




Analysis of secondary phenotypes in multigroup association studies

Fan Zhou¹  | Haibo Zhou¹  | Tengfei Li^{2,3} | Hongtu Zhu^{1,3} 

¹Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

²Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

³Biomedical Research Imaging Center, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

Correspondence

Hongtu Zhu, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599.

Email: htzhu@email.unc.edu

Funding information

National Institute of Mental Health, Grant/Award Numbers: 086633, 116527; National Institute of Environmental Health Sciences, Grant/Award Number: P30ES010126; National Cancer Institute, Grant/Award Number: P01 CA142538

Abstract

Although case-control association studies have been widely used, they are insufficient for many complex diseases, such as Alzheimer's disease and breast cancer, since these diseases may have multiple subtypes with distinct morphologies and clinical implications. Many multigroup studies, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI), have been undertaken by recruiting subjects based on their multiclass primary disease status, while extensive secondary outcomes have been collected. The aim of this paper is to develop a general regression framework for the analysis of secondary phenotypes collected in multigroup association studies. Our regression framework is built on a conditional model for the secondary outcome given the multigroup status and covariates and its relationship with the population regression of interest of the secondary outcome given the covariates. Then, we develop generalized estimation equations to estimate the parameters of interest. We use both simulations and a large-scale imaging genetic data analysis from the ADNI to evaluate the effect of the multigroup sampling scheme on standard genome-wide association analyses based on linear regression methods, while comparing it with our statistical methods that appropriately adjust for the multigroup sampling scheme. Data used in preparation of this article were obtained from the ADNI database.

KEYWORDS

ascertainment, genome-wide association study, multigroup, secondary trait, selection bias

1 | INTRODUCTION

To motivate the proposed methodology, we consider a large database with imaging, genetic, and clinical data from 1737 subjects collected through the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. The overall design of the ADNI is a longitudinal study of various biomarkers at baseline and their longitudinal profiles. ADNI has gone through four phases from ADNI1, GO, 2 to ADNI3 from 2004 until 2016. ADNI1 began with 204 cognitively normal controls (NC), 362 subjects with mild cognitive impairment (MCI), and 179

subjects with Alzheimer's disease (AD), and was extended by three follow-up phases with a different number of subjects in each category. ADNI is a typical example of multigroup studies. Similar to the case-control design, the multigroup sample is usually not a random sample from the whole population because of the unequal selection probabilities between different disease groups. The proportions of AD and MCI in ADNI are much bigger than their global prevalences in the age-matched general population (Kim *et al.*, 2015). In this paper, we focus on the brain regions of the left and right hippocampi of each ADNI subject and a large genetic

data set with over 6 000 000 genotyped and imputed single-nucleotide polymorphisms (SNPs) on all 22 human chromosomes. Since the hippocampus is critical for learning and memory and is vulnerable to damage in the early stages of AD (Schuff *et al.*, 2009), the volume and shape of the hippocampi may be effective phenotypes that facilitate the identification of causal genes and the mechanistic understanding of pathophysiological processes of AD. Our primary goal is to search for genetic patterns associated with local hippocampal changes, while correcting for the selection bias associated with ascertainment in multigroup studies.

In many genetic association studies, some variables of interest are the marker genotype(s), G , secondary (or intermediate) traits Y , the primary phenotype (multigroup status) D , clinical variables C , and the ascertainment (sampling) indicator S . For instance, various imaging measures (eg, subcortical volumes) have been widely used as secondary traits that may be directly associated with a specific disease outcome for most brain-related diseases. A statistical challenge arises from the fact that the main target of interest is the population model of Y given G , whereas both secondary traits Y and marker genotype(s) G are collected depending upon the grouping phenotype D . In genetic epidemiology, standard statistical methods that either ignore ascertainment or naively adjust for ascertainment by conditioning on the disease status (eg, meta-analysis of subjects in different subgroups) can lead to estimation bias, inflated false-positive rate, and decreased statistical power. Therefore, adjusting for D is critical when one models Y given G in genetic association studies.

There is a large literature on the development of statistical methods for eliminating the selection bias associated with ascertainment in case-control (or two-group) studies. The simplest method is to fit a regression model to all subjects in a single group (eg, cases or controls, or each subgroup in multigroup study). It requires a strong assumption that no group difference exists in the genetic effects regarding the corresponding secondary traits. Moreover, dropping a certain number of observations can substantially decrease the estimation efficiency and statistical power. Another simple method, called LRegD (Potkin *et al.*, 2010), is to include the case-control status D as an additional covariate in the regression models. However, LRegD may yield flawed conclusions, since the associations between a secondary outcome and an exposure of interest in the case and control groups can be quite different from that in the underlying target population (Tchetgen Tchetgen, 2014). Various weighted likelihoods, such as the inverse probability weighting (IPW) approach, have been widely used (Richardson *et al.*, 2007; Schifano *et al.*, 2013; Sofer *et al.*, 2017), but they do not utilize the information collected on the primary outcome D . Lee *et al.* (1997) and Jiang *et al.*

(2006) develop a maximum likelihood estimate of the regression coefficients assuming that the sampling rates for cases and controls are known. Lin and Zeng (2009) introduce a retrospective likelihood function by explicitly conditioning on the sampling scheme. He *et al.* (2012) use a Gaussian copula approach, allowing more flexible distributions of the secondary outcome Y compared to Lin and Zeng (2009). Wei *et al.* (2013) propose a robust estimation method for secondary analysis of case-control data by assuming that the secondary trait Y follows a homoscedastic regression model given X . Breslow *et al.* (2000) apply the semiparametric inference method through building an augmented estimation equation to improve the efficiency of IPW. Song *et al.* (2016) introduce a set of counterfactual estimation functions under an alternative disease status and combine the observed and counterfactual estimation functions into a set of weighted estimation equations. However, all these approaches focus on the case-control design.

The aim of this paper is to develop a general regression framework for the analysis of secondary phenotypes collected in multigroup association studies, called MGLREG. There are two major contributions in this paper.

- (I) To the best of our knowledge, this is the first paper that systematically discusses the secondary trait analysis in multigroup studies, while allowing the multiphase design.
- (II) We have developed companion software, called MGLREG, along with its documentation and released it to the public through the link from github (see reference MGLREG).

2 | METHODS

In Section 2.1, we introduce the data structure and some notations. In Sections 2.2 and 2.3, we build the conditional model for Y given D and X and derive its associated estimation equations for the three-group study, that is, $J = 3$. Our approach can be easily extended from the basic $J = 3$ case to the more general setting of $J > 3$ (details for general J discussed in supplements). In Section 2.4, we discuss how to extend our regression framework from continuous secondary outcomes to binary ones. In Section 2.5, we further consider the extension to multiple phases scenario.

2.1 | Data structure and notation

Suppose that we consider N independent subjects from a multigroup study. For each subject, given the group

status $D_i \in \{0, 1, \dots, J-1\}$, we denote S_i as the ascertainment (sampling) indicator and observe the secondary phenotype Y_i of interest, the clinical factors C_i , as well as the genotype score G_i for $i = 1, \dots, N$, where J is a positive integer. For instance, $J=2$ corresponds to the case-control design, whereas $J>2$ corresponds to the multi-group design. Without loss of generality, we focus on continuous secondary traits, while the group 0 corresponds to the control group. Suppose there are n_j subjects in the j -th group for $j = 0, \dots, J-1$ such that N is equal to $n_0 + n_1 + \dots + n_{J-1}$. An important assumption is that the prevalence of each subgroup j is known to be $\tilde{p}_j = P(D=j)$ in the target population and $\tilde{\pi}_j = P(D=j|S=1) = n_j/N$ in the sample for $j = 0, 1, \dots, J-1$. Although the true value of \tilde{p}_j is required, our method still works for an approximated value of \tilde{p}_j . To demonstrate this point, we allow misspecification of \tilde{p}_j in the simulation studies and find that our method performs acceptably stable with varied \tilde{p}_j 's combinations.

2.2 | Model setup

The main target of inference is the population mean model for Y given \mathbf{X} , denoted as $\mu(\mathbf{X}) = E(Y|\mathbf{X})$. We focus on the three-group case with $J=3$ from now on, but all derivations given below are valid when we replace 2 by $J-1$. By using the law of conditional expectations, we have

$$\mu(\mathbf{X}) = \sum_{j=0}^2 \tilde{\mu}(\mathbf{X}, D=j) \times P(D=j|\mathbf{X}), \quad (1)$$

where $\tilde{\mu}(\mathbf{X}, D) = E(Y|\mathbf{X}, D)$. A sufficient condition for estimating $\mu(\mathbf{X})$ is to estimate both $\tilde{\mu}(\mathbf{X}, D)$ and $P(D|\mathbf{X})$. Since we observe Y and \mathbf{X} conditional on D and $S=1$, we can consistently estimate $E(Y|\mathbf{X}, D, S=1)$ and $P(D|\mathbf{X}, S=1)$ instead of $\tilde{\mu}(\mathbf{X}, D)$ and $P(D|\mathbf{X})$.

