Journal of Biopharmaceutical Statistics, 20: 334–350, 2010

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400903572787

Taylor & Francis
Taylor & Francis Group

HAPLOTYPE-BASED PHARMACOGENETIC ANALYSIS FOR LONGITUDINAL QUANTITATIVE TRAITS IN THE PRESENCE OF DROPOUT

Jung-Ying Tzeng[1,*], Wenbin Lu[1,*], Mark W. Farmen[2], Youfang Liu[3], and Patrick F. Sullivan[3]

[1]*Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA*
[2]*Department of Statistics, Eli Lilly and Company, Indianapolis, Indiana*
[3]*Department of Genetics, Psychiatry and Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA*

We propose a variety of methods based on the generalized estimation equations to address the issues encountered in haplotype-based pharmacogenetic analysis, including analysis of longitudinal data with outcome-dependent dropouts, and evaluation of the high-dimensional haplotype and haplotype–drug interaction effects in an overall manner. We use the inverse probability weights to handle the outcome-dependent dropouts under the missing-at-random assumption, and incorporate the weighted L_1 penalty to select important main and interaction effects with high dimensionality. The proposed methods are easy to implement, computationally efficient, and provide an optimal balance between false positives and false negatives in detecting genetic effects.

Key Words: Adaptive LASSO; Haplotype-based association analysis; Haplotype–treatment interaction; Missing data; Repeated outcome; Variable selection.

1. INTRODUCTION

Pharmacogenetics aims to understand the genetic differences among individuals in drug response. It shares a large amount of overlap with disease genetics, except that the trait of interest is drug response instead of disease predisposition. As in disease association studies, haplotype-based analysis provides an attractive option for understanding genetic effects on drug response. From a statistical perspective, haplotype-based analysis is asymptotically more powerful than single-marker analysis in detecting association of latent causal variants (Zaitlen et al., 2007). From a biological point of view, haplotypic polymorphisms are more informative for studying genetic association, as they preserve the joint linkage disequilibrium (LD) structure among multiple adjacent markers. Haplotypes also

Received June 30, 2008; Accepted September 20, 2009
*These authors contributed equally to this work.
Address correspondence to Jung-Ying Tzeng, Department of Statistics, North Carolina State University, Campus Box 7566, Raleigh, NC 27695, USA; E-mail: jytzeng@stat.ncsu.edu

incorporate the joint nonadditive effect of multimarkers, and therefore can better capture the combined effects of *cis*-acting causal alleles (Clark, 2004; Schaid, 2004).

However, typical association techniques may not be directly applicable to pharmacogenetic studies. First, instead of a typical cross-sectional case-control approach, pharmacogenetic study designs tend to use clinical trials, with repeatedly measured outcomes and important covariates. Second, missingness patterns tend to be more complex and nonrandom (e.g., patient dropout may depend on an outcome like drug response). Finally, research interests tend to focus more on gene–drug interactions than genetic main effects. For haplotype analysis, the challenges just described are further complicated by additional issues. The unknown phase creates missing values in covariates aside from the problem of response-dependent dropouts. Moreover, the high dimensionality of haplotype–drug interaction ($H \times D$) often causes unstable inference and decreased power. As a result, while there exist methods that allow for evaluating haplotype main and interaction effects in theory, the practical implementation is limited to certain prespecified haplotypes, and an overall exploration of $H \times D$ effects in an unprejudiced manner tends not to be applicable in reality.

In this work, we propose a variety of approaches that are based on inverse probability weighted (IPW) generalized estimation equations (GEE) to address these issues encountered in haplotype-based pharmacogenetic analysis. We adapt the GEE framework (Liang and Zeger, 1986) to bypass the full specification of the likelihood. We use the IPW estimation methods (Robins et al., 1995) to account for the response-dependent dropouts under the missing-at-random (MAR) assumption (Little and Rubin, 2002), which refers to the scenario that the dropout or non-dropout probability depends only on the past observed outcomes and covariates. For those who remain in a study at a particular time, the IPW methods weight each subject's contribution to the estimation equations at that time by the inverse of the non-dropout probability. Next, to facilitate the evaluation of $H \times D$ effects in an overall manner, we couple the IPWGEE framework with variable selection techniques. Specifically, we consider two commonly used penalizing approaches: LASSO (Tibshirani, 1996) and adaptive LASSO (Zhang and Lu, 2007; Zou, 2006). The former applies the equal-weight L_1 penalty to all variables (i.e., $\sum |\beta_k|$ where β_k is the regression coefficient). The latter inversely weights the variables by their consistent estimates (i.e., $\sum |\beta_k|/|\hat{\beta}_k|$ with $\hat{\beta}_k$ terms being the IPWGEE estimates without penalty), which makes unimportant variables receive larger penalties than important variables. These penalized approaches can simultaneously select important variables and estimate their effect sizes. We are particularly interested to learn their performance relative to the ordinal IPWGEE method, in which effect estimation and variable selection are done in two separate steps.

Alternatively, several likelihood-based approaches are also available for carrying out haplotype-based longitudinal data analysis, such as HAPSTAT (Lin and Zeng, 2006) and SimHap (Carter et al., 2008). These methods use mixed-effects models to study haplotype effect on the longitudinal outcomes, and can also handle the response-dependent dropouts under the MAR missingness (Jansen et al., 2006). Compared to the IPWGEE method, the likelihood-based approaches are efficient and do not need to model the dropout probability. However, they require specification of the joint distribution of the longitudinal outcome process, which can be difficult in practice and sensitive to model misspecification. In addition,

when incorporating penalty terms, the optimization of mixed-effects models can become computational demanding. In contrast, the IPWGEE method can be expressed as a weighted least-square problem, which makes its penalized extensions computationally convenient.

We focus this work on quantitative longitudinal traits measured during regular visits, which is the scenario encountered in our motivating study, the Clinical Antipsychotic Trails of Intervention Effectiveness (CATIE; Lieberman et al., 2005; Stroup et al., 2003). The CATIE study examined whether atypical antipsychotics can reduce morbidity and resource use compared to a conventional antipsychotic drug for patients suffering from chronic schizophrenia. Recently, the CATIE participants were also genotyped genome-wide for about 500 K single-nucleotide polymorphisms (SNP; Sullivan et al., 2008). The availability of the CATIE genetic and clinical data makes it possible to evaluate individual differences in treatment response. However, such evaluation is intricate, as only a proportion of patients respond to a specific antipsychotic, and nonresponse and dropouts are key indicators for individual differences in drug treatment. We aim to tackle these challenges with the methods constructed in the work.

This article is organized as follows. Section 2 describes the regression model for studying the haplotype effects on drug response and the proposed inference procedure based on the IPWGEE and its penalized variations. Section 3 examines the performance of these methods using simulation, and Section 4 showcases the proposed methods by applying them on the CATIE data. Finally, Section 5 concludes the work with summary and discussion.

