JIRSS (2011)
Vol. 10, No. 2, pp 109-123

Marginal Analysis of A Population-Based Genetic Association Study of Quantitative Traits with Incomplete Longitudinal Data

Baojiang Chen[1], Zhijian Chen[2], Longyang Wu[3], Lihua Wang[4], Grace Y. Yi[3]

[1]Department of Biostatitics, University of Nebraska Medical Center, U.S.
[2]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada.
[3]Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada.
[4]Cancer Care Ontario, Toronto, Canada.

Abstract. A common study to investigate gene-environment interaction is designed to be longitudinal and population-based. Data arising from longitudinal association studies often contain missing responses. Naive analysis without taking missingness into account may produce invalid inference, especially when the missing data mechanism depends on the response process. To address this issue in the analysis concerning gene-environment interaction effects, in this paper, we adopt an inverse probability weighted generalized estimating equations (IPWGEE) approach to conduct statistical inference. This approach is attractive because it does not require full model specification yet it can provide consistent estimates under the missing at random (MAR) mechanism. We utilize this method to analyze data arising from a cardiovascular disease study.

Keywords. Generalized estimating equations, genetic association, longitudinal data, missing at random.

MSC: 62J12.

Baojiang Chen(baojiang.chen@unmc.edu), Zhijian Chen(zhijian@lunenfeld.ca), Longyang Wu(lwu@math.uwaterloo.ca), Lihua Wang(li.wang@cancercare.un.ca), Grace Y. Yi (✉)(yyi@uwaterloo.ca)
Received: February, 2011; Accepted: July, 2011

1 Introduction

Many complex human traits are outcomes of interplay of genetic and environmental factors, and individual genetic contributions to these traits and disorders are unlikely to be large. A recent review of several completed genome wide association studies indicates that nine confirmed loci for type 2 diabetes only account for 3% of the genetic variation, and 14 loci identified for Crohn diseases only account for less than 10% of the genetic variation (Estivill and Armengol, 2007). These loci are usually identified via large-scale case-control studies, exemplified by the Wellcome trust case control consortium (WTCCC, 2007), similar attempts to identify new associations likely yield similar results. Understanding the influences of environmental factors and the interactions of gene and environment ($G \times E$) would then become important to unravel the mechanism of complex human traits. Environmental factors usually refer to the total influences of non-genetic factors, including gender, physical, psychological, and cultural factors. It is well known that quantitative traits (QTs), such as fasting serum glucose, total plasma cholesterol and high density lipoprotein, are heavily influenced by non-genetic factors such as diet and physical exercise. Incorporation of $G \times E$ in either linkage or association studies of QTs would certainly be useful.

To investigate gene-environment interaction effects on binary trait of disease status, retrospective designs, including case-control and case-only designs are commonly adopted, and logistic regression is usually invoked for analysis (e.g., Kraft *et al.*, 2007). Large longitudinal population-based cohort studies, with extensive clinical information and ongoing follow-up, for example, the Framingham Heart Study (*www.framinghamheartstudy.org*) and the UK Biobank (*www.ukbiobank.ac.uk*), provide useful sources to study the effects of non-genetic factors and $G \times E$ interactions on QTs. A common feature of longitudinal studies is that phenotypes of interest are repeatedly measured for a subject over time, yielding correlated measurements. To accommodate various association structures, marginal methods based on generalized estimation equations (GEE) are widely used in practice (Liang and Zeger, 1986).

Missing observations occur frequently in longitudinal studies. Problems arise if the mechanism leading to the missing data is related to the response process. Little and Rubin (2002) and Laird (1988) presented a general treatment of statistical analysis of missing data mechanisms. A missing-data mechanism is called missing completely at random (MCAR) if the missing data process is independent of responses, and missing at random (MAR) if the missing data process does not

depend on unobserved responses. In contrast, data are called missing not at random (MNAR) or nonignorable if the missing data process is related to the unobserved responses.