The sampling design of the multigroup study depends on D only, and therefore (Y, \mathbf{X}) is randomly sampled within each group D . Accordingly, we could characterize a relationship between $E(Y|\mathbf{X}, D, S=1)$ and $\tilde{\mu}(\mathbf{X}, D)$ as

$$\tilde{\mu}(\mathbf{X}, D) = E(Y|\mathbf{X}, D) = E(Y|\mathbf{X}, D, S=1). \quad (2)$$

It then follows from (2) that $\tilde{\mu}(\mathbf{X}, D)$ can be consistently estimated.

Second, we characterize a relationship between $P(D|\mathbf{X}, S=1)$ and $P(D|\mathbf{X})$. Let $\Pi_j(\mathbf{X}) = P(D=j|\mathbf{X}, S=1)$ denote the risk function of $D=j$ at \mathbf{X} in the multigroup sample and $P_j(\mathbf{X}) = P(D=j|\mathbf{X})$ be the probability of D given \mathbf{X} in the whole population. For

each $j = 0, 1, 2$, $\Pi_j(\mathbf{X})$ and $P_j(\mathbf{X})$ satisfy the following relationship:

$$\frac{\Pi_j(\mathbf{X})}{\Pi_0(\mathbf{X})} \cdot \frac{\tilde{\pi}_0}{\tilde{\pi}_j} = \frac{P_j(\mathbf{X})}{P_0(\mathbf{X})} \cdot \frac{\tilde{p}_0}{\tilde{p}_j}. \quad (3)$$

We assume that $\Pi_j(\mathbf{X})$ follows a multinomial logistic regression model as follows:

$$\log \left\{ \frac{\Pi_j(\mathbf{X})}{\Pi_0(\mathbf{X})} \right\} = \log \left\{ \frac{P_j(\mathbf{X})}{P_0(\mathbf{X})} \right\} + \eta_j = \mathbf{X}^T \boldsymbol{\varphi}_j \quad (4)$$

for $j = 0, 1$, and 2, where $\eta_j = \log(\tilde{p}_0 \tilde{\pi}_j) - \log(\tilde{p}_j \tilde{\pi}_0)$. If the η_j 's are known and the ratio of $\Pi_j(\mathbf{X})$ over $\Pi_0(\mathbf{X})$ can be consistently estimated, then the ratio of $P_j(\mathbf{X})$ over $P_0(\mathbf{X})$ can be consistently estimated.

We derive a conditional model of $\tilde{\mu}(\mathbf{X}, D)$ based on (2). Specifically, it follows from the equality $\sum_{j=0}^2 P(D=j|\mathbf{X}) = 1$ and (2) that $\tilde{\mu}(\mathbf{X}, j)$ is given by

$$\begin{aligned} \tilde{\mu}(\mathbf{X}, j) = \mu(\mathbf{X}) + \sum_{k \neq j} P(D=k|\mathbf{X}) \{ \tilde{\mu}(\mathbf{X}, j) \\ - \tilde{\mu}(\mathbf{X}, k) \}. \end{aligned} \quad (5)$$

Furthermore, we define $\gamma_1(\mathbf{X}) = \tilde{\mu}(\mathbf{X}, 1) - \tilde{\mu}(\mathbf{X}, 0)$ and $\gamma_2(\mathbf{X}) = \tilde{\mu}(\mathbf{X}, 2) - \tilde{\mu}(\mathbf{X}, 0)$. With some algebraic calculations, we can rewrite (5) as follows:

$$\tilde{\mu}(\mathbf{X}, j) = \mu(\mathbf{X}) + \sum_{k=1}^2 \{1(j=k) - P(D=k|\mathbf{X})\} \gamma_k(\mathbf{X}) \quad (6)$$

for $j = 0, 1$, and 2. The term besides $\mu(\mathbf{X})$ on the right-hand side of (6) encodes the selection bias by modeling the group difference of Y given different D statuses with fixed \mathbf{X} (Tchetgen Tchetgen, 2014).

Equation (6) has several important implications. If the selection bias is absent, then we have $\gamma_1(\mathbf{X}) = \gamma_2(\mathbf{X}) = 0$ and $\tilde{\mu}(\mathbf{X}, i)$ reduces to $\mu(\mathbf{X})$ regardless of the status of D . If the disease is rare, then both $P(D=1|\mathbf{X})$ and $P(D=2|\mathbf{X})$ are close to zero in the whole population and (6) reduces to

$$\tilde{\mu}(\mathbf{X}, j) = \mu(\mathbf{X}) + \sum_{k=1}^2 1(j=k) \times \gamma_k(\mathbf{X}). \quad (7)$$

Furthermore, if we set $\gamma_1(\mathbf{X}) = \mathbf{X}^T \boldsymbol{\Gamma}_1$, $\gamma_2(\mathbf{X}) = \mathbf{X}^T \boldsymbol{\Gamma}_2$, and $\mu(\mathbf{X}) = \mathbf{X}^T \boldsymbol{\beta}$, where $\boldsymbol{\Gamma}_1$, $\boldsymbol{\Gamma}_2$, and $\boldsymbol{\beta}$ are three vectors of regression coefficients, then model (7) reduces to

$$\tilde{\mu}(\mathbf{X}, j) = \mathbf{X}^T \boldsymbol{\beta} + \sum_{k=1}^2 1(j=k) \mathbf{X}^T \boldsymbol{\Gamma}_k, \quad (8)$$

in which β represents the main effects of \mathbf{X} on Y and Γ_1 and Γ_2 represent the interaction effects of D and \mathbf{X} on Y . However, if the disease is not rare, then the selection bias can be substantial when $\tilde{\mu}(\mathbf{X}, D)$ varies dramatically across D .

2.3 | Estimation

Our conditional model consists of three key components including (2), (4), and (6). We can develop a two-stage estimation procedure to estimate the parameters of interest in $\mu(\mathbf{X})$, $\{\gamma_j(\mathbf{X}): j = 1, 2\}$, and $\{P_j(\mathbf{X}): j = 1, 2\}$ as follows:

- Stage I: Based on (4), we can construct a set of estimation equations to estimate the unknown parameters in $P_j(\mathbf{X})$ in order to obtain its estimate, denoted as $\hat{P}_j(\mathbf{X})$.
- Stage II: We can substitute $\hat{P}_j(\mathbf{X})$ in (6) and then construct the other set of estimation equations to estimate the parameters in $\mu(\mathbf{X})$, $\gamma_1(\mathbf{X})$, and $\gamma_2(\mathbf{X})$ based on (6).

In stage I, we assume that $\log\{P_j(\mathbf{X})\} - \log\{P_0(\mathbf{X})\} = f_j(\mathbf{X}; \varphi_j, \eta_j)$ holds for $j = 1, 2$, where $f_j(\cdot; \cdot, \cdot)$ is a known parametric function. For instance, in (4), we set $f_1(\mathbf{X}; \varphi_j, \eta_j) = \mathbf{X}^T \varphi_j - \eta_j$ for each j . Since $\eta_j = \log(\bar{p}_0 \tilde{\pi}_j) - \log(\bar{p}_j \tilde{\pi}_0)$ is known, we can construct a log pseudo-likelihood function, denoted as $L(\varphi)$, to estimate unknown parameters $\varphi = (\varphi_1^T, \varphi_2^T)^T$ in $\{\Pi_j(\mathbf{X})\}$ based on N observations in the sample $\{(\mathbf{X}_i, D_i, S_i = 1): i = 1, \dots, N\}$. Specifically, the log pseudo-likelihood function $L(\varphi)$ is given by

$$\sum_{i=1}^N \left[\sum_{j=1}^2 \{1(D_i = j) \mathbf{X}_i^T \varphi_j\} - \log \left\{ 1 + \sum_{j=1}^2 \exp(\mathbf{X}_i^T \varphi_j) \right\} \right]. \quad (9)$$

We can calculate the maximum pseudo-likelihood estimate, $\hat{\varphi} = (\hat{\varphi}_1^T, \hat{\varphi}_2^T)^T = \arg\max_{\varphi} L(\varphi)$ or equivalently, $\partial L(\hat{\varphi}) / \partial \varphi^T = \mathbf{0}$. Then, we compute

$$\hat{P}_j(\mathbf{X}) = \exp\{f_j(\mathbf{X}; \hat{\varphi}_j, \eta_j)\} / [1 + \exp\{f_1(\mathbf{X}; \hat{\varphi}_1, \eta_1)\} + \exp\{f_2(\mathbf{X}; \hat{\varphi}_2, \eta_2)\}]$$

as a consistent estimate of $P_j(\mathbf{X})$ for $j = 1$ and 2 .