2. MODEL AND ESTIMATION METHOD

2.1. Model

We consider a follow-up study conducted over a fixed interval. Assume for each subject i ($i = 1, \ldots, n$), a sequence of the outcome variables $Y_{i,t}$ are designed to be measured at visit time $t = 1, \ldots, T$. In practice, some patients may quit the study or may only miss some visits but resume at a later time. These two missing data patterns are referred to as monotone missingness (or dropouts) and nonmonotone (or intermittent) missingness, respectively. See Tsiatis and Davidian (2004) for a good review on analyzing longitudinal data with these two types of missingness. To fix the idea, we consider the monotone missingness. However, the method described here can also accommodate nonmonotone missingness.

For each visit time t, define $\delta_{i,t}$ an indicator that equals 1 if $Y_{i,t}$ is observed and 0 otherwise. Note that for monotone missingness, $\delta_{i,t} = 0$ implies $\delta_{i,s} = 0$ for $\forall s > t$. Let $\mathbf{Y}_i = (Y_{i,1}, \ldots, Y_{i,T_i})'$ denote the observed outcome vector of subject i, where $T_i \in [1, T]$ is the number of visits that subject i had. Also, for subject i, let D_i denote the treatment indicator vector; H_i the haplotype design vector; $Y_{i,0}$ the baseline outcome value (occurred at $t = 0$); and $\mathbf{Z}_{i,t}$ a vector of covariates for patients' characteristics and other environmental exposures that are measured at visit time t ($t = 0, \ldots, T_i$) and may be time dependent. Assume (h_{i1}, h_{i2}) is the haplotype pair for subject i; then $H_{i,h}$, the hth element of H_i, is set to be $I(h_{i1} = h) + I(h_{i2} = h)$ with $I(\cdot)$ an indicator function. This particular choice of coding represents an additive-effect model. We note that other types of coding

can be used to represent recessive effect (i.e., $H_{i,h} = I(h_{i1} = h) \times I(h_{i2} = h)$) or dominant effect (i.e., $H_{i,h} = I(h_{i1} = h) + I(h_{i2} = h) - I(h_{i1} = h) \times I(h_{i2} = h)$). Lastly, we denote a patient's information up to time t using $\overline{Y}_{i,t} = (Y_{i,0}, Y_{i,1}, \ldots, Y_{i,t})'$, $\overline{Z}_{i,t} = (Z_{i,0}, Z_{i,1}, \ldots, Z_{i,t})'$, and $\overline{\delta}_{i,t} = (\delta_{i,0}, \delta_{i,1}, \ldots, \delta_{i,t})'$ with $\delta_{i,0} \equiv 1$ (i.e., assuming no missing data at baseline).

To account for potential outcome-dependent dropouts, we assume that the probability of a subject dropping out at time t may depend on his past observed outcomes and covariates, and we posit a model for the non-dropout probability $P_{i,t}(\gamma)$ at time t as

$$\begin{aligned}&\Pr(\delta_{i,t} = 1 \mid \overline{\delta}_{i,t-1} = \mathbf{1}_t, D_i, \overline{Y}_{i,t-1}, \overline{Z}_{i,t-1}) \\ &\quad = \frac{\exp(\gamma_I + \gamma_D' D_i + \gamma_Y' g_1(\overline{Y}_{i,t-1}) + \gamma_Z' g_2(\overline{Z}_{i,t-1}))}{1 + \exp(\gamma_I + \gamma_D' D_i + \gamma_Y' g_1(\overline{Y}_{i,t-1}) + \gamma_Z' g_2(\overline{Z}_{i,t-1}))}\end{aligned} \tag{1}$$

where $\gamma = (\gamma_I, \gamma_D', \gamma_Y', \gamma_Z')'$ is the coefficients vector, $\mathbf{1}_t$ is a $t \times 1$ vector of 1, and g_1 and g_2 are pre-specified functions of past observed outcomes and covariates, respectively. For example, one may use $g_1(\overline{Y}_{i,t-1}) = Y_{i,t-1}$ and $g_2(\overline{Z}_{i,t-1}) = Z_{i,t-1}$ to represent an assumption that the dropout probability only depends on the most recent observed data.

For longitudinal outcomes, we assume a linear model that includes the main effects of drugs and haplotypes, and their interactions. That is, for $t = 1, \ldots, T$,

$$E(Y_{i,t} \mid Y_{i,0}, D_i, H_i, \mathbf{Z}_{i,t}) = \beta_I + \beta_Y Y_{i,0} + \beta_T' b(t) + \beta_Z' \mathbf{Z}_{i,t} + \beta_D' D_i + \beta_H' H_i + \beta_{H\times D}' D_i \otimes H_i \tag{2}$$

where $\otimes$ is the Kronecker product and $b(t)$ is a vector of functions of the time t. For example, we may choose $b(t) = t$ if the longitudinal outcomes $Y_{i,t}$ are linear in t, but the high-order polynomials or more flexible spline basis functions can be used for nonlinear trend of the longitudinal outcomes.

In reality, haplotype H_i is usually not available and only genotype G_i is observed. Therefore, the model posited on $E(Y_{i,t} \mid Y_{i,0}, D_i, H_i, \mathbf{Z}_{i,t})$ [i.e., Eq. (2)] cannot be used to construct the estimating equations, and instead we use $E(Y_{i,t} \mid Y_{i,0}, D_i, G_i, \mathbf{Z}_{i,t})$, the conditional expected trait values given the observed genotypes, for this purpose. Let $\widehat{H}_i = E(H_i \mid G_i)$, and as shown next, we see this genotype-conditioned trait expectation is a linear function of $\widehat{H}_i$ for quantitative traits:

$$\begin{aligned}&E(Y_{i,t} \mid Y_{i,0}, D_i, G_i, \mathbf{Z}_{i,t}) \\ &\quad = E\{E(Y_{i,t} \mid Y_{i,0}, D_i, G_i, H_i, \mathbf{Z}_{i,t}) \mid Y_{i,0}, D_i, G_i, \mathbf{Z}_{i,t}\} \\ &\quad = \beta_I + \beta_Y Y_{i,0} + \beta_T' b(t) + \beta_Z' \mathbf{Z}_{i,t} + \beta_D' D_i + \beta_H' \widehat{H}_i + \beta_{H\times D}' D_i \otimes \widehat{H}_i.\end{aligned} \tag{3}$$

The hth element of $\widehat{H}_i$ is $E(H_{i,h} \mid G_i)$ and is equal to $\sum_{(a,b)\in S(G_i)}[I(a = h) + I(b = h)] \times P((a, b) \mid G_i)$, where $S(G_i)$ denotes the set of haplotype pairs that are consistent with the observed genotype G_i, and $P((a, b) \mid G_i)$ is the conditional distribution of haplotype pair (a, b) given genotype G_i, which is equal to $\pi_a \pi_b / \sum_{(c,d)\in S(G_i)} \pi_c \pi_d$ with π_a the frequency of haplotype a.