Likelihood methods and marginal methods such as GEE are two powerful tools that have been developed to accommodate missing data for longitudinal data analysis (e.g., Chen, Yi and Cook, 2009; Chen, Yi and Cook, 2010a; Liang and Zeger, 1986; Chen, Yi and Cook, 2010b). Under MCAR or MAR, a valid analysis can be obtained based on available data when using a likelihood-based approach. Difficulties often associated with likelihood-based methods are that they require specification of the joint distributions of longitudinal responses. Inference based on GEE is attractive because it does not require full model specification for longitudinal response processes. Under the MCAR mechanism, the GEE approach yields consistent estimates for regression parameters. Robins, Rotnitzky and Zhao (1994, 1995) developed a class of estimators based on inverse probability weighted generalized estimating equations (IPWGEE) in a regression setting when incomplete data are MAR. This approach involves modeling the missing data process and weighting the estimating equations by the inverse of a probability that is calculated based on the models for the missing data process. If the models for both the marginal mean of the response and the missing data process are correctly formulated, the IPWGEE approach gives consistent estimates under the MAR mechanism.

Methods concerning missing data have mainly focused on monotone missing data patterns (e.g., Fitzmaurice, Molenberghs and Lipsitz, 1995; Fitzmaurice *et al.*, 2001; Yi, Cook and Chen, 2010), but relatively little work has been done for intermittently missing data with marginal methods. In this paper we explore a marginal method that handles longitudinal data with intermittently missing responses, and we apply this method to analyze data arising from a population-based genetic association study of quantitative traits.

The remainder of this paper is organized as follows. In Section 2, we describe a weighted estimating equation to address the missing data problem. In Section 3, we discuss modeling of the missing data process. Details on estimation and inference are given in Section 4. Data arising from a cardiovascular disease study are analyzed in Section 5. Section 6 includes the concluding remarks.

2 Weighted Estimating Equation

Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$ be a response vector of subject i observed at time points $t_1, t_2, \ldots, t_J$, and X_{ij} be the covariate vector recorded for subject i at the jth time point, $j = 1, \ldots, J$, $i = 1, \ldots, n$. Let $X_i = (X_{i1}', X_{i2}', \ldots, X_{iJ}')'$. Define $\mu_{ij} = E(Y_{ij}|X_i)$, and let $\mu_i = (\mu_{i1}, \ldots, \mu_{iJ})'$. Provided that the mean structure of Y_{ij} depends on the covariate vector for subject i at time j (e.g., Pepe and Anderson, 1994; Robins, Greenland and Hu, 1999), i.e., $E(Y_{ij}|X_i) = E(Y_{ij}|X_{ij})$, we consider generalized linear regression models

$$g(\mu_{ij}) = X_{ij}'\beta, \quad j = 1, \ldots, J,$$

where $g(\cdot)$ is a monotone differentiable link function, and β is a vector of regression parameters.

Here we only consider the missing response problem, and assume all the covariates are observed. Define the indicator random variable R_{ij}, which equals 1 if response Y_{ij} is observed and 0 if Y_{ij} is missing. Let $R_i = (R_{i1}, R_{i2}, \ldots, R_{iJ})'$, and $r_i = (r_{i1}, r_{i2}, \ldots, r_{iJ})'$ be a realization of R_i. For ease of exposition, we write $y_i = (y_i^{(o)}, y_i^{(m)})$, where $y_i^{(o)}$ and $y_i^{(m)}$ denote the observed data and missing data parts of y_i, a realization of Y_i. Note that arbitrary, nonmonotone patterns of missing data in y_i are considered.

With missing data, a common method of estimation is the naive data analysis estimate, $\hat{\beta}^*$, obtained as the solution to the generalized estimating equations (Liang and Zeger, 1986) for the available data, $U^*(\hat{\beta}^*) = 0$, where

$$U^*(\beta) = \sum_{i=1}^{n} D_i \Delta_i^* V_i^{-1} (Y_i - \mu_i), \tag{1}$$

$D_i = \partial \mu_i'/\partial\beta$, $\Delta_i^* = \text{diag}(r_{ij}\ , j = 1, \ldots, J)$, and V_i is the covariance matrix for the response Y_i. In actual implementation, a "working" covariance matrix is used to replace V_i, which is often decomposed as

$$V_i = a(\phi) A_i^{1/2} G_i(\rho) A_i^{1/2},$$

where $a(\cdot)$ is a known function, ϕ is a scaled parameter, A_i is a $J \times J$ diagonal matrix with elements $v_{ij} = Var(Y_{ij})$, $G_i(\rho)$ is a $J \times J$ "working" correlation matrix that is fully specified by a vector of parameters ρ. Note that only data with $r_{ij} = 1$ contribute to (1). When data are missing completely at random, (1) has expectation 0 when $\beta = \beta_{true}$.

