Published in final edited form as:
Stat Med. 2008 November 20; 27(26): 5471–5483. doi:10.1002/sim.3344.

Genotype adjusted familial correlation analysis using three generalized estimating equations

Hye-Seung Lee[1], **Myunghee Cho Paik**[2,*], and **Joseph H. Lee**[3]

[1]Pediatrics Epidemiology Center, University of South Florida, 3650 Spectrum Blvd., Suit 100, Tampa, FL33612, U.S.A.

[2]Department of Biostatistics, Columbia University, 722 W. 168th Street, New York, NY 10032, U.S.A.

[3]Sergievsky Center, Columbia University, 630 W. 168 St, New York, NY 10032, U.S.A.

SUMMARY

We analyze familial correlation of a memory score from Caribbean Hispanic families that have multiple family members affected with Alzheimer's disease, adjusting for having at least 1 APOE-ε4 allele, as well as other confounders. To enhance the efficiency of correlation model, this paper proposes an alternative approach for three generalized estimating equations proposed by Yan and Fine [1]. The efficiency of correlation model is evaluated through the asymptotic relative efficiency computation and simulations.

1. INTRODUCTION

Researchers investigating the genetics of common diseases have been examining subcomponents of the clinical entity of their interest to better understand the underlying mechanism. This is possible because investigators routinely obtain information on many subcomponents to make accurate diagnosis. In our example, investigators conducting a genetic study of familial Alzheimer disease (AD) measure many cognitive functions, including memory scores, to help the diagnosis of AD. Given that AD is a multifactorial disease, there are likely to be multiple genetic and environmental factors involved in the disease process, and each factor by itself will have limited influence on the clinical endpoint. On the contrary, memory impairment precedes the onset of AD by several years and is a strong predictor of subsequent AD, which can be considered closer to the action of the gene influencing on AD; moreover, since it is measured continuously, a wide range of memory performance in unaffected as well as affected individuals can be used to measure the influence of genetic variants with weak to modest effects [2-5]. Thus, one can expect to enhance power to detect the relationship between genes and AD by examining memory scores, and there has been increasing attention to characterize genetic component in memory to better understand the genetics of AD [6-7].

Genetic studies of memory can start from investigating whether or not pairwise correlations for memory scores among family members (familial correlation) are greater than those for genetically unrelated people. Like other genetic analyses, the estimation of familial correlation is complicated by multiple non-genetic factors within each individual. For example, regardless

*Correspondence to: Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 W. 168th Street, New York, NY 10032, U.S.A.

of sharing genes, memory scores decline with age and are higher for individuals with higher levels of education, and greater differences in age or education between two family members would make lower familial correlations in a memory score. Moreover, apolipoprotein E $\varepsilon 4$ (APOE-$\varepsilon 4$) allele has been shown to be associated with AD consistently across different populations, which is only one genetic variant known to date [8-9]. Since our data are from AD families with multiple affected individuals, elevated familial correlations in memory could have resulted from having excess APOE-$\varepsilon 4$ alleles. Therefore, using a regression model, we want to analyze familial correlations of a memory score after controlling for the influence of APOE-$\varepsilon 4$ allele, as well as other non-genetic confounders.

For a regression model for correlations, Ziegler *et al.* [10] proposed a joint modelling of mean and correlation adopting the Generalized Estimating Equation (GEEs) for covariance parameters proposed by Prentice and Zhao [11], which extends the GEEs for mean parameters by Liang and Zeger [12]. Yan and Fine [1] added a variance model and proposed a joint modelling of mean, variance and correlation. However, as described in Reference [1], a misspecified correlation structure among 3GEEs can reduce the efficiency of the GEE estimators. In this paper, we first propose an alternative 3GEEs improving efficiency for the correlation model, and then compare its performance with Yan and Fine's 3GEEs. We then analyze familial correlations of total recall memory score to examine the genetic influence remained after controlling for the APOE gene along with other non-genetic confounders. Section 2 presents the efficient 3GEEs, specifically for the correlation model. In Section 3, we evaluate the efficiency of the proposed 3GEEs through the asymptotic relative efficiency (ARE) and the relative efficiency (RE) of Yan and Fine's 3GEEs with respect to the proposed 3GEEs. Section 4 shows the familial correlation analysis of a memory score from Hispanic AD families, adjusting for APOE-$\varepsilon 4$ allele, as well as age, education time and gender. A discussion concludes in Section 5.

2. Proposed 3GEEs

In this section, we propose an alternative 3GEEs by specifying the covariance structure among 3GEEs.

For the i^{th} unit, Y_i is the $n_i \times 1$ outcome vector $Y_i = (Y_{i1} \cdots Y_{in_i})$. Let X_{1i}, X_{2i} and X_{3i} be the $n_i \times p$ covariate matrix for the mean, $n_i \times q$ covariate matrix for the scale, and $\frac{n_i(n_i-1)}{2} \times r$ covariate matrix for the correlation of outcomes, respectively. Note that while X_{1i} and X_{2i} are determined for each individual, X_{3i} is determined for pairs in the i^{th} unit. The three models for the mean, scale and correlation of unit i are as follows:

$$g_1(\mu_i) = X_{1i}\beta$$
$$g_2(\phi_i) = X_{2i}\gamma$$
$$g_3(\rho_i) = X_{3i}\alpha$$

where $E(Y_{ij}|X_{1ij}) = \mu_{ij}$ and $Var(Y_{ij}|X_{2ij}) = \phi_{ij}v_{ij}$, $j = 1 \cdots n_i$. The ϕ_{ij} is the scale of outcome given X_{2ij}; the variance function v_{ij} is a function of μ_{ij}, which is also a function of X_{1ij}. The $Corr(Y_{ij}, Y_{ij'}|X_{3ijj'}) = \rho_{ijj'}$, that is, $\rho_{ijj'}$ is the correlation of outcomes given $X_{3ijj'}$ for $j \neq j' = 1 \cdots n_i$. The $g(\cdot)$ is the link function for each model. While $E(s_{ij}|X_{2ij}) = \phi_{ij}$ when $s_{ij} = \frac{(Y_{ij}-\mu_{ij})^2}{v_{ij}}$ for the j^{th} member, $E(z_{ijj'}|X_{3ijj'})$ is also asymptotically $\rho_{ijj'}$, when $z_{ijj'} = \frac{(Y_{ij}-\mu_{ij})(Y_{ij'}-\mu_{ij'})}{\sqrt{v_{ij}\phi_{ij}v_{ij'}\phi_{ij'}}}$ for a pair (j, j').