2.2. IPWGEE

Define $\beta = (\beta_I, \beta_J')'$, where $\beta_J = (\beta_Y, \beta_T', \beta_Z', \beta_D', \beta_H', \beta_{H\times D}')'$ with length p. To estimate the parameters β in model (2), we propose to use the IPWGEE method (Robins et al., 1995). Specifically, we first need to obtain the estimated non-dropout probabilities $P_{i,t}(\hat{\gamma})$, where $\hat{\gamma}$ is an estimator of γ in model (1). The estimator $\hat{\gamma}$ can be obtained by fitting a logistic model for the observed non-dropout indicators $\delta_{i,t}$ on the past observed outcomes and covariates. Then based on Eq. (3) and $P_{i,t}(\hat{\gamma})$, we obtain the IPWGEE estimator $\hat{\beta} = (\hat{\beta}_I, \hat{\beta}_J')'$ of β by solving

$$\sum_{i=1}^{n} (\mathbf{1}_{T_i}, \mathbf{X}_i)' R_i V_i^{-1} (\mathbf{Y}_i - \beta_I \mathbf{1}_{T_i} - \mathbf{X}_i \beta_J) = 0 \tag{4}$$

where $\mathbf{Y}_i = (Y_{i,1}, \ldots, Y_{i,T_i})'$, $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,T_i})'$ is the $T_i \times p$ design matrix with $X_{i,t}' = (Y_{i,0}, b(t)', Z_{i,t}', D_i', \widehat{H}_i', (D_i \otimes \widehat{H}_i)')$ for $t = 1, \ldots, T_i$, $R_i = \text{diag}\{\frac{1}{P_{i,1}(\hat{\gamma})}, \ldots, \frac{1}{P_{i,T_i}(\hat{\gamma})}\}$, and V_i is a prespecified $T_i \times T_i$ weight matrix characterizing the conditional covariance of the traits $\mathbf{Y}_i$. One practical choice for V_i is the identity matrix, i.e., assuming the working independence. Such choice for V_i can give the consistent estimator of β, but note that a correct specification or a good estimate of V_i will improve the efficiency of the IPWGEE estimator (Liang and Zeger, 1986). Finally, solving Eq. (4) is equivalent to minimizing the following weighted least squares:

$$\sum_{i=1}^{n} (\mathbf{Y}_i - \beta_I \mathbf{1}_{T_i} - \mathbf{X}_i \beta_J)' R_i V_i^{-1} (\mathbf{Y}_i - \beta_I \mathbf{1}_{T_i} - \mathbf{X}_i \beta_J). \tag{5}$$

As a result, the minimization of Eq. (5) can be easily accomplished using standard software, such as *R*, for weighted least-squares regression.

Let β_0 denote the true values of β. As shown in Robins et al. (1995), if the non-dropout model (1) and the longitudinal model (2) are correctly specified, then the IPWGEE estimator $\hat{\beta}$ is consistent and $\sqrt{n}(\hat{\beta} - \beta_0)$ converges in distribution to a normal random vector with mean 0 and variance–covariance matrix Σ as $n \to \infty$. The variance–covariance matrix Σ involves complicated expressions since it needs to take into account of the variations in the estimation of $\hat{\gamma}$ and $\widehat{H}_i$. Here instead, to directly estimate Σ, we propose to use the bootstrap method (Davison and Hinkley, 1997) to estimate Σ. Then we can use Wald tests to select important covariates.

2.3. Penalized IPWGEE

To facilitate the evaluation of the main and interaction effects in an overall manner, we propose a penalized IPWGEE method that can simultaneously estimate the model parameters and select important variables. The penalization term in this method shrinks the coefficients of unimportant variables to exactly zero. To be specific, we consider the following penalized weighted least-squares estimation:

$$\sum_{i=1}^{n} (\mathbf{Y}_i - \beta_I \mathbf{1}_{T_i} - \mathbf{X}_i \beta_J)' R_i V_i^{-1} (\mathbf{Y}_i - \beta_I \mathbf{1}_{T_i} - \mathbf{X}_i \beta_J) + n\lambda \sum_{k=1}^{p} w_k |\beta_{J,k}|, \tag{6}$$

where $\beta_{J,k}$ is the kth element of β_J; the w_k terms are the weights that are preselected nonnegative constants and could be data dependent, and $\lambda > 0$ is the tuning parameter. When the w_k terms are set to 1, it becomes the LASSO penalty (Tibshirani, 1996); when $w_k = 1/|\hat{\beta}_{J,k}|$, the penalty term becomes the adaptive LASSO penalty (Zhang and Lu, 2007; Zou, 2006). Following the techniques of Zhang and Lu (2007), it can be shown that proposed adaptive LASSO IPWGEE estimator has the selection consistency property; i.e., as $\sqrt{n}\lambda \to 0$ and $n\lambda \to \infty$, the probability of estimating the nonzero coefficients as nonzero and zero coefficients as zero converges to 1. Moreover, the estimates of nonzero coefficients are consistent and asymptotically normal.

For each of the possible values of the tuning parameter λ, we minimize Eq. (6) to obtain the estimates for β and calculate the BIC for the corresponding model. The minimization of Eq. (6) can be achieved using standard LASSO packages, such as the shooting algorithm (Fu, 1998), the algorithm proposed by Osborne et al. (2000), and the *lars* algorithm (Efron et al., 2004). We use the *lars* algorithm in our numerical studies since it can give the whole solution path. We choose the model with the λ value that results in the smallest BIC (Bayesian Information Criterion), as the optimal λ chosen by the BIC criterion can identify the true model consistently (Wu et al., 2007). In other words, we choose λ to minimize $\mathrm{BIC}(\lambda) = \sum_{i=1}^{n}(\mathbf{Y}_i - \beta_I \mathbf{1}_{T_i} - \mathbf{X}_i\beta_J)' R_i V_i^{-1}(\mathbf{Y}_i - \beta_I \mathbf{1}_{T_i} - \mathbf{X}_i\beta_J) + \log n \cdot \mathrm{df}_\lambda$, where df_λ is the number of nonzero coefficients in $\hat{\beta}_J^{aLASSO}(\lambda)$, a simple estimate for the degree of freedom (Zou et al., 2007).

3. SIMULATION STUDY

3.1. Setup

We perform simulation studies to examine the performance of the proposed IPWGEE methods. We also carried out analysis using the likelihood-based method SimHap (Carter et al., 2008) as a benchmark using the "haplo.long" function in the R package "SimHap" provided by the authors. We simulate data akin to the CATIE study, including 500 individuals with outcomes measured at time points of 1, 3, 6, 9, 12, 15, and 18 months. We consider 5 drugs with equal probability of assignment, and a 3-SNP haplotype region forming 8 haplotypes: 000, 001, 010, 011, 100, 101, 110, and 111, with frequencies of 0.22, 0.09, 0.19, 0.09, 0.10, 0.11, 0.11, and 0.08. For each individual, we randomly sample a pair of haplotypes, assign one of the drugs, and generate the baseline response value from Normal(0, 1). We evaluate each method's ability to detect the causal effects under nine scenarios (Table 1), regarding whether the interacting treatment and haplotype also exhibit main effects, and whether the involved haplotype is of high or low frequency. Given the causal haplotypes and drugs for each scenario, we generate $Y_{i,t}$ based on model (2) by the following steps. First, we set $\mu_{i,t} = E(Y_{i,t} \mid Y_{i,0}, D_i, H_i, \mathbf{Z}_{i,t})$ as given in model (2) with $\beta_I = \beta_Y = 1$, $\beta_T = 0.1$, and $b(t) = t$. The drug effect (i.e., β_D) is set to 1 for the causal-effect drug and 0 for the rest. The same effect size (i.e., 1) is used for the causal haplotype and the causal $H \times D$, which leads to a heritability ranging from 0.07 to 0.20. Next, to create additional correlation among the outcomes values for subject i, we generate $Y_{i,t}$ from Normal($\mu_{i,t} + \alpha_i$, 1) where $\alpha_i \sim$ Normal(0, 0.5). To simulate the dropout process, we assume that the dropout