However, when data are missing at random or missing not at random, (1) no longer has expectation 0 at $\beta = \beta_{true}$, hence, the resulting parameter estimates may be inconsistent.

Robins, Rotnitzky and Zhao (1995) replace r_{ij} in the naive analysis with r_{ij}/π_{ij}, where $\pi_{ij} = P(R_{ij} = 1|Y_i^{(o)}, X_i)$. Specifically, we solve

$$\sum_{i=1}^{n} U_i(\beta) = 0, \tag{2}$$

where $U_i(\beta) = D_i \Delta_i V_i^{-1}(Y_i - \mu_i)$, and $\Delta_i = \text{diag}(r_{ij}/\pi_{ij}\ , j = 1, \ldots, J)$. The estimating equation given by (2) is unbiased for 0 at the true β because

$$\begin{aligned} E[U_i(\beta)] &= E_{(R_i,Y_i,X_i)}[D_i \Delta_i V_i^{-1}(Y_i - \mu_i)] \\ &= E_{X_i} E_{(Y_i|X_i)} E_{(R_i|Y_i,X_i)}[D_i \Delta_i V_i^{-1}(Y_i - \mu_i)] \\ &= E_{X_i}[D_i V_i^{-1}\{E_{(Y_i|X_i)}(Y_i - \mu_i)\}] \\ &= 0, \end{aligned}$$

where a key component in the derivation above is that $E_{(R_i|Y_i,X_i)}[r_{ij}/\pi_{ij}] = 1$. An intuitive explanation of this method is that the weight has eliminated the bias, by "reconstructing" the full population by upweighting the data from subjects who have a small chance of being observed.

If π_{ij} is either known or consistently estimated, then a consistent estimate of β can be obtained as the solution to $\sum_{i=1}^{n} U_i(\beta) = 0$. However, in practice, π_{ij} is unknown and we often estimate it by modeling the missing data process, and this is discussed in the next section.

3 Modeling the Missing Data Process

In this paper, we consider the case that missing data follow a MAR mechanism satisfying

$$P(R_{ij} = 1|Y_i, X_i, H_{ij}^r) = P(R_{ij} = 1|Y_i^{(o)}, X_i, H_{ij}^r),$$

where $H_{ij}^r = \{r_{i1}, r_{i2}, \ldots, r_{i,j-1}\}$. Let $\lambda_{ij} = P(R_{ij} = 1|Y_i, X_i, H_{ij}^r)$ be the probability that the response is observed at the jth time point. In practice, a logistic regression model is commonly employed with

$$\text{logit}(\lambda_{ij}) = Z_{ij}'\alpha, \tag{3}$$

where Z_{ij} is a vector featuring various missingness, which may include a function of $\{H_{ij}^r, Y_i, X_i\}$ or their interactions, and α is the corresponding parameter vector. Different specifications of Z_{ij} may facilitate different missing data models. In particular, here we consider

$$\text{logit}(\lambda_{ij}) = \alpha_0 + \alpha_1 \cdot r_{i,j-1} + \alpha_2 \cdot r_{i,j-1} y_{i,j-1} + \alpha_3 \cdot r_{ij} y_{ij} + \alpha_x' \cdot X_{ij}.$$

To estimate α, one may employ the maximum likelihood method. That is, consider the log-likelihood for the parameter α

$$\ell(\alpha) = \sum_{i=1}^{n} \ell_i(\alpha) = \sum_{i=1}^{n} \sum_{j=2}^{J} \lambda_{ij}^{r_{ij}} (1 - \lambda_{ij})^{1 - r_{ij}}, \tag{4}$$

then maximizing (4) with respect to α yields the maximum likelihood estimate, say $\hat{\alpha}$, of α. Consequently, the marginal probability π_{ij} of missingness can be estimated by $\pi_{ij}(\hat{\alpha})$, where $\pi_{ij}(\alpha)$ is given by