Following the framework by Reference [11], we generalize the GEE for the parameter $\theta = (\beta, \gamma, \alpha)$ as follows:

$$U(\theta) = \sum_{i=1}^{n} D_i' V_i^{-1} f_i = 0 \tag{1}$$

where $D_i = \begin{pmatrix} \frac{\partial \mu_i}{\partial \beta} & 0 & 0 \\ \frac{\partial \phi_i}{\partial \beta} & \frac{\partial \phi_i}{\partial \gamma} & 0 \\ \frac{\partial \rho_i}{\partial \beta} & \frac{\partial \rho_i}{\partial \gamma} & \frac{\partial \rho_i}{\partial \alpha} \end{pmatrix}$, $V_i = \begin{pmatrix} V_{11i} & V_{12i} & V_{13i} \\ V_{21i} & V_{22i} & V_{23i} \\ V_{31i} & V_{32i} & V_{33i} \end{pmatrix}$, and $f_i = \begin{pmatrix} Y_i - \mu_i \\ s_i - \phi_i \\ z_i - \rho_i \end{pmatrix}$. The $\frac{\partial \mu_i}{\partial \beta}$ is the $n_i \times p$ matrix whose component is $\frac{\partial \mu_{ij}}{\partial \beta_k}$, $j = 1, \cdots, n_i$, $k = 1, \cdots, p$, $\frac{\partial \phi_i}{\partial \gamma}$ is the $n_i \times q$ matrix whose component is $\frac{\partial \phi_{ij}}{\partial \gamma_k}$, $j = 1, \cdots, n_i$, $k = 1, \cdots, q$, and $\frac{\partial \rho_i}{\partial \alpha}$ is the $\frac{n_i(n_i-1)}{2} \times r$ matrix whose component is $\frac{\partial \rho_{ijj'}}{\partial \alpha_k}$, $j \neq j' = 1, \cdots, n_i$, $k = 1, \cdots, r$. The $n_i \times 1$ vector s_i has the j^{th} component, $s_{ij} = \frac{(Y_{ij} - \mu_{ij})^2}{v_{ij}}$, and the $\frac{n_i(n_i-1)}{2} \times 1$ vector z_i has the component for (j, j'), $z_{ijj'} = \frac{(Y_{ij} - \mu_{ij})(Y_{ij'} - \mu_{ij'})}{\sqrt{v_{ij}\phi_{ij}v_{ij'}\phi_{ij'}}}$.

As discussed in Reference [11], $\frac{\partial \phi_i}{\partial \beta}$, $\frac{\partial \rho_i}{\partial \beta}$ and $\frac{\partial \rho_i}{\partial \gamma}$ in the equation (1) are reasonable to be a zero matrix to make the parameter estimators robust to covariance misspecification. However, off diagonal blocks of V_i do not have to be set to zero since s_i and z_i are asymptotically unbiased estimators for ϕ_i and ρ_i, respectively. As a working covariance, the choice of V_i may not affect the consistent estimation of parameters, but the efficiency of estimators can be improved by choosing an appropriate form. An intuitive approach for the choice of V_i is to consider the normality of $Y_i = (Y_{i1}, \cdots, Y_{in_i})$. This assumption does not mean that the outcomes should be distributed as a multivariate normal for the analysis. Rather, this anticipates that the variance structure from a multivariate normal distribution may increase the efficiency by incorporating the covariance structure among Y_i, s_i and z_i. Appendix I.1 includes the covariance structure of (Y_i, s_i, z_i) from the multivariate normality.

Incorporating that V_{12i} and V_{13i} are zero matrices, a generalized 3GEEs can be written as follows:

$$U(\theta) = \sum_{i=1}^{n} \begin{pmatrix} U_{1i}(\theta) \\ U_{2i}(\theta) \\ U_{3i}(\theta) \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} \left(\frac{\partial \mu_i}{\partial \beta}\right)' V_{11i}^{-1}(Y_i - \mu_i) \\ \left(\frac{\partial \phi_i}{\partial \gamma}\right)' V_{22i}^{*}(s_i - \phi_i) + \left(\frac{\partial \phi_i}{\partial \gamma}\right)' V_{23i}^{*}(z_i - \rho_i) \\ \left(\frac{\partial \rho_i}{\partial \alpha}\right)' V_{32i}^{*}(s_i - \phi_i) + \left(\frac{\partial \rho_i}{\partial \alpha}\right)' V_{33i}^{*}(z_i - \rho_i) \end{pmatrix} = 0 \tag{2}$$

where $V_{22i}^{*} = \left(V_{22i} - V_{23i}V_{33i}^{-1}V_{32i}\right)^{-1}$, $V_{23i}^{*} = -\left(V_{22i} - V_{23i}V_{33i}^{-1}V_{32i}\right)^{-1}V_{23i}V_{33i}^{-1}$, $V_{32i}^{*} = V_{23i}'^{*}$ and $V_{33i}^{*} = V_{33i}^{-1}V_{32}\left(V_{22i} - V_{23i}V_{33i}^{-1}V_{32i}\right)^{-1}V_{23i}V_{33i}^{-1} + V_{33i}^{-1}$. While the 3GEEs by reference [1] is a special case of the equation (2) when V_{23i} is chosen as zero matrix, V_{22i} as the diagonal matrix with the diagonal elements of $Cov(s_i, s_i)$ and V_{33i} as the identity matrix, our proposed 3GEEs

implements $V_{23i}=\frac{Cov(s_i, z_i)}{2}$, $V_{22i} = Cov(s_i; s_i)$ and V_{33i} as a diagonal matrix with the diagonal elements of $Cov(z_i, z_i)$, where $Cov(s_i, s_i)$, $Cov(s_i, z_i)$ and $Cov(z_i, z_i)$ are from the normality as obtained in Appendix I.1. In our proposed 3GEEs, as opposed to Yan and Fine's, V_{23i} was chosen to suffice the condition $Var(\alpha) > 0$, and V_{33i} to improve the efficiency and simplify the estimation procedure. Note that the choice of V_{11i} does not affect the efficiency for correlation model.

The sandwich variance of the proposed 3GEEs is shown in Appendix I.2. The variance of α is a set of corresponding diagonal elements of $\left\{W_0^{-1}\right\} W_1 \left\{W_0^{-1}\right\}'$:

$$Var(\alpha) = \left(G - FC^{-1}D\right)^{-1} \left[FC^{-1}eC^{-1}D - hC^{-1}D - FC^{-1}f + i\right] \left(G - FC^{-1}D\right)^{-1}$$

where W_0, W_1, C, D, F, G, e, f, h and i are as defined in Appendix I.2. However, implementing sandwich variance can be a drawback when the number of samples is less than 50, especially with binary outcomes, as discussed elsewhere [14].