In stage II, we need to assume an explicit form of $\mu(\mathbf{X})$, $\gamma_1(\mathbf{X})$, and $\gamma_2(\mathbf{X})$ as follows:

$$\begin{aligned} \mu(\mathbf{X}) &= \mu(\mathbf{X}; \beta), \quad \gamma_1(\mathbf{X}) = g_1(\mathbf{X}; \Gamma_1), \quad \text{and} \\ \gamma_2(\mathbf{X}) &= g_2(\mathbf{X}; \Gamma_2), \end{aligned} \quad (10)$$

where $\mu(\cdot, \cdot)$, $g_1(\cdot, \cdot)$, and $g_2(\cdot, \cdot)$ are known functions and β , Γ_1 , and Γ_2 are unknown parameter vectors. Suppose that $\theta = (\beta^T, \Gamma_1^T, \Gamma_2^T)^T$ and $\mu(\cdot, \cdot)$, $g_1(\cdot, \cdot)$, and $g_2(\cdot, \cdot)$ are all in the linear form as described in last section. In this case, (6) can be rewritten as

$$\begin{aligned} \tilde{\mu}(\mathbf{X}, D; \theta, \hat{\varphi}) &= \mu(\mathbf{X}; \beta) + \sum_{j=1}^2 \{1(D = j) \\ &\quad - \hat{P}_j(\mathbf{X}; \hat{\varphi})\} g_j(\mathbf{X}; \Gamma_j). \end{aligned} \quad (11)$$

We construct consistent estimation equations based on N observations $\{(y_i, \mathbf{X}_i, D_i, S_i = 1): i = 1, \dots, N\}$ as follows:

$$U(\theta; \hat{\varphi}) = \sum_{i=1}^N \frac{\partial \tilde{\mu}(\mathbf{X}_i, D_i; \theta, \hat{\varphi})}{\partial \theta^T} \varepsilon_i(\theta, \hat{\varphi}) = \mathbf{0}, \quad (12)$$

where $\varepsilon_i(\theta, \hat{\varphi}) = y_i - \tilde{\mu}(\mathbf{X}_i, D_i; \theta, \hat{\varphi})$ for $i = 1, \dots, N$. Let $\hat{\theta}$ be the solution to $U(\theta; \hat{\varphi}) = \mathbf{0}$ such that $U(\hat{\theta}; \hat{\varphi}) = \mathbf{0}$.

The algorithm which jointly solves $U(\hat{\theta}; \hat{\varphi}) = \mathbf{0}$ and $\partial L(\hat{\varphi}) / \partial \varphi^T = \mathbf{0}$ is denoted as “MGLReg” throughout the paper. We can show that

$$\sqrt{n} \begin{pmatrix} \hat{\theta} - \theta_* \\ \hat{\varphi} - \varphi_* \end{pmatrix} \rightarrow^L N(\mathbf{0}, \Sigma), \quad (13)$$

where \rightarrow^L denotes the convergence in distribution and θ_* and φ_* are the true value of θ and φ , respectively. Moreover, Σ as a covariance matrix can be approximated by $\hat{\Sigma}$, which is given by

$$\begin{aligned} &\begin{pmatrix} \frac{1}{N} \partial_{\theta} U(\hat{\theta}, \hat{\varphi}) & \frac{1}{N} \partial_{\varphi} U(\hat{\theta}, \hat{\varphi}) \\ \mathbf{0} & \frac{1}{N} \partial_{\varphi^2} L(\hat{\varphi}) \end{pmatrix}^{-1} \widehat{\text{Cov}} \begin{pmatrix} \frac{U(\hat{\theta}, \hat{\varphi})}{\sqrt{N}} \\ \frac{\partial_{\varphi} L(\hat{\varphi})}{\sqrt{N}} \end{pmatrix} \\ &\times \begin{pmatrix} \frac{1}{N} \partial_{\theta} U(\hat{\theta}, \hat{\varphi}) & \frac{1}{N} \partial_{\varphi} U(\hat{\theta}, \hat{\varphi}) \\ \mathbf{0} & \frac{1}{N} \partial_{\varphi^2} L(\hat{\varphi}) \end{pmatrix}^{-T}, \end{aligned} \quad (14)$$

where $\partial_{\theta} = \partial / \partial \theta$ and $\partial_{\varphi} = \partial / \partial \varphi$.

We discuss an extension of the semiparametric locally efficient estimation (“SLEE”) method of Tchetgen Tchetgen (2014). Specifically, the joint density of the observed data in the multigroup case can be written as

$$f(Y|\mathbf{X}, D) f(\mathbf{X}|D) \prod_{j=0}^2 \tilde{\pi}_j^{1(D=j)} \propto f(Y|\mathbf{X}, D) f^*(D|\mathbf{X}) f^*(\mathbf{X}) \quad (15)$$

where $f^*(\mathbf{X}) \propto f(\mathbf{X})f(D=0|\mathbf{X})/f^*(D=0|\mathbf{X})$ and

$$\text{logit}(f^*(D=j|\mathbf{X})) = \text{logit}(\Pi_j(\mathbf{X})) = \text{logit}(P_j(\mathbf{X})) - \log \left\{ \frac{\tilde{p}_j(1 - \tilde{\pi}_j)}{\tilde{\pi}_j(1 - \tilde{p}_j)} \right\}$$

for $j = 1, 2$. We can derive the efficient score of (θ, φ) as

$$R(\theta, \varphi) = (R_\theta(\theta, \varphi)^T, R_\varphi(\theta, \varphi)^T)^T, \quad (16)$$

where $R_\theta = \partial_\theta \tilde{\mu}(\mathbf{X}, D; \theta, \varphi) \{\text{var}(\varepsilon(\theta, \varphi|\mathbf{X}, D))\}^{-1} \varepsilon(\theta, \varphi)$ and

$$R_\varphi = \partial_\varphi L(\varphi) + \partial_\varphi \tilde{\mu}(\mathbf{X}, D; \theta, \varphi) \times \{\text{var}(\varepsilon(\theta, \varphi|\mathbf{X}, D))\}^{-1} \varepsilon(\theta, \varphi).$$

The SLEE method by solving (16) is theoretically more efficient than MRLReg, but it is computationally much more difficult. However, simulations in the next section demonstrates that “MRLReg” is competitive in comparison of estimation efficiency compared with “SLEE.”

2.4 | Extension to binary secondary outcome

Our framework can be easily extended to the case when Y is binary. Assume that $\tilde{\mu}(\mathbf{X}, D) = E(Y|\mathbf{X}, D) = P(Y=1|\mathbf{X}, D)$ and $\mu(\mathbf{X}) = P(Y=1|\mathbf{X})$ on the logit scale. Let $\text{Odds}(\mathbf{X}, D) = (Y=1|\mathbf{X}, D)/P(Y=0|\mathbf{X}, D)$ and $\text{Odds}(\mathbf{X}) = P(Y=1|\mathbf{X})/P(Y=0|\mathbf{X})$. Following the derivation of (3.1) in Tchetgen Tchetgen (2014), we can get

$$\text{Odds}(\mathbf{X}, D) = \exp[\log\{\text{Odds}(\mathbf{X})\} + \nu(\mathbf{X}, D) - \bar{\nu}(\mathbf{X})], \quad (17)$$

where $\nu(\mathbf{X}, D) = \log(\text{Odds}(\mathbf{X}, D)/\text{Odds}(\mathbf{X}, D=0))$ and

$$\bar{\nu}(\mathbf{X}) = \sum_{j=1}^2 \exp\{\nu(\mathbf{X}, D=j)\} P(D=j|\mathbf{X}, Y=0) + P(D=0|\mathbf{X}, Y=0).$$

If (3) holds, we have

$$\log \left\{ \frac{\Pi_j^*(\mathbf{X})}{\Pi_0^*(\mathbf{X})} \right\} = \log \left\{ \frac{P_j^*(\mathbf{X})}{P_0^*(\mathbf{X})} \right\} = m(\mathbf{X}; \varphi_j), \quad (18)$$

where $\Pi_j^*(\mathbf{X})$ and $P_j^*(\mathbf{X})$ here correspond to $P(D=j|\mathbf{X}, Y=0, S=1)$ and $P(D=j|\mathbf{X}, Y=0)$, respectively.

By setting $\log\{\text{Odds}(\mathbf{X})\} = \mu(\mathbf{X}; \beta)$ and $\nu(\mathbf{X}, D=j) = \sum_j 1(D=j)g_j(\mathbf{X}; \gamma_j)$, we have

$$\text{logit}\{P(Y=1|D, \mathbf{X}; \theta, \varphi)\} = \mu(\mathbf{X}; \beta) + \sum_j 1(D=j)g_j(\mathbf{X}; \gamma_j) - \bar{\nu}(\mathbf{X}; \gamma_1, \gamma_2, \varphi) \quad (19)$$

with $\theta^T = (\beta^T, \gamma_1^T, \gamma_2^T)$. Similar to $L(\varphi)$, we solve the log-likelihood function given by

$$\sum_{i=1}^N (1 - Y_i) \left[\sum_{j=1}^2 \left\{ 1(D_i=j) \mathbf{X}_i^T \varphi_j \right\} - \log \left\{ 1 + \sum_{j=1}^2 \exp(\mathbf{X}_i^T \varphi_j) \right\} \right]. \quad (20)$$

Finally, estimating θ can be done by solving estimation equations based on (19).

2.5 | Extension to multiphase scenario

In this subsection, we extend our regression framework to large-scale multigroup studies with multiple phases. In practice, some studies (eg, ADNI) collect data across multiple phases, while different phases may follow different sampling schemes. We only consider the case that each subject participates in a single phase, which agrees with the study design of ADNI. For notational simplicity, we consider a three-group study with two phases.