$$\begin{aligned}
\pi_{ij}(\alpha) &= \sum_{r_{i1},\ldots,r_{i,j-1}} P(R_{i1} = r_{i1}, \ldots, R_{i,j-1} = r_{i,j-1}, R_{ij} = 1 | Y_i^{(o)}, X_i) \\
&= \sum_{r_{i1},\ldots,r_{i,j-1}} \Big\{ P(R_{ij} = 1 | H_{ij}^r, Y_i^{(o)}, X_i) \\
&\qquad \times \prod_{\ell=2}^{j-1} P(R_{i\ell} = r_{i\ell} | H_{i\ell}^r, Y_i^{(o)}, X_i) \cdot P(R_{i1} = r_{i1} | Y_i^{(o)}, X_i) \Big\} \\
&= \sum_{r_{i1},\ldots,r_{i,j-1}} \Big\{ \lambda_{ij} \cdot \prod_{\ell=1}^{j-1} (\lambda_{i\ell})^{r_{i\ell}} (1 - \lambda_{i\ell})^{1 - r_{i\ell}} \Big\},
\end{aligned}$$

where $\lambda_{i1} = P(R_{i1} = 1 | Y_i^{(o)}, X_i)$, and the dependence of $\lambda_{i\ell}$ on α is suppressed in the notation.

4 Estimation and Inference

Our primary interest lies in estimating the parameter β. Using the Fisher-scoring algorithm, we solve for β from (2) with π_{ij} replaced by $\pi_{ij}(\hat{\alpha})$. Let

$$M(\beta, \hat{\alpha}) = -\sum_{i=1}^{n} D_i V_i^{-1} \cdot \Delta_i(\hat{\alpha}) \cdot D_i'.$$

For an initial value $\beta = \beta^{(0)}$, update β by the iterative equations

$$\beta^{(t)} = \beta^{(t-1)} - [M(\beta^{(t-1)}, \hat{\alpha})]^{-1} \cdot \sum_{i=1}^{n} U_i(\beta^{(t-1)}, \hat{\alpha}), \quad , t = 1, 2, \ldots$$

until $\beta^{(t)}$ converges to $\hat{\beta}$, say.

We conclude this section with a discussion on the asymptotic distribution of the estimator $\hat{\beta}$. Let $U(\beta, \alpha) = n^{-1/2} \sum_{i=1}^{n} U_i(\beta, \alpha)$. When α is specified to be α_0, under standard regularity conditions for estimating functions, $U(\beta, \alpha_0)$ is asymptotically normal with mean 0 and covariance matrix $E(U_i(\beta, \alpha_0)U_i'(\beta, \alpha_0))$ and

$$n^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Gamma_0^{-1} E(U_i(\beta, \alpha_0)U_i'(\beta, \alpha_0))[\Gamma_0^{-1}]'), \quad n \to \infty,$$

where $\Gamma_0 = E(\partial U_i(\beta, \alpha_0)/\partial \beta')$. When α is unkown and estimated, the variation in the estimator $\hat{\alpha}$ must be taken into account. Under the regularity conditions stated by Robins, Rotnitzky and Zhao (1995), $U(\beta, \hat{\alpha})$ and $n^{1/2}(\hat{\beta} - \beta)$ are asymptotically normal with mean 0 and asymptotic variance Σ and $\Gamma^{-1}\Sigma[\Gamma^{-1}]'$, respectively, where

$$\begin{aligned} \Gamma &= E[\partial U_i(\beta, \alpha)/\partial \beta'], \\ \Sigma &= E[Q_i(\beta, \alpha)Q_i'(\beta, \alpha)], \\ Q_i(\beta, \alpha) &= U_i(\beta, \alpha) - E(\partial U_i(\beta, \alpha)/\partial \alpha') \cdot [E(\partial S_i(\alpha)/\partial \alpha')]^{-1} \cdot S_i(\alpha), \end{aligned}$$

and $S_i(\alpha) = \partial \ell_i(\alpha)/\partial \alpha'$.

