Published in final edited form as:

Front Math China. 2013 June 1; 8(3): 731–743. doi:10.1007/s11464-012-0256-8.

A nonparametric regression method for multiple longitudinal phenotypes using multivariate adaptive splines

Wensheng ZHU^{1,2} and Heping ZHANG²

¹Key Laboratory for Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China

²Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520-8034, USA

Abstract

In genetic studies of complex diseases, particularly mental illnesses, and behavior disorders, two distinct characteristics have emerged in some data sets. First, genetic data sets are collected with a large number of phenotypes that are potentially related to the complex disease under study. Second, each phenotype is collected from the same subject repeatedly over time. In this study, we present a nonparametric regression approach to study multivariate and time-repeated phenotypes together by using the technique of the multivariate adaptive regression splines for analysis of longitudinal data (MASAL), which makes it possible to identify genes, gene-gene and gene-environment, including time, interactions associated with the phenotypes of interest. Furthermore, we propose a permutation test to assess the associations between the phenotypes and selected markers. Through simulation, we demonstrate that our proposed approach has advantages over the existing methods that examine each longitudinal phenotype separately or analyze the summarized values of phenotypes by compressing them into one-time-point phenotypes. Application of the proposed method to the Framingham Heart Study illustrates that the use of multivariate longitudinal phenotypes enhanced the significance of the association test.

Keywords

Multivariate phenotypes; longitudinal data analysis; genetic association test; multivariate adaptive regression splines

1 Introduction

In genetic studies of many complex diseases, particularly mental illnesses, and behavior disorders, it is common that researchers collect multiple phenotypes related to a disorder as well as other disorders that are potentially correlated with the primary disorder of interest [21]. On one hand, these phenotypes may be mutually correlated, and some or all of them may share a common genetic mechanism. On the other hand, these phenotypes may also be time-dependent, since some phenotypes are sometimes collected from the same subject

repeatedly over time, which is known as longitudinal study design. A well-known example is the Framingham Heart Study (FHS), where hundreds of phenotypes related to the Cardiovascular disease (CVD) were collected repeatedly every two to four years over time. Hence, it is useful to analyze multiple phenotypes with repeated measures simultaneously in the genetic association studies.

Several methods have been developed to test the association between multivariate phenotypes and genetic markers. Lange et al. [8] proposed a multivariate extension of family-based association tests based on generalized estimation equation (FBAT-GEE) to analyze multiple phenotypes simultaneously [9]. Xu et al. [13] suggested combining all univariate trait-specific test statistics based on individual phenotypes for a global association test. Furthermore, Zhu and Zhang [21] performed a series of simulations to enhance the understanding on how beneficial analyzing multiple phenotypes jointly is and when it makes the most sense to analyze them simultaneously. Their simulation results demonstrated that it is advantageous to analyze multiple phenotypes jointly, especially when the genes affect more than one phenotype and when there exist causal relations between multivariate phenotypes. Recently, Zhang et al. [18] constructed a nonparametric test based on the generalized Kendall's tau to accommodate any combination of dichotomous, ordinal, and quantitative traits. Zhu et al. [20] extended their nonparametric test to adjust for covariate effects.

However, some existing methods cannot deal with the longitudinal data directly, and they require repeated measures to be summarized into one-time-point trait by taking the average of several measures or by other transformations [19]. Although the FBAT-GEE can be applied to longitudinal studies in theory, it is conveniently designed to include time-varying covariates and gene-time interaction. Furthermore, most methods only considered the main effects of genes but did not accommodate interactions among genes and between genes and environmental factors when analyzing multiple phenotypes simultaneously; yet, it has been documented that considering those interactions will increase the chance of identifying susceptible loci in the association studies [1,3].

Zhang [15] proposed a nonparametric regression method to analyze longitudinal data, called multivariate adaptive splines for analysis of longitudinal data (MASAL). MASAL is an automatic procedure for fitting longitudinal data without specifying the full regression surface, and it can directly be used to establish the relationship between univariate longitudinal phenotype and genetic markers while adjusting for environmental and time-dependent covariates.

