Published in final edited form as:

Biometrics. 2010 March ; 66(1): 89–96. doi:10.1111/j.1541-0420.2009.01235.x.

A Partially Linear Tree-based Regression Model for Multivariate Outcomes

Kai Yu¹, William Wheeler², Qizhai Li^{1,3}, Andrew W. Bergen⁴, Neil Caporaso¹, Nilanjan Chatterjee¹, and Jinbo Chen⁵

¹Division of Cancer Epidemiology and Genetics, NCI, Rockville MD, USA

²Information Management Services Inc, Rockville, MD, USA

³Academy of Mathematics and Systems Science, CAS, Beijing, P. R. China

⁴Molecular Genetics Program, Center for Health Sciences, SRI International, Menlo Park, CA, USA

⁵Department of Biostatistics and Epidemiology, School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Summary

In the genetic study of complex traits, especially behavior related ones, such as smoking and alcoholism, usually several phenotypic measurements are obtained for the description of the complex trait, but no single measurement can quantify fully the complicated characteristics of the symptom because of our lack of understanding of the underlying etiology. If those phenotypes share a common genetic mechanism, rather than studying each individual phenotype separately, it is more advantageous to analyze them jointly as a multivariate trait in order to enhance the power to identify associated genes. We propose a multilocus association test for the study of multivariate traits. The test is derived from a partially linear tree-based regression model for multiple outcomes. This novel tree-based model provides a formal statistical testing framework for the evaluation of the association between a multivariate outcome and a set of candidate predictors, such as markers within a gene or pathway, while accommodating adjustment for other covariates. Through simulation studies we show that the proposed method has an acceptable type I error rate and improved power over the univariate outcome analysis, which studies each component of the complex trait separately with multiple-comparison adjustment. A candidate gene association study of multiple smoking-related phenotypes is used to demonstrate the application and advantages of this new method. The proposed method is general enough to be used for the assessment of the joint effect of a set of multiple risk factors on a multivariate outcome in other biomedical research settings.

Keywords

Generalized estimating equation; Genetic association study; Model selection; Multiple-comparison adjustment; Tree-based model

Address for Correspondence: Kai Yu, Ph.D., 6120 Executive Boulevard, EPS 8050, Rockville, MD 20852, E-mail: yuka@mail.nih.gov.

Supplementary Materials Web Appendices, Tables, and Figures referenced in Sections 3, 4 and 5 are available under the Paper Information link at the *Biometrics* website <http://www.biometrics.tibs.org>.

1. Introduction

In genetic association studies of many complex traits, particularly those related to behavior, such as alcoholism and smoking, we usually do not have a single phenotypic measurement that can quantify fully the complicated characteristics of the trait because of our lack of understanding of the underlying etiology. More often, we obtain multiple phenotypic measurements, with each phenotype variable reflecting one aspect of the complex trait. For example, in a genetic study of asthma, we usually have several pulmonary function related measurements since asthma is associated with a variety of abnormalities in pulmonary function (Lange et al., 2003). If several related phenotypes share a common genetic mechanism, rather than studying each individual phenotype separately, it may be advantageous to analyze them jointly in order to utilize efficiently all of the relevant information.

Xu et al. (2003) provided a brief summary of approaches for testing the association between a multivariate phenotype and genetic variants. A commonly used one is to combine univariate analyses based on individual phenotypes for a global assessment of the association in the framework of the generalized linear model (GLM) by applying the generalized estimating equations (GEE) method (Liang and Zeger, 1986). The GLM is parsimonious for examining main effects of a single or several genetic risk factors, but it is not suitable for modeling complex joint effects (such as high-order gene-gene interactions) of a relatively large number of risk factors. For example, to detect the association between an outcome variable and a candidate gene, it is often desirable to model the joint effect of multiple markers within that candidate gene. The tree-based modeling approach (Breiman et al., 1984; Zhang and Singer, 1999), has an advantage over GLM in that it is more flexible in modeling complicated joint effects, and it has increasingly been applied in genetic studies of complex traits (see, e.g., Cook et al., 2003; Zhang et al., 2000; Yu et al., 2005; Yu et al., 2007).

We recently proposed a new class of semi-parametric regression models for the analysis of univariate phenotypes, termed partially linear tree-based regression (PLTR) model, that integrates the advantages of the generalized linear regression and tree-structure models (Chen et al., 2007). This model includes both linear terms and a tree term, which are assumed to be additive on an appropriate scale, e.g., the logit scale for a binary outcome. The linear part is used to model main effects of some genetic/environmental exposures and to control for confounders. The nonparametric tree part is used for approximating the complex joint effect of multiple genes and environmental exposures.