It is assumed that all subjects from different phases follow the same population-level models in terms of $\mu(\mathbf{X}) = E(Y|\mathbf{X})$, $\tilde{\mu}(\mathbf{X}, D) = E(Y|\mathbf{X}, D)$, and $P(D=j|\mathbf{X})$, and (2) holds for both phases. Similar to (5), we have

$$\tilde{\mu}(\mathbf{X}, j) = \mu(\mathbf{X}) + \sum_{k \neq j} P(D=k|\mathbf{X}) \{\tilde{\mu}(\mathbf{X}, j) - \tilde{\mu}(\mathbf{X}, k)\} \quad (21)$$

for both phases and each $j = 0, 1, 2$. We still use $\gamma_1(\mathbf{X}) = \mathbf{X}^T \Gamma_1$, $\gamma_2(\mathbf{X}) = \mathbf{X}^T \Gamma_2$, and $\mu(\mathbf{X}) = \mathbf{X}^T \beta$ to characterize the group difference and target the model at the population level. However, it is assumed that different sampling schemes are used for phases 1 and 2. Let A be the phase from now on, and denote $\Pi_j^{(m)}(\mathbf{X}) = P(D=j|\mathbf{X}, A=m, S=1)$ for phase $m = 1, 2$ and group $j = 0, 1, 2$. Thus, (3) is given by

$$\frac{\Pi_j^{(m)}(\mathbf{X})}{\Pi_0^{(m)}(\mathbf{X})} \cdot \frac{\tilde{\pi}_0}{\tilde{\pi}_j} = \frac{P_j(\mathbf{X})}{P_0(\mathbf{X})} \cdot \frac{\bar{p}_0^{(m)}}{\bar{p}_j^{(m)}} \quad \text{for } m = 1, 2 \text{ and } j = 0, 1, 2, \quad (22)$$

where $\bar{p}_j^{(m)} = P(D = j | S = 1, A = m)$ corresponds to the proportion of group j in the sample at phase m . Subsequently, by assuming a multinomial logistic regression model for $P_j(\mathbf{X})$, we have

$$\log \left\{ \frac{\Pi_j^{(m)}(\mathbf{X})}{\Pi_0^{(m)}(\mathbf{X})} \right\} = \log \left\{ \frac{P_j(\mathbf{X})}{P_0(\mathbf{X})} \right\} + \eta_j^{(m)} = \mathbf{X}^T \boldsymbol{\varphi}_j + \eta_j^{(m)}, \quad (23)$$

where $\eta_j^{(m)} = \log(\bar{p}_0^{(m)} \tilde{\pi}_j) - \log(\bar{p}_j^{(m)} \tilde{\pi}_0)$ for $m = 1, 2$.

We use a slightly different two-stage estimation procedure to estimate all the parameters of interest. Specifically, in stage I, we estimate $P_j(\mathbf{X})$ for the two phases by combining the observations from both phases. Afterwards, we use the same estimation method in stage II to estimate additional parameters in $\mu(\mathbf{X})$, $\gamma_1(\mathbf{X})$, and $\gamma_2(\mathbf{X})$. The log pseudo-likelihood function $L(\boldsymbol{\varphi})$ in stage I is given by

$$\sum_{i=1}^N \sum_{m=1}^2 \left[\sum_{j=1}^2 \left\{ 1(D_i = j) (\mathbf{X}_i^T \boldsymbol{\varphi}_j + \eta_j^{(m)}) \right\} - \log \left\{ 1 + \sum_{j=1}^2 \exp(\mathbf{X}_i^T \boldsymbol{\varphi}_j + \eta_j^{(m)}) \right\} \right] 1(A_i = m).$$

Under some mild conditions, it can be shown that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*, \hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}_*) \rightarrow^L N(\mathbf{0}, \boldsymbol{\Sigma}_*)$, where the covariance matrix $\boldsymbol{\Sigma}_*$ can be approximated by $\hat{\boldsymbol{\Sigma}}_*$, which is given in the supplements.

3 | SIMULATION STUDIES

We carry out Monte Carlo simulations to evaluate the finite sample performance of five methods including (a) LReg: linear regression without bias correction; (b) LRegD: linear regression method adjusted for the group status $\mathbf{X}_s = (1(D = 1), 1(D = 2))^T$; (c) IPW: inverse probability weighting approach (Richardson *et al.*, 2007); (d) SPREG: the retrospective likelihood method in (Lin and Zeng, 2009); (e) MGLReg; and (f) SLEE: the semiparametric locally efficient estimation method.

3.1 | Two-SNP setup

We consider two parts of the simulation. The first part assumes that group difference exists in the genetic effects on the secondary trait. The second part assumes an incorrect specification of the conditional model and a misspecification of the $\gamma_1(\mathbf{X})$, $\gamma_2(\mathbf{X})$ (Lin and Zeng, 2009;

Song *et al.*, 2016; Zhu *et al.*, 2017). In this setup, one SNP has significant effect on the secondary trait, whereas the other is unrelated.

3.1.1 | Setting 1

The details of the first part are described as follows:

- (i) Generate a nongenetic covariate $C \sim N(0, 1)$ for each subject.
- (ii) Generate two SNP-level genetic variables G_1, G_2 with minor allele frequency (MAF) = 0.3 following a multinomial distribution with frequencies $(p_A^2, 2p_A(1 - p_A), (1 - p_A)^2)$ for (AA, Aa, aa) respectively, with the Hardy-Weinberg equilibrium assumption under the additive mode of inheritance.
- (iii) Generate the primary trait D according to the following multinomial logistic model:

$$\log \left\{ \frac{P(D = j | \mathbf{X})}{P(D = 0 | \mathbf{X})} \right\} = \mathbf{X}^T \boldsymbol{\varphi}_j \quad \text{for } j = 1, 2,$$

where $\mathbf{X}^T = (1, C, G_1, G_2)$. Subsequently, we can calculate the two dummy variables $1(D = 1)$ and $1(D = 2)$. Moreover, we choose $\boldsymbol{\varphi}_1 = \boldsymbol{\varphi}_2$ so that the global prevalence of groups 0, 1, and 2 are respectively 10%, 15%, and 75%. We also consider a rare disease case with the global prevalence of groups 0, 1, and 2 being 5%, 5%, and 90%, respectively.

- (iv) Generate the secondary phenotype Y for each subject according to (6) as follows:

$$Y = \beta_0 + \boldsymbol{\beta}_1^T \mathbf{X} + \sum_{j=1}^2 \{1(D = j) - P_j(\mathbf{X})\} \gamma_j(\mathbf{X}) + \varepsilon, \quad (24)$$

where $\varepsilon \sim N(0, \delta)$, $\boldsymbol{\beta}_1^T = (1, 2, 0)$. β_0 and δ are equal to the sample mean and variance of left hippocampi volume from ADNI, respectively. We also set $\gamma_j(\mathbf{X}_i) = \mathbf{X}_i^T \boldsymbol{\Gamma}_j$ for $j = 1, 2$ with $\boldsymbol{\Gamma}_1 = (-2, -1, -1, -1)^T$ and $\boldsymbol{\Gamma}_2 = (1, 1, 1, 1)^T$.

- (v) Repeat steps (i) to (iv) to generate (Y, \mathbf{X}, D) until we obtain a total of $N = 50\,000$ observations as the whole population. Then, we randomly select 500, 1000, and 500 subjects from the $D = 0$, $D = 1$, and $D = 2$ groups to build a nonrandom three-group sample.

3.1.2 | Setting 2

- (i) Generate $\mathbf{X}^T = (1, C, G_1, G_2)$ as setting 1.
- (ii) Generate the secondary phenotype Y for each subject according to

$$Y = \beta_0 + \beta_1^T \mathbf{X} + \varepsilon, \quad (25)$$

and we still have $\varepsilon \sim N(0, \delta)$, $\beta_1^T = (1, 2, 0)$, and the same (β_0, δ) as setting 1.

- (iii) Simulate the primary trait D using a multinomial model given by

$$\log \left\{ \frac{P(D = j | \mathbf{X}, Y)}{P(D = 0 | \mathbf{X}, Y)} \right\} = (\mathbf{X}^T, Y) \tilde{\varphi}_j \text{ for } j = 1, 2,$$

and we also vary $\tilde{\varphi}_1, \tilde{\varphi}_2$ to get the global group prevalences to be (10%, 15%, 75%) and (5%, 5%, 90%) for the rare case, respectively.

- (iv) Repeat steps (i) to (iii) until the sample size reaches 500 000 and then sample 500 ($D = 0$), 1000 ($D = 1$), and 500 ($D = 2$) observations from the above large pool of subjects.

