Received: 19 March 2019 | Accepted: 13 April 2020

DOI: 10.1111/biom.13286

BIOMETRIC PRACTICE

A penalized structural equation modeling method accounting for secondary phenotypes for variable selection on genetically regulated expression from PrediXcan for Alzheimer's disease

Ting-Huei Chen[1,2] | Hanaa Boughal[3]

[1]Département de mathématiques et de statistique, Université Laval, Québec, Canada

[2]Cervo Brain Research Centre, Québec, Canada

[3]École d'Actuariat, Université Laval, Québec, Canada

Correspondence
Ting-Huei Chen, Département de mathématiques et de statistique, Université Laval, Québec, Canada; and Cervo Brain Research Centre, Québec, Canada.
Email: ting-huei.chen@mat.ulaval.ca

Funding information
Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: Discovery Grants

Abstract

As the global burden of mental illness is estimated to become a severe issue in the near future, it demands the development of more effective treatments. Most psychiatric diseases are moderately to highly heritable and believed to involve many genes. Development of new treatment options demands more knowledge on the molecular basis of psychiatric diseases. Toward this end, we propose to develop new statistical methods with improved sensitivity and accuracy to identify disease-related genes specialized for psychiatric diseases. The qualitative psychiatric diagnoses such as case control often suffer from high rates of misdiagnosis and oversimplify the disease phenotypes. Our proposed method utilizes endophenotypes, the quantitative traits hypothesized to underlie disease syndromes, to better characterize the heterogeneous phenotypes of psychiatric diseases. We employ the structural equation modeling using the liability-index model to link multiple genetically regulated expressions from PrediXcan and the manifest variables including endophenotypes and case-control status. The proposed method can be considered as a general method for multivariate regression, which is particularly helpful for psychiatric diseases. We derive penalized retrospective likelihood estimators to deal with the typical small sample size issue. Simulation results demonstrate the advantages of the proposed method and the real data analysis of Alzheimer's disease illustrates the practical utility of the techniques. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative database.

KEYWORDS

Alzheimer's disease, Alzheimer's Disease Neuroimaging Initiative, penalized estimation, structural equation model

1 | INTRODUCTION

Late-onset Alzheimer's disease (LOAD) is the most common cause of age-related dementia. Despite the significant investments in therapeutic drug discovery, no effective drugs are available to cure Alzheimer's disease (AD). Apparently, identifying the molecular mechanisms fundamentally involved in AD's pathogenesis is the primary step for the discovery of drug targets. However, due to ethical reasons, it is nearly impossible to conduct experiments on human subjects to evaluate causation. Therefore, we normally rely on human epidemiological studies to compare the molecular profiles in cases versus controls such as genome-wide association studies (GWAS) and differential expression analysis to identify disease-associated variants and genes. However, qualitative diagnosis such as AD patients (cases) and non-AD individuals (controls) often suffers from high rates of misdiagnosis and oversimplifies the disease phenotypes due to the complex and highly heterogeneous psychiatric syndromes. Unlike physical diseases that are diagnosed mainly by medical tests, the diagnosis of psychiatric diseases is majorly symptom-based using the criteria in a diagnosis manual (Diagnostic and Statistical Manual of Mental Disorders). In other words, the AD or non-AD diagnosis is an aggregate variable of various cognitive and physical measurements. This oversimplification of the disease phenotypes causes a loss of information and then a loss of statistical power to identify important variables associated with the disease.

Endophenotypes, introduced into the behavioral sciences by Dr. Irv Gottesman (Gottesman and Shields, 1982), are the quantitative traits hypothesized to underlie disease syndromes. They are believed to better capture the human behavioral abnormalities than the imprecise categorical psychiatric diagnoses (Almasy and Blangero, 2001). Allen *et al.* (2009) provided a review of literatures on endophenotypes for schizophrenia and Reitz and Mayeux (2009) summarized several endophenotypes genetic studies for AD. However, there has been no statistical framework available to model a group of endophenotypes (Kendler and Neale, 2010) and the disease status (case-control) appropriately. Furthermore, the sampling design of genetic studies causes another potential issue on endophenotypes analysis. In practice, most genetic studies employ the case-control design for a particular disease, which consists of a sample of cases (ie, diseased individuals) and a sample of controls (ie, disease-free individuals). Under the sampling design based on the dichotomic classification on a particular disease, the case-control status is defined as the primary phenotype and the measurements of endophenotypes are considered as secondary phenotypes. Most publications associate endophenotypes with genetic variables by a linear regression analysis. This approach may lead to bias estimation for the genetic effect because the population association between genetic variables and endophenotypes can be distorted in the case-control sample. Therefore, similar to the work of Lin and Zeng (2009), we use a retrospective likelihood method to address this issue.

AD has a relatively high heritability, estimated in a range of 58-79% (Gatz *et al.*, 2006), suggesting that the genetic components highly involve in AD's pathogenesis. However, those identified genetic variants (single nucleotide polymorphism [SNP]) from large-scale GWAS on LOAD are very difficult to evaluate their functional mechanisms. The methodology PrediXcan (Gamazon *et al.*, 2015) makes it possible to estimate the tissue-dependent gene expression profiles driven by SNP variations, ie, genetically regulated expression (GReX). It also evaluates the association between the disease and GReX to provide more insights on the functional mechanism of the disease-associated variants. Moreover, the disease-associated GReX, if exist, are more informative than those from the traditional differential gene expression analysis since they do not suffer from the issue of reverse causation, ie, expression level is altered by disease status. Our method exploits the advantages of gene-based analysis. Unlike the method of Gamazon *et al.* (2015) using a univariate analysis for each gene individually, our proposed method has a better statistical power through the modeling of correlations among GReXs and utilizes a penalized estimation approach to deal with the issue of small sample size relative to the number of genes.

Kendler and Neale (2010) provided two types of models to relate an endophenotype and a psychiatric disease to genetic components as mediational and liability-index models. A liability-index model specifies that both the disease and endophenotype share some genetic components, while a mediational model specifies that the effect of genetic components on the disease pass through the endophenotype. In practice, it is difficult to design experimental studies on humans to distinguish them validly. However, as argued by Kendler and Neale (2010), the mediational model poses a stronger assumption than the liability-index one in that the former assumes that the genetic effects on a psychiatric disease are exclusively via the endophenotype. Therefore, we choose to model the relationship among multiple endophenotypes and a psychiatric disease based on the liability-index model.