Furthermore, this asymptotic covariance matrix can be consistently estimated by $\hat{\Gamma}^{-1}\hat{\Sigma}\hat{\Gamma}^{-1'}$ with

$$\begin{aligned} \hat{\Gamma} &= n^{-1} \sum_{i=1}^{n} \left\{ \frac{\partial U_i(\hat{\beta}, \hat{\alpha})}{\partial \beta'} \right\}, \\ \hat{\Sigma} &= n^{-1} \sum_{i=1}^{n} \hat{Q}_i(\hat{\beta}, \hat{\alpha}) \hat{Q}_i'(\hat{\beta}, \hat{\alpha}), \end{aligned}$$

where $\hat{Q}_i(\hat{\beta}, \hat{\alpha}) = U_i(\hat{\beta}, \hat{\alpha}) - \sum_{i=1}^{n} \partial U_i(\hat{\beta}, \hat{\alpha})/\partial \alpha' \cdot [\sum_{i=1}^{n} \partial S_i(\hat{\alpha})/\partial \alpha']^{-1} \cdot S_i(\hat{\alpha})$. Inference about β is conducted by replacing Σ and Γ with these consistent estimates in the expression of the asymptotic covariance matrix.

5 Application to a Cardiovascular Disease Study

Cardiovascular disease is a common cause of death in Canada. Levels of high-density lipoprotein (HDL) and low-density lipoprotein (LDL) are two important risk factors of cardiovascular disease; that is, low level of HDL cholesterol and high level of LDL cause the increasing of risk. In order to reduce the level of LDL cholesterol and increase the level of HDL cholesterol, dietary recommendations are made on the consumption of different types of fat. They include reducing the intake of saturated fatty acids (SFA) and increasing the intake of polyunsaturated fatty acids (PUFA). The presence of tumour necrosis factor-α (TNF-α) represented by a single nuclear polymorphism (SNP) whit two alleles of G and A, a chemical called inflammatory cytokines, may modify the concentrations of HDL.

Apolipoprotein-A1, denoted as apo-A1, is a major protein component of HDL in plasma. A longitudinal clinical trial was carried out to examine (1) the effect of randomly assigned dietary changes on HDL and apo-A1, and (2) whether the diet-cholesterol relationship differs among the polymorphisms associated with the gene for TNF-α (Fontaine-Bisson *et al.*, 2007; Wolever *et al.*, 2008). Patients with type 2 diabetes were assigned to one of three diets that differed in the types and amounts of carbohydrates and fat. Patients were followed on their dietary treatment for one year, and cholesterol measurements were taken at six time points: 0, 4, 12, 26, 39 and 52 weeks. The responses are HDL level and apo-A1; covariates include: Age, Gender, BMI, Weeks, Centre, Statins (1–the subject was on cholesterol lowering medication; 0–the subject was not on cholesterol lowering medication), PUFA intake (during trial), TNF-α-238 (dominant effect for A), TNF-α-308 (dominant effect for A). However, the collected measurements for the response HDL and apo-A1 are incomplete. Genotype data were available for 112 subjects, but complete data for HDL and apo-A1 were only available for 79 of them.

We first conducted some exploratory analyses for the two genetic loci. The results are reported in Table 1. The frequency of the minor allele A, defined as $f_A = 1/2 * P(GA) + P(AA)$, is estimated to be 14% at TNF-α-238 and 18% at TNF-α-308, where $P(GA)$ and $P(AA)$ are the population relative frequencies, i.e., marginal distributions, of the genotypes GA and AA. The Fisher's exact test shows no significant departure from Hardy-Weinberg equilibrium (HWE) at both TNF-α-238 and TNF-α-308 (p-values = 0.70 and 0.36, respectively), indicating that the sampling of subjects from the population was independent of genotypes. No linkage disequilibrium (LD) is found between TNF-α-238 and

TNF-α-308, suggesting that the two loci can be treated as independent. We also conducted analysis of variance and found no significant difference in baseline characteristics (age, BMI, HDL, apo-A1 and PUFA) among the three genotype groups at both TNF-α-238 and TNF-α-308.

Now we apply the method discussed in the preceding sections to analyze this data set with missingness accommodated. Let Y_{ij} denote the response (e.g., the HDL level), and X_{ij} be the covariate vector listed above. We are interested in modeling the mean of the response

$$\mu_{ij} = E(Y_{ij}|X_{ij}) = X'_{ij}\beta.$$

For the missing data models, we model $\lambda_{ij} = P(R_{ij} = 1|R_{i,j-1}, X_i, Y_i^{(o)})$ through a logistic regression

$$\text{logit}(\lambda_{ij}) = Z'_{ij}\alpha,$$

where Z_{ij} is a vector that may include the previous observed response, $r_{i,j-1}$, Age, BMI, Center, Statins, TNF-α-238, TNF-α-308, PUFA, interaction between TNF-α-238 and PUFA, interaction between TNF-α-308 and PUFA, and Week.