Table 1 List of causal effects of drug, haplotype, and haplotype–drug interactions considered in the simulation

Scenario	Causal drug	Causal haplotype	Interactions	Heritability
NULL: no effect	NA	NA	NA	0
A: The haplotype in the interactions has no main effect				
A1	Drug 3	010 (*0.19*)	Drug 3 × 101 (*0.11*)	0.14
A2	Drug 3	111 (*0.08*)	Drug 3 × 101 (*0.11*)	0.07
B: The treatment in the interactions has no main effect				
B1	Drug 2	010 (*0.19*)	Drug 3 × 010 (*0.19*)	0.20
B2	Drug 2	111 (*0.08*)	Drug 3 × 111 (*0.08*)	0.10
C: Both the haplotype and treatment in the interactions have main effects				
C1	Drug 3	010 (*0.19*)	Drug 3 × 010 (*0.19*)	0.13
C2	Drug 3	111 (*0.08*)	Drug 3 × 111 (*0.08*)	0.10
D: Both the haplotype and treatment in the interactions have no main effects				
D1	Drug 2	010 (*0.19*)	Drug 3 × 101 (*0.11*)	0.14
D2	Drug 2	111 (*0.08*)	Drug 3 × 101 (*0.11*)	0.08

Note. Values in parentheses indicate the haplotype frequencies.

status depends on drug 2 and the previous outcome values. Here we only consider the monotone missingness, and generate $\delta_{i,t}$ from the binomial distribution with the success probability specified in model (1), where $\gamma_I = 3$, $\gamma'_D = (1, 0, 0, 0)$, $\gamma_Y = -1$, $g_1(\overline{Y}_{i,t-1}) = Y_{i,t-1}$, and $\gamma_Z = 0$. The γ terms were set to obtain similar dropout rates observed in CATIE, and the resulting simulated non-dropout probabilities are about 0.86, 0.74, 0.78, 0.76, 0.75, 0.75, and 0.72 at visits 1 to 7, respectively.

3.2. Results

In the simulation analyses, we use only unphased genotypes and implement the IPWGEE methods under a working independence assumption (i.e., set $V_i = I_{T_i \times T_i}$). To investigate the potential power loss that is attributed to the working independence assumption, we repeat our analysis using the true variance–covariance structure for V_i. We also use the true covariance/correlation structure in SimHap analysis. For the scenario "ALL NULL," we run 5000 replications in the IPWGEE analysis and 1000 replications in the SimHap analysis. We run 1000 replications for scenarios A to D in all analyses. We summarize the results by reporting the frequencies of each variable being identified as significant. For the ordinary IPWGEE method (referred to as oIPWGEE), the significance of a variable is determined by the Wald test of 5% level based on the asymptotic normal distribution of the β estimates, with the variances of the β estimates obtained from 100 bootstrap samples. For the IPWGEE combined with the LASSO (referred to as LASSO) and adaptive LASSO (referred to as aLASSO), the significance is determined by whether the regression coefficient is estimated as exactly zero: nonzero means significant and zero otherwise. Finally, for the SimHap method (referred to as SimHap), the significance of a variable is determined at the 5% level using the p values obtained from 100 simulations. The simulations are conducted to account for uncertainty in haplotype assignment when phase is unknown.

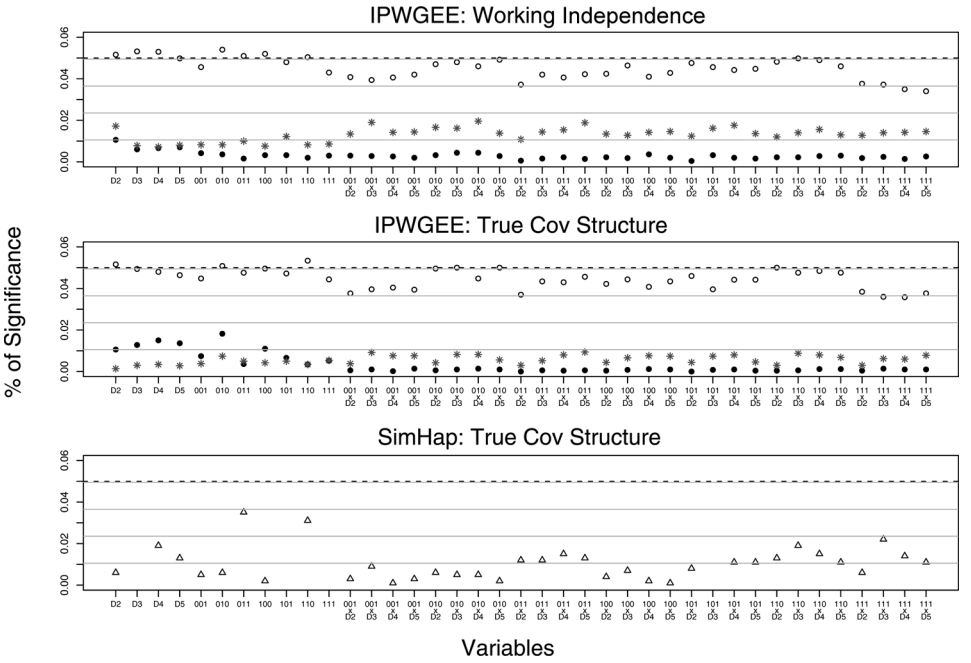

Figure 1 Proportion of significance under the Scenario of ALL NULL (i.e., type I error rates). The top two panels are the results of the proposed IPWGEE methods (i.e., oIPWGEE, LASSO, and aLASSO) with different covariance structures. The open circles, filled circles, and the star signs indicate the results for the oIPWGEE, aLASSO, and LASSO, respectively. The bottom panel is the results of the SimHap, where the type I error rates are indicated by triangles. In all panels, the horizontal dashed line indicates the nominal level of 0.05 used in the oIPWGEE and SimHap methods.