3. Efficiency of 3GEEs

In this section, we evaluate the efficiency of 3GEEs through the ARE and simulations.

3.1. Asymptotic relative efficiency

The ARE of Yan and Fine's 3GEEs with respect to the proposed 3GEEs is defined as the ratio of the set of diagonal elements corresponding to correlation parameter α of $\left\{W_0^{-1}\right\} W_1 \left\{W_0^{-1}\right\}'$ from the efficient 3GEEs to that from Yan and Fine's 3GEEs. Considering outcome vector from a bivariate normal distribution with zero means and unit variances, figure 1 displays the relationship between AREs and correlations distributing from 0 to 1, when true covariances among 3GEEs are from the given bivariate normal distribution. This shows that the efficiency gain is greater in stronger correlations. For example, when correlation is 0.5, the efficiency gain from using the proposed 3GEEs is greater than 50% when the outcome is from the given bivariate normal distribution.

3.2. Simulation study

Through simulations, we evaluate the efficiency of the proposed 3GEEs in the extended length of outcome vector and additional pair specific covariate adjustments to mimic our data.

Outcome vector was generated from a multivariate normal of length 6 with zero means and marginal variances of 1. The correlation matrix is given as follows: $\rho_1 = Corr(Y_{i3}, Y_{i6}) = Corr(Y_{i4}, Y_{i6}) = Corr(Y_{i5}, Y_{i6}) = 0$; $\rho_2 = Corr(Y_{i1}, Y_{i2}) = Corr(Y_{i1}, Y_{i3}) = Corr(Y_{i1}, Y_{i4}) = Corr(Y_{i2}, Y_{i3}) = 0.125$; $\rho_3 = Corr(Y_{i1}, Y_{i5}) = Corr(Y_{i1}, Y_{i6}) = Corr(Y_{i2}, Y_{i4}) = Corr(Y_{i2}, Y_{i5}) = 0.25$; $\rho_4 = Corr(Y_{i2}, Y_{i6}) = Corr(Y_{i3}, Y_{i4}) = Corr(Y_{i3}, Y_{i5} = Corr(Y_{i4}, Y_{i5}) = 0.5$. Sharing of genes identical by descent (i.e., genes from some common ancestor) is expected to be, on average, 0.5 for first-degree relative pairs, 0.25 for second-degree relative pairs, 0.125 for third-degree relative pairs and 0 for unrelated pairs. Hence, the data generation implies that we generate a family with 6 members, which includes 4 pairs each from the first to the third degree relatives and 3 unrelated pairs. We considered two types of outcomes: 1) continuous outcome from multivariate normal distribution with the above correlation matrix and 2) binary outcome indicating an underlying variable from normality less than 1. For 2), to have the above correlation matrix, we generated the underlying data from multivariate normal distribution with replacing the correlations 0, 0.125, 0.25 and 0.5 with 0.05, 0.26, 0.46 and 0.77, respectively. Moreover, to account for the

influence of shared and non-shared environment, we considered two types of additional pair specific covariates for the correlation model: 1) Continuous X_1 from N(0,1); 2) Dichotomous X_2 (0 or 1) with $P(X_2 = 1) = 0.3$. Four models to estimate each correlation parameter were investigated: Model1 unadjusted; Model2 adjusted for X_1; Model3 adjusted for X_2; and Model4 adjusted for X_1 and X_2. Since this paper focuses on estimating correlations for continuous outcomes, the following link functions were used: $g_1(\mu_i) = \mu_i$; $g_2(\phi_i) = \phi_i$; and $g_3(\rho_i) = tanh^{-1}(\rho_i)$. Intercept model was used for the mean and scale. In each model, we ran 500 times with 100 families.

We examined the RE of Yan and Fine's 3GEEs with respect to the proposed 3GEEs. The REs were obtained from the ratio of the simulation variance of estimates from the efficient 3GEEs to that from Yan and Fine's 3GEEs for given correlations. Table I presents the results. For continuous data from multivariate normal distribution, the proposed 3GEEs was efficient in almost all situations. The efficiency gain was greater for stronger correlations or in bigger models. We also notice that the efficiency gain can be greater in smaller dimension of outcomes. For example, for 0.5 correlation, the efficiency gain was 25% in unadjusted model1 for this 6 × 1 outcome vector, as opposed to 50% gain in ARE for 2 × 1. For binary data, there was no big difference between two different 3GEEs. In unadjusted Model1, Yan and Fine's 3GEEs performed a little better, but in bigger models, the proposed 3GEEs retained the advantage.

4. Familial correlation analysis using 3GEEs

In this section, for Hispanic AD families, we analyze familial correlations for sibling, second degree (e.g., grand parent-grandchild, half sibs and uncle-nephews), third degree (e.g., first cousins) and other types of relative pairs.

Detailed description of the neuropsychology battery was presented elsewhere [5-6]. Selective Reminding Test (SRT) to assess verbal memory and dementia was used to help the diagnosis of AD. The tests were performed on all family members in Spanish. In this test, subjects were administered six trials in which they were given a list of 12 unrelated words to memorize. After each attempt at recalling the list, the subject was reminded only of the words that were not recalled and then asked to recall the entire list. Total recall score was measured at this stage (max score=72, failure<25).

The study of familial AD in Caribbean Hispanics included 210 families and 1078 individuals with memory score data [13]. Each family had at least 2 AD patients. These families were collected from multiple sources, including clinics in the Dominican Republic and in Puerto Rico, the Alzheimer's Disease Research Center-Memory Disorders Center, and doctor's private offices in the Department of Neurology and the General Medical Services at Columbia University. In this sample, women were 66%. The mean age was 71 years (SD=13), and the mean time of education was 7 years (SD=6). The 40% of them had 1 APOE-ε4 and 10% had 2 APOE-ε4 alleles. As shown in Table II, the distribution of total recall score is skewed to the right, so the normality assumption may not be reasonable for the analysis. Unlike other likelihood approaches, since our model can be relaxed from the normality assumption, employing 3GEEs can take advantage to estimate familial correlations. The mean age difference in pairs was 12 years (SD=10), and the mean difference in years of education was 4 years (SD=4).