Tables 1 and 2 present the simulation results under the first and second simulation setups. They include the mean absolute biases and the variances of $\hat{\beta}_G$ and, their 95% confidence interval coverage rates based on the 1000 Monte Carlo samples for all six methods. Both LReg and LRegD perform poorly in correcting the sampling bias for both settings. Under the first setting, MGLReg and SLEE introduced in this paper have the smallest estimation bias. The SLEE performs slightly better than MGLReg, but the difference is not substantial. The IPW achieves a comparable performance with MGLReg, whereas our method is more efficient under both settings. The likelihood-based approach SPREG does not work in the first part, since it highly depends on the correct specification of the conditional model. For the second part, MGLReg and SLEE provide competitive estimation results with SPREG, especially in the rare disease case. On the other hand, as we misspecify $(\tilde{p}_0, \tilde{p}_1)$, both MGLReg and SLEE perform acceptably stable under

TABLE 1 Estimation biases, variances, and 95% coverage rates of $\hat{\beta}_G$ for $p_A = 0.3$

	Setting 1			Setting 2		
	Absolute bias	Variance	Coverage	Absolute bias	Variance	Coverage
$\beta_{G_1} = 2$						
LReg	0.8526	1.06×10^{-2}	0.012	0.1774	9.22×10^{-3}	0.572
LRegD	0.5639	2.79×10^{-2}	0.066	0.8633	7.59×10^{-3}	0.000
IPW	0.0848	1.94×10^{-2}	0.945	0.1180	2.14×10^{-2}	0.945
SPREG	1.2001	1.78×10^{-1}	0.000	0.0889	1.13×10^{-2}	0.946
MGLReg ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.15$)	0.0615	2.69×10^{-3}	0.969	0.0987	1.49×10^{-2}	0.946
SLEE ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.15$)	0.0613	2.63×10^{-3}	0.970	0.0986	1.48×10^{-2}	0.948
MGLReg ($\tilde{p}_0 = 0.05, \tilde{p}_1 = 0.15$)	0.0633	3.21×10^{-3}	0.954	0.1014	1.37×10^{-2}	0.936
SLEE ($\tilde{p}_0 = 0.05, \tilde{p}_1 = 0.15$)	0.0631	3.20×10^{-3}	0.956	0.1006	1.36×10^{-2}	0.935
MGLReg ($\tilde{p}_0 = 0.15, \tilde{p}_1 = 0.15$)	0.0671	2.75×10^{-3}	0.960	0.1053	1.77×10^{-2}	0.926
SLEE ($\tilde{p}_0 = 0.15, \tilde{p}_1 = 0.15$)	0.661	2.69×10^{-3}	0.961	0.1048	1.74×10^{-2}	0.928
MGLReg ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.1$)	0.1008	6.26×10^{-3}	0.884	0.1065	1.58×10^{-2}	0.914
SLEE ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.1$)	0.0993	6.15×10^{-3}	0.886	0.1029	1.56×10^{-2}	0.918
MGLReg ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.2$)	0.0955	3.15×10^{-3}	0.854	0.0982	1.10×10^{-2}	0.956
SLEE ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.2$)	0.0942	3.07×10^{-3}	0.886	0.0972	1.04×10^{-2}	0.960
$\beta_{G_2} = 0$						
LReg	0.8483	1.08×10^{-2}	0.000	0.1478	1.11×10^{-2}	0.776
LRegD	0.9744	2.04×10^{-2}	0.000	0.3744	6.79×10^{-3}	0.014
IPW	0.0752	1.73×10^{-2}	0.944	0.1137	2.55×10^{-2}	0.950
SPREG	0.7418	1.04×10^{-1}	0.112	0.0994	1.53×10^{-2}	0.954
MGLReg ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.15$)	0.0655	6.80×10^{-3}	0.954	0.1050	1.99×10^{-2}	0.952
SLEE ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.15$)	0.0644	6.55×10^{-3}	0.954	0.1036	1.92×10^{-2}	0.952
MGLReg ($\tilde{p}_0 = 0.05, \tilde{p}_1 = 0.15$)	0.0868	9.86×10^{-3}	0.868	0.1050	1.99×10^{-2}	0.952
SLEE ($\tilde{p}_0 = 0.05, \tilde{p}_1 = 0.15$)	0.0851	9.75×10^{-3}	0.870	0.1036	1.92×10^{-2}	0.952
MGLReg ($\tilde{p}_0 = 0.15, \tilde{p}_1 = 0.15$)	0.0714	5.62×10^{-3}	0.924	0.1070	1.76×10^{-2}	0.948
SLEE ($\tilde{p}_0 = 0.15, \tilde{p}_1 = 0.15$)	0.0706	5.53×10^{-3}	0.928	0.1049	1.99×10^{-2}	0.950
MGLReg ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.1$)	0.0945	7.53×10^{-3}	0.846	0.0987	2.61×10^{-2}	0.930
SLEE ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.1$)	0.0947	7.38×10^{-3}	0.848	0.1043	2.54×10^{-2}	0.932
MGLReg ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.2$)	0.0938	6.39×10^{-3}	0.844	0.1023	1.91×10^{-2}	0.946
SLEE ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.2$)	0.0897	5.95×10^{-3}	0.850	0.1019	1.86×10^{-2}	0.946

TABLE 2 Estimation biases, variances, and 95% coverage rates of $\hat{\beta}_G$ for rare disease case

	Setting 1			Setting 2		
	Absolute bias	Variance	Coverage	Absolute bias	Variance	Coverage
$\beta_{G_1} = 2$						
LReg	1.5466	7.14×10^{-3}	0.000	0.6095	7.49×10^{-3}	0.102
LRegD	0.7638	2.37×10^{-2}	0.004	1.0535	6.54×10^{-3}	0.000
IPW	0.0686	5.68×10^{-3}	0.832	0.1859	3.25×10^{-2}	0.640
SPREG	0.1546	1.22×10^{-1}	0.896	0.1486	9.82×10^{-2}	0.891
MGLReg ($\tilde{p}_0 = 0.05, \tilde{p}_1 = 0.05$)	0.0552	4.86×10^{-3}	0.916	0.1139	2.14×10^{-2}	0.928
SLEE ($\tilde{p}_0 = 0.05, \tilde{p}_1 = 0.05$)	0.0552	4.85×10^{-3}	0.920	0.1081	1.92×10^{-2}	0.930
MGLReg ($\tilde{p}_0 = 0.05, \tilde{p}_1 = 0.1$)	0.0726	4.42×10^{-3}	0.868	0.1313	2.84×10^{-2}	0.911
SLEE ($\tilde{p}_0 = 0.05, \tilde{p}_1 = 0.1$)	0.0720	4.39×10^{-3}	0.872	0.1308	2.59×10^{-2}	0.912
MGLReg ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.05$)	0.0709	3.83×10^{-3}	0.880	0.1293	2.47×10^{-2}	0.912
SLEE ($\tilde{p}_0 = 0.01, \tilde{p}_1 = 0.05$)	0.0714	3.81×10^{-3}	0.884	0.1252	2.38×10^{-2}	0.916
$\beta_{G_2} = 0$						
LReg	1.0773	1.34×10^{-2}	0.000	0.3857	8.91×10^{-3}	0.390
LRegD	1.0959	8.33×10^{-3}	0.000	0.5367	7.35×10^{-3}	0.004
IPW	0.0751	7.88×10^{-3}	0.850	0.1536	3.42×10^{-2}	0.950
SPREG	0.1376	1.38×10^{-1}	0.884	0.1349	5.10×10^{-2}	0.921
MGLReg ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.15$)	0.0712	6.80×10^{-3}	0.970	0.1270	2.30×10^{-2}	0.946
SLEE ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.15$)	0.0710	6.55×10^{-3}	0.972	0.1240	2.18×10^{-2}	0.950
MGLReg ($\tilde{p}_0 = 0.05, \tilde{p}_1 = 0.1$)	0.0806	9.94×10^{-3}	0.926	0.1448	3.37×10^{-2}	0.938
SLEE ($\tilde{p}_0 = 0.05, \tilde{p}_1 = 0.1$)	0.0797	9.70×10^{-3}	0.932	0.1399	3.18×10^{-2}	0.940
MGLReg ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.05$)	0.0795	8.87×10^{-3}	0.930	0.1432	3.01×10^{-2}	0.942
SLEE ($\tilde{p}_0 = 0.1, \tilde{p}_1 = 0.05$)	0.0793	8.85×10^{-3}	0.932	0.1429	2.96×10^{-2}	0.945

different global prevalence settings. More details are given in Tables 1 and 2. In terms of the computation efficiency, MGLReg is about 10 times faster than SLEE. Therefore, we choose MGLReg to do the large-scale ADNI data analysis.

3.2 | Multiple-SNP setup

To better mimic the real-world genome-wide association study (GWAS) analysis, we use the same simulation settings as those for the two-SNP setup except adopting a multiple-SNP setup within total 500 SNPs and randomly sampling 10 SNPs as causal SNPs with effect size being 0.5. For details, please refer to the supplementary document.

Table 3 presents the mean absolute biases, the mean estimation variances and their 95% confidence interval coverage rates based on 100 Monte Carlo samples of both the causal and noncausal SNPs for all methods. Table 3 shows that our method MGLReg can detect more causal SNPs (higher mean coverage rates) compared to the other methods in both settings, demonstrating that our method is more robust against biased sampling and less sensitive to model misspecification. Compared to the two-SNP setup, IPW is more biased especially for setting 2, whereas our method is much more stable. SPREG does

not perform well in this case even for setting 2, which confirms our conclusion that SPREG highly depends on the correct specification of the conditional model. For SNPs not associated with secondary phenotype, MGLReg performs similar to others. It means that it does not overestimate the genetic effects of noncausal SNPs even with higher model complexity.

4 | THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE DATA

We apply the MGLReg method to the ADNI data set. The main goal of this data analysis is to search for genetic patterns that are associated with local hippocampal changes, while correcting for the selection bias associated with ascertainment in multigroup studies.