In Section 2, we introduce our proposed methods and their implementation. In Section 3, simulation studies are presented to demonstrate that the information loss resulting from the dichotomization of quantitative disease liability reduces statistical power considerably. The proposed methods show significantly improved performances over conventional methods. Finally, the proposed methods are illustrated in a real data analysis using the datasets from the Alzheimer's Disease Neuroimaging Initiative (ADNI) in Section 4 and conclude in Section 5 with a discussion.

2 | METHODS

2.1 | Model specification

For N independent samples, let $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{iJ})$ denote a vector of J observed quantitative endophenotypes for individual i, where $1 \leq i \leq N$, and each of the endophenotype variables is assumed to be standardized with mean 0 and variance equal to 1. Let $\boldsymbol{D}_i$ and ξ_i denote the disease status ($\boldsymbol{D}_i = 1$ for case and $\boldsymbol{D}_i = 0$ for control) and the latent disease genetic liability, respectively, for each sample i. For gene expression (GReX) data, let $\boldsymbol{X}_i = (X_{i1}, \dots, X_{iP})$ denote the GReXs for individual i, where P is the number of genes; they are assumed to be normalized and jointly follow a multivariate normal distribution with mean 0 and variance-covariance matrix Σ_X. Without loss of generality, we assume that the first S variables of $\boldsymbol{X}_i$ are important (ie, they have nonzero effects on the response variable) and denote this by $\boldsymbol{X}_{i1}$ and let the remaining $P - S$ variables be $\boldsymbol{X}_{i2}$, such that $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2})$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\top}, \boldsymbol{\beta}_2^{\top} = \boldsymbol{0}^{\top})$ be the associated coefficients vector for $\boldsymbol{X}_i$. In addition, let $\boldsymbol{Z}_i = (Z_{i1}, \dots, Z_{iQ})$ denote the Q nongenetic variables. It is assumed that the number of nongenetic variables Q, the number of disease-related GReX S, and the number of manifest variables J are much smaller than N, while the number of genetic variables P is usually larger than N.

As mentioned in Section 1, we extend the liability-index model in Kendler and Neale (2010) to relate the manifest variables including multiple endophenotypes and the disease status to a latent disease genetic liability. We use a linear regression model, which specifies that the conditional distribution of $\boldsymbol{Y}_i$ given $(\xi_i, \boldsymbol{Z}_i)$ is a multivariate normal with mean $\boldsymbol{\mu}_{\boldsymbol{Y}_i} = \boldsymbol{\Lambda}\xi_i + \sum_{l=1}^{Q} \boldsymbol{\Gamma}_l z_{il}$, where $\boldsymbol{\Lambda} = (\lambda_1, \dots, \lambda_J)^{\top}$ is a factor loading vector and $\boldsymbol{\Gamma}_l = (\gamma_{l1}, \dots, \gamma_{lJ})^{\top}$ is a vector of regression coefficients for lth variable of $\boldsymbol{Z}$, and the matrix of variance-covariance $\Sigma_Y = \text{diag}(\sigma^2_{Y_1}, \dots, \sigma^2_{Y_J})$. In other words, the endophenotypes $(Y_{i1}, \dots, Y_{iJ})$ are conditionally independent given the latent disease genetic liability ξ_i and the nongenetic covariates $\boldsymbol{Z}_i$. The latent disease genetic liability is further modeled as a normal variable with mean as a linear combination of GReXs $\mu_{\xi_i} = \boldsymbol{X}_i\boldsymbol{\beta}$ and with variance fixed to be one to ensure the identifiability, a typical strategy in factor analysis. In addition, we use a logistic regression model for D_i: $P(D_i = 1|\boldsymbol{Z}_i, \xi_i) = \frac{\exp^{\{\alpha_0 + \sum_{l=1}^{Q} Z_{il}\alpha_l + \alpha_\xi \xi_i\}}}{1+\exp^{\{\alpha_0 + \sum_{l=1}^{Q} Z_{il}\alpha_l + \alpha_\xi \xi_i\}}}$, where α_0 is the intercept, $(\alpha_1, \dots, \alpha_Q)$ are the regression coefficients of variable $\boldsymbol{Z}$ and α_ξ is the regression coefficient of ξ. A graphical representation of this structural equation model is given in Figure 1.

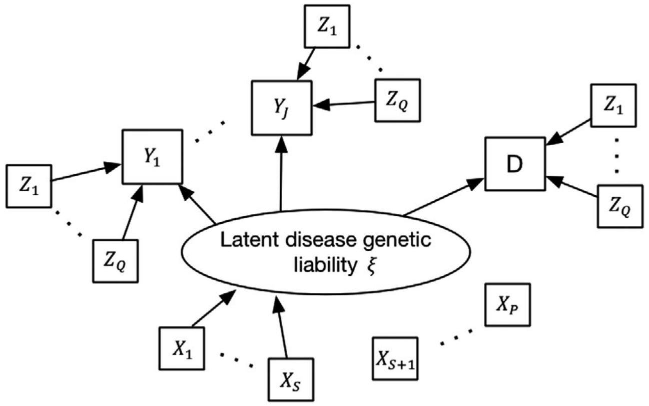

FIGURE 1 The specified structural equation model for the relationship among the manifest variables $\boldsymbol{Y}$ and $\boldsymbol{D}$ and genetic variables $\boldsymbol{X}$ and nongenetic covariates $\boldsymbol{Z}$. The latent genetic disease liability is drawn as circle, while the manifest or measured variables are shown as squares. In addition, the arrows represent the regression relationship, and for simplicity, the symbols for regression coefficients are omitted