In this paper, we extend the univariate PLTR model to the study of multivariate outcomes. As in the univariate PLTR model, its multivariate extension consists of linear terms and a tree term. We adopt the GEE method to accommodate the multivariate outcome in the construction of the tree model while allowing the linear term in the model to be appropriately accounted for. We provide a formal testing procedure for assessing the joint effect of multiple risk factors, such as genotypes on a panel of genetic markers within a gene or pathway, modeled as the tree term, on a multivariate outcome while adjusting for a set of covariates. The presence of the linear term in the new model distinguishes our method from existing multivariate tree models (Segal, 1992; Zhang, 1998; Larsen and Speckman, 2004). Also, existing multivariate tree models focus mainly on prediction, while ours emphasizes hypothesis testing. Compared with the standard GEE approach for multivariate outcome, our model is more flexible in modeling complicated joint effects. The new test derived from our proposed model can be used as a multilocus genetic association test for the study of multivariate phenotypes, and it is general enough to be used for the assessment of the joint effect of a set of multiple risk factors on a multivariate outcome in other biomedical research settings.

2. Review of the Partially Linear Tree-based Regression (PLTR) Model

We provide a brief review of the PLTR model and refer readers to Chen et al. (2007) for more details. Let Y be the outcome, X be covariates to be modeled linearly, and G be the set of risk factors whose joint effect is to be approximated by a tree-structural model. The data $\{(Y_i, X_i, G_i): i = 1, \dots, N\}$ are observed for a sample of N subjects. The main purpose of this PLTR model is to test the joint effect of G while adjusting for covariates X . In this paper, we always assume that G consists of genotypes measured on a panel of biallelic genetic markers, with each genotype coded as 0, 1 or 2 representing the copy number of the given allelic type. We used the GLM model for the relationship between Y and (X, G) . The mean of Y , $E(Y) = \mu$, is specified as

$$h(\mu) = \beta' \mathbf{x} + \gamma' \mathbf{z}(\mathbf{T}), \quad (1)$$

where $h(\bullet)$ is the link function, for example, we can use the logit link function

$h(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ for a binary outcome, and the identity link function for a continuous outcome. In (1), $\mathbf{x}$ is the standard coding vector for the set of covariates X , and $\mathbf{z}(\mathbf{T})$ is the coding vector for the risk factor derived from the tree model $\mathbf{T}$. For example, the tree model shown in Figure 1 has five terminal nodes (nodes without offspring nodes) and defines a risk factor (still denoted as $\mathbf{T}$) with 5 categories. To test the joint effect of G , we can use either the standard likelihood ratio test statistic, which compares the model with the tree term $\mathbf{z}(\mathbf{T})$ to the one without the tree term, or the Wald test statistic, which tests the coefficient vector γ . Its significance level is evaluated through a resampling procedure that accounts for any model selection bias introduced by the choice of $\mathbf{T}$.

3. Extension of the PLTR Model to Multivariate Outcomes

In this section, we extend the PLTR model to study a multivariate outcome with M components, $(y_1, \dots, y_M)'$, which we again denote by Y . As with the univariate PLTR model, the goal of its multivariate extension is to assess the joint effect of G on the multivariate outcome Y . We still use the PLTR model for the marginal distribution of each individual outcome. That is, we relate the mean of the m th ($1 \leq m \leq M$) outcome, $E(y_m) = \mu_m$, to covariates (X, G) by the following model,

$$h(\mu_m) = \beta_m' \mathbf{x} + \gamma_m' \mathbf{z}(\mathbf{T}), \quad 1 \leq m \leq M, \quad (2)$$

where $h(\bullet)$ is the link function. In (2), β_m and γ_m are the vectors of coefficients of $\mathbf{x}$ and $\mathbf{z}(\mathbf{T})$ for the m th outcome y_m . In model (2), we allow all coefficients to be outcome specific, i.e., the effect of a predictor can be different for different outcomes. We can certainly impose a common effect from a predictor on different outcomes if we believe a predictor has the same effect on all outcomes.

The tree model $\mathbf{T}$ is formed by splitting rules based on G and defines a risk factor with each level represented by a terminal node. In model (2), we assume the tree structure is the same across different outcomes so that the same risk factor (derived from the tree model $\mathbf{T}$) is defined for all outcomes, but we allow the effect to be outcome-specific. This strategy is reasonable when there is a common latent risk factor (either continuous or categorical) underlying all the considered phenotypes. For example, in the association study of smoking-related phenotypes

and the gene CYP2B6 (described later), if all three considered phenotypes are related to the same unmeasured functional marker within the gene (so we have a latent risk factor with three categories), we could use the tree model in (2) as a surrogate for the unobserved functional marker.