Figure 1 shows the results from the scenario of ALL NULL, where the proportion of significance corresponds to the type I error rates. The top panel is for the IPWGEE methods when a working independence covariance is used. The type I error rates for oIPWGEE are around the nominal level 0.05. It is observed that the type I error rates for the interaction terms are a little conservative, which agrees with our expectation since there is less information for the gene–drug interaction effects than for their main effects. The type I error rates for LASSO and aLASSO are more conservative since these two methods are not test-based and shrink coefficients of unimportant variables to exact zeros. Indeed, the error rates will be closer to zero as the sample size increases, especially for aLASSO. This is because the aLASSO estimates have the variable selection consistency property (Zhang and Lu, 2007; Zou, 2006)—i.e., when the sample size increases, the procedure will estimate the zero coefficients as exact zero with probability converging to 1. With a sample of size 500, the range of the type I error rates across all variables for LASSO is 0.0072 to 0.0198, with mean equal to 0.0132, and the range for aLASSO is 0.0004 to 0.0106, with mean error rate equal to 0.0030. The middle panel shows the results of the IPWGEE methods when the true covariance is used. The type I error rates are compatible to those in the top panel. Finally, the bottom panel shows the type I error rates of the SimHap method. We see that the type I error rates are conservative, which is somewhat expected as the dimension of the parameter space is big.

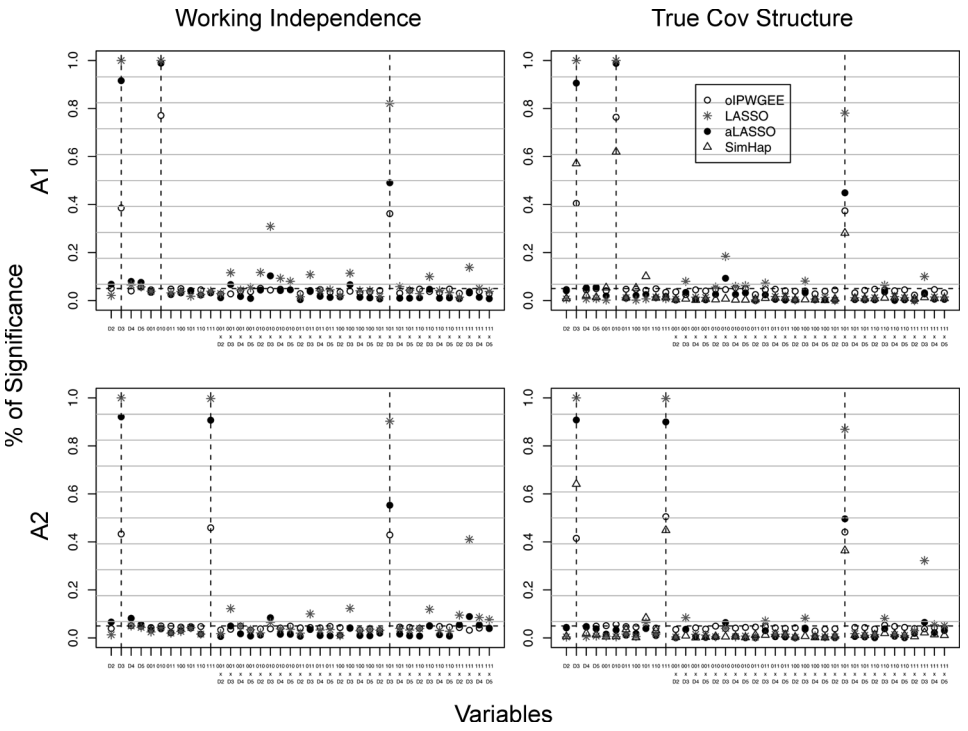

Figure 2 Proportion of significance under scenario A as defined in Table 1. The left panel shows the results of the proposed IPWGEE methods (i.e., oIPWGEE, LASSO, aLASSO) when a working independence covariance is used, and the right panel shows the results of the IPWGEE and SimHap analysis when the true covariance structure is used. The horizontal dashed line indicates the nominal level of 0.05 used for the Wald test in oIPWGEE and SimHap. The vertical dashed lines indicate the causal effects.

Figures 2–5 show the power (true positive) and type I error (false positive) rates for scenarios A to D. The left panels present the results of the IPWGEE methods (i.e., oIPWGEE, LASSO, and aLASSO) with a working independence covariance structure. Across all these scenarios, the oIPWGEE method offers the lowest power in detecting an effect. The LASSO method exhibits the greatest power but at a cost of high false positive rates. This is not too surprising because the LASSO method does not have the variable selection consistency property and it tends to select more variables than necessary (Zou, 2006). The simulation result suggests that all false positive detections tend to involve an interaction term that contains at least one of the main effects. On the other hand, the aLASSO method has power nearly as high as the LASSO but it does not have as many false positives. The aLASSO method achieves a better balance between true and false positives than the oIPWGEE and the LASSO, which is again a result from the nice theoretical properties of the adaptive LASSO method.

The right panels of Figs. 2–5 show the results of the IPWGEE methods and SimHap when the true covariance structure is used in the analysis. Focusing on the IPWGEE methods (oIPWGEE, LASSO and aLASSO), we note that the left and

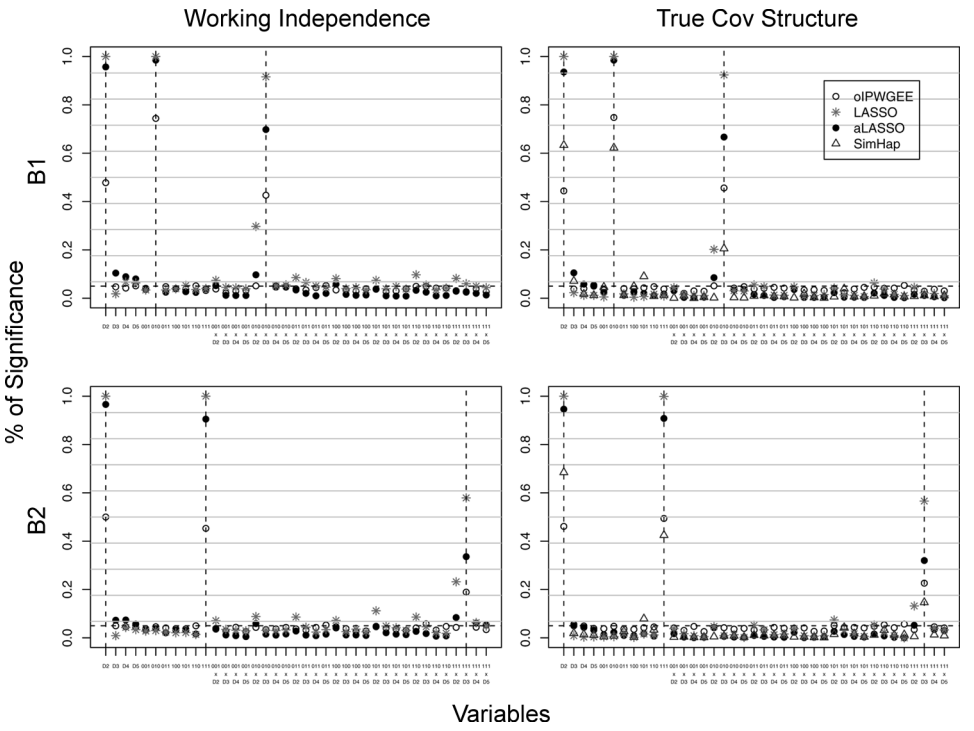

Figure 3 Proportion of significance under scenario B as defined in Table 1. The left panel shows the results of the proposed IPWGEE methods (i.e., oIPWGEE, LASSO, aLASSO) when a working independence covariance is used, and the right panel shows the results of the IPWGEE and SimHap analysis when the true covariance structure is used. The horizontal dashed line indicates the nominal level of 0.05 used for the Wald test in oIPWGEE and SimHap. The vertical dashed lines indicate the causal effects.

right panels provide similar results, indicating that there is only marginal power loss when the independence covariance matrix is used instead of the true covariance structure. For the comparisons between the proposed IPWGEE methods with the benchmark SimHap method, we see that compared to oIPWGEE, the SimHap method gives higher power for detecting drug main effects, but lower power for haplotype main effects and $H \times D$ interactions. On the other hand, other proposed methods (i.e., LASSO and aLASSO) result in much higher power than the SimHap method. A similar trend holds for all scenarios of A, B, C, and D.