Since the sampling is conditional on the categorical diagnoses (case-control), the likelihood function for complete data $(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{X}_i, D_i, \xi_i)$ will have a retrospective form: $\prod_{i=1}^{N} P(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{X}_i, \xi_i | D_i)$, which is

$$\prod_{i=1}^{N} \left\{ \frac{P(D_i = 1, \boldsymbol{Y}_i | \boldsymbol{Z}_i, \boldsymbol{X}_i, \xi_i) P(\xi_i | \boldsymbol{Z}_i, \boldsymbol{X}_i) P(\boldsymbol{Z}_i, \boldsymbol{X}_i)}{P(D_i = 1)} \right\}^{D_i}$$

$$\times \left\{ \frac{P(D_i = 0, \boldsymbol{Y}_i | \boldsymbol{Z}_i, \boldsymbol{X}_i, \xi_i) P(\xi_i | \boldsymbol{Z}_i, \boldsymbol{X}_i) P(\boldsymbol{Z}_i, \boldsymbol{X}_i)}{P(D_i = 0)} \right\}^{1-D_i},$$

where $P(D_i, \boldsymbol{Y}_i | \boldsymbol{Z}_i, \boldsymbol{X}_i, \xi_i) = P(D_i | \boldsymbol{Z}_i, \boldsymbol{X}_i, \xi_i) P(\boldsymbol{Y}_i | \boldsymbol{Z}_i, \boldsymbol{X}_i, \xi_i)$ and $P(\boldsymbol{Z}_i, \boldsymbol{X}_i) = P(\boldsymbol{Z}_i) P(\boldsymbol{X}_i)$ by the specified liability-index model assumptions. In addition,

$$P(D_i = 1) =$$

$$\iiint P(D_i = 1 | \boldsymbol{Z}_i, \boldsymbol{X}_i, \xi_i) P(\xi_i | \boldsymbol{Z}_i, \boldsymbol{X}_i) P(\boldsymbol{Z}_i, \boldsymbol{X}_i)\, d\boldsymbol{Z}_i\, d\boldsymbol{X}_i\, d\xi_i.$$

2.2 | EM algorithm for parameters estimation

For parameters estimation, let $\boldsymbol{\Gamma}$ denote $(\boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_Q)$, and let $\boldsymbol{\theta}$ denote the vector of parameters $(\boldsymbol{\Lambda}, \Sigma_Y, \boldsymbol{\Gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \alpha_\xi)$ and let $(\boldsymbol{D}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\xi})$ denote the data for N individuals. The penalized log likelihood function for complete data $pl(\boldsymbol{\theta} | \boldsymbol{D}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\xi})$ is defined as

$$\log \prod_{i=1}^{N} P(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{X}_i, \xi_i | D_i) - P_\vartheta(\boldsymbol{\beta}),$$

where $P_\vartheta(\boldsymbol{\beta})$ denotes a penalty function imposed on $\boldsymbol{\beta}$, and ϑ denotes the tuning parameter(s) of the penalty function.

When the disease is rare, $P(D_i = 0 | \boldsymbol{Z}_i, \xi_i) \approx P(D_i = 0) \approx 1$ and $P(D_i = 1 | \boldsymbol{Z}_i, \xi_i) \approx \exp^{\{\alpha_0 + \sum_{l=1}^{Q} Z_{il}\alpha_l + \alpha_\xi \xi_i\}}$. This condition applies to our motivating example since the prevalence of AD is less than 2%. In addition, the simulation study in Section 3 is based on the prevalence as 5% and the proposed methods show good performances. Thus, the following estimation procedures are based on the rare disease assumption using a penalized log likelihood function as

$$pl(\boldsymbol{\theta}|\boldsymbol{D},\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{X},\boldsymbol{\xi}) = \log\prod_{i=1}^{N}\left\{P(\boldsymbol{Y}_i|\boldsymbol{Z}_i,\boldsymbol{X}_i,\xi_i)P(\xi_i|\boldsymbol{Z}_i,\boldsymbol{X}_i)P(\boldsymbol{Z}_i,\boldsymbol{X}_i)\right\} + \log\prod_{i=1}^{N}\left\{\frac{\exp^{\{\alpha_0+\sum_{l=1}^{Q} Z_{il}\alpha_l+\alpha_\xi\xi_i\}}}{P(D_i=1)}\right\}^{D_i} - P_\vartheta(\boldsymbol{\beta}),$$

where

$$P(D_i=1) = \left\{e^{\alpha_0+\frac{1}{2}\alpha_\xi^2}e^{\frac{1}{2}(\alpha_\xi\beta)^\top\Sigma_X(\alpha_\xi\beta)}\right\}\boldsymbol{M}_{\boldsymbol{Z}}(\boldsymbol{\alpha}),$$

and $\boldsymbol{M}_{\boldsymbol{Z}}(\boldsymbol{\alpha})$ is the moment-generating function for $\boldsymbol{Z}$, and $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_Q)^\top$.

Since the parameter α_0 cancels in the numerator and denominator of the likelihood function, it is unidentifiable. For the remaining parameters, we use a expectation-maximization with a combination of the coordinate descent algorithm (Friedman *et al.*, 2010a) for parameters estimation. Define the Q function as $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E_{\xi|(\theta^{(t)},Y,Z,X)}(pl(\boldsymbol{\theta}|\boldsymbol{D},\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{X},\boldsymbol{\xi}))$. For simplicity, we use E_ξ to denote $E_{\xi|(\theta^{(t)},Y,Z,X)}$ for the following presentations.

Initialization:

To initialize $\boldsymbol{\Lambda}$ and Σ_Y, one latent factor model is applied to $\boldsymbol{Y}$ to obtain $(\hat{\lambda}_1^{(0)},\dots,\hat{\lambda}_J^{(0)})$ and $(\hat{\sigma}_{Y_1}^{2(0)},\dots,\hat{\sigma}_{Y_J}^{2(0)})$. The remaining parameters are initialized as zero.

E-step:

$$\begin{cases} E_\xi^{(t)}(\xi_i) = \dfrac{\left(\sum_{j=1}^{J} y_{ij}\frac{\hat{\lambda}_j^{(t)}}{\hat{\sigma}_{Y_j}^{2(t)}}+X_i^\top\hat{\boldsymbol{\beta}}^{(t)}\right)}{\left(\sum_{j=1}^{J}\frac{\hat{\lambda}_j^{2(t)}}{\hat{\sigma}_{Y_j}^{2(t)}}+1\right)} & \text{if } D_i = 0 \\ E_\xi^{(t)}(\xi_i) = \dfrac{\left(\hat{\alpha}_\xi^{(t)}+\sum_{j=1}^{J} y_{ij}\frac{\hat{\lambda}_j^{(t)}}{\hat{\sigma}_{Y_j}^{2(t)}}+X_i^\top\hat{\boldsymbol{\beta}}^{(t)}\right)}{\left(\sum_{j=1}^{J}\frac{\hat{\lambda}_j^{2(t)}}{\hat{\sigma}_{Y_j}^{2(t)}}+1\right)} & \text{if } D_i = 1 \end{cases},$$

and $E_\xi^{(t)}(\xi_i^2) = (\sum_{j=1}^{J}\frac{\hat{\lambda}_j^{2(t)}}{\hat{\sigma}_{Y_j}^{2(t)}}+1)^{-1} + E_\xi^{(t)}(\xi_i)^2$.