In the present paper, we extend the MASAL to establish the relationship between multivariate, repeated measured phenotypes and genetic markers as well as other non-genetic covariates in the study of complex disorders. To analyze multiple phenotypes jointly, we must acknowledge the fact that different phenotypes of the same subject share the common non-time covariates (say, the common genotypes and the common gender), but these common covariates may have different effects on different phenotypes. In addition, in order for the MASAL to distinguish the different effects of the common covariates, we present a suitable way to re-code covariates for different phenotypes. Moreover, based on

the fitted MASAL model, we provide a formal testing procedure for assessing the significance of gene, gene-gene, gene-environment, and gene-time interactions in association with multiple phenotypes. To demonstrate the benefit of our proposed method, we carry out simulation studies and analyze the data from the Framingham Heart Study (FHS).

2 Method

2.1 MASAL for univariate phenotype (univariate MASAL)

We first briefly review MASAL presented by Zhang [15] and demonstrate its application in genetic association studies. Let Y be the univariate phenotype, and let $X = (x_1, \dots, x_p)^T$ be the p covariates that consist of the measured environmental factors E and genotypes G measured on a series of SNPs, with each genotype coded as 0, 1, and 2 representing the number of copies of one allele at this marker. For the i th individual ($i = 1, \dots, n$), suppose that Y and X are repeatedly measured at T_i different time points $t_{i1}, \dots, t_{iT_i}$, where $x_{l,ij}$ and y_{ij} are respectively the measurements of the l th covariate x_l ($l = 1, \dots, p$) and the observed value of the phenotype Y at time t_{ij} , $j = 1, \dots, T_i$. We note, however, that the genotypes of each individual are identical over time.

To establish the relationship between phenotype Y and the p environmental and genotypic covariates, we consider a nonparametric model

$$y_{ij} = f(x_{1,ij}, \dots, x_{p,ij}, t_{ij}) + e_{ij}, \quad (1)$$

where f is an unknown smooth function and e_{ij} is the error term for $j = 1, \dots, T_i$, $i = 1, \dots, n$.

To fit model (1), we adopt the MASAL method presented by Zhang [15–17]. MASAL selects a model from the following class of functions:

$$\left\{ f; f(\mathbf{x}) = \sum_{k=0}^K \beta_k B_k(\mathbf{x}), K=0, 1, \dots \right\}, \quad (2)$$

where $B_k(\mathbf{x})$ is a special basis function of the p covariates $\mathbf{x} = (x_1, \dots, x_p)^T$, and β_k is the regression coefficient of $B_k(\mathbf{x})$ ($k = 0, 1, \dots, K$). All basis functions are made of the following two functions:

$$(x_l - \tau)^+, x_l \quad (l=1, \dots, p), \quad (3)$$

where τ is called a knot and $a^+ = \max(a, 0)$ for any number of a . Because $B_k(\mathbf{x})$ can either be one of the functions in equation (3) or the product of those functions involving distinct covariates such as

$$(x_2 - \tau_2)^+ \times x_3 (x_4 - \tau_3)^+,$$

the MASAL model can accommodate not only gene-gene interactions but also gene-environment and gene-time interactions in the genetic association studies.

To select a model from (2), MASAL uses a forward step first and then a backward step. In the forward step, all knots are found and terms are added to minimize the (weighted) sum of squared residuals

$$WLS = \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T \mathbf{W}_i^{-1} (\mathbf{y}_i - \hat{\mathbf{y}}_i),$$

where $\mathbf{y}_i = (y_{i1}, \dots, y_{iT_i})^T$ and $\hat{\mathbf{y}}_i$ is the predicted value of $\mathbf{y}_i$, and $\mathbf{W}_i$ is the within-subject covariance matrix for

$$\mathbf{e}_i = (e_{i1}, \dots, e_{iT_i})^T, \quad i = 1, \dots, n.$$

In the backward step, based on generalized cross-validation (GCV), MASAL deletes one least significant term at a time from the large model obtained after the forward step. The final model is the one that yields the smallest

$$GCV = WLS_k \cdot \left[1 - \frac{\lambda k + 1}{\sum_{i=1}^n T_i} \right]^{-2},$$

where WLS_k is the WLS of a reduced model with k terms and λ is the penalizing parameter for model complexity, which is usually chosen between three and five [2,16].