Table III includes simple estimates of familial correlation for each relative type, using Pearson's correlation. If genes were to influence memory, we expect familial correlations to be significantly different from zero, and we may further expect the correlations to be higher for genetically closer relatives compared with the correlations for more distant relatives. In overall estimations, correlations were not strong other than sibling pairs, but they increased

when the age or education difference was small. When stratified by whether or not they have APOE-ε4 allele identity by state (i.e., APOE-ε4 allele not necessarily from the same ancestor), correlations were similar for sibling and third degree relative pairs, but greater for second degree relative pairs in the group sharing APOE-ε4 alleles.

We implement the 3GEEs to estimate familial correlations adjusting for these subject and pair specific confounders. Both Yan and Fine's and proposed 3GEEs were applied. For the mean model, we adjusted for gender, age in years, years of education time and the number of copies of APOE-ε4 alleles. Intercept model was used for scale parameter, and three models were used for correlation parameters. Model1 examined familial correlations adjusting for subject specific confounders only. Model2 adjusted for differences in age, education and gender, in addition to subject specific confounders in the Model 1. Hence, the Model2 produces familial correlations if there were no differences in age, education level and gender in pairs. In addition to adjusting for subject and pair specific confounders in the Model2, Model3 adjusted for the indicator of having at least 1 APOE-ε4 allele in identity by state. We used the Model3 to investigate the residual genetic influence in memory remained after considering the effect of APOE-ε4 allele. Table IV presents the regression familial correlation analysis. From the mean model, we found that older age, lower level of education and presence of the APOE-ε4 allele significantly reduced total recall memory score. The estimates for scale parameters were almost identical for both 3GEEs. The proposed 3GEEs was efficient for correlation models, but for third degree relative pairs, Yan and Fine's 3GEEs produced smaller variance estimates. Efficiency gain for correlation parameters was greater when we included additional pair specific confounders in the model. However, the efficiency gain was not very striking, which it can be due to relatively small correlations or non-normality in this data. We indeed found stronger p-values in using efficient 3GEEs, but it could be mostly because the correlation estimates were greater than those from Yan and Fine's 3GEEs. In the Model1, familial correlation for sib pairs was significantly different from zero and decreased as relationships became more distant. When non-genetic differences in pairs were adjusted in the Model2, familial correlations for second degree relative pairs became stronger than ones for sib pairs and became significant. This is likely to be due to the fact that second degree relative pairs include grandparent-grandchild and uncle-nephews, where the differences in age or education level are greater than those for other types of relationships. In the Model3, familial correlations for sibling and second degree relative pairs remained significant and we found the same tendency for all relationships, although the effect sizes were somewhat smaller than ones from the Model2. Therefore, this regression familial correlation analysis suggests that additional genetic factors-independent of the APOE gene-are likely to be present for the memory measured by total recall score.

5. Discussion

In this paper, we proposed an alternative 3GEEs for correlation model from a generalization of 3GEEs. Both the proposed 3GEEs and Yan and Fine's 3GEEs are obtained by specifying certain types of working covariances in the generalized 3GEEs. The proposed model considered correlations among 3GEEs, and the choice of working correlations was motivated from the multivariate normality. However, the proposed 3GEEs do not require the normality of an outcome but just employ the properties from the normality. Through ARE and simulations, we showed that the efficiency gain from the proposed 3GEEs was greater for stronger correlations and models including additional pair specific covariates, but the efficiency gain was not impressive with binary outcomes.

Using both 3GEEs, we analyzed familial correlations adjusting for non-genetic differences, as well as the APOE gene in Hispanic AD families. Since correlation estimates were not strong in our data, using the proposed 3GEEs over Yan and Fine's 3GEEs did not seem to be

significantly beneficial. However, when the model includes additional pair specific confounders, our proposed model is recommended. In our analysis, the sandwich variance estimates were used. Under 3GEEs model, we observed that there is a support for presence of additional genetic factors, other than APOE gene, after controlling for subject and pair specific confounders.

Further studies are needed to yield unbiased estimates. The parameter estimates using the proposed 3GEEs are less likely to be comparable to population based parameter estimates, given that these families were chosen because they have multiple affected family members. Hence, an adjustment to the 3GEEs should be considered to produce an unbiased estimator. In addition, since familial correlation is one of the fundamental analytic tools in quantitative genetics, this regression approach can be utilize in various types of genetic analyses.

ACKNOWLEDGEMENTS

The first author is grateful to Dr. Jeffrey Krischer for his support. We thank the editor and reviewers for their careful reading and constructive suggestions.

Contract/grant sponsor: Funding for this study was provided by the National Institutes on Aging, National Institutes of Health R37 AG15473 and National Institute Of Neurological Disorders And Stroke R01 NS036928; contract/grant number: 98-1846389

Appendix

APPENDIX

I.1. Covariance structure among three GEEs from multivariate normality

Each component of V_i has the following correspondence: $V_{11i} = Cov(Y_i, Y_i)$, $V_{12i} = Cov(Y_i, s_i)$, $V_{13i} = Cov(Y_i, z_i)$, $V_{21i}=V'_{12i}$, $V_{22i} = Cov(s_i, s_i)$, $V_{23i} = Cov(s_i, z_i)$, $V_{31i}=V'_{13i}$, $V_{32i}=V'_{23i}$, and $V_{33i} = Cov(z_i, z_i)$. The factors for the calculation of each component of V_i are summarized by assuming $Y_{ij} \sim N(\mu_{ij}, \phi_{ij})$ and $Corr(Y_{ijj}, Y_{ijj'}) = \rho_{ijj'}$.

1. The $Cov(Y_i, s_i)$ is a zero matrix, since $Cov(Y_{ij}, s_{ij}) = E(Y_{ij} - \mu_{ij})^3 = 0$, and $Cov(Y_{ij}, s_{ij'} = E(Y_{ij} - \mu_{ij})(Y_{ij'} - \mu_{ij'})^2 = 0$, for $j \neq j' = 1, \cdots, n_i$.