M-step:

The maximum likelihood estimation (MLE) for λ_j is then the solution of

$$\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\lambda_j} = \sum_{i=1}^{N}\left\{-Y_{ij}E_\xi^{(t)}(\xi_i)+\lambda_j E_\xi^{(t)}\left(\xi_i^2\right)\right\} = 0,$$

that is, $\hat{\lambda}_j = \frac{\sum_{i=1}^{N} Y_{ij}E_\xi^{(t)}(\xi_i)}{\sum_{i=1}^{N} E_\xi^{(t)}(\xi_i^2)}$. The MLE for σ_j^2 is $\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{N}\{Y_{ij}^2-2Y_{ij}\hat{\lambda}_j E_\xi^{(t)}(\xi_i)+\hat{\lambda}_j^2 E_\xi^{(t)}(\xi_i^2)\}}{N}$ and the MLE of α_ξ is $\hat{\alpha}_\xi = \frac{\sum_{i=1}^{n} D_i E_\xi(\xi_i)}{N_1(1+\hat{\boldsymbol{\beta}}^T\hat{\Sigma}_X\hat{\boldsymbol{\beta}})}$.

To estimate $\boldsymbol{\beta}$, the coordinate descent algorithm is used to update each β_m sequentially. We provide two different estimators for $\boldsymbol{\beta}$ using the penalty functions Lasso (Tibshirani, 1996) and Log (Sun *et al.*, 2010; Friedman, 2012) respectively. Lasso is one of the most popular penalty functions for variable selection in high-dimensional variables problem. We first present the Lasso estimator β_m^{Lasso} using the Lasso penalty function $P_\vartheta(\boldsymbol{\beta}) = P_{(\lambda)}(\boldsymbol{\beta}) = \sum_{m=1}^{P}\lambda|\beta_m|$, where $\lambda > 0$ is a tuning parameter. Letting $\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})/\partial\beta_m = 0$, the estimate of β_m^{Lasso} is

$$\begin{cases} \hat{\beta}_m = 0 & \text{if } \left|\omega_m^{(t)}\right| \le \lambda \\ \hat{\beta}_m = \operatorname{sgn}\left(\hat{\beta}_m^{(t)}\right)\left[\left|\omega_m^{(t)}\right| - \lambda\right] & \text{if } \left|\omega_m^{(t)}\right| > \lambda, \end{cases}$$

where $\omega_m^{(t)} = \frac{\sum_{i=1}^{N} E_\xi^{(t)}(\xi_i)X_{im}-\sum_{m'\neq m}\hat{\beta}_{m'}^{(t)}\hat{\sigma}_{mm'}^2(N+N_1\hat{\alpha}_\xi^{2(t)})}{\hat{\sigma}_{mm}^2(N+N_1\hat{\alpha}_\xi^{2(t)})}$, where $N_1 = \sum_{i=1}^{N} D_i$.

Although Lasso has demonstrated its utility in various applications, the variable selection consistency of Lasso requires the irrepresentable condition (Zhao and Yu, 2006) that there is no strong correlation between the "*important covariates*," which have nonzero effects and the "*unimportant covariates*," which have zero effects. Since some of the gene expressions are highly correlated, this condition may not be satisfied in our study. Therefore, we introduce another estimator for $\boldsymbol{\beta}$ using the Log penalty. Compared to some existing penalty functions, the Log penalty has shown its advantages in variable selection in genetic studies, where the variables are highly correlated (Sun *et al.*, 2010; Chen *et al.*, 2016). The Log estimator β_m^{Log} using the Log penalty function $P_\vartheta(\boldsymbol{\beta}) = P_{(\alpha,\tau)}(\boldsymbol{\beta}) = \sum_{m=1}^{P}\alpha\log(|\beta_m|+\tau)$, where $\alpha > 0$ and $\tau > 0$ are tuning parameters. Since the Log penalty is nonconcave, a local linear approximation (LLA) (Zou and Li, 2008) is applied to it: $p_{(\alpha,\tau)}(|\beta_m|) \approx \frac{\alpha|\beta_m|}{|\hat{\beta}_m|+\tau}$, where $\hat{\beta}_m$ is the estimate in the previous iteration. Letting $\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})/\partial\beta_m = 0$, the estimate of β_m^{Log} is

$$\begin{cases} \hat{\beta}_m = 0 & \text{if } \left|\omega_m^{(t)}\right| \le \frac{\alpha}{\left|\hat{\beta}_m^{(t)}\right|+\tau} \\ \hat{\beta}_m = \operatorname{sgn}\left(\hat{\beta}_m^{(t)}\right)\left[\left|\omega_m^{(t)}\right| - \frac{\alpha}{\left|\hat{\beta}_m^{(t)}\right|+\tau}\right] & \text{if } \left|\omega_m^{(t)}\right| > \frac{\alpha}{\left|\hat{\beta}_m^{(t)}\right|+\tau} \end{cases}.$$

Since $\boldsymbol{\Gamma}$ is not the parameter of interest, they can be eliminated through the residualization of $\boldsymbol{Y}$ using the variables of $\boldsymbol{Z}$ to simplify the estimation procedure.

The estimation of $(\alpha_1, \ldots, \alpha_Q)$ depends on the characteristics of the variables that are included in the model. For instance, in the section of real data analysis, the model includes two variables for $\boldsymbol{Z}$: age for $\boldsymbol{Z}_1$ and gender for $\boldsymbol{Z}_2$. For a continuous variable like $\boldsymbol{Z}_1$, it is assumed to be standardized and follow a standard normal distribution; the MLE of α_1 is $\hat{\alpha}_1 = \frac{\sum_{i=1}^{N} D_i Z_{i1}}{N_1}$. For the categorical variable that has only two categories like $\boldsymbol{Z}_2$, it is assumed to follow a Bernoulli distribution with probability P_{α_2}; the MLE of α_2 is $\hat{\alpha}_2 = \log\{\frac{\sum_{i=1}^{N} D_i Z_{i2}(1-\hat{P}_{\alpha_2})}{\hat{P}_{\alpha_2}(N_1-\sum_{i=1}^{N} D_i Z_{i2})}\}$.

We then iteratively apply the estimation procedure until convergence is reached. Empirically, we consider that convergence is achieved if the maximum difference in the coefficient estimates between consecutive iterations is less than a predefined threshold, eg, 10^{-4}.