4. APPLICATION TO CATIE DATA

We apply the proposed IPWGEE methods to the CATIE study of schizophrenia, and perform a search on chr22 for potential genetic variants related to drug response. We choose chr22 because it contains the most published candidate genes (i.e., 91 genes) for schizophrenia etiology according to the Schizophrenia Research Forum (http://www.schizophreniaforum.org/res/sczgene/dbindex.asp) as of June 2009.

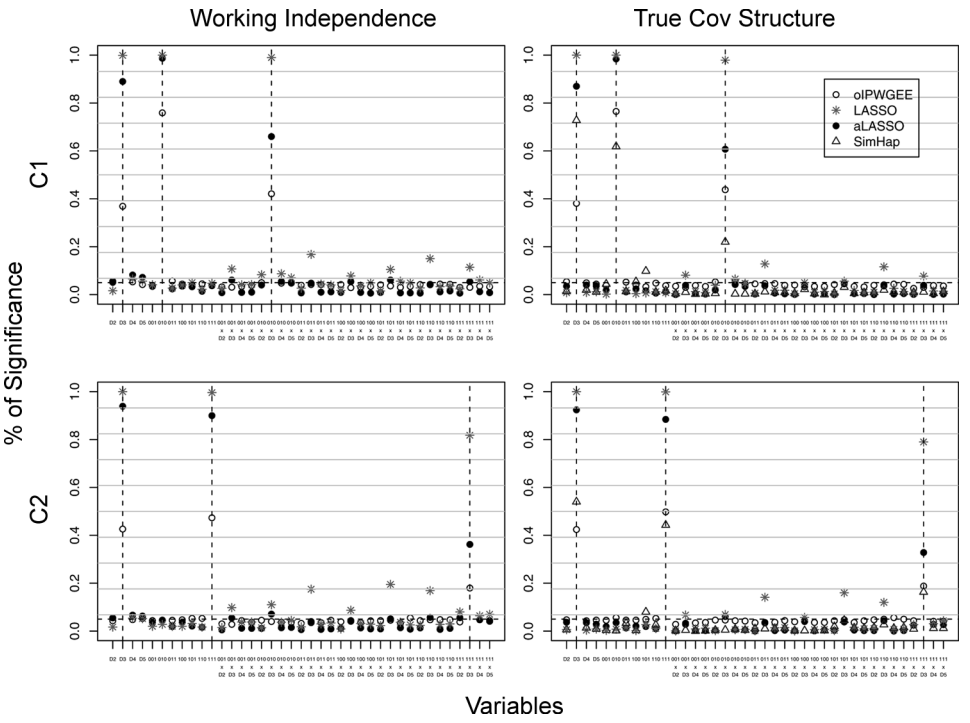

Figure 4 Proportion of significance under scenario C as defined in Table 1. The left panel shows the results of the proposed IPWGEE methods (i.e., oIPWGEE, LASSO, aLASSO) when a working independence covariance is used, and the right panel shows the results of the IPWGEE and SimHap analysis when the true covariance structure is used. The horizontal dashed line indicates the nominal level of 0.05 used for the Wald test in oIPWGEE and SimHap. The vertical dashed lines indicate the causal effects.

4.1. Data

CATIE uses a multiphase design to study the effect of antipsychotic medications for schizophrenia. In phase 1, 1460 patients are randomly assigned to double-blinded treatment with either the conventional drug perphenazine or one of the new-generation drugs olanzapine, quetiapine, risperidone, or ziprasidone. The patients are followed up for up to 18 months or until treatment was discontinued for any reason. Patients whose assigned treatment is discontinued could receive other treatments in phases 1B, 2, and 3 (see Stroup et al., 2003, for further details). Our analysis focused on the phase 1 data only.

About 51% of the 1460 CATIE participants provided DNA samples, and in total 738 patients are genotyped after further inclusion and exclusion criteria. Genotyping is conducted using the Affymetrix 500 K platform and a custom 164 K chip created by Perlegen, which, after quality control, led to 6378 SNPs on chr22. Among the 738 patients, 2 individuals do not have treatment information, 1 individual does not have baseline Positive and Negative Symptom Scale (PANSS) scores, and 86 individuals only have baseline PANSS scores but no follow-up information. Excluding these subjects results in a sample of 649 patients in our analysis. Among the 649 patients, 60% of the subjects have missing outcome values,

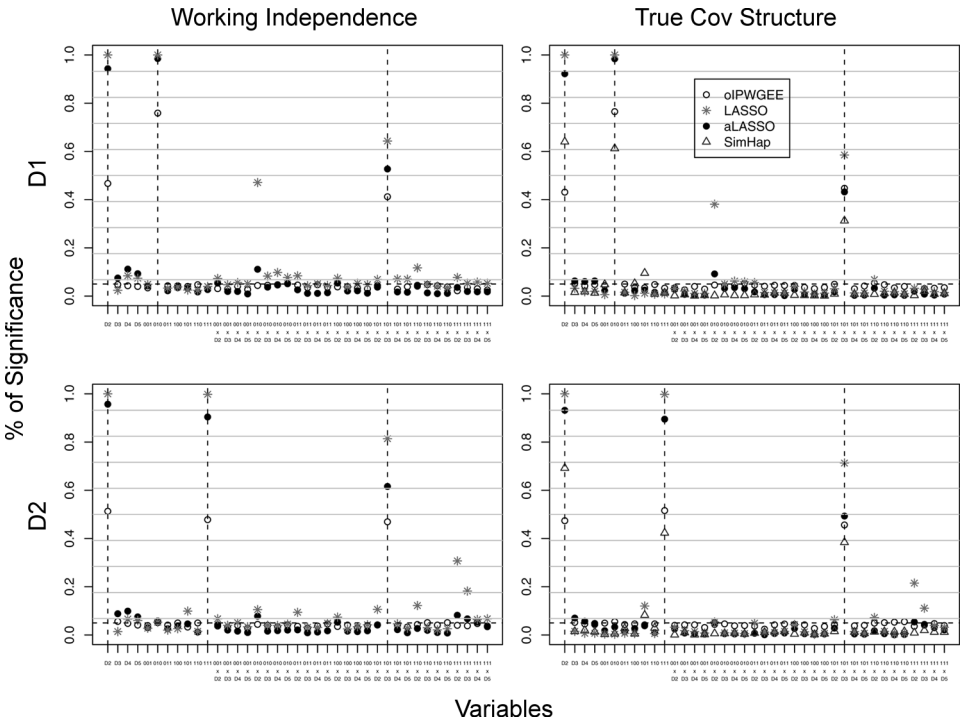

Figure 5 Proportion of significance under scenario D as defined in Table 1. The left panel shows the results of the proposed IPWGEE methods (i.e., oIPWGEE, LASSO, aLASSO) when a working independence covariance is used, and the right panel shows the results of the IPWGEE and SimHap analysis when the true covariance structure is used. The horizontal dashed line indicates the nominal level of 0.05 used for the Wald test in oIPWGEE and SimHap. The vertical dashed lines indicate the causal effects.

and the monotone missingness accounts for a large percentage of missing (i.e., 96% of the 60% individuals).