Fitting the missing data model, we see that, without reporting the results here, the coefficients of previous observed response, Gender, Age, BMI, Center, Statins, TNF-α-308, PUFA, Week, and the interaction between TNF-α-308 and PUFA are statistically significant, suggesting that a MAR mechanism may be reasonable, thus the IPWGEE method may be applicable.

Table 2 reports the results, where we report the naive analysis and the IPWGEE estimates. It is seen that both methods produce close results, which might particularly be attributed to the small missing proportion (about 10%). For the response of HDL, gender is statistically significant, indicating that male patients are easier to reduce the HDL level compared with female; BMI is significant in the naive analysis but moderately significant from the proposed method, indicating that patients with high BMI are more likely to reduce the HDL level; TNF-α-308 is significant, indicating that it has a positive effect on increasing HDL; the interaction of TNF-α-308 and PUFA is moderately significant, indicating that it has a moderately negative effect on increasing HDL level; for the time covariates, Week26 and Week39 are significant, indicating that the HDL level changes as the time changes; other covariates are not statistically significant. For the response of apo-A1, only the gender and Week39 are significant.

Table 1: Baseline characteristics and dietary intake by TNF-α genotypes

	TNF-α-238				TNF-α-308			
	GG	GA	AA	p-value	GG	GA	AA	p-value
# of men/ # of women	36/47	14/12	2/1	0.50	36/40	13/16	2/3	0.93
Age (y)	60.0	58.2	69.3	0.07	59.3	61.3	65.2	0.14
BMI (kg/m^2)	30.97	30.35	30.27	0.80	31.23	29.38	33.18	0.07
HDL ($mmol/L$)	1.18	1.18	0.97	0.39	1.16	1.23	1.23	0.39
apo-A1 (g/L)	1.57	1.58	1.36	0.36	1.54	1.61	1.65	0.33
PUFA (% of energy)	6.17	5.60	6.41	0.26	5.95	6.19	6.81	0.43

Except for the first row, the entries under "GG", "GA" and "AA" record the mean of the baseline measurements; the p-values are obtained from testing for their difference across genotype groups.

6 Discussion

In this paper, we discuss a marginal method for a population-based genetic association study of the quantitative trait with incomplete observations. Under the missing at random mechanism, the IPWGEE method can provide consistent estimators for the model parameters when the missing data model and the marginal model for the response are correctly specified. Applications to the cardiovascular disease demonstrate that TNF-α-308 interacts with PUFA intake to affect HDL level; the presence of allele A at TNF-α-238 has no significant effects on the relationship between PUFA intake and HDL (or apo-A1) level; in contrast, the presence of allele A at TNF-α-308 has a negative effect on the relationship between PUFA intake and HDL level. In this analysis here, we employ separate modeling for the HDL and apo-A1 variables. A more comprehensive strategy is to incorporate possible association between them in analysis, and an addition model for the association structure is typically required.

As is known, the IPWGEE method is sensitive to the misspecification of the missing data model. So, use of model diagnostics for the missing data process, perhaps most easily carried out in the MAR setting through model expansion, is warranted. The appealing feature of inverse weighting is that the models for the missing data processes can be made as elaborative as necessary by introducing a considerable amount of information on previous responses or covariates. Empirical evidence shows that there is often little price to pay for introducing additional

Table 2: Analysis results for the cardiovascular disease data

	HDL					
	Naive Analysis			IPWGEE		
Parameter	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	0.345	0.225	0.125	0.348	0.232	0.134
GenderM	-0.150	0.042	<.001	-0.161	0.044	<.001
Age	0.002	0.002	0.343	0.002	0.002	0.317
BMI	-0.009	0.004	0.037	-0.009	0.005	0.072
Center2	-0.040	0.057	0.479	-0.063	0.058	0.277
Center3	-0.002	0.052	0.972	-0.022	0.054	0.684
Center4	0.045	0.094	0.630	0.056	0.098	0.568
Center5	-0.019	0.048	0.696	-0.058	0.050	0.246
Statins	-0.044	0.041	0.279	-0.027	0.042	0.520
TNF-α-238	0.068	0.170	0.688	0.083	0.174	0.633
TNF-α-308	0.343	0.163	0.035	0.361	0.168	0.032
PUFA	0.005	0.015	0.759	0.005	0.015	0.739
TNF-α-238 : PUFA	-0.005	0.028	0.856	-0.007	0.028	0.803
TNF-α-308 : PUFA	-0.044	0.023	0.055	-0.044	0.024	0.067
Week4	0.015	0.012	0.202	0.015	0.013	0.249
Week12	0.008	0.012	0.525	0.008	0.013	0.538
Week26	0.029	0.014	0.038	0.030	0.015	0.046
Week39	0.039	0.012	0.002	0.038	0.014	0.007
Week52	0.019	0.014	0.155	0.019	0.015	0.205

