although the proposed model is described for time-invariant covariates, it can be easily generalized to incorporate time-varying covariates for which model (2.1) becomes

$$\Lambda_{ij}(t|G_{ij}, X_{ij}(s), b_i, r_{ij}; s \leq t) = H\left\{\int_0^t I(s \leq T_{ij}) \exp(\beta G_{ij} + \gamma^T X_{ij}(s) + b_i + r_{ij}) d\Lambda(s)\right\}. \quad (6.1)$$

The approaches dealing with time-varying covariates in (Zeng and Lin, 2007) can be adapted to make inferences for model (6.1). However, if b_i or r_{ij} is also time-varying, the estimation would become more complicated. Currently, there is a lack of efficient method for handling such scenario.

When the family size is large, a computational challenge is how to handle multidimensional integral in the proposed method, since the dimension of quadratures in the EM algorithm increases as the family size increases. One alternative is to consider sampling manageable number of family members from the families with a big size. The proposed method can then be applied to this subset of the data. This process can be repeated for multiple times and proper combination of the results, after accounting for sampling weights and data overlapping, will be used to obtain the final estimates. Moreover, using the data from relatives should always improve efficiency in theory, but with a large σ_b^2 , such efficiency gain may not be significant. Thus, we would also recommend randomly selecting a small

pedigree from each family for analysis, which greatly reduces computation cost without much efficiency loss.

In some cases, disease prevalence in the non-carrier group is available on the entire population (e.g., age-specific population risk of dementia). Thus, such information can be used to further increase efficiency and numeric stability in the analysis of the familial data. As a future work, we will calibrate the parameter estimation and prediction using information available from the population.

Supplemental Materials Proof of Theorem 1 and additional simulation studies mimicking the real data are presented in the supplemental file. The simulations show an adequate performance of selecting transformation parameter based on the profile likelihood method and little efficiency loss of our method as compared to the standard frailty model when true $\sigma_r^2 = 0$.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is supported by U.S. National Institute of Health grants (NS073671) and National Natural Science Foundation of China (No. 11371062). WHICAP study was supported by AG037212.

References

- Abramowitz, M, Stegun, IA. Handbook of Mathematical Functions. 9. Dover Publications; New York: 1972.
- Antoniou AC, Cunningham AP, Peto J, Evans DG, Lalloo F, Narod SA, Risch HA, Eyfjord JE, Hopper JL, Southey MC, Olsson H, Johannsson O, Borg A, Pasini B, Radice P, Manoukian S, Eccles DM, Tang N, Olah E, Anton-Culver H, Warner E, Lubinski J, Gronwald J, Gorski B, Tryggvadottir L, Syrjakoski K, Kallioniemi OP, Eerola H, Nevanlinna H, Pharoah PD, Easton DF, Easton DF. 2008; The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. Br. J. Cancer. 98:1457–1466. [PubMed: 18349832]
- Burdick, RK, Graybill, FA. Confidence Intervals on Variance Components. Marcel Dekker; New York: 1992.
- Chen L, Hsu L, Malone K. 2009; A frailty-model-based approach to estimating the age-dependent penetrance function of candidate genes using population-based case-control study designs: an application to data on the BRCA1 gene. Biometrics. 65:1105–1114. [PubMed: 19210733]
- Cheng SC, Wei LJ, Ying Z. 1995; Analysis of transformation models with censored data. Biometrika. 82:835–845.
- Devi G, Ottman R, Tang MX, Marder K, Stern Y, Mayeux R. 2000; Familial aggregation of Alzheimer disease among whites, African Americans, and Caribbean Hispanics in northern Manhattan. Arch. Neurol. 57:72–77. [PubMed: 10634451]
- Evans, M, Swartz, T. Approximating Integrals via Monte Carlo and Deterministic Methods. Oxford University Press; 2000.
- Heiss F, Winschel V. 2008; Likelihood approximation by numerical integration on sparse grids. J. Econometrics. 144:62–80.
- Garcia T, Ma Y, Marder K, Wang Y. 2017; Robust mixed-effects model for clustered failure time data: Application to Huntington's disease event measures. Ann. Appl. Stat. 11:1085–1116. [PubMed: 29399240]