2.2 MASAL for multivariate phenotypes (multivariate MASAL)

In this subsection, we will discuss how to use MASAL to analyze multiple phenotypes. Let $\mathbf{Y} = (Y_1, \dots, Y_M)^T$ denote the M phenotypes of interest. Our goal is to assess the effect of X on the multiple phenotypes $\mathbf{Y}$. Similar to the case of univariate phenotype, each phenotype of $Y_1, \dots, Y_M$ is measured repeatedly over time. Although measurements may be taken at different times for different phenotypes for each subject, for clarity, we assume that M phenotypes are measured repeatedly over the same time points $t_{i1}, \dots, t_{iT_i}$ for the i th individual, where T_i is the number of repeated measurements, $i = 1, \dots, n$.

Let y_{ijm} denote the observed value of the m th phenotype Y_m at time t_{ij} for the i th individual, and $X = (x_1, \dots, x_p)^T$ still denote the p covariates that consist of the measured environmental factors E and genotypes G measured on a series of SNPs. Although different phenotypes of each subject share the common environmental factors as well as the common genotypes, these common factors may have different effects on different phenotypes. To distinguish the different effects of X for different phenotypes, we should re-code different covariates for each one. For the m th phenotype Y_m , we define the corresponding covariates as

$$X_m = \mathbf{1}_m \otimes (x_1, \dots, x_p),$$

where $\otimes$ represents the Kronecker product and $\mathbf{1}_m$ is an M -dimensional row-vector consisting of $(M - 1)$ 0 and one 1 for the m th element. After recoding, the new covariates corresponding to each phenotype actually become a pM -dimensional row-vector. For example, if we consider two phenotypes ($M = 2$) and three covariates ($p = 3$), X_1 and X_2 could be coded as

$$(x_1, x_2, x_3, 0, 0, 0)^T, \quad (0, 0, 0, x_1, x_2, x_3)^T,$$

respectively. In general, let $x_{l,ijm}$ denote the value of the l th covariate ($l = 1, \dots, pM$) at time t_{ij} for the m th phenotype of the i th individual.

Now, the nonparametric model of interest becomes

$$y_{ijm} = f_m(x_{1,ijm}, \dots, x_{pM,ijm}, t_{ij}) + e_{ijm}, \quad (4)$$

where $m = 1, \dots, M, j = 1, \dots, T_i, i = 1, \dots, n$.

The procedure that is used by MASAL to fit model (4) is the same as the univariate-trait case, except the following modifications. First, the $\mathbf{x}$ in (2) now is $\mathbf{x} = (x_1, \dots, x_{pM})^T$ and $B_k(\mathbf{x})$ is a special basis function of the pM covariates $\mathbf{x} = (x_1, \dots, x_{pM})^T$. Second, the $\mathbf{y}_i$ in WLS should be

$$\mathbf{y}_i = (y_{i11}, \dots, y_{i1M}, \dots, y_{iT_i1}, \dots, y_{iT_iM})^T$$

and $\hat{\mathbf{y}}_i$ is the predicted value of $\mathbf{y}_i$. Furthermore, $\mathbf{W}_i$ should be the within-subject covariance matrix for

$$\mathbf{e}_i = (e_{i11}, \dots, e_{i1M}, \dots, e_{iT_i1}, \dots, e_{iT_iM})^T, \quad i = 1, \dots, n.$$

2.3 Fitted model and test procedure

According to above discussion, all the terms in the fitted MASAL model can be summarized as three categories: intercept, genetic covariate terms, and non-genetic covariate terms, where the genetic terms include main effect of genes, gene-gene, gene-environment, and gene-time interactions, and the non-genetic terms include environmental factors, time-dependent covariate, as well as interactions among them. In general, by re-arranging the order of terms, the final MASAL model can be summarized as

$$\hat{f}(\mathbf{x}) = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{B}^G(\mathbf{x}) + \hat{\boldsymbol{\gamma}}^T \mathbf{B}^{ET}(\mathbf{x}), \quad (5)$$

where

$$\mathbf{B}^G(\mathbf{x}) = (B_1^G(\mathbf{x}), \dots, B_{k_1}^G(\mathbf{x}))^T$$

represents k_1 genetic covariate terms, and

$$\mathbf{B}^{ET}(\mathbf{x}) = (B_1^{ET}(\mathbf{x}), \dots, B_{k_2}^{ET}(\mathbf{x}))^T$$

represents k_2 non-genetic covariate terms. $\hat{\alpha}$ is the estimate of intercept, and