2. The $Cov(Y_i, z_i)$ is also a zero matrix, since

$$Cov\left(Y_{ij}, z_{ijj'}\right)=\frac{1}{\sqrt{\phi_{ij}\phi_{ij'}}}E\left(Y_{ij}-\mu_{ij}\right)^2\left(Y_{ij'}-\mu_{ij'}\right)=0 \text{ and}$$

$$Cov\left(Y_{ij}, z_{ij'k}\right)=\frac{1}{\sqrt{\phi_{ij'}\phi_{ik}}}E\left(Y_{ij}-\mu_{ij}\right)\left(Y_{ij'}-\mu_{ij'}\right)(Y_{ik}-\mu_{ik})=0 \text{ for } j \neq j' \neq k = 1, \cdots, n_i.$$

3. For the $Cov(s_i, s_i)$ where $j \neq j' = 1, \cdots, n_i$, $Cov\left(s_{ij}, s_{ij}\right)=E\left(Y_{ij}-\mu_{ij}\right)^4-\phi_{ij}^2=2\phi_{ij}^2$, where $E\left(Y_{ij}-\mu_{ij}\right)^4=3\phi_{ij}^2$,
$Cov\left(s_{ij}, s_{ij'}\right)=E\left(Y_{ij}-\mu_{ij}\right)^2\left(Y_{ij'}-\mu_{ij'}\right)^2-\phi_{ij}\phi_{ij'}=2\rho_{ijj'}^2\phi_{ij}\phi_{ij'}$, where
$E\left(Y_{ij}-\mu_{ij}\right)^2\left(Y_{ij'}-\mu_{ij'}\right)^2=\phi_{ij}\phi_{ij'}\left(1+2\rho_{ijj'}^2\right)$.

4. For the $Cov(s_i, z_i)$ where $j \neq j' \neq k = 1, \cdots, n_i$,

$$Cov\left(s_{ij}, z_{ijj'}\right)=\frac{1}{\sqrt{\phi_{ij}\phi_{ij'}}}E\left(Y_{ij}-\mu_{ij}\right)^3\left(Y_{ij'}-\mu_{ij'}\right)-\phi_{ij}\rho_{ijj'}=2\rho_{ijj'}\phi_{ij}, \text{ where}$$

$$E\left(Y_{ij}-\mu_{ij}\right)^3\left(Y_{ij'}-\mu_{ij'}\right)=3\rho_{ijj'}\phi_{ij}^{1.5}\sqrt{\phi_{ij'}};$$

$$Cov\left(s_{ij}, z_{ij'k}\right)=\frac{1}{\sqrt{\phi_{ij'}\phi_{ik}}}E\left(Y_{ij}-\mu_{ij}\right)^2\left(Y_{ij'}-\mu_{ij'}\right)(Y_{ik}-\mu_{ik})-\phi_{ij}\rho_{ij'k}=2\rho_{ijj'}\rho_{ijk}\phi_{ij},$$

where $E\left(Y_{ij}-\mu_{ij}\right)^2\left(Y_{ij'}-\mu_{ij'}\right)(Y_{ik}-\mu_{ik})=\left(\rho_{ij'k}+2\rho_{ijj'}\rho_{ijk}\right)\phi_{ij}\sqrt{\phi_{ij'}\phi_{ik}}$.

5. For the $Cov(z_i, z_i)$ where $j \neq j' \neq k \neq l = 1, \cdots, n_i$,

$$Cov\left(z_{ijj'}, z_{ijj'}\right)=\frac{1}{\phi_{ij}\phi_{ij'}}E\left(Y_{ij}-\mu_{ij}\right)^2\left(Y_{ij'}-\mu_{ij'}\right)^2-\rho_{ijj'}^2=1+2\rho_{ijj'}^2;$$

$$Cov\left(z_{ijj'}, z_{ijk}\right)=\frac{1}{\phi_{ij}\sqrt{\phi_{ij'}\phi_{ik}}}E\left(Y_{ij}-\mu_{ij}\right)^2\left(Y_{ij'}-\mu_{ij'}\right)^2(Y_{ik}-\mu_{ik})-\rho_{ijj'}\rho_{ijk}=\rho_{ij'k}+\rho_{ijj'}\rho_{ijk}$$

;

$$Cov\left(z_{ijj'}, z_{ikl}\right)=\frac{1}{\sqrt{\phi_{ij}\phi_{ij'}\phi_{ik}\phi_{il}}}E\left(Y_{ij}-\mu_{ij}\right)\left(Y_{ij'}-\mu_{ij'}\right)(Y_{ik}-\mu_{ik})(Y_{il}-\mu_{il})-\rho_{ijj'}\rho_{ikl}=\rho_{ijk}\rho_{ij'l}+\rho_{ij'k}\rho_{ijl}$$

, where

$$E\left(Y_{ij}-\mu_{ij}\right)\left(Y_{ij'}-\mu_{ij'}\right)(Y_{ik}-\mu_{ik})(Y_{il}-\mu_{il})=\sqrt{\phi_{ij}\phi_{ij'}\phi_{ik}\phi_{il}}\left(\rho_{ijk}\rho_{ij'l}+\rho_{ij'k}\rho_{ijl}+\rho_{ijj'}\rho_{ikl}\right)$$

.

Hence, when the covariance structure from the normality assumption for $Y_i = (Y_{i1}, \cdots, Y_{in_i})$ is incorporated, the V_i can have the following form:

$$V_i=\begin{pmatrix} V_{11i} & 0 & 0 \\ 0 & V_{22i} & V_{23i} \\ 0 & V_{32i} & V_{33i} \end{pmatrix}$$

I.2. Sandwich Variance

The sandwich variance of the parameter estimates is formulated as $\left\{W_0^{-1}\right\}W_1\left\{W_0^{-1}\right\}'$ for the efficient 3GEEs. Note that the s_i for the scale model depends on the mean parameter β, while the z_i for the correlation model depends on the mean parameter β and scale parameter γ.

The W_0 is defined as follows:

$$W_0=-\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \frac{\partial U_{1i}}{\partial\beta} & \frac{\partial U_{1i}}{\partial\gamma} & \frac{\partial U_{1i}}{\partial\alpha} \\ \frac{\partial U_{2i}}{\partial\beta} & \frac{\partial U_{2i}}{\partial\gamma} & \frac{\partial U_{2i}}{\partial\alpha} \\ \frac{\partial U_{3i}}{\partial\beta} & \frac{\partial U_{3i}}{\partial\gamma} & \frac{\partial U_{3i}}{\partial\alpha} \end{pmatrix}$$