- Gorfine M, Zucker DM, Hsu L. 2009; Case-control survival analysis with a general semiparametric shared frailty model-a pseudo full likelihood approach. *Ann. Statist.* 37:1489–1517.
- Gorfine M, Hsu L, Parmigiani G. 2013; Frailty models for familial risk with application to breast cancer. *J. Am. Statist. Ass.* 108:1205–1215.
- Graber-Naidich A, Gorfine M, Malone KE, Hsu L. 2011; Missing genetic information in case-control family data with general semi-parametric shared frailty model. *Lifetime Data Anal.* 17:175–194. [PubMed: 21153764]
- Khoury, MJ, Beaty, TH, Cohen, BH. *Fundamentals of Genetic Epidemiology (Monographs in Epidemiology and Biostatistics, vol. 22)*. Oxford University Press; 1993.
- Liu Q, Pierce DA. 1994; A note on Gauss-Hermite quadrature. *Biometrika.* 81:624–629.
- Louis TA. 1982; Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc. B.* 44:226–233.
- Maestre G, Ottman R, Stern Y, Gurland B, Chun M, Tang MX, Shelanski M, Tycko B, Mayeux R. 1995; Apolipoprotein E and Alzheimer's disease: ethnic variation in genotypic risks. *Ann. Neurol.* 37:254–259. [PubMed: 7847867]
- Parmigiani G, Berry DA, Aguilar O. 1998; Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *Am. J. Hum. Genet.* 62:145–158. [PubMed: 9443863]
- Satterthwaite FE. 1946; An approximate distribution of estimates of variance components. *Biometrics.* 2:110–114. [PubMed: 20287815]
- Tang MX, Cross P, Andrews H, Jacobs DM, Small S, Bell K, Merchant C, Lantigua R, Costa R, Stern Y, Mayeux R. 2001; Incidence of AD in African-Americans, Caribbean hispanics, and caucasians in northern Manhattan. *Neurology.* 56:49–56. [PubMed: 11148235]
- Whittemore AS. 1995; Logistic regression of family data from case-control studies. *Biometrika.* 82:57–67.
- Whittemore AS, Halpern J. 2003; Logistic regression of family data from retrospective study designs. *Genet. Epidemiol.* 25:177–189. [PubMed: 14557986]
- Zeng D, Lin DY. 2007; Maximum likelihood estimation in semiparametric regression models with censored data. *J. R. Statist. Soc. B.* 69:507–564.

Case I

Case II

Figure 1.

The pedigree structures used for simulating genotype data: black solid icons indicate mutation gene carriers for one particular mutation example.

Figure 2.

The first row presents the barplots of relative biases of the estimated β , γ under scenarios of Case I (left) and Case II (right), the second row shows the barplots of relative efficiency improvement in terms of the root mean integrated squared errors for estimating Λ , the left panel and right panel correspond to $\alpha = 0$ and $\alpha = 1$, respectively.

Figure 3.
The log-profile-likelihood function along with different values of α calculated from the combined data of probands and relatives.

Figure 4.

Estimated cumulative risk of dementia for APOE- ϵ 4 carriers and non-carriers: the top two plots are the estimates from proband data only and the bottom two are from both probands and relative data.

Table 1

Simulation results under Case I scenario

$H(x)$	σ_r^2	n	Par	True	Proposed Model				Misspecified Model			
					Bias	SD	SE	CP%	Bias	SD	SE	CP%
$\alpha = 0$	0.5	300	σ_b^2	0.25	0.001	0.121	0.129	92.0	0.080	0.099	0.097	77.0
			σ_r^2	0.50	0.001	0.328	0.341	93.4	—	—	—	—
			β	0.50	-0.028	0.335	0.332	95.0	-0.033	0.328	0.305	93.6
			γ	-0.50	-0.005	0.119	0.113	94.2	0.056	0.098	0.095	89.0
			$\Lambda(\pi/4)$	1.25	0.044	0.216	0.208	93.6	0.020	0.207	0.186	91.0
			$\Lambda(\pi/2)$	1.50	0.089	0.437	0.425	93.6	-0.099	0.365	0.327	86.4
		500	σ_b^2	0.25	-0.007	0.091	0.101	94.2	0.072	0.071	0.074	72.6
			σ_r^2	0.50	-0.011	0.265	0.272	95.2	—	—	—	—
			β	0.50	-0.009	0.258	0.256	95.6	-0.010	0.254	0.236	93.6
			γ	-0.50	-0.001	0.085	0.087	95.6	0.056	0.073	0.073	87.0
			$\Lambda(\pi/4)$	0.75	0.017	0.156	0.156	94.2	-0.006	0.149	0.139	91.6
			$\Lambda(\pi/2)$	1.50	0.031	0.315	0.314	95.2	-0.146	0.257	0.242	81.8
	1.0	300	σ_b^2	0.25	0.017	0.146	0.154	89.9	0.136	0.110	0.106	56.8
			σ_r^2	1.00	-0.076	0.468	0.492	97.0	—	—	—	—
			β	0.50	-0.022	0.387	0.378	94.1	-0.023	0.379	0.335	91.4
			γ	-0.50	0.001	0.128	0.124	94.7	0.093	0.098	0.097	83.2
			$\Lambda(\pi/4)$	0.75	0.047	0.249	0.237	93.5	0.003	0.234	0.199	88.0
			$\Lambda(\pi/2)$	1.50	0.073	0.503	0.481	93.4	-0.211	0.393	0.331	75.2