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_{k_1})^T, \quad \hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \dots, \hat{\gamma}_{k_2})^T$$

are the estimates of the corresponding regression coefficients.

To assess the significance of gene, gene-gene, gene-environment, and gene-time interactions in association with traits, we propose a Wald-type statistic to test whether or not $\boldsymbol{\beta} = 0$ based on the final MASAL model (5). The Wald-type statistic can be written as

$$W = \hat{\boldsymbol{\beta}}^T \left(\hat{\Sigma}_{\boldsymbol{\beta}, \boldsymbol{\beta}} \right)^{-1} \hat{\boldsymbol{\beta}}, \quad (6)$$

where $\hat{\Sigma}_{\boldsymbol{\beta}, \boldsymbol{\beta}}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$, which can be obtained with the MASAL model establishment.

As Zhang [15] pointed out, MASAL requires the identification of the adaptive knot allocation, and therefore, W no longer follows the chi-square distribution under the null hypothesis. We use a permutation procedure to establish the null distribution of W . The permutation is done by randomly assigning the genotypes while keeping the phenotype and non-genetic covariates intact for each individual. This permutation generates the data under the null hypothesis of no association between genetic covariates and the multiple phenotypes. For each permuted data set, we use MASAL to find the fitted model and then calculate the Wald-type statistic as we have done by using the original data set. We repeat this procedure 1000 times to generate the distribution of W under the null hypothesis of no association between genetic covariates and the multiple phenotypes.

3 Simulation

We carried out a series of simulations to demonstrate the promise of our method in genetic association studies of multiple phenotypes with repeated measures. To consider the linkage

disequilibrium (LD) in our simulations, we simulated multi-locus genotypes at 11 SNPs by randomly drawing pairs of haplotypes according to the haplotype frequencies given by Yeager et al. [14, Fig. 3]. Then, we assumed the fifth SNP as the disease locus that affects the multiple phenotypes under study. This disease causing SNP is used to generate the phenotypes but is not used for association studies. In addition to the genotypes, a non-genetic binary covariate, X , was generated with

$$\Pr(X=1)=0.3, \quad \Pr(X=0)=0.7.$$

To simulate the repeated-measure values of multiple phenotypes, for the m th phenotype, we consider a 3-dimensional function

$$f_m(X, g, t) = \alpha_m X + \beta_m g + \gamma_m t,$$

where $g = 0, 1$, or 2 represents the number of copies of the minor allele at the disease locus, $m = 1, \dots, M$. For simplicity, we only consider two phenotypes (i.e., $M = 2$) and each of them was measured at three different time points $t = 1, 2$, and 3 . For each subject, based on the obtained X and g , we generated the repeated-measure values of multivariate phenotypes

$$Y = (y_{11}, y_{21}, y_{12}, y_{22}, y_{13}, y_{23})^T$$

from the multivariate normal distribution $N_6(\mu, \Sigma)$, where

$$\mu = (f_1(X, g, 1), f_2(X, g, 1), f_1(X, g, 2), f_2(X, g, 2), f_1(X, g, 3), f_2(X, g, 3))^T,$$

$\Sigma = (\sigma_{mt, m't'}^2)_{6 \times 6}$ is the covariance matrix, and $\sigma_{mt, m't'}^2$ represents the covariance between y_{mt} and $y_{m't'}$, $m, m' = 1, 2$, $t, t' = 1, 2, 3$. In our simulations, we let the variance $\sigma_{mt, mt}^2 = 1$. Then the covariance

$$\sigma_{mt, m't'}^2 = \rho_{mt, m't'}.$$

For ease of interpretation, we refer to $\rho_{m1, m2}$, $\rho_{m2, m3}$, $\rho_{m1, m3}$ ($m = 1, 2$) as the within-trait correlation coefficients ρ_{WT} , and refer to the rest of coefficients as the between-trait correlation coefficients ρ_{BT} . For all simulation studies, we let $\alpha_m = 0.2$, $\gamma_m = 0.1$ for $m = 1, 2$, and let the within-trait correlation coefficients $\rho_{WT} = 0.2$.