2.3 | Tuning parameters and model selection

The estimator of parameters of interest $\boldsymbol{\theta}$ depends on the values of tuning parameters ϑ. In practice, a set of numerical tuning values are provided for model fitting. Then a model selection criterion is applied to select the optimal model. To ensure a good performance of a penalization method, proper numerical tuning values have to be provided to cover the optimal one that can penalize the estimates of the coefficients of the unimportant variables to zero, and keep those of the important variables to be nonzero. With failure to do so, the penalization method would perform poorly, either having too strong a penalization threshold so that too few important variables are being selected, or having too weak a threshold so that too many unimportant variables are being selected.

Generally, the scales of the tuning values depend on the scales of the estimates of the parameters to be penalized, that is, $\hat{\boldsymbol{\beta}}$ in this study. However, since the parameter $\boldsymbol{\beta}$ is associated with the latent unobserved factor $\boldsymbol{\xi}$, to identify a vector of proper tuning values is less straightforward than the usual regression analysis on observed variables. We propose to fit a model without any penalty on a set of genes whose number is much smaller than the sample size to identify the scales of the estimates of the parameters. To choose such set of genes, we use the marginal association analysis on each of the genes and each of the observed variables including the endophenotypes and the disease case-control status. Each gene m is associated with a vector of $(J+1)$ P-values from the J marginal linear regression analyses on J endophenotypes and from the logistic regression analysis on the case-control status. Then the genes are ordered based on the minimum P-value across the $(J+1)$ marginal P-values from the most significant to the least. Let a denote the number of the smallest integer value that is greater than or equal to 20% of the sample size N, and the first a gene based on the ordering by the minimum P-values is used for model fitting without penalty. Finally, the maximum of the obtained unpenalized estimates of the regression coefficients in absolute value can be used to generate the tuning values. Specifically, let $\hat{b}_{\max}$ denote this maximum value. For the Lasso penalty, a vector of tuning values for λ is generated as a sequence of a particular length, say 100, from $\hat{b}_{\max} \times 0.1$ to $\hat{b}_{\max} \times 1.5$. For the Log penalty, a vector of the tuning values for τ is set as $(1, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001) \times \hat{b}_{\max}$. For each value of τ, a vector of the tuning values for α is generated as a sequence of a particular length, say 100, from $(\tau \times \hat{b}_{\max} \times 0.005)$ to $(\tau \times \hat{b}_{\max} \times 1.5)$.

For model selection, according to the empirical results from the simulation studies, where the number of variables is comparable to the sample size, the regular Bayesian information criterion (BIC) (Schwarz *et al.*, 1978) performs well. When the number of variables is much larger than the sample size, the extended BIC (Chen and Chen, 2008) can be used for model selection.

3 | SIMULATIONS

3.1 | Procedure for data simulation and methods for comparison

The motivating application is to select important GReXs associated with AD using the datasets from ADNI. Apparently, the influential factors on the empirical performances are sample size, the numbers of candidate genes ($\boldsymbol{P}$) and truly associated ones ($\boldsymbol{S}$), the magnitude of effect sizes, and the correlation structure of GReXs. To evaluate the performances of our proposed methods subject to these factors, the number of genes and sample sizes considered in this simulation study is based on the datasets of ADNI. In addition, to faithfully reproduce the correlation structure of GReXs, we simulate datasets based on the estimates of the correlation matrix of GReX obtained in the real data analysis.

For data simulation, the variable $\boldsymbol{X}$ (gene expression) is simulated from a multivariate normal distribution with mean $\boldsymbol{0}$ and a correlation matrix using the estimates of the correlation matrix of GReXs. Then the latent disease liability ξ is simulated from a linear model: $\boldsymbol{\xi} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}_{N\times N})$ and $\boldsymbol{I}_{N\times N}$ is an identity matrix. The sample size N is specified as 600 or 1200, and the numbers of genes $\boldsymbol{P}$ are specified as 2498 or 5980, the smallest and the largest number of GReXs generated by PrediXcan from a brain tissue. To specify the nonzero components in $\boldsymbol{\beta}$, we consider S as 20 or 40, and the effect sizes for these variables to be the replication of a vector of $b = (0.35, 0.40, 0.45, 0.50, 0.55, -0.35, -0.40, -0.45, -0.50, -0.55)$.

To simulate manifest variable Y (endophenotypes), we consider two scenarios: *Ideal* and *Realistic* ones. For *Ideal* scenario, Y are simulated based on the specified model in Figure 1: $\boldsymbol{Y}_{N\times J} = \boldsymbol{\xi}_{N\times 1}\Lambda_{1\times J} + \boldsymbol{E}_{N\times J}$, where $\boldsymbol{E} \sim MVN(\mathbf{0}, \boldsymbol{I}_{N\times N})$. On the other hand, for *Realistic* scenario, Y are allowed to have their own unique set of associated genes that are not involved with the studied disease. $\boldsymbol{Y} = (y_{ij})$ are simulated from: $y_{ij} = \xi_i\lambda_j + \boldsymbol{X}_{(s_j)i}\boldsymbol{b}_j + e_{ij}$, where $\boldsymbol{X}_{(s_j)i}$ denotes the unique set of associated genes for manifest variable, j, $\boldsymbol{b}_j$ are the corresponding coefficients, and e_{ij} is a standard normal variable. For both scenarios, we set $\Lambda = (0.30, 0.35, 0.40)$ for $J = 3$. For the *Realistic* scenario, the number of genes that are uniquely associated with each of the manifest variables is set as 10% of S and the coefficient for each of them is set as 0.35.

The case-control status D is simulated from a logistic regression model: $P(D_i = 1|\xi_i) = \frac{\exp^{\{\alpha_0+\alpha_\xi\xi_i\}}}{1+\exp^{\{\alpha_0+\alpha_\xi\xi_i\}}}$, where α_ξ is set as .5, and α_0 is the set for each setting, respectively, to make the prevalence about 5%. To generate random samples using the case-control sampling scheme, we first simulate a large population based on the aforementioned model structure. Then a random case-control sample is generated from a random selection of $N/2$ cases from the sample units with $D_i = 1$ and $N/2$ controls from the sample units with $D_i = 0$ in the simulated population.