4.1 | GWAS analysis

The 299 subjects with normal cognition (NC), 553 with MCI and 185 with AD build the final sample data, where 712 of them are from ADNI1 with the other 325 from ADNI2 and GO. The secondary outcome Y used in the

TABLE 3 Mean estimation biases, variances, and 95% coverage rates of causal and noncausal single-nucleotide polymorphisms (SNPs)

	Setting 1			Setting 2		
	Absolute bias	Variance	Coverage	Absolute bias	Variance	Coverage
Causal SNPs						
LReg	0.2996	1.16×10^{-2}	0.128	0.2032	9.02×10^{-3}	0.512
LRegD	0.3042	1.02×10^{-1}	0.220	0.3691	6.38×10^{-3}	0.000
IPW	0.0693	7.61×10^{-3}	0.902	0.1772	5.46×10^{-2}	0.648
SPREG	0.2998	1.17×10^{-1}	0.132	0.1534	3.57×10^{-2}	0.904
MGLReg	0.0557	4.83×10^{-3}	0.944	0.1075	1.14×10^{-2}	0.956
Noncausal SNPs						
LReg	0.0674	7.07×10^{-3}	0.923	0.1464	1.01×10^{-2}	0.929
LRegD	0.0576	4.01×10^{-3}	0.943	0.0961	6.91×10^{-2}	0.933
IPW	0.0700	7.59×10^{-3}	0.907	0.1898	5.68×10^{-2}	0.645
SPREG	0.0693	8.39×10^{-3}	0.940	0.1302	3.39×10^{-2}	0.937
MGLReg	0.0549	5.06×10^{-3}	0.951	0.0896	1.67×10^{-2}	0.947

experiment are the logarithm of the left and right hippocampi volumes divided by the whole brain volume. The 6 017 259 SNPs after quality control are analyzed, and the genetic factor at each individual SNP is coded as 0, 1, and 2. To correct for the population stratification, the top three principal components (PCs) of the whole-genome data are included as covariates (Price *et al.*, 2006). We also add a dummy variable for distinguishing ADNI1 from (ADNI2, ADNIGO), since different imaging protocols were used in ADNI1 and (ADNI2, ADNIGO), which may affect the volume segmentation results. We apply two-sample *t* test to test the difference between ADNI1 and (ADNI2, ADNIGO), whose *p*-value is smaller than $2e - 16$. Thus, a significant difference exists between the distribution of *Y* for ADNI1 and that for (ADNI2, ADNIGO) according to the boxplot in the supplements. The details of data description and processing procedures are discussed in Supplementary Material.

In this data analysis, $D = 0, 1$, and 2 represent AD, MCI, and NC, respectively. The global prevalence of AD within people older than 65 is more than 10% (Alzheimer's Association, 2012) while MCI is between 10% and 20% (Kim *et al.*, 2015). We compare four different combinations of $(\tilde{p}_0, \tilde{p}_1)$, $(0.1, 0.15)$, $(0.1, 0.2)$, $(0.15, 0.15)$, and $(0.15, 0.2)$ for our proposed method, since the prevalences of AD and MCI vary with patients getting old, and the chance of developing MCI and AD increases as adults age.

4.2 | Results

Table 4 presents the most significant pairs of SNPs combined with the regions of interest detected by LReg, where significant SNPs are selected according to the 5×10^{-8} *p*-value threshold for both the left and right

hippocampi. The *p*-values of these SNPs by MGLReg with different $(\tilde{p}_0, \tilde{p}_1)$ selections are also provided. Those *p*-values smaller than 5×10^{-8} are marked by bold italic in Table 4.

The SNP rs429358, related to gene APOE, is detected as the most significant SNP for both left and right hippocampi by both LReg and MGLReg. Specifically, rs429358 has significant genetic effects on the volume size of left hippocampi since its *p*-value is consistently smaller than the $5e^{-8}$ threshold with different combinations of $(\tilde{p}_0, \tilde{p}_1)$. This result agrees with the previous findings (Kim *et al.*, 2002; Shen *et al.*, 2010; Lu *et al.*, 2011; Kim *et al.*, 2015). Another significant SNP rs769449, also in APOE region, has competitive significance with rs429358 for both left and right hippocampi, which was found to be associated with cerebrospinal fluid (CSF) tau (Cruchaga *et al.*, 2013) and verbal memory (Arpawong *et al.*, 2017). Therefore, our results may prove that rs769449 may have potential effects on the hippocampi volumes. Other significant SNPs detected by LReg are not stably significant when the population rates vary according to the results of our approach. For example, rs59007384 (associated with gene TOMM40) is related to the progression from MCI status to AD (Cervantes *et al.*, 2011). The higher group proportion of AD in the sample data may result in the significant *p*-value by LReg. However, our method MGLReg indicates that rs59007384 may not be significantly related with the hippocampi volume sizes in the whole population, especially the group of normal people.

Figure 1 presents the heatmaps of $\log_{10}(p)$ -value for SNPs rs429358, rs769449, and rs59007384 using MGLReg, with \tilde{p}_0 and \tilde{p}_1 varying within $[0.1, 0.35]$ and $[0.1, 0.65]$, respectively, demonstrating a dynamic

TABLE 4 Top single-nucleotide polymorphisms (SNPs) and p -values for association tests with the left and right hippocampus volumes

Common effect				Interaction								
				MGLReg			MGLReg					
		Chromosomes	LReg	(0.2, 0.15)	(0.15, 0.15)	(0.2, 0.1)	(0.15, 0.1)	LReg	(0.2, 0.15)	(0.15, 0.15)	(0.2, 0.1)	(0.15, 0.1)
Left hippocampus												
rs429358	19		1.76e-11	3.79e-11	2.00e-10	5.01e-09	3.48e-08	0.797	0.938	0.883	0.766	0.732
rs769449	19		5.21e-10	1.38e-09	5.15e-09	6.09e-08	2.96e-07	0.874	0.718	0.642	0.615	0.554
rs10414043	19		6.34e-10	4.44e-09	1.72e-08	1.68e-07	8.18e-07	0.827	0.700	0.633	0.595	0.542
rs73052335	19		1.39e-09	1.55e-08	5.70e-08	5.45e-07	2.47e-06	0.751	0.643	0.582	0.529	0.484
rs59007384	19		3.77e-08	1.38e-05	4.96e-05	6.56e-04	1.86e-03	0.406	0.771	0.742	0.661	0.655
Right hippocampus												
rs429358	19		1.17e-10	3.82e-09	1.77e-08	4.69e-08	3.04e-06	0.089	0.325	0.324	0.223	0.239
rs769449	19		2.37e-09	4.99e-10	1.38e-09	3.24e-08	1.20e-07	0.109	0.286	0.287	0.205	0.221
rs10414043	19		2.35e-09	9.55e-10	2.70e-08	5.70e-08	2.09e-07	0.105	0.297	0.302	0.204	0.221
rs73052335	19		3.76e-09	3.82e-09	1.08e-08	2.01e-07	7.21e-07	0.100	0.260	0.263	0.171	0.185
rs6857	19		5.20e-09	4.31e-07	1.86e-06	3.18e-05	1.33e-04	0.253	0.730	0.701	0.511	0.516
rs283812	19		2.92e-08	2.24e-06	7.29e-06	1.23e-04	3.89e-04	0.116	0.121	0.139	0.124	0.150
rs59007384	19		7.81e-09	1.89e-05	6.51e-05	7.98e-04	2.42e-03	0.106	0.747	0.768	0.653	0.708

change of significance over various MCI and AD prevalence rates in the whole population. We introduce the significance prevalence heatmap (SPH) by using ellipse contours corresponding to different p -value thresholds to determine the population prevalence range for the significance of a specific SNP. For instance, if $\tilde{p}_0 + \tilde{p}_1$ is smaller than 0.5, then within the given $(\tilde{p}_0, \tilde{p}_1)$ range, rs429358 is significant for the left hippocampi as $\tilde{p}_0 + 2.625 \times \tilde{p}_1 > 0.4045$ and for the right hippocampi as $\tilde{p}_0 + 3.138 \times \tilde{p}_1 > 0.596$; rs769449 is significant for the left hippocampi as $\tilde{p}_0 + 2.70 \times \tilde{p}_1 > 0.478$ and for the right hippocampi as $\tilde{p}_0 + 3.5 \times \tilde{p}_1 > 0.534$.

To more clearly show how the global prevalence rate $(\tilde{p}_0, \tilde{p}_1)$ influences the genetic effects, we plot the density curves of the $-\log_{10}(p)$ -values of 50 SNPs in the APOE region by LReg and MGLReg with different $(\tilde{p}_0, \tilde{p}_1)$ combinations (Figure 2). The curves shift to left as $(\tilde{p}_0, \tilde{p}_1)$ decreases. It indicates that most significant SNPs in this region detected by LReg are considered unimportant in normal people. Only those SNPs jointly detected by both LReg and MGLReg with all $(\tilde{p}_0, \tilde{p}_1)$ settings have significant population-level genetic effects on the hippocampi volume size.