From the equation (2), the W_0 can be expressed as follows:

$$W_0=-\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} A & 0 & 0 \\ -B & C & D \\ -E & F & G \end{pmatrix}$$

where $A=\frac{\partial\mu_i}{\partial\beta}'V_{11i}^{-1}\frac{\partial\mu_i}{\partial\beta}$, $B=\frac{\partial\phi_i}{\partial\gamma}'V_{22i}^{*}\frac{\partial s_i}{\partial\beta}+\frac{\partial\phi_i}{\partial\gamma}'V_{23i}^{*}\frac{\partial z_i}{\partial\beta}$, $C=\frac{\partial\phi_i}{\partial\gamma}'V_{22i}^{*}\frac{\partial\phi_i}{\partial\gamma}-\frac{\partial\phi_i}{\partial\gamma}'V_{23i}^{*}\frac{\partial z_i}{\partial\gamma}$, $D=\frac{\partial\phi_i}{\partial\gamma}'V_{23i}^{*}\frac{\partial\rho_i}{\partial\alpha}$, $E=\frac{\partial\rho_i}{\partial\alpha}'V_{32i}^{*}\frac{\partial s_i}{\partial\beta}+\frac{\partial\rho_i}{\partial\alpha}'V_{33i}^{*}\frac{\partial z_i}{\partial\beta}$, $F=\frac{\partial\rho_i}{\partial\alpha}'V_{32i}^{*}\frac{\partial\phi_i}{\partial\gamma}-\frac{\partial\rho_i}{\partial\alpha}'V_{33i}^{*}\frac{\partial z_i}{\partial\gamma}$, and $G=\frac{\partial\rho_i}{\partial\alpha}'V_{33i}^{*}\frac{\partial\rho_i}{\partial\alpha}$. Also, the j^{th} component of $\frac{\partial s_i}{\partial\beta}$ is

$$\frac{\partial s_{ij}}{\partial \beta}=\frac{1}{v_{ij}^2}\left[2\left(y_{ij}-\mu_{ij}\right)\left(-\frac{\partial \mu_{ij}}{\partial \beta}\right)v_{ij}-\left(y_{ij}-\mu_{ij}\right)^2\left(\frac{\partial v_{ij}}{\partial \mu_{ij}}\right)\left(\frac{\partial \mu_{ij}}{\partial \beta}\right)\right],$$

the component for the pair (j, j') of $\frac{\partial z_i}{\partial \beta}$ is

$$\begin{aligned}\frac{\partial z_{ijj'}}{\partial \beta}= & \frac{1}{\sqrt{\phi_{ij}\phi_{ij'}v_{ij}v_{ij'}}}\left[-\frac{\partial \mu_{ij}}{\partial \beta}\left(y_{ij'}-\mu_{ij'}\right)-\frac{\partial \mu_{ij'}}{\partial \beta}\left(y_{ij}-\mu_{ij}\right)\right.\\ & \left.-\frac{\sqrt{s_{ij}s_{ij'}}}{2\sqrt{v_{ij}v_{ij'}}}\left(\frac{\partial v_{ij}}{\partial \mu_{ij}}\frac{\partial \mu_{ij}}{\partial \beta}v_{ij'}+\frac{\partial v_{ij'}}{\partial \mu_{ij'}}\frac{\partial \mu_{ij'}}{\partial \beta}v_{ij}\right)\right]\end{aligned}$$

and the component for the pair (j, j') of $\frac{\partial z_i}{\partial \gamma}$ is

$$\frac{\partial z_{ijj'}}{\partial \gamma}=-\frac{1}{2}\sqrt{\frac{s_{ij}s_{ij'}}{\phi_{ij}\phi_{ij'}}}\left(\phi_{ij}^{-1}\frac{\partial \phi_{ij}}{\partial \gamma}+\phi_{ij'}^{-1}\frac{\partial \phi_{ij'}}{\partial \gamma}\right)$$

The W_1 is defined as follows:

$$W_1=\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} Cov(U_{1i},U_{1i}) & Cov(U_{1i},U_{2i}) & Cov(U_{1i},U_{3i}) \\ Cov(U_{2i},U_{1i}) & Cov(U_{2i},U_{2i}) & Cov(U_{2i},U_{3i}) \\ Cov(U_{3i},U_{1i}) & Cov(U_{3i},U_{2i}) & Cov(U_{3i},U_{3i}) \end{pmatrix}$$

From this, the W_1 can be expressed as follows:

$$W_1=\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$

where $a=\left(\frac{\partial \mu_i}{\partial \beta}\right)' V_{11i}^{-1}Var(Y_i)\left\{V_{11i}^{-1}\right\}'\frac{\partial \mu_i}{\partial \beta}$,

$b=\left(\frac{\partial \mu_i}{\partial \beta}\right)' V_{11i}^{-1}Cov(Y_i,s_i)\left\{V_{22i}^{*}\right\}'\frac{\partial \phi_i}{\partial \gamma}+\left(\frac{\partial \mu_i}{\partial \beta}\right)' V_{11i}^{-1}Cov(Y_i,z_i)\left\{V_{32i}^{*}\right\}'\frac{\partial \phi_i}{\partial \gamma}$,

$c=\left(\frac{\partial \mu_i}{\partial \beta}\right)' V_{11i}^{-1}Cov(Y_i,s_i)\,V_{23i}^{*\prime}\frac{\partial \rho_i}{\partial \alpha}+\left(\frac{\partial \mu_i}{\partial \beta}\right)' V_{11i}^{-1}Cov(Y_i,z_i)\left\{V_{33i}^{*}\right\}'\frac{\partial \rho_i}{\partial \alpha}$, $d = b'$,

$e=\left(\frac{\partial \phi_i}{\partial \gamma}\right)' V_{22i}^{*}Var(s_i)\left\{V_{22i}^{*}\right\}'\frac{\partial \phi_i}{\partial \gamma}+2\left(\frac{\partial \phi_i}{\partial \gamma}\right)' V_{22i}^{*}Cov(s_i,z_i)\,V_{32i}^{*}\frac{\partial \phi_i}{\partial \gamma}+\left(\frac{\partial \phi_i}{\partial \gamma}\right)' V_{23i}^{*}Var(z_i)\,V_{32i}^{*}\frac{\partial \phi_i}{\partial \gamma}$,

$f=\left(\frac{\partial \phi_i}{\partial \gamma}\right)' V_{22i}^{*}Var(s_i)\,V_{23i}^{*}\frac{\partial \rho_i}{\partial \alpha}+\left(\frac{\partial \phi_i}{\partial \gamma}\right)' V_{22i}^{*}Cov(s_i,z_i)\left\{V_{33i}^{*}\right\}\frac{\partial \rho_i}{\partial \alpha}+\left(\frac{\partial \phi_i}{\partial \gamma}\right)' V_{23i}^{*}Cov(s_i,z_i)'V_{23i}^{*}\frac{\partial \rho_i}{\partial \alpha}$
$+\left(\frac{\partial \phi_i}{\partial \gamma}\right)' V_{23i}^{*}Var(z_i)\left\{V_{33i}^{*}\right\}'\frac{\partial \rho_i}{\partial \alpha}$,