For estimating the type I error, under the null hypothesis, the multiple phenotypes are not associated with the disease susceptibility locus. Then we let $\beta_m = 0$, $m = 1, 2$. To accommodate different covariance structures among multiple traits, we let the between-trait

correlation coefficients $\rho_{BT} = 0, 0.1, 0.2$, and 0.3 . To evaluate power, we let $\beta_m = 0.2$, $m = 1, 2$, and let the between-trait correlation coefficients

$$\rho_{BT} = -0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3.$$

The simulations were replicated 2000 times (2000 data sets) for type I error estimation and power evaluation, and each consisting of 400 subjects.

To demonstrate the advantages of analyzing the multivariate phenotypes with repeated measures simultaneously, we also used the following two methods to analyze the simulated data and compared the results of them with that of our proposed method (Multivariate MASAL). One method is that we first used Univariate MASAL to analyze each longitudinal phenotype separately, tested the associations between genetic markers and each of the phenotype using the Wald-type statistic, and then adjusted for multiple testing by using Bonferroni correction. The other method is that we first took the average value of each phenotype over all the time points and then used Multivariate MASAL to analyze the multiple average phenotypes without repeated measures. For the purpose of illustration, we refer to these two methods as Univariate MASAL and One-Time-Point MASAL (OTP-MASAL), respectively.

Table 1 provides the comparison results of type I error for Multivariate MASAL, Univariate MASAL, and OTP-MASAL. It is clear from Table 1 that the empirical type I errors of Multivariate MASAL are close to the nominal significance level. OTP-MASAL slightly inflates the type I error rate but it is still acceptable. For Univariate MASAL, the tests performed on each phenotype are not independent due to the correlations between the phenotypes, and this violation of the independence assumption should limit the ability of the Bonferroni adjustment to control type I error effectively. However, it is interesting to see that Univariate MASAL controls the type I error rates well according to the results of Table 1. We can also see from Table 1 that all three methods are very robust to different correlations among multiple phenotypes for the type I error estimation.

Table 2 provides the power comparison for three methods at the significance level 0.05. It appears that analyzing multiple longitudinal phenotypes jointly with Multivariate MASAL has the best power among all the cases shown in Table 2. This indicates that there is a substantial loss of power if we analyze each longitudinal phenotype separately by using Univariate MASAL or we analyze all the average value of each phenotype over all the time points when multiple longitudinal phenotypes are available in our research. The explanation is that Multivariate MASAL outperforms Univariate MASAL by avoiding Bonferroni correction. As we pointed out above, correlations among the phenotypes result in the dependence of the tests performed on each phenotype, which limits the ability of the Bonferroni adjustment for Univariate MASAL. In our simulation studies, two phenotypes are really correlated even when $\rho_{BT} = 0$, because they share the common factors (age, gene, and time) in the linear model used to simulate the data. For OTP-MASAL, the loss of power is mostly due to the change of correlation information between phenotypes when taking the average value of each phenotype over all the time points. However, the advantage is not

one-sided between Univariate MASAL and OTP-MASAL based on our simulation results. Furthermore, we can see from Table 2 that the performance of the three methods varies across different correlations between phenotypes, especially for Multivariate MASAL and OTP-MASAL. As expected, Multivariate MASAL is more powerful when the correlation between the two phenotypes becomes stronger. Interestingly, the power of OTP-MASAL increases as the correlation between the two phenotypes changes from positive to negative. One possible explanation is that the correlation between the average phenotypes is not the same as the correlation between the original phenotypes any more.