Since the genetic measurements were on different platforms, we do an interaction analysis to test its potential differences and consequences on inference. Specifically, we repeat the experiment above, but adding an interaction term between phase status and genetic factor into the covariates set. We include the p -values of testing the interaction term for the top SNPs in Table 4. We observe that the genetic data acquired at the two phases do not have a significant difference based on the p -values. Figures 3 and 4 present the Manhattan plots of the GWAS results based on the left and right hippocampi by all the 6 017 259 SNPs to give a global view of the genetic effects and their variation as the global prevalence rate varies.

5 | DISCUSSION

The aim of this article is to develop a general regression framework based on the conditional model for the secondary outcome given the multigroup status and covariates and its relationship with the population regression of interest of the secondary outcome given covariates. It allows us to reduce the effect of sampling bias on the association between a certain genetic factor G and secondary trait Y in multigroup studies. Our method shares a similar idea with the traditional weighted likelihoods method such as IPW in correcting the weights of subjects in multiple groups, but it

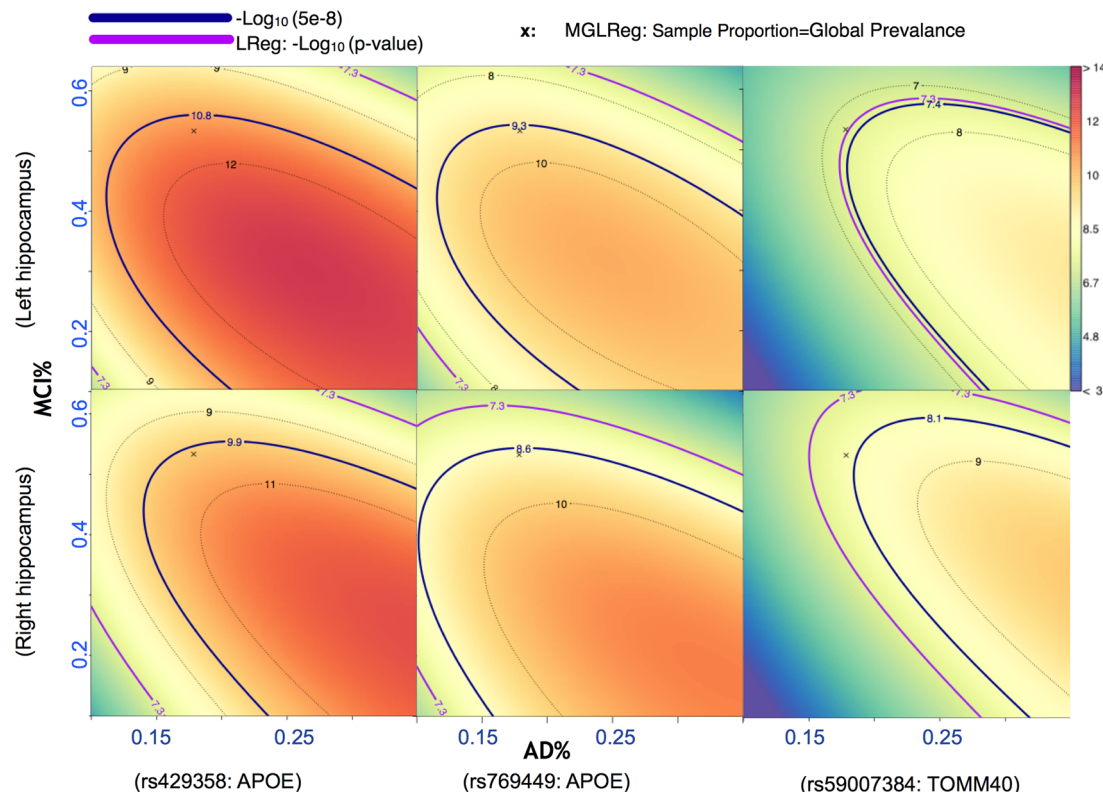


FIGURE 1 The heatmaps of $-\log_{10}(p)$ -value for three selected single-nucleotide polymorphisms by MGLReg with different global Alzheimer's disease (AD) and mild cognitive impairment (MCI) prevalence rates in the whole population [This figure appears in color in the electronic version of this article, and any mention of color refers to that version]

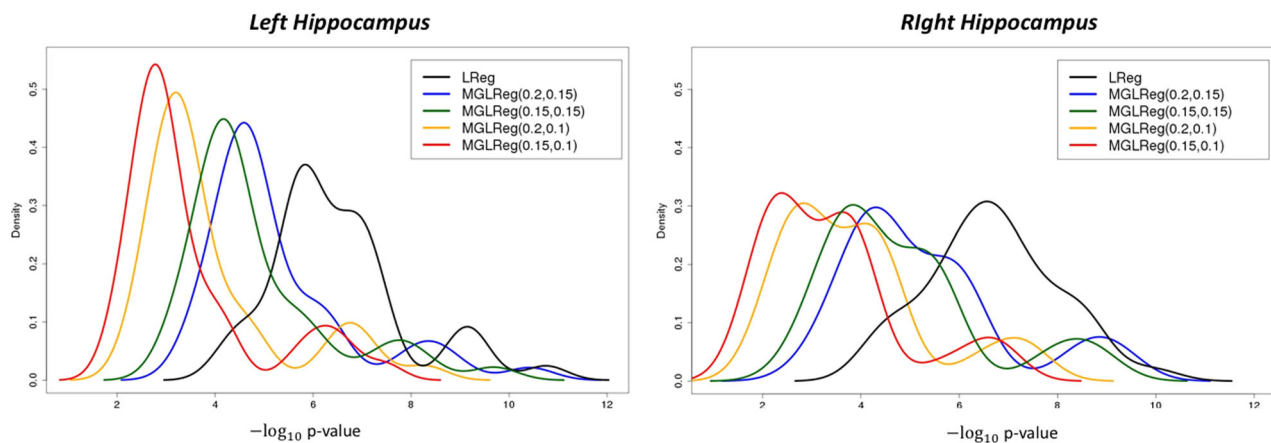


FIGURE 2 The density curves of $-\log_{10}(p)$ -values of top 50 APOE region single-nucleotide polymorphisms by each method for the left and right hippocampus volumes [This figure appears in color in the electronic version of this article, and any mention of color refers to that version]

outperforms IPW in terms of smaller estimation bias and type-I error rate. The GWAS experiment clearly demonstrates how the global prevalence rates influence the effects of covariates on the secondary outcome. Our experiment provides more evidence that

rs429358 and rs769449 have whole-population-level genetic effects on the volume sizes of left and right hippocampi. On the other hand, other top SNPs detected by LReg may be caused by the sampling bias by our method.

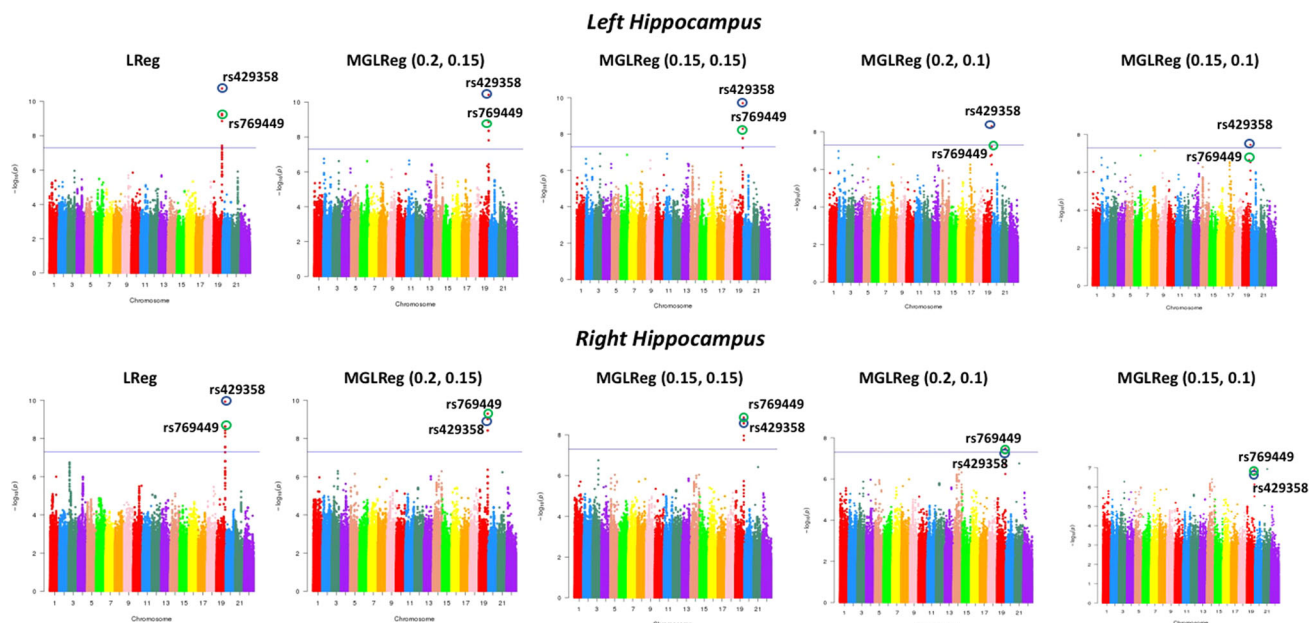


FIGURE 3 The Manhattan plots of the $-\log_{10}(p)$ -values by LReg and MGLReg on all 22 chromosomes for the left and right hippocampus volumes [This figure appears in color in the electronic version of this article, and any mention of color refers to that version]