We evaluate our methods in addition to several other approaches for variable selection. As described in Section 1, the conventional way to identify the disease-associated genes is differential expression analysis to run association analysis for each gene separately, denoted by DEA. Another approach is univariate regression analysis using the Lasso penalty on the case-control status to select disease-associated genes. It has been implemented in the R/GLMNET package (Friedman *et al.*, 2010b), denoted by ULasso. For studies with endophenotypes, the conventional approach is to use marginal linear regression on each of the endophenotypes or to use marginal logistic regression on the disease case-control status to identify the associated genes. Since there are $(J + 1)$ marginal P-values corresponding to the J endophenotypes and the disease case-control status, respectively, the Fisher's combined probability test (Fisher, 2006) is used to compute the combined P-value for each gene, denoted by Fisher. Another considered method is based on a group-wise penalized estimation. Under the model in Figure 1, all observed responses $\boldsymbol{Y}$ share the same genetic components, and thus, one can use a multivariate regression model to identify the disease-associated genes. When the number of variables is larger than the sample size, the group-wise penalized estimation is a natural choice, which, by borrowing information from multiple response variables, can discover genes that are weakly associated with multiple response variables. Group Lasso (Yuan and Lin, 2006) is used for methods comparison. The R package R/GLMNET also implements the Group Lasso penalty for multivariate Gaussian responses, denoted by GLasso.

3.2 | Simulation results

We use the false discoveries rate (FDR) and true positive rate (TPR) to evaluate the performance of the methods, where FDR and TPR are calculated as FD/D and TD/S, respectively (D, TD, and FD are the numbers of discoveries, true discoveries, and false discoveries). Our proposed methods (LLasso and LLog), ULasso, and GLasso employ tuning parameters for model fitting, and 1000 tuning values are given for each of the methods to obtain 1000 models, respectively. The R package implemented for ULasso and GLasso by GLMNET uses cross-validation (CV) for model selection. Therefore, we present the results for ULasso and GLasso using CV for model selection in the main text and present the results using BIC in the Supporting Information. For the methods Fisher and DEA, we use several q-value cutoffs for variable selection and the model with relatively similar values of TPR to other methods is presented in the main text and the results for all q-value cutoffs are presented in the Supporting Information.

Table 1 shows the results for the eight settings varying the parameters of S, P, and N in the *Ideal* scenario. As the signals contained in simulated data are relatively complicated, that is, the number of variables (P) is large or the number of true associated variables (S) is large, LLog performs better than LLasso. The performances of our proposed methods are in general better than the alternative methods. For instance, for the setting with $N = 1200$, $P = 2498$, and $S = 40$, LLog and LLasso have TPR higher than 98%, which is higher than the alternative methods, and they also have much lower false discovery rates. DEA has a lower power and a higher false discovery rate compared to the proposed methods. With endophenotypes data, the combined Fisher method can improve the power to detect the disease-related genes compared to DEA. However, it has a much higher false discovery rate than the proposed methods due to the failure of not taking into account the correlation among the genes and the fact that the observed endophenotypes come from a conditional sampling on case-control status. ULasso and GLasso have similar issues to DEA and Fisher.

The conclusions are similar for the *Realistic* scenario as shown in Table 2, where the endophenotypes are associated with some genes that are not involved with the genetic risk for AD. The performances of our proposed methods are still much better than the alternative methods. In summary, the

TABLE 1 Simulation results for *Ideal* scenario. The first column indicates the number of important covariates (S), the number of candidate covariates (P), and the sample size (N). For each method, we present the median of the true positive rate (TPR) × 100% and its standard error (in parentheses), and the median of the false discovery rate (FDR) × 100% (in brackets) and its standard error (in parentheses) across 100 simulations

Setting	LLog[a]	LLasso[a]	Fisher[b]	DEA[b]	ULasso[c]	GLasso[c]
$S = 20; P = 2498$	88 (16)	90 (21)	70 (8.8)	0 (3.3)	45 (17)	100 (0.86)
$N = 600$	[0 (2.5)]	[13.6 (7.7)]	[38.1 (11)]	[0 (36)]	[82.1 (8.7)]	[84.8 (3.8)]
$S = 20; P = 2498$	100 (2.5)	100 (0)	95 (3.6)	30 (9.8)	85 (7.2)	100 (0)
$N = 1200$	[0 (1.1)]	[9.09 (5.9)]	[60.4 (5.8)]	[20 (15)]	[81.9 (4.4)]	[83.9 (3.6)]
$S = 20; P = 5980$	75 (19)	55 (31)	60 (8.7)	0 (3.6)	35 (14)	95 (4)
$N = 600$	[5.72 (10)]	[11.1 (12)]	[40 (8.7)]	[0 (35)]	[86.4 (11)]	[87.7 (3.0)]
$S = 20; P = 5980$	100 (4.1)	100 (2.4)	90 (4.2)	35 (11)	72.5 (8.7)	100 (0.98)
$N = 1200$	[0 (3.6)]	[18.7 (7.7)]	[75 (4.3)]	[28.6 (16)]	[85.9 (3.9)]	[86.5 (2.9)]
$S = 40; P = 2498$	65 (19)	5 (19)	62.5 (5.3)	12.5 (8.2)	52.5 (8.2)	100 (1.6)
$N = 600$	[3.18 (3.8)]	[0 (5.5)]	[58.1 (9.9)]	[28.6 (21)]	[79 (4.6)]	[82.2 (2.8)]
$S = 40; P = 2498$	98.8 (2.3)	100 (1.8)	88.8 (3.5)	50 (7.6)	80 (5.3)	100 (0.25)
$N = 1200$	[2.44 (1.8)]	[20 (5.4)]	[79.4 (3.2)]	[53.7 (11)]	[78.5 (3)]	[80.9 (2.5)]
$S = 40; P = 5980$	50 (22)	2.5 (9.4)	45 (5.8)	2.5 (3.1)	32.5 (8.1)	95 (2.9)
$N = 600$	[6.16 (19)]	[0 (7.8)]	[59.5 (8.7)]	[0 (36)]	[86.1 (5)]	[87.2 (2.0)]
$S = 40; P = 5980$	100 (3.3)	95 (17)	77.5 (4.7)	25 (7.8)	67.5 (6.9)	100 (0.25)
$N = 1200$	[0 (2)]	[26.5 (8.9)]	[80.7 (2.7)]	[38.9 (14)]	[83.5 (2.7)]	[86.3 (1.7)]

[a] Using BIC for the model selection criterion.

[b] Using q-values for the model selection criterion.

[c] Using cross-validation for the model selection criterion.