Table 2-Continued.

	apo-A1					
	Naive Analysis			IPWGEE		
Parameter	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	0.490	0.148	0.001	0.501	0.150	0.031
GenderM	-0.121	0.030	<.001	-0.120	0.030	0.006
Age	0.002	0.002	0.320	0.002	0.002	0.317
BMI	-0.002	0.003	0.404	-0.003	0.003	0.549
Center2	-0.010	0.037	0.780	-0.019	0.037	0.743
Center3	0.048	0.034	0.156	0.045	0.035	0.405
Center4	-0.016	0.056	0.772	-0.009	0.057	0.927
Center5	-0.000	0.027	0.990	-0.003	0.028	0.952
Statins	-0.011	0.024	0.643	-0.013	0.025	0.757
TNF-α-238	-0.023	0.118	0.846	-0.027	0.119	0.877
TNF-α-308	0.165	0.103	0.108	0.162	0.105	0.335
PUFA	-0.004	0.011	0.677	-0.004	0.011	0.790
TNF-α-238 : PUFA	0.006	0.020	0.751	0.007	0.020	0.803
TNF-α-308 : PUFA	-0.022	0.015	0.147	-0.021	0.015	0.382
Week4	0.021	0.009	0.023	0.021	0.010	0.106
Week12	0.011	0.011	0.289	0.011	0.013	0.397
Week26	0.018	0.012	0.115	0.019	0.013	0.205
Week39	0.040	0.010	<.001	0.040	0.011	0.004
Week52	0.011	0.009	0.221	0.011	0.009	0.463

covariates into the missing data regression models.

The IPWGEE method is valid when data are missing at random. Actually, it is generally not possible to check formally for the presence of a MNAR mechanism from a MAR mechanism, so sensitivity analysis is required if this is a serious concern. Scharfstein, Rotnitzky and Robins (1999), Scharfstein and Irizarry (2003), Robins, Rotnitzky and Scharfstein (2000), and Robins and Rotnitzky (2001) each discuss strategies for conducting sensitivity analyses for marginal semiparametric methods for incomplete data. For other approaches, several authors have proposed the use of global and local influence tools to do sensitivity analyses in missing data contexts (e.g., Verbeke *et al.*, 2001; Molenberghs, Kenward and Goetghebeur, 2001).

Acknowledgements

The authors thank Dr. Thomas Wolever for providing the data. Yi's research was supported by the National Sciences and Engineering Research Council of Canada. Wu's research was supported by a Postgraduate Scholarship from Natural Sciences and Engineering Research Council of Canada.

References

Chen, B., Yi, G. Y., and Cook, R. J. (2009), Likelihood Analysis of Joint Marginal and Conditional Models for Longitudinal Categorical Data. Canadian Journal of Statistics, **37**, 182–205.

Chen, B., Yi, G. Y., and Cook, R. J. (2010a), Analysis of Interval-Censored Disease Progression Data via Multi-State Models under a Nonignorable Inspection Process. Statistics in Medicine, **29**, 1175–1189.

Chen, B., Yi, G. Y., and Cook, R. J. (2010b), Weighted Generalized Estimating Functions for Incomplete Longitudinal Response and Covariate Data that are Missing at Random. Journal of the American Statistical Association, **105**, 336–353.

Estivill, X. and Armengol, L. (2007), Copy Number Variants and Common Disorders: Filling the Gaps and Exploring Complexity in Genome-Wide Association Studies. PLoS Genet, **3**(10), e190.

Fitzmaurice, G. M., Molenberghs, G., and Lipsitz, S. R. (1995), Regression Models for Longitudinal Binary Data Responses with Informative Drop-outs. Journal of Royal Statistical Society, Series. B, **57**, 691–704.