4 Application on Framingham Heart Study (FHS)

4.1 Background

Cardiovascular disease (CVD) is the leading cause of death and serious illness in the US. The Framingham Heart Study, founded in 1948 and now under the direction of the National Heart, Lung, and Blood Institute (NHLBI), was aimed to identify the common factors or characteristics that contribute to CVD. The participants in the FHS belong to three different cohort studies (Original Cohort, Offspring Cohort, and Gen3 Cohort) and were recruited and had been examined every two to four years over a long period of time. It belongs to the longitudinal cohort study.

Many studies pointed out that blood lipid levels, such as low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and triglycerides (TG), are heritable risk factors for CVD [4–6,10–12]. However, some of studies did not directly deal with longitudinal data but required the longitudinal measures to be summarized into one-time-point trait [5]. Furthermore, few of studies analyzed several blood lipid levels together. On the contrary, most of methods examined each blood lipid level one at a time. The goal of the present study is to illustrate the practical relevance of the proposed method for analyzing multiple blood lipid levels jointly in the longitudinal cohort studies.

4.2 Data analysis

Phenotypes of our primary interest are the blood lipid levels measured in main exams of the FHS. Specifically, we considered three quantitative traits: total cholesterol (CHOL), HDL-C, and TG. Although LDL-C can be calculated according to Friedewald's formula from CHOL, HDL-C, and TG, which is only valid if the TG level is less than 400 mg/dL. As a result, we used CHOL instead of LDL-C here. In the main clinical exams, since blood lipid levels only from a single exam are available for Gen3, and HDL-C levels are not available at all for original cohort, these datasets are not suitable for longitudinal analyses. We restricted our analyses to participants from the offspring cohort, where the blood lipid levels were measured from up to four exam visits (exams 1, 3, 5, and 7). For each family, we randomly selected one subject and obtained an independent sample consisting of 702 subjects. Besides blood lipid levels, we also included sex, age, body mass index (BMI), smoking status, and alcohol intake and treated them as covariates in our analyses, which are all repeatedly measured variables.

We focused on the gene LPL and studied the associations between the candidate gene and three blood lipid traits jointly. The gene LPL, especially, the SNP rs328 (S447X) in LPL,

was previously found to be associated with TG and HDL-C [5–7]. In the FHS 50K SNP data set, 29 SNPs were genotyped around the LPL gene as well as its neighboring genes. In order to minimize false positive associations due to rarer SNPs and genotyping artifact, we limited our analyses to SNPs with minor allele frequency not less than 1% and the p -value for testing Hardy-Weinberg Equilibrium greater than 0.001 by using Pearson's chi-squared test. Thus, there are a total of 22 SNPs remained in our analyses.

We first performed an association analysis for each of the three traits separately by using Univariate MASAL. None of these three analyses revealed any evidence of association between a single trait and the underlying SNPs. Actually, MASAL did not put any SNP terms into the final fitted model when it handled each trait separately. In this situation, it is not necessary to perform the following testing procedure to test the associations between traits and SNPs. However, when three traits were analyzed jointly, not only the main effects of SNPs but also SNP-SNP and SNP-covariate interactions were selected in the final MASAL model. Based on the final model, we listed all the terms that are related to TG trait as well as the corresponding coefficients in Table 3. For the CHOL and HDL-C traits, only BMI covariate has effect on them. We further performed a permutation study to test the associations between genetic covariates and the three traits jointly, where the value of the corresponding Wald statistic is 370.53, and the result revealed a strong association (p -value = 0.027). Thus, the power of the association test was increased by analyzing the three traits jointly.

From Table 3, we can see that SNP rs6988757 was selected by MASAL and had an interaction effect on TG trait. This SNP is strong linkage to SNP rs328 ($D' = 1$), the latter was previously found to be associated with TG and HDL-C [5–7]. Furthermore, an interaction between SNP rs328 and SNP rs6988825 was also found by MASAL, where SNP rs6988825 lies in the CSGALNACT1 that plays a role in the initiation and elongation in the synthesis of chondroitin sulfate. Besides the gene-gene interactions, the fitted models also provided us gene-covariate interactions (e.g., rs4922086 and alcohol, rs6988825, rs3923920 and alcohol, etc.).