TABLE 2 Simulation results for *Realistic* scenario. The first column indicates the number of important covariates (S), the number of candidate covariates (P), and the sample size (N). For each method, we present the median of the true positive rate (TPR) × 100% and its standard error (in parentheses), and the median of the false discovery rate (FDR) × 100% (in brackets) and its standard error (in parentheses) across 100 simulations

Setting	LLog[a]	LLasso[a]	Fisher[b]	DEA[b]	ULasso[c]	GLasso[c]
$S = 20; P = 2498$	90 (16)	85 (34)	60 (9.8)	0 (4.2)	45 (16)	100 (1.4)
$N = 600$	[14.3 (17)]	[26.1 (21)]	[45 (8.4)]	[0 (32)]	[82.1 (8.9)]	[87.9 (2.1)]
$S = 20; P = 2498$	100 (7.3)	100 (0)	95 (5)	35 (12)	85 (7.6)	100 (0)
$N = 1200$	[13.6 (5.9)]	[33.3 (6.7)]	[64 (3.8)]	[26.1 (16)]	[81.5 (5)]	[87.6 (2.3)]
$S = 20; P = 5980$	75 (19)	65 (27)	60 (8)	0 (4.5)	40 (14)	95 (3.8)
$N = 600$	[10.3 (23)]	[16.7 (25)]	[51.9 (8)]	[0 (29)]	[86.8 (7.5)]	[90.2 (1.9)]
$S = 20; P = 5980$	95 (9.9)	95 (6.1)	85 (6)	30 (9.9)	70 (9)	100 (0.98)
$N = 1200$	[17.4 (8.6)]	[37.3 (8.2)]	[77.4 (4)]	[30 (16)]	[85 (5.2)]	[90.0 (1.7)]
$S = 40; P = 2498$	85 (22)	10 (33)	55 (6.5)	10 (6.6)	52.5 (7.7)	100 (1.9)
$N = 600$	[25 (33)]	[0 (31)]	[58.7 (7.9)]	[20 (20)]	[78.1 (5.1)]	[86.2 (1.6)]
$S = 40; P = 2498$	95 (7.5)	95 (5.8)	85 (4.2)	53.8 (7.9)	82.5 (5.3)	100 (0.49)
$N = 1200$	[15.9 (6.9)]	[38.8 (10)]	[82.8 (3)]	[55.7 (12)]	[77.1 (3.4)]	[85.2 (1.6)]
$S = 40; P = 5980$	50 (25)	5 (26)	37.5 (6.6)	2.5 (3.2)	35 (8)	95 (3.7)
$N = 600$	[30.3 (34)]	[0 (31)]	[55.9 (11)]	[0 (26)]	[85.2 (4.7)]	[90.2 (1.10)]
$S = 40; P = 5980$	97.5 (9.9)	55 (25)	70 (5.8)	25 (6.9)	62.5 (6.4)	100 (0.43)
$N = 1200$	[16.1 (5.9)]	[21.6 (18)]	[79.1 (3.3)]	[44.7 (14)]	[84.4 (3.2)]	[89.8 (0.88)]

[a] Using BIC for the model selection criterion.

[b] Using q-values for the model selection criterion.

[c] Using cross-validation for the model selection criterion.

simulation analysis demonstrates that our proposed methods have much more advantageous empirical performances than the alternative ones. In addition, as described above, we also include the simulation results using BIC for model selection criterion for ULasso and GLasso, and conclusions are also consistent to those using CV.

4 | REAL DATA ANALYSIS

Data used in preparation of this article were obtained from the ADNI database (see Acknowledgements section for more details). The list of secondary phenotypes in this analysis is cognitive endophenotypes that have been used in several AD genetic studies summarized by Reitz and Mayeux (2009). They are not only highly correlated with the disease status but also can further reveal the variation of subject's cognitive function within each disease-status. Most of studies employed univariate regression analysis on each of the endophenotype separately. As our proposed method makes it possible to model multiple endophenotypes simultaneously to improve statistical power, we selected the cognitive endophenotypes that have been considered in the literature and were measured for most of the subjects with percentages of missing value less than 15% in ADNI dataset. This gave a list of clinical attributes measured at baseline included in the model: Clinical Dementia Rating Sum of Boxes, properties of the 11-item ADAS-cog, properties of the 13-item ADAS-cog, Mini-Mental State Examination, immediate score of Rey Auditory Verbal Learning Test, learning score of Rey Auditory Verbal Learning Test, and the percent forgetting score of Rey Auditory Verbal Learning Test, where ADAS-cog refers to Alzheimer's Disease Assessment Schedule - Cognition. For the imputation of missing values, we used the multiple imputation method based on fully conditional specification implemented in the R package MICE (van Buuren and Groothuis-Oudshoorn, 2011).

To obtain the gene expression (GReX) in brain tissues, we used the method PrediXcan. The number of generated GReXs varies by brain tissue from 2002 to 5980. The number of subjects with both the endophenotypes and GReX is 812. In addition, since there are only 48 AD subjects in this dataset, we group these subjects with 483 mild cognitive impairment (MCI) subjects together as 531 AD-MCI subjects versus 281 normal control subjects. We applied all the methods mentioned in the simulation study to each set of GReXs from the 13 brain regions to identify the genes associated with the AD-MCI phenotypes while accounting for the effects of unpenalized covariates, age, and gender. As the simulation results show that LLog performs better than LLasso and outperforms the rest methods, we describe its findings in the followings, while the remaining results are shown in Tables S6-S14 in Supporting Information. Among the 13 brain tissues, the number of identified GReXs by LLog ranges from 9 to 24. This analysis may answer research questions such as "what are the genes associated with the AD-MCI phenotype in a particular brain tissue?" or "what are the genes associated with the AD-MCI phenotype across multiple brain tissues?" The list of identified genes in any particular brain tissue can be found in Table S9 in Supporting Information. Here, we report the list of identified genes across brain tissues using the approach of majority vote. In total, there are seven AD-MCI-associated genes in more than three brain tissues. The identified gene in most of the brain issues is VPS11, which plays an important role in autophagy (Zhang *et al.*, 2016) that has been shown to involve extensively in AD (Nixon *et al.*, 2005). Another identified gene is NPHP3, which is involved with canonical Wnt-signaling, and the activation of Wnt signaling has shown to be able to protect against Aβ neurotoxicity and to improve cognitive performance in AD patients (Vallée and Lecarpentier, 2016). The gene LRRC37A3 has shown to be associated with frontotemporal dementia and immune-mediated disease in microglia (Broce *et al.*, 2018). Another identified gene is SP1, which regulates the expression of several AD-related proteins and is considered to be one of the therapeutic targets in AD (Citron *et al.*, 2008). The remaining genes are NDE1, BIN3, and NUDT14, where the first gene has an essential role in the cerebral cortex neurogenesis (Bakircioglu *et al.*, 2011), but the rest have not shown to be related to AD.