Fitzmaurice, G. M., Lipsitz, S. R., Molenberghs, G., and Ibrahim, J. G. (2001), Bias in Estimating Association Parameters for Longitudinal Binary Responses with Drop-outs. Biometrics, **57**, 15–21.

Fontaine-Bisson, B., Wolever, T. M. S., Chiasson, J-L., Rabasa-Lhoret, R., Maheux, P., Josse, R. G., Leiter, L. A, Rodger, N. W., Ryan, E. A, Connelly, P. W., Corey, P. N., and El-Sohemy, A. (2007), Genetic Polymorphisms of Tumor Necrosis Factor-Modify the Association Between Dietary Polyunsaturated Fatty Acids and Fasting HDL-cholesterol and Apo A-1 Concentrations. The American Journal of Clinical Nutrition, **86**, 768–74.

Kraft, P., Yen, Y. C., Stram, D. O., and Morrison, J. (2007), Exploiting Gene-Environment Interaction to Detect Genetic Associations. Human Heredity, **63**, 111-119.

Laird, N. M. (1988), Missing Data in Longitudinal Studies. Statistics in Medicine, **7**, 305–315.

Liang, K. Y. and Zeger, S. L. (1986), Longitudinal Data Analysis Using Generalized Linear Models. Biometrika, **73**, 13–22.

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

Molenberghs, G., Kenward, M. G., and Goetghebeur, E. (2001), Sensitivity Analysis for Incomplete Contingency Tables. Applied Statistics, **50**, 15–29.

Pepe, M. S. and Anderson, G. L. (1994), A Cautionary Note on Inference for Marginal Regression Models with Longitudinal Data and General Correlated Response Data. Communications in Statistics, Simulation and Computation, **23**, 939–951.

Robins, J. M., Greenland, S., and Hu, F. C. (1999), Estimation of the Causal Effect of a Time-varying Exposure on the Marginal Mean of a Repeated Binary Outcome (with discussion), Journal of the American Statistical Association, **94**, 687–712.

Robins, J. M. and Rotnitzky, A. (2001), Comment on "Inference for Semiparametric Models: Some Questions and Answer," by P. J. Bickel and J. Kwon. Statistical Sinica, **11**, 920–936.

Robins, J. M., Rotnitzky, A. and Scharfstein, D. O. (2000), Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. In Statistical Models in Epidemiology, the Enviroment, and Clinical Trials, eds. E. M. Halloran and D. Berry, New York: Springer-Verlag, 1–94.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994), Estimation of Regression Coefficients When Some Regressor are not Always Observed. Journal of the American Statistical Association, **89**, 846–866.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995), Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing data. Journal of the American Statistical Association, **90**, 106–121.

Scharfstein, D. O. and Irizarry, R. A. (2003), Generalized Additive Selection Models for the Analysis of Studies With Potentially Nonignorable Missing Outcome Data. Biometrics, **59**, 601–613.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), Adjusting for Nonignorable Drop-out Using Semiparametric Nonresponse Models (with discussion). Journal of the American Statistical Association, **94**, 1096–1120.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., and Kenward, M. G. (2001), Sensitivity Analysis for Non-random Dropout: A Local Influence Approach. Biometrics, **57**, 43-50.

The Welcome Trust Case Control Consortium (2007), Genome-wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls. Nature, **447**, 661–678.

Wolever, T. M. S., Gibbs, A. L., Mehling, C., Chiasson, J-L, Connelly, P. W., Josse, R. G., Leiter, L. A., Maheux, P., Rabasa-Lhoret, R., Rodger, N. W., and Ryan, E. A. (2008), The Canadian Trial of Carbohydrates in Diabetes (CCD), a 1-y Controlled Trial of Low-glycemic-index Dietary Carbohydrate in Type 2 Diabetes: no Effect on Glycated Hemoglobin but Reduction in C-reactive Protein. The American Journal of Clinical Nutrition, **87**, 114–125.

Yi, G. Y., Cook, R. J., and Chen, B. (2010), Estimating Functions for Evaluating Treatment Effects in Cluster-randomized Longitudinal Studies in the Presence of Drop-out and Non-compliance. Canadian Journal of Statistics, **38**, 232–255.