$H(x)$	σ_r^2	n	Par	True	Proposed Model				Misspecified Model			
					Bias	SD	SE	CP%	Bias	SD	SE	CP%
		500	σ_b^2	0.25	0.003	0.110	0.123	92.8	0.129	0.079	0.081	45.0
			σ_r^2	1.00	-0.068	0.433	0.392	94.9	—	—	—	—
			β	0.50	-0.014	0.293	0.292	94.9	-0.007	0.293	0.258	92.6
			γ	-0.50	0.006	0.094	0.096	95.2	0.096	0.074	0.074	74.8
			$\Lambda(\pi/4)$	0.75	0.021	0.178	0.177	94.2	-0.025	0.167	0.148	89.6
			$\Lambda(\pi/2)$	1.50	0.032	0.369	0.360	94.9	-0.251	0.274	0.245	67.8
$\alpha = 1$	0.5	300	σ_b^2	0.25	0.003	0.171	0.211	91.9	0.117	0.163	0.163	75.3
			σ_r^2	0.50	0.017	0.386	0.527	96.2	—	—	—	—
			β	0.50	-0.007	0.439	0.438	94.7	-0.008	0.436	0.422	94.5
			γ	-0.50	-0.012	0.155	0.148	94.7	0.020	0.140	0.135	93.6
			$\Lambda(\pi/4)$	0.75	0.040	0.280	0.276	93.6	0.045	0.276	0.264	92.6
			$\Lambda(\pi/2)$	1.50	0.087	0.569	0.554	93.2	0.015	0.527	0.498	90.9
		500	σ_b^2	0.25	-0.013	0.130	0.167	94.3	0.110	0.122	0.127	73.0
			σ_r^2	0.50	0.033	0.334	0.417	94.2	—	—	—	—
			β	0.50	-0.002	0.334	0.339	95.2	-0.003	0.333	0.327	95.3
			γ	-0.50	-0.005	0.112	0.114	95.4	0.026	0.104	0.105	94.3
			$\Lambda(\pi/4)$	0.75	0.022	0.204	0.207	94.3	0.019	0.202	0.199	93.6
			$\Lambda(\pi/2)$	1.50	0.049	0.406	0.415	94.5	-0.030	0.376	0.374	90.6
	1.0	300	σ_b^2	0.25	0.045	0.193	0.245	88.7	0.210	0.179	0.178	52.5

$H(x)$	σ_r^2	n	Par	True	Proposed Model				Misspecified Model			
					Bias	SD	SE	CP%	Bias	SD	SE	CP%
			σ_r^2	1.00	-0.095	0.563	0.734	97.2	—	—	—	—
			β	0.50	-0.007	0.474	0.471	95.1	-0.025	0.471	0.447	94.0
			γ	-0.50	-0.006	0.164	0.156	95.1	0.043	0.141	0.137	93.6
			$\Lambda(\pi/4)$	0.75	0.055	0.301	0.299	93.4	0.004	0.295	0.279	92.5
			$\Lambda(\pi/2)$	1.50	0.097	0.618	0.598	92.6	-0.109	0.549	0.511	88.5
		500	σ_b^2	0.25	0.027	0.153	0.190	89.6	0.206	0.131	0.137	42.8
			σ_r^2	1.00	-0.075	0.480	0.561	96.7	—	—	—	—
			β	0.50	0.000	0.359	0.365	95.6	0.002	0.362	0.346	94.1
			γ	-0.50	0.004	0.116	0.120	95.6	0.050	0.105	0.106	91.5
			$\Lambda(\pi/4)$	0.75	0.024	0.218	0.223	94.6	0.018	0.217	0.209	93.0
			$\Lambda(\pi/2)$	1.50	0.035	0.433	0.444	93.8	-0.083	0.390	0.381	88.6