5 | DISCUSSION

In this work, we address several important issues mentioned in Section 1 for genetic association analysis on psychiatric diseases; the numerical analyses have demonstrated its advantages compared to the conventional approaches. Our method provides reliable results when the endophenotypes carry similar genetic components for disease risk. In practice, this condition can be obtained from subjective knowledge on the properties of the endophenotypes. Even when the data deviate from this ideal scenario such that each endophenotype has its own associated genetic variables that are not disease-related, our method still provides the most satisfactory results compared to the conventional approaches as demonstrated by the simulation analysis. As our method relates the effect of genetic variables to the latent disease liability or phenotypes, we cannot interpret their coefficients in the same way as in the regular regression analysis on observed variables. However, since the objective is to select the disease-associated genes for wet-lab validations, our method is able to provide such information since it gives good TPR and low false discovery rate. Note that in order to have an easier interpretation on the identified genetic expression to either increase or decrease the disease risk with the increment of their expression levels, the measurements of endophenotypes will either keep their original values or be multiplied by -1 so that they will be all positively correlated with the severity of the disease syndromes. For future studies, it will be of interest to extend the framework to a more complex model such as a two-factor model to relate the gene expressions and the manifest variables and/or to model the nondisease-related gene expression associated with each endophenotype explicitly to improve the performances in variable selection.

ACKNOWLEDGMENTS

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC)(RGPIN/05880-2016).

Data used in preparation of this article were obtained from the 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.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are available from ADNI. Restrictions apply to the availability of these data, which were used under license for this study. Data are available http://adni.loni.usc.edu with the permission of ADNI.

ORCID

Ting-Huei Chen https://orcid.org/0000-0002-7731-2374

REFERENCES

Allen, A.J., Griss, M.E., Folley, B.S., Hawkins, K.A. and Pearlson, G.D. (2009) Endophenotypes in schizophrenia: a selective review. *Schizophrenia research*, 109, 24–37.

Almasy, L. and Blangero, J. (2001) Endophenotypes as quantitative risk factors for psychiatric disease: rationale and study design. *American Journal of Medical Genetics*, 105, 42–44.

Bakircioglu, M., Carvalho, O.P., Khurshid, M., Cox, J.J., Tuysuz, B., Barak, T. et al. (2011) The essential role of centrosomal nde1 in human cerebral cortex neurogenesis. *The American Journal of Human Genetics*, 88, 523–535.

Broce, I., Karch, C.M., Wen, N., Fan, C.C., Wang, Y., Tan, C.H. et al. (2018) Immune-related genetic enrichment in frontotemporal dementia: an analysis of genome-wide association studies. *PLoS Medicine*, 15, e1002487.

Chen, J. and Chen, Z. (2008) Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95, 759–771.

Chen, T.-H., Sun, W. and Fine, J.P. (2016) Designing penalty functions in high dimensional problems: the role of tuning parameters. *Electronic Journal of Statistics*, 10, 2312–2328.

Citron, B.A., Dennis, J.S., Zeitlin, R.S. and Echeverria, V. (2008) Transcription factor sp1 dysregulation in Alzheimer's disease. *Journal of Neuroscience Research*, 86, 2499–2504.

Fisher, R.A. (2006) *Statistical Methods for Research Workers*. Cosmo Publications: Genesis Publishing Pvt Ltd 354.

Friedman, J. (2012) Fast sparse regression and classification. *International Journal of Forecasting*, 28, 722–738.

Friedman, J., Hastie, T. and Tibshirani, R. (2010a) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1.

Friedman, J., Hastie, T. and Tibshirani, R. (2010b) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.

Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J. et al. (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47, 1091–1098.

Gatz, M., Reynolds, C.A., Fratiglioni, L., Johansson, B., Mortimer, J.A., Berg, S. et al. (2006) Role of genes and environments for explaining Alzheimer disease. *Archives of General Psychiatry*, 63, 168–174.

Gottesman, I.I. and Shields, J. (1982) *Schizophrenia*. CUP Archive.

Kendler, K.S. and Neale, M.C. (2010) Endophenotype: a conceptual analysis. *Molecular Psychiatry*, 15, 789–797.

Lin, D. and Zeng, D. (2009) Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33, 256–265.

Nixon, R.A., Wegiel, J., Kumar, A., Yu, W.H., Peterhoff, C., Cataldo, A. et al. (2005) Extensive involvement of autophagy in alzheimer disease: an immuno-electron microscopy study. *Journal of Neuropathology & Experimental Neurology*, 64, 113–122.

Reitz, C. and Mayeux, R. (2009) Endophenotypes in normal brain morphology and Alzheimer's disease: a review. *Neuroscience*, 164, 174–190.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.

Sun, W., Ibrahim, J. and Zou, F. (2010) Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. *Genetics*, 185, 349.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Vallée, A. and Lecarpentier, Y. (2016) Alzheimer disease: crosstalk between the canonical Wnt/Beta-catenin pathway and PPARS alpha and gamma. *Frontiers in Neuroscience*, 10, 459.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011) mice: multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45, 1–67.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49–67.

Zhang, J., Lachance, V., Schaffner, A., Li, X., Fedick, A., Kaye, L.E. et al. (2016) A founder mutation in vps11 causes an autosomal recessive leukoencephalopathy linked to autophagic defects. *PLoS Genetics*, 12, e1005848.

Zhao, P. and Yu, B. (2006) On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7, 2541–2563.

Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36, 1509–1533.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 3.2 and 4 are available with this paper at the Biometrics website on Wiley Online Library. Our code is also available at the Biometrics website on Wiley Online Library. The R package for the proposed methods can be found in https://github.com/THstat22/PENLatent.

How to cite this article: Chen T-H, Boughal H. A penalized structural equation modeling method accounting for secondary phenotypes for variable selection on genetically regulated expression from PrediXcan for Alzheimer's disease. *Biometrics*. 2020;1–10. https://doi.org/10.1111/biom.13286