$g=c'$, $h=f'$, and

$i=\left(\frac{\partial \rho_i}{\partial \alpha}\right)' V_{32i}^{*}Var(s_i)\,V_{23i}^{*}\frac{\partial \rho_i}{\partial \alpha}+2\left(\frac{\partial \rho_i}{\partial \alpha}\right)' V_{32i}^{*}Cov(s_i,z_i)\left\{V_{33i}^{*\prime}\right\}\frac{\partial \rho_i}{\partial \alpha}+\left(\frac{\partial \rho_i}{\partial \alpha}\right)' V_{33i}^{*}Var(z_i)\left\{V_{33i}^{*}\right\}'\frac{\partial \rho_i}{\partial \alpha}$.

All components for W_0 and W_1 are replaced by the estimates from 3GEEs. The estimators of $Var(a)$ and $Cov(a, b)$ are $\{a - E(a)\}\{a - E(a)\}'$ and $\{a - E(a)\}\{b - E(b)\}'$, respectively.

REFERENCES

1. Yan J, Fine J. Estimating equations for association structures. Statistics in medicine 2004;23:859–874. [PubMed: 15027075]
2. Masur DM, Sliwinski M, Lipton RB, Blau AD, Crystal HA. Neuropsychological prediction of dementia and the absence of dementia in healthy elderly persons. Neurology 1994;44:1427–1432. [PubMed: 8058143]
3. Schofield PW, Jacobs D, Marder K, Sano M, Stern Y. The validity of new memory complaints in the elderly. Arch Neurol 1997;54:756–759. [PubMed: 9193211]
4. Tierney MC, Yao C, Kiss A, McDowell I. Neuropsychological tests accurately predict incident Alzheimer disease after 5 and 10 years. Neurology 2005;64:1853–1859. [PubMed: 15955933]
5. Jacobs DM, Sano M, Dooneief G, Marder K, Bell KL, Stern Y. Neuropsychological detection and characterization of preclinical Alzheimer's disease. Neurology 1995;45:957–962. [PubMed: 7746414]
6. Lee JH, Flaquer A, Stern Y, Tycko B, Mayeux R. Genetic influences on memory performance in familial Alzheimer disease. Neurology 2004;62:414–421. [PubMed: 14872023]
7. McClearn GE, Johansson B, Berg S, Pedersen NL, Ahern F, Petrill SA, Plomin R. Substantial Genetic Influence on CognitiveAbilities in Twins 80 or More Years Old. Science 1997;276:1560–1563. [PubMed: 9171059]
8. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA. Gene Dose of Apolipoprotein E Type 4 Allele and the Risk of Alzheimer's Disease in Late Onset Families. Science 1993;261:921–923. [PubMed: 8346443]
9. Farre LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, Myers RH, Pericak-Vance MA, Risch N, van Duijn CM. Effects of Age, Sex, and Ethnicity on the Association Between Apolipoprotein E Genotype and Alzheimer Disease: A Meta-analysis. JAMA 1997;278:1349–1356. [PubMed: 9343467]
10. Ziegler A, Kastner C, Brunner D, Blettner M. Familial association of lipid profile: a generalized estimating equations approach. Statistics in Medicine 2000;19:3345–3357. [PubMed: 11122500]
11. Prentice RL, Zhao LP. Estimating equations for parameters in mean and covariances of multivariate discrete and continuous responses. Biometrics 1991;47:825–839. [PubMed: 1742441]
12. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.
13. Lee JH, Cheng R, Santana V, Williamson J, Lantigua R, Medrano M, Arriaga A, Stern Y, Tycko B, Rogaeva E, Wakutani Y, Kawarai T, St. George-Hyslop P, Mayeux R. Expanded genome-wide scan implicates a novel locus at 3q28 among Caribbean Hispanics with familial Alzheimers disease. Archives of Neurology 2006;63:1591–1598. [PubMed: 17101828]
14. Mancle LA, DeRouen TA. A covariance estimator for GEE with improved small-sample Properties. Biometrics 2001;57:126–134. [PubMed: 11252587]