5 Discussion

Understanding comorbidity is a very important and challenging issue in the genetic studies of complex diseases, especially for mental health conditions. A common approach is to analyze all the phenotypes of interest together instead of analyzing one at a time. In this article, we proposed a nonparametric regression method to perform associations between genes and multivariate phenotypes with repeated measures simultaneously, which is based on a proven method, MASAL [15]. The major advantage of our proposed method is that it can handle the complex data set that contains two distinguishing characteristics: multivariate phenotypes and longitudinal data, which are hard to be handled simultaneously by the existing methods. Besides, the new method can also accommodate gene-gene, gene-environment, and time-covariate interactions in the association studies.

On the basis of our simulations, we found that our proposed method gained more power when we analyze multivariate phenotypes with longitudinal measurements jointly, as

opposed to analyzing single longitudinal phenotype one at a time or analyzing multivariate phenotypes by summarizing repeated measures into one-time point value. By applying our method into the FHS data focusing on the blood lipid levels, we demonstrated that the analysis of multivariate longitudinal phenotypes enhanced the significance of the association test. Furthermore, our method did identify the gene-gene and gene-environment interactions (e.g., SNP rs328 and SNP rs6988825, alcohol and SNP rs4922086). Previous studies have shown that rs328 was associated the TG and HDL-C [5–7]. Our method not only verified their findings, but also drew us attention to the interactions between rs328 and other SNPs on TG trait, which was never reported by other methods to our best knowledge. Thus, further investigation is warranted to confirm our findings.

We used a permutation procedure to establish the null distribution of the Wald-type statistic because it no longer follows the chi-square distribution under the null hypothesis as Zhang [15] pointed out. The computation time of permutation is reasonable for a real data set, but can be intensive for simulation studies. Moreover, to permute the data, we randomly assigned the genotypes (G) only, while keeping the phenotype (Y) and non-genetic covariates (E) intact for each individual. However, this may destroy the correlation between E and G although our method works well for simulation studies and real data analysis. Theoretical studies for exploring the asymptotic distribution of the proposed statistic will be useful.

Acknowledgments

The authors thank two anonymous referees for their constructive comments and suggestions. This work was supported by grant R01 DA016750-09 from the National Institute on Drug Abuse. Zhu's work was also supported by the National Natural Science Foundation of China (Grant No. 11001044), the Fundamental Research Funds for the Central Universities (11CXPY007, 10JCXK001), the Natural Science Foundation of Jilin Province (Grant No. 201215007), the Scientific Research Foundation for Returned Scholars, MOE of China, and the Program for Changjiang Scholars and Innovative Research Team in University. The Framingham Heart Study project is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (N01 HC25195). The Framingham data used for the analyses described in this manuscript were obtained through dbGaP (phs000128.v3.p3).

References

1. Carlborg O, Haley CS. Epistasis: too often neglected in complex trait studies? *Nat Rev Genet.* 2004; 5:618–625. [PubMed: 15266344]
2. Friedman JH. Multivariate adaptive regression splines. *Ann Stat.* 1991; 19:1–141.
3. Kallberg H, Padyukov L, Plenge RM, et al. Gene-gene and gene-environment interactions involving HLA-DRB1, PTPN22, and smoking in two subsets of rheumatoid arthritis. *Am J Hum Genet.* 2007; 80:867–875. [PubMed: 17436241]
4. Kannel WB, Dawber TR, Kagan A, et al. Factors of risk in the development of coronary heart disease-six year follow-up experience. The Framingham Study. *Ann Intern Med.* 1961; 55:33–50. [PubMed: 13751193]
5. Kathiresan S, Manning AK, Demissie S, et al. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med Genet.* 2007; 8(Suppl 1):S17. [PubMed: 17903299]
6. Kathiresan S, Melander O, Guiducci C, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet.* 2008; 40:189–197. [PubMed: 18193044]
7. Kooner JS, Chambers JC, Aguilar-Salinas CA, et al. Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *Nat Genet.* 2008; 40:149–151. [PubMed: 18193046]