Table 2

Simulation results under Case II scenario

$H(x)$	σ_r^2	σ_b^2	n	Par	True	Proposed Model					Misspecified Model				
						Bias	SD	SE	CP%		Bias	SD	SE	CP%	
$\alpha = 0$	0.5	σ_b^2	300		0.25	-0.001	0.106	0.105	91.8		0.075	0.079	0.075	71.6	
					0.50	-0.002	0.283	0.311	94.8		—	—	—	—	
					0.50	0.010	0.207	0.204	95.2		-0.027	0.191	0.183	94.8	
					-0.50	0.001	0.100	0.097	94.6		0.061	0.085	0.081	86.6	
					0.75	0.004	0.124	0.120	92.8		0.000	0.113	0.107	92.4	
	1.0	σ_b^2	500		1.50	0.002	0.260	0.249	93.2		-0.141	0.198	0.187	83.6	
					0.25	-0.001	0.083	0.084	94.0		0.075	0.060	0.058	63.0	
					0.50	0.011	0.238	0.257	94.7		—	—	—	—	
					0.50	-0.004	0.163	0.159	94.3		-0.049	0.149	0.142	91.8	
					-0.50	0.000	0.074	0.076	96.0		0.065	0.062	0.062	82.4	
		σ_b^2	300		0.75	0.004	0.095	0.094	95.1		0.005	0.086	0.084	93.8	
					1.50	0.012	0.198	0.198	96.0		-0.126	0.150	0.147	82.4	
					0.25	0.011	0.127	0.128	89.8		0.129	0.086	0.081	47.2	
					1.00	-0.062	0.430	0.452	95.6		—	—	—	—	
					0.50	0.013	0.227	0.229	95.8		-0.049	0.201	0.192	92.0	
		σ_b^2	500		-0.50	0.007	0.113	0.107	93.4		0.100	0.085	0.082	73.6	
					0.75	0.003	0.136	0.134	93.4		-0.006	0.117	0.112	91.6	
					1.50	-0.015	0.294	0.280	92.0		-0.231	0.197	0.184	68.8	
					0.25	0.004	0.102	0.104	92.0		0.129	0.065	0.063	28.6	

$H(x)$	σ_r^2	n	α	Par	True	Proposed Model					Misspecified Model				
						Bias	SD	SE	CP%		Bias	SD	SE	CP%	
				σ_r^2	1.00	-0.017	0.357	0.369	94.4		—	—	—	—	—
				β	0.50	-0.020	0.181	0.178	95.2		-0.076	0.157	0.149	90.8	
				γ	-0.50	0.006	0.082	0.083	95.2		0.105	0.062	0.063	62.8	
				$\Lambda(\pi/4)$	0.75	0.013	0.108	0.106	95.4		0.001	0.090	0.087	94.4	
				$\Lambda(\pi/2)$	1.50	0.027	0.236	0.224	94.4		-0.216	0.150	0.144	61.2	
				σ_b^2	0.25	-0.004	0.163	0.192	89.1		0.118	0.137	0.126	68.9	
		300			0.50	0.047	0.372	0.566	93.1		—	—	—	—	—
				σ_r^2	0.50	0.002	0.268	0.269	95.0		-0.027	0.254	0.252	94.5	
				β	-0.50	0.001	0.132	0.130	95.0		0.050	0.121	0.116	93.0	
				γ	0.75	0.002	0.165	0.159	93.5		0.001	0.157	0.151	93.4	
				$\Lambda(\pi/4)$	1.50	0.012	0.328	0.325	93.7		-0.140	0.294	0.286	90.9	
				$\Lambda(\pi/2)$	0.25	-0.008	0.124	0.149	93.9		0.120	0.102	0.097	58.1	
		500		σ_b^2	0.50	0.081	0.317	0.446	93.9		—	—	—	—	—
				σ_r^2	0.50	-0.015	0.209	0.208	94.8		-0.035	0.198	0.195	93.5	
				β	-0.50	-0.001	0.095	0.101	96.0		0.052	0.088	0.090	92.8	
				γ	0.75	0.011	0.125	0.125	94.1		0.021	0.120	0.118	94.5	
				$\Lambda(\pi/4)$	1.50	0.024	0.253	0.254	94.1		-0.023	0.228	0.224	92.5	
				$\Lambda(\pi/2)$	0.25	0.033	0.190	0.206	85.4		0.215	0.149	0.137	42.3	
		300		σ_b^2	1.00	-0.086	0.554	0.672	95.8		—	—	—	—	—
				σ_r^2	0.50	0.019	0.287	0.288	95.4		-0.029	0.274	0.261	94.2	
				β	-0.50	0.011	0.140	0.136	94.9		0.057	0.116	0.117	93.7	
				γ	0.75	0.009	0.176	0.171	93.7		0.027	0.168	0.159	92.0	
				$\Lambda(\pi/4)$	1.50	-0.007	0.347	0.342	92.4		-0.076	0.293	0.290	89.3	
				$\Lambda(\pi/2)$											