4.2. Analysis

The primary outcome variable was the PANSS total scores measured at months 1, 3, 6, 9, 12, 15, and 18. The effects to be assessed include: (a) the relative effects of the four new-generation antipsychotic drugs to the conventional drug perphenazine; (b) the genetic effects; and (c) the interactions between the genes and drugs. In addition to the effects of interest, we also incorporate baseline PANSS score, drug–time interaction, age, sex, and ancestry in model (2), and set $b(t) = t$ and $V_i = I_{n\times n}$ for the analysis. The ancestry is approximated using the first seven principle components identified in the CATIE genome-wise association study (Sullivan et al., 2008) using EigenSoft (Price et al., 2006). The missing mechanism is modeled by a logistic regression, in which the dropout status was regressed on the previous PANSS scores and the drugs. When previous PANSS scores are not available, we use the most recent observed value instead. The chromosomal scan is carried out using a sliding window of four SNPs. One challenge for the haplotype

sliding-window scan is the multiple testing problem, as the tests can be highly correlated due to the use of overlapping SNPs. While this issue may be bypassed in the LASSO or aLASSO since they do not involve any test procedure for selecting significant variables, a chromosome-wide significant threshold would still be needed for the Wald test in the oIPWGEE method. Because our analysis is for exploratory purposes, and because it is beyond our focus to address this unsettled issue here, we use an ad hoc way to determine a less stringent threshold: We set the total number of tests as 6378/4 (i.e., as if non-overlapping windows were used), and treat the tests as perfectly correlated when they are next to each other, and uncorrelated otherwise. This leads to a Bonferroni threshold of $0.05/(6378/4/2) = 6.3 \times 10^{-5}$ for the p values from chr22.

4.3. Result

We focus on the oIPWGEE and aLASSO methods, as we see in the simulation that the LASSO method tends to select more variables than necessary. First, the dropout rates seem to depend on drug olanzapine (p value 0.026): Patients with olanzapine have higher odds to stay in the study relative to the baseline drug perphenazine (OR = 1.36). Next, for PANSS analysis, it appears that the PANSS scores depend positively on the baseline PANSS score, negatively on time, but not on the drugs or the drug–time interactions. The mean estimate across all regions for $(\beta_Y, \beta_T, \beta'_D, \beta'_{D\times T}) = (\beta_Y, \beta_T, \beta_{\text{Ola}}, \beta_{\text{Que}}, \beta_{\text{Ris}}, \beta_{\text{Zip}}, \beta_{\text{Ola}\times T}, \beta_{\text{Que}\times T}, \beta_{\text{Ris}\times T}, \beta_{\text{Zip}\times T})$ is (0.59, −0.36, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00) by aLASSO, and is (0.58, −0.59, 1.66, 0.81, 2.32, 1.03, 0.01, 0.04, 0.00, 0.34) by oIPWGEE, with the corresponding mean Z statistics of the Wald test (obtained based on 100 bootstrap samples) as (21.00, 4.68, 0.91, 0.71, 1.04, 0.70, 0.38, 0.42, 0.11, 1.94).

For genetic effect detections, while there are quite a few overlaps in regions identified by both methods, there are also regions where the findings from the two methods do not agree. The inconsistent results may be due to various reasons. For example, one possible reason is that the oIPWGEE selects significant genetic effects using Wald tests adjusting for the multiple testing issue, while the aLASSO selects important genetic factors and estimates their effect sizes at the same time avoiding the multiple testing. Another possible reason is that the dropout status may depend on other covariates than what we have incorporated, and hence the assumed logistic model for the non-dropout probability cannot catch the complete missingness mechanism. Because there are no known positive controls for the data analysis, we report those regions that are identified by both methods. We see that some of the significant regions are adjacent with each other. We list and annotate our findings in Table 2, where the annotations are based on HapMap genome browser (http://www.hapmap.org/cgi-perl/gbrowse/hapmap27_B36), UCSC genome browser (http://www.genome.ucsc.edu/cgi-bin/hgGateway), NCBI (http://www.ncbi.nlm.nih.gov/sites/entrez), Sullivan Lab Evidence Project (SLEP, https://slep.unc.edu/evidence/?tab=GeneName), and Schizophrenia Research Forum (http://www.schizophreniaforum.org/res/sczgene/chromo.asp?c=22). As with many findings in complex trait genetics, there are intriguing suggestions in the results that require replication to understand more fully.

Table 2 Bioinformatics information on the haplotype regions identified by both oIPWGEE and aLASSO methods

Region	Effect found	1st SNP (rs)	Starting position (hg17)	Gene (HapMap)	Putative function (NCBI)	Brain expression (SLEP)	Published candidate genes for SCZ
1936	TTAT	rs738575	26160947				
2718	TCGG	rs2076036	31098683				
2795	CCGC	rs5749468	31316943	SYN3	Synapsin gene family, encode neuronal phosphoproteins	>50th percentile of intensity values	SCZ gene
2890	GGGT × olanzapine	rs135029	31564844	SYN3	Synapsin gene family, encode neuronal phosphoproteins	>50th percentile of intensity values	SCZ gene
				TIMP3	Tissue inhibitor of metalloproteinase-3, located within an intron of SYN3	>75th percentile of intensity values	SCZ gene
3621	CCTT	rs133425	34139516	MCM5	Minichromosome maintenance 5		
3624	TCTC	rs8141025	34150011				
4074	AGTA × quetiapine	rs5750547	36871200	PLA2G6	A2 phospholipase, a class of enzyme that catalyzes the release of fatty acids from phospholipids		SCZ gene
4075	GTAG × quetiapine	rs6001031	36880760	PLA2G6			SCZ gene
4076	TAGC × quetiapine	rs132966	36886062	PLA2G6			SCZ gene
5210	AGGC × quetiapine	rs6007127	44363072				
5211	GGCT × quetiapine	rs5765558	44363516				
5841	GCGG	rs12627816	47148210				

5. DISCUSSION

We introduce an ordinary IPWGEE method and its penalized extensions to facilitate haplotype-based pharmacogenetic analysis for longitudinal quantitative data. It allows for outcome-dependent missingness, and permits an overall evaluation of the high-dimensional haplotype–drug interaction in an unbiased manner. By re-expressing the IPWGEE as a weighted least-square problem, the proposed method is easy to implement and computationally efficient. Our simulations show that the IPWGEE combined with the adaptive LASSO penalty can improve the power to identify important genetic effects while retaining the false positive rates at a desired level. The R code that implements this method is available from the corresponding author's website at http://www4.stat.ncsu.edu/~jytzeng/Software/HapWGEE/R

In our numerical studies, we set V_i to be the identity matrix in the estimation Eq. (4) to obtain the parameter estimates, which treats the outcome values from the same subject as independent after conditioning on the covariates incorporated in model (2). Under the content of genetic studies, this working independence assumption might not be completely incorrect, as the within-subject correlation might be removed after conditioning on the genetic factors of an individual. Nevertheless, the GEE does not require a correctly specified working variance–covariance matrix in order to obtain consistent estimates. The use of a more precisely specified V_i can improve detecting power.