8. Lange C, Silverman E, Xu X, et al. A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics*. 2003; 4:195–206. [PubMed: 12925516]
9. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 3:13–22.
10. Miller NE, Miller GJ. Letter: high-density lipoprotein and atherosclerosis. *Lancet*. 1975; 1:10–33. [PubMed: 46335]
11. Namboodiri KK, Kaplan EB, Heuch I, et al. The Collaborative Lipid Research Clinics Family Study: biological and cultural determinants of familial resemblance for plasma lipids and lipoproteins. *Genet Epidemiol*. 1985; 2:227–254. [PubMed: 4054601]
12. Pollin TI, Damcott CM, Shen H, et al. A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science*. 2008; 322:1702–1705. [PubMed: 19074352]
13. Xu X, Tian L, Wei LJ. Combining dependent tests for linkage or association across multiple phenotypic traits. *Biostatistics*. 2003; 4:223–229. [PubMed: 12925518]
14. Yeager M, Orr N, Hayes RB, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet*. 2007; 39:645–649. [PubMed: 17401363]
15. Zhang HP. Multivariate adaptive splines for analysis of longitudinal data. *J Comput Graph Stat*. 1997; 6:74–91.
16. Zhang HP. Analysis of infant growth curves using multivariate adaptive splines. *Biometrics*. 1999; 55:452–459. [PubMed: 11318199]
17. Zhang HP. Mixed effects multivariate adaptive splines model for the analysis of longitudinal and growth curve data. *Stat Methods Med Res*. 2004; 13:63–82. [PubMed: 14746441]
18. Zhang HP, Liu C-T, Wang XQ. An association test for multiple traits based on the generalized Kendall's tau. *J Amer Stat Assoc*. 2010; 105:473–481. [PubMed: 20711441]
19. Zhang HP, Zhong X. Linkage analysis of longitudinal data and design consideration. *BMC Genet*. 2006; 7:37. [PubMed: 16768806]
20. Zhu WS, Jiang Y, Zhang HP. Nonparametric covariate-adjusted association tests based on the generalized Kendall's tau. *J Amer Stat Assoc*. 2012; 107:1–11. [PubMed: 22745516]
21. Zhu WS, Zhang HP. Why do we test multiple traits in genetic association studies? (with discussion). *J Korean Stat Soc*. 2009; 38:1–10. [PubMed: 19655045]

Table 1

Estimated type I errors of Multivariate MASAL, Univariate MASAL, and OTP-MASAL at significance level of 0.05

ρ_{BT}	method		
	Multivariate MASAL	Univariate MASAL	OTP-MASAL
0.0	0.048	0.052	0.057
0.1	0.052	0.051	0.056
0.2	0.050	0.048	0.058
0.3	0.047	0.046	0.053

Table 2

Power comparisons of Multivariate MASAL, Univariate MASAL, and OTP-MASAL at significance level of 0.05

ρ_{BT}	method		
	Multivariate MASAL	Univariate MASAL	OTP-MASAL
0.3	0.892	0.592	0.551
0.2	0.791	0.613	0.583
0.1	0.715	0.627	0.615
0.0	0.724	0.601	0.629
-0.1	0.771	0.629	0.632
-0.2	0.787	0.642	0.684
-0.3	0.816	0.657	0.690

Table 3

Main and interaction effect on TG trait selected by MASAL

effect	selected terms	coefficients
Main effect	rs6988825	-62.4857
	Sex	-108.4280
	Alcohol	-33.6991
	BMI	+8.9254
Gene-gene interaction	rs328 $\times$ rs6988825	+13.7025
	rs6988825 $\times$ rs6988757 $\times$ rs2975418	+22.1655
Gene-environment interaction	rs6988825 $\times$ (Alcohol-1.6556) ⁺	+16.2935
	rs4922086 $\times$ Alcohol	+3.4890
	rs3923920 $\times$ Alcohol	+10.3054
	rs6988825 $\times$ rs3923920 $\times$ (Alcohol-1.6556) ⁺	-8.4261
Environment-environment interaction	Sex $\times$ Age	+1.4097
	BMI $\times$ (Age-23) ⁺	-0.2104
	BMI $\times$ (Age-53.4752) ⁺	+0.1683
	Alcohol $\times$ Smoking	+2.8496

A $\times$ B denotes interaction term between A and B