$H(x)$	σ_r^2	n	Par	True	Proposed Model					Misspecified Model				
					Bias	SD	SE	CP%		Bias	SD	SE	CP%	
	σ_b^2	500		0.25	0.021	0.145	0.164	88.1		0.216	0.110	0.106	26.5	
				1.00	-0.028	0.468	0.533	94.3		—	—	—	—	
	σ_r^2			0.50	-0.021	0.225	0.222	95.3		-0.052	0.209	0.202	92.9	
				-0.50	0.006	0.101	0.106	95.8		0.058	0.090	0.091	91.0	
			$\lambda(\pi/4)$	0.75	0.019	0.136	0.134	94.9		0.029	0.126	0.123	94.7	
				1.50	0.025	0.273	0.270	94.7		-0.066	0.229	0.225	90.0	

Table 3

Results of mean of square root integrated squared error ($\times 10^{-2}$) for estimated cumulative distribution functions with carrier and non-carrier groups

Case	$H(x)$	σ_r^2	n	Proposed Model		Misspecified Model		Ratio	
				Carrier	Non-Carr	Carrier	Non-Carr	Carrier	Non-Carr
I	$\alpha = 0$	0.5	300	6.472	13.276	6.736	14.715	1.041	1.109
			500	5.092	10.163	5.339	11.665	1.049	1.148
	1.0	300	6.583	14.124	7.215	17.378	1.096	1.231	
			500	5.142	10.619	5.734	13.822	1.115	1.301
	$\alpha = 1$	0.5	300	8.234	17.465	8.975	19.284	1.090	1.104
			500	6.587	13.126	7.495	14.561	1.138	1.109
II	1.0	300	8.361	17.592	9.862	20.181	1.180	1.147	
			500	6.653	13.138	8.641	15.510	1.299	1.180
	$\alpha = 0$	0.5	300	5.426	8.713	5.548	9.204	1.022	1.056
			500	4.302	6.668	4.416	7.192	1.026	1.079
	1.0	300	5.523	8.870	5.758	9.841	1.043	1.109	
			500	4.488	6.664	4.614	7.540	1.028	1.131
	$\alpha = 1$	0.5	300	6.876	10.980	7.703	11.710	1.120	1.067
			500	5.625	8.320	6.274	9.103	1.115	1.094
	1.0	300	6.856	10.985	8.736	12.220	1.274	1.113	
			500	5.618	8.451	7.325	9.792	1.304	1.159

Table 4

Estimated log hazard ratios and variance components from WHICAP combined probands and relatives data

Parameters	Estimate	Standard error	<i>p</i> -value
APOE- ϵ 4	0.930	0.193	< 0. 0001
gender	0.009	0.180	0.961
education	-0.687	0.091	< 0. 0001
σ_b^2	0.132	0.152	0.383
σ_r^2	0.224	0.197	0.254

Table 5
Estimated cumulative risk of dementia ($\times 10^{-2}$) from combined probands and relatives in WHICAP study

Age	Proband and relative data						Proband data only					
	Carrier			Non-Carrier			Carrier			Non-Carrier		
	Est.	SE	95% CI	Est.	SE	95% CI	Est.	SE	95% CI	Est.	SE	95% CI
70	0.32	0.14	(0.14, 0.75)	0.13	0.05	(0.05, 0.29)	0.22	0.13	(0.07, 0.70)	0.11	0.07	(0.04, 0.36)
75	1.47	0.37	(0.90, 2.39)	0.58	0.14	(0.36, 0.93)	1.13	0.33	(0.64, 1.98)	0.59	0.17	(0.34, 1.02)
80	6.54	1.13	(4.61, 9.24)	2.63	0.41	(1.94, 3.57)	6.86	1.11	(4.98, 9.41)	3.64	0.53	(2.72, 4.84)
85	15.93	2.37	(11.66, 21.57)	6.62	0.83	(5.15, 8.48)	16.07	2.24	(12.19, 21.03)	8.73	1.05	(6.89, 11.03)
90	32.31	4.20	(24.07, 42.48)	14.26	1.59	(11.33, 17.87)	32.57	4.12	(25.24, 41.37)	18.58	2.03	(14.96, 22.95)