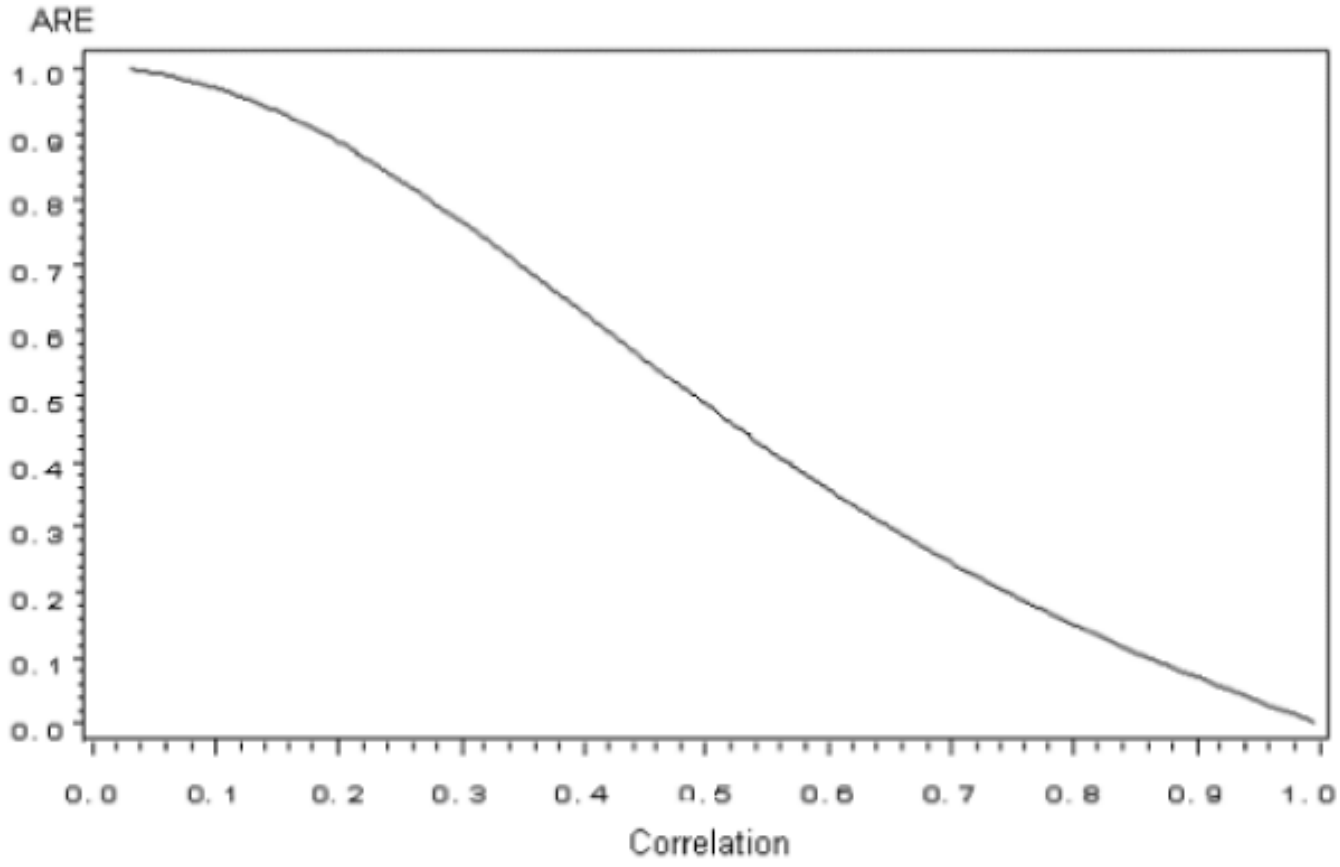

Figure 1.
Asymptotic relative efficiencies (AREs) for parameter estimates from correlation model of Yan and Fine's 3GEEs with respect to efficient 3GEEs

Table I

Relative Efficiencies (REs) for parameter estimates from Yan and Fine's 3GEEs with respect to efficient 3GEEs

Data	Correlation	Model1	Model2	Model3	Model4
Multivariate Normal	0	0.986	0.985	0.974	0.944
	0.125	1.005	0.995	1.003	0.968
	0.25	0.908	0.909	0.915	0.938
	0.5	0.778	0.756	0.778	0.747
	X_1	-	0.870	-	0.858
	X_2	-	-	0.910	0.881
Binary	0	1.017	1.004	1.017	0.991
	0.125	1.017	1.003	0.990	1.006
	0.25	0.966	0.973	0.953	0.967
	0.5	1.011	1.001	1.004	0.979
	X_1	-	0.920	-	0.920
	X_2	-	-	0.905	0.937

X_1 $N(0, 1)$; $X_2 = 0$ or 1 with $P(X_2 = 1) = 0.3$

Table II

Distribution of total recall score from Hispanic Alzheimer's disease families

	Mean	STD	MIN	Q1	MEDIAN	Q3	MAX
Total recall	22.275	17.680	0	0	23	36	66
age difference	12.126	9.939	0	4	9	18	52
education difference	4.274	4.037	0	1	3	6	23

Table III

Simple familial correlations of total recall score from Hispanic Alzheimer's disease families

Relationship		sibling	2^{nd}	3^{rd}	other
Pairs		1383	594	331	607
Overall		0.306	0.037	0.161	-0.156
age difference	<4	0.456	0.373	0.122	0.122
	4-9	0.293	0.405	0.349	0.304
	9-18	0.211	0.196	0.096	-0.138
	≥18	0.140	0.013	0.165	-0.182
education difference	< 1	0.342	0.320	0.506	0.093
	1-3	0.361	0.021	0.156	0.027
	3-6	0.238	0.062	0.239	-0.303
	≥6	0.204	-0.023	-0.067	-0.232
indicator of having at least 1 APOE-ε4 allele	0	0.310	-0.028	0.154	-0.109
	1	0.301	0.156	0.218	-0.308

2^{nd}: 2^{nd} degree relative pairs; 3^{rd}: 3^{rd} degree relative pairs; other: other types of relative pairs

Table IV

Results from 3GEEs for Total recall

			Efficient 3GEEs			Yan and Fine's 3GEEs		
			PE	**SE**	**p**	**PE**	**SE**	**p**
Mean		int	71.081	3.976	0.000	71.081	3.976	0.000
		male	-0.126	0.846	0.882	-0.126	0.846	0.882
		age	-0.716	0.046	0.000	-0.716	0.046	0.000
		educ	0.668	0.105	0.000	0.668	0.105	0.000
		apoe4	-2.984	0.733	0.000	-2.984	0.733	0.000
Scale		int	182.756	10.521	0.000	182.254	10.509	0.000
Corr	Model1	sibling	0.163	0.061	0.008	0.147	0.062	0.017
		2^{nd}	0.119	0.088	0.176	0.102	0.088	0.247
		3^{rd}	0.051	0.113	0.652	0.062	0.107	0.563
		other	-0.001	0.099	0.993	-0.001	0.099	0.988
	Model2	sibling	0.276	0.085	0.001	0.244	0.090	0.007
		2^{nd}	0.371	0.137	0.007	0.354	0.141	0.012
		3^{rd}	0.201	0.133	0.130	0.217	0.116	0.062
		other	0.247	0.132	0.061	0.211	0.136	0.119
		dage	-0.008	0.005	0.092	-0.005	0.004	0.206
		dedu	-0.029	0.012	0.017	-0.028	0.011	0.012
		dsex	0.077	0.064	0.232	0.059	0.066	0.371
	Model3	sibling	0.269	0.090	0.003	0.230	0.096	0.016
		2^{nd}	0.362	0.140	0.010	0.340	0.143	0.017
		3^{rd}	0.197	0.134	0.140	0.211	0.117	0.071
		other	0.241	0.134	0.072	0.203	0.137	0.139
		dage	-0.008	0.005	0.093	-0.005	0.004	0.207
		dedu	-0.028	0.012	0.017	-0.028	0.011	0.011
		dsex	0.076	0.064	0.234	0.059	0.066	0.373

	Efficient 3GEEs			Yan and Fine's 3GEEs		
	PE	**SE**	**p**	**PE**	**SE**	**p**
sapoe4	0.019	0.087	0.825	0.041	0.092	0.655

PE: parameter estimate; SE: standard error from the sandwich variance estimate; p: p-value; Corr: correlation model; Model1: adjusted for subject speciffic confounders only; Model2: additionally adjusted for age, education and gender difference; Model3: additionally adjusted for having at least 1 APOE-ϵ4 allele