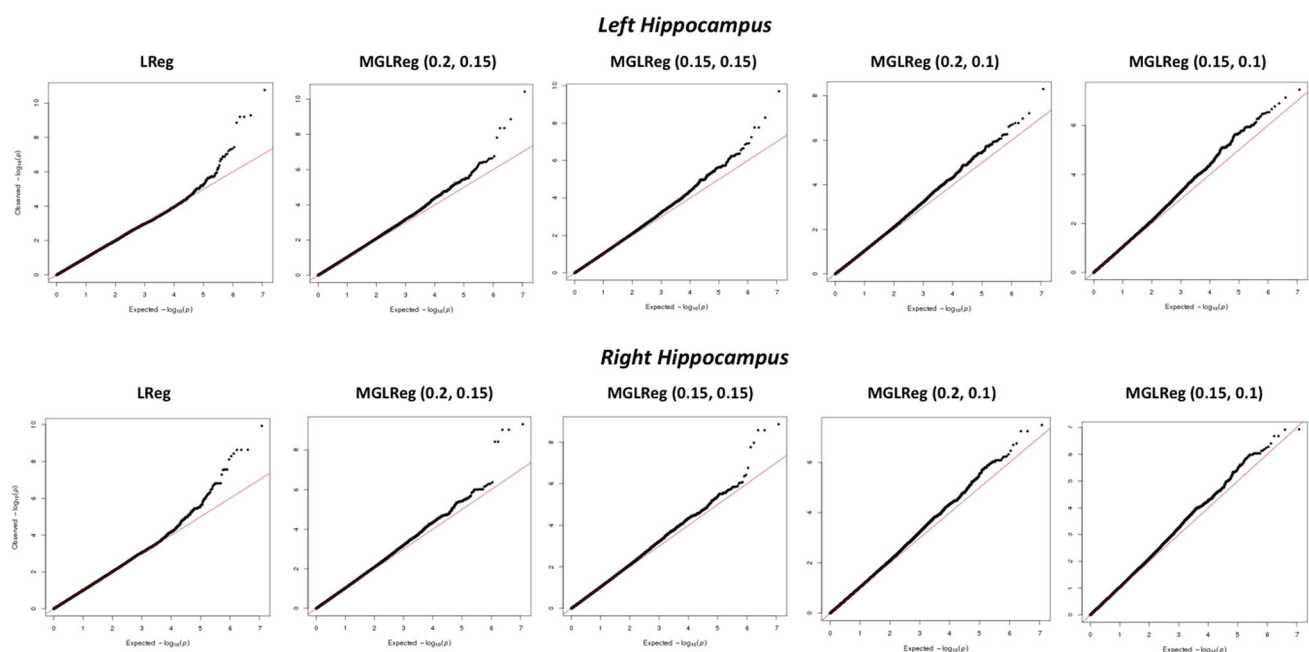


FIGURE 4 The QQ plots of the $-\log_{10}(p)$ -values by LReg and MGLReg on all 22 chromosomes for the left and right hippocampus volumes [This figure appears in color in the electronic version of this article, and any mention of color refers to that version]

ACKNOWLEDGMENTS

This work was partially supported by U.S. NIH grants MH086633 and MH092335, and a grant from the Cancer Prevention Research Institute of Texas, and grants P01 CA142538 (P01) and P30ES010126 (CEHS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Data used in the preparation of this article were obtained from

the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (<http://adni.loni.usc.edu>). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

ORCID

Fan Zhou  <http://orcid.org/0000-0003-2874-2148>

Haibo Zhou  <http://orcid.org/0000-0002-2468-0093>

Hongtu Zhu  <http://orcid.org/0000-0002-6781-2690>

REFERENCES

- ADNI. *Alzheimer's disease neuroimaging initiative*. Available at: <http://www.adni-info.org/> or <http://adni.loni.usc.edu>. [Accessed March 5, 2019].
- Alzheimer's Association. (2012) 2012 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 8, 131–168.
- Arpawong, T.E., Pendleton, N., Mekli, K., McArdle, J.J., Gatz, M., Armoskus, C. *et al.* (2017) Genetic variants specific to aging-related verbal memory: insights from GWASS in a population-based cohort. *PLoS One*, 12, e0182448.
- Breslow, N.E., Robins, J.M. and Wellner, J.A. (2000) On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli*, 6, 447–455.
- Cervantes, S., Samaranch, L., Vidal-Taboada, J.M., Lamet, I., Bullido, M.J., Frank-Garca, A. *et al.* (2011) Genetic variation in APOE cluster region and Alzheimer's disease risk. *Neurobiology of Aging*, 32, 2107–e7.
- Cruchaga, C., Kauwe, J.S., Harari, O., Jin, S.C., Cai, Y., Karch, C.M. *et al.* (2013) GWAS of cerebrospinal fluid tau levels identifies risk variants for Alzheimer's disease. *Neuron*, 78, 256–268.
- He, J., Li, H., Edmondson, A.C., Rader, D.J. and Li, M. (2012) A Gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies. *Biostatistics*, 13, 1–8.
- Jiang, Y., Scott, A.J. and Wild, C.J. (2006) Secondary analysis of case-control data. *Statistics in Medicine*, 25, 1323–1339.
- Kim, D.H., Payne, M.E., Levy, R.M., MacFall, J.R. and Steffens, D.C. (2002) Apoe genotype and hippocampal volume change in geriatric depression. *Biological Psychiatry*, 51, 426–429.
- Kim, J., Pan, W. and Alzheimer's Disease Neuroimaging Initiative. (2015) A cautionary note on using secondary phenotypes in neuroimaging genetic studies. *NeuroImage*, 121, 136–145.
- Lee, A., McMurchy, L. and Scott, A. (1997) Re-using data from case-control studies. *Statistics in Medicine*, 16, 1377–1389.
- Lin, D.Y. and Zeng, D. (2009) Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology*, 33, 256–265.
- Lu, P.H., Thompson, P.M., Leow, A., Lee, G.J., Lee, A., Yanovsky, I. *et al.* (2011) Apolipoprotein e genotype is associated with temporal and hippocampal atrophy rates in healthy elderly adults: a tensor-based morphometry study. *Journal of Alzheimer's Disease*, 23, 433–442.
- MGLREG. *R package*. Available at: <https://github.com/BIG-S2/MGLREG/>. [Accessed March 5, 2019].
- Potkin, S.G., Macciardi, F., Guffanti, G., Fallon, J.H., Wang, Q., Turner, J.A. *et al.* (2010) Identifying gene regulatory networks in schizophrenia. *NeuroImage*, 53, 839–847.
- Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38, 904–909.
- Richardson, D.B., Rzehak, P., Klenk, J. and Weiland, S.K. (2007) Analyses of case-control data for additional outcomes. *Epidemiology*, 18, 441–445.
- Schifano, E.D., Li, L., Christiani, D.C. and Lin, X. (2013) Genome-wide association analysis for multiple continuous secondary phenotypes. *The American Journal of Human Genetics*, 92, 744–759.
- Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L., Trojanowski, J. *et al.* (2009) Mri of hippocampal volume loss in early Alzheimer's disease in relation to APOE genotype and biomarkers. *Brain*, 132, 1067–1077.
- Shen, L., Kim, S., Risacher, S.L., Nho, K., Swaminathan, S., West, J.D. *et al.* (2010) Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: a study of the ADNI cohort. *NeuroImage*, 53, 1051–1063.
- Sofer, T., Cornelis, M.C., Kraft, P. and Tchetgen Tchetgen, E. (2017) Control function assisted IPW estimation with a secondary outcome in case-control studies. *Statistica Sinica*, 27, 785–804.
- Song, X., Ionita-Laza, I., Liu, M., Reibman, J. and We, Y. (2016) A general and robust framework for secondary traits analysis. *Genetics*, 202, 1329–1343.
- Tchetgen Tchetgen, E. (2014) A general regression framework for a secondary outcome in case-control studies. *Biostatistics*, 5, 117–128.
- Wei, J., Carroll, R.J., Müller, U.U., Keilegom, I.V. and Chatterjee, N. (2013) Robust estimation for homoscedastic regression in the secondary analysis of case-control data. *Journal of the Royal Statistical Society: Series B*, 75, 185–206.
- Zhu, W., Yuan, Y., Zhang, J., Zhou, F., Knickmeyer, R.C., Alzheimer's Disease Neuroimaging Initiative *et al.* (2017) Genome-wide association analysis of secondary imaging phenotypes from the Alzheimer's disease neuroimaging initiative study. *NeuroImage*, 146, 983–1002.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 2, 3, and 4, proofs, and detailed description of the real data are available with this paper at the Biometrics website on Wiley Online Library. We developed the R package MGLREG as our companion software, which is publicly available from github (see reference MGLREG).

How to cite this article: Zhou F, Zhou H, Li T, Zhu H. Analysis of secondary phenotypes in multigroup association studies. *Biometrics*. 2019;1–13. <https://doi.org/10.1111/biom.13157>