In this work, the IPWGEE-based approaches are constructed for quantitative traits, but the framework can be also extended to binary traits. For quantitative traits, the expected outcome values $E(Y_{i,t} \mid Y_{i,0}, D_i, H_i, \mathbf{Z}_{i,t})$ is a linear function of the unobserved haplotype H_i, and consequently, its genotype-conditioned expectation, $E(Y_{i,t} \mid Y_{i,0}, D_i, G_i, \mathbf{Z}_{i,t})$, is also linear in $E(H_i \mid G_i)$. With binary traits, the same principle of Eq. (3) applies, but $E(Y_{i,t} \mid Y_{i,0}, D_i, H_i, \mathbf{Z}_{i,t}) = \{1 + \exp[-(\beta_I \mathbf{1}_{T_i} + \mathbf{X}_i \beta_J)]\}^{-1}$ is no longer linear in H_i. As a result, the mean effect model conditioning on genotypes require additional work in the computation, and we plan to continue the work in our future study.

Finally, modeling longitudinal quantitative traits has also drawn big attention in the field of quantitative trait locus (QTL) mapping. Among the many methods proposed, functional mapping emerges to be a powerful and promising tool (Wu and Lin, 2006). Functional mapping uses mathematical equations to describe the profile of the response values, such as using logistic equations for the growth trajectories, and using bi-exponential equations for HIV dynamics. These mathematical functions are typically governed by a few parameters that have biological interpretation and can be further expressed in terms of genetic effects. In recent years, the framework of functional mapping has also been extended from controlled crosses to natural population (Lin et al., 2007; Ma et al., 2004; Wu et al., 2007). It will be of great interest to incorporate such mechanistic mathematical modeling in our IPWGEE methods and consider it in the CATIE data analysis.

ACKNOWLEDGMENTS

J.Y.T. was supported by NSF grant DMS-0504726 and NIH grants R01 MH074027 and R01 MH084022. W.L. was supported by NSF grant DMS-0504269 and NIH grant R01 CA140632. P.S.F. was supported by NIH grants R01

MH074027, R01 MH080403, and R01 MH084022. The CATIE project was funded by NIMH contract N01 MH90001 (PIs Drs. Jeffrey Lieberman and Scott Stroup). Jung-Ying Tzeng and Wenbin Lu are contributed equally to this work.

REFERENCES

Carter, K. W., McCaskie, P. A., Palmer, L. J. (2008). SimHap GUI: an intuitive graphical user interface for genetic association analysis. *BMC Bioinformatics.* 25:557.

Clark, A. G. (2004). The role of haplotypes in candidate gene studies. *Genet. Epidemiol.* 27:321–333.

Davison, A. C., Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* 32:407–451.

Fu, W. J. (1998). Penalized regression: the bridge versus the lasso. *J. Comput. Graph. Stat.* 7:397–416.

Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G., Mallinckrodt, C. (2006). Analyzing incomplete discrete longitudinal clinical trial data. *Stat. Sci.* 21:52–69.

Liang, K. Y., Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*. 73:13–22.

Lieberman, J. A., Stroup, T. S., McEvoy, J. P., Swartz, M. S., Rosenheck, R. A., Perkins, D. O., Keefe, R. S., Davis, S. M., Davis, C. E., Lebowitz, B. D., Severe, J., Hsiao, J. K. (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *N. Engl. J. Med.* 22:1209–1223.

Lin, D. Y., Zeng, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies. *J. Am. Stat. Assoc.* 101:89–104.

Lin, M., Li, H., Hou, W., Johnson, J. A., Wu, R. (2007). Modeling sequence–sequence interactions for drug response. *Bioinformatics* 23:1251–1257.

Little, R. J. A., Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd ed. New York: Wiley.

Ma, C. X., Wu, R., Casella, G. (2004). FunMap: functional mapping of complex traits. *Bioinformatics* 20:1808–1811.

Osborne, M. R., Presnell, B., Turlach, B. A. (2000). On the LASSO and its dual. *J. Comput. Graph. Stat.* 9:319–337.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38:904–909.

Robins, J. M., Rotnitzky, A., Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Stat. Assoc.* 90:106–121.

Schaid, D. J. (2004). Evaluating associations of haplotypes with traits. *Genet. Epidemiol.* 27:348–364.

Stroup, T. S., McEvoy, J. P., Swartz, M. S., Byerly, M. J., Glick, I. D., Canive, J. M., McGee, M. F., Simpson, G. M., Stevens, M. C., Lieberman, J. A. (2003). The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: schizophrenia trial design and protocol development. *Schizophr Bull.* 29:15–31.

Sullivan, P. F., Lin, D., Tzeng, J. Y., van den Oord, E., Perkins, D., Stroup, T. S., Wagner, M., Lee, S., Wright, F. A., Zou, F., Liu, W., Downing, A. M., Lieberman, J., Close, S. L. (2008). Genomewide association for schizophrenia in the CATIE study: results of stage 1. *Mol. Psychiatry.* 13:570–584.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B.* 58:267–288.

Tsiatis, A. A., Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Stat. Sin.* 14:809–834.

Wang, H., Li, R., Tsai, C. L. (2007). Tuning parameter selector for SCAD. *Biometrika* 94:553–568.

Wu, S., Yang, J., Wu, R. (2007). Semiparametric functional mapping of quantitative trait loci governing long-term HIV dynamics. *Bioinformatics* 23:i569–i576.

Wu, R., Lin, M. (2006). Functional mapping—how to map and study the genetic architecture of dynamic complex traits. *Nat. Rev. Genet.* 7:229–237.

Zaitlen, N., Kang, H. M., Eskin, E., Halperin, E. (2007). Leveraging the HapMap correlation structure in association studies. *Am. J. Hum. Genet.* 80:683–691.

Zhang, H. H., Lu, W. (2007). Adaptive-LASSO for Cox's proportional hazards model. *Biometrika* 94:1–13.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* 101:1418–1429.

Zou, H., Hastie, T., Tibshirani, R. (2007). On the degrees of freedom of the Lasso. *Ann. Stat.* 35:2173–2192.