Our method consists of the following three steps with details provided in the next two sections:

Step I: Fit the model (2) by finding the tree model $\mathbf{T}$ as well as estimates for (β, γ) .

Step II: Based on $\mathbf{T}$, find a sequence of nested subtrees $\mathbf{T}_f, f = 2, \dots, F$, with $\mathbf{T}_f$ having f terminal nodes, and F being a pre-specified value, say, 10.

Step III: Find an optimal tree model $\mathbf{T}_{\hat{f}}$ from $\mathbf{T}_f, f = 2, \dots, F$, and evaluate the significance level of this optimal tree with respect to the joint effect of G .

We implemented the method by means of an R program using functions from the R packages Rpart (Therneau and Atkinson, 1997) and geepack (Halekoh et al., 2006).

3.1 Details on Step I

To account for the correlation among multiple outcomes when choosing the splitting variable in the tree-building procedure and when estimating coefficients $(\beta_m, \gamma_m), m = 1, \dots, M$, we use the GEE method and assume a working correlation matrix $R(\theta)$ for $Y = (y_1, \dots, y_M)'$, with unknown parameter θ (Liang and Zeger, 1986). We propose to use the following iteration algorithm to find the parsimonious tree model (usually over-sized) $\mathbf{T}$ as well as the estimates for $(\beta_m, \gamma_m), m = 1, \dots, M$. This algorithm is similar to the one for fitting the univariate PLTR model (Chen et al., 2007), except that the GEE method is used for fitting the linear model and constructing the tree.

1. Fit the linear model $h(\mu_m) = (\beta_m^{(1)})' \mathbf{x}, 1 \leq m \leq M$, by means of the GEE method.
2. Iterate between the following steps starting with $k = 1$.
 - 2.1 Build a tree model $\mathbf{T}^{(k)}$ based on G , using $(\beta_m^{(k)})' \mathbf{x}, 1 \leq m \leq M$, as the offset, with $\beta_m^{(k)}$ being the current estimate for β_m at the k th iteration.
 - 2.2 Use GEE to fit the multivariate GLM $h(\mu_m) = \gamma_m' \mathbf{z}(\mathbf{T}^{(k)}), 1 \leq m \leq M$, using $(\beta_m^{(k)})' \mathbf{x}_m$ as the offset. Let the new estimate be $\gamma^{(k)}$.
 - 2.3 Use GEE to fit the multivariate GLM $h(\mu_m) = \beta_m' \mathbf{x}_m, 1 \leq m \leq M$, with $(\gamma_m^{(k)})' \mathbf{z}(\mathbf{T}^{(k)})$ being the offset. Update the estimate $\beta_m^{(k)}$ to $\beta_m^{(k+1)}$.
3. The iteration stops when $\beta_m^{(k)}$ converges.

In step 2.1, the tree $\mathbf{T}$ is constructed recursively by splitting a node into two. Assume that the current estimation for β_m is $\beta_m^{(k)}, m = 1, \dots, M$, and that node A is to be split. With a candidate splitting rule Γ , subjects in node A are separated into two offspring nodes A_L and A_R . Let the indicator variable $I(Z \in A_L)$ take value 1 if the subject goes into A_L and 0 otherwise, and $I(Z \in A_R)$ be similarly defined. We fit the model

$$h(\mu_m) = \eta_{m,1} I(Z \in A_L) + \eta_{m,2} I(Z \in A_R) + \text{offset} \left\{ (\beta_m^{(k)})' \mathbf{x} \right\}, 1 \leq m \leq M,$$

where $\eta_m = (\eta_{m,1}, \eta_{m,2})$ is the vector of regression coefficients for outcome y_m , and $(\beta_m^{(k)})' \mathbf{x}$ is used as an offset when using the GEE to estimate η_m . Let the GEE estimate of $\eta = (\eta_1, \dots, \eta_M)'$ be $\hat{\eta}$, and let $\hat{\Sigma}_A$ be the sandwich estimate of the covariance matrix of $\hat{\eta}$. To assess the “goodness-of-split” (LeBlanc and Crowley, 1993) of this candidate splitting rule, we use the Wald test statistic designed for testing the null hypothesis $C\eta = \mathbf{0}$, where $C = \mathbf{I}_{M \times M} \otimes (1 - 1)$, with $\otimes$ being the Kronecker product. The null hypothesis is equivalent to $\eta_{m,1} = \eta_{m,2}$, for $m = 1, \dots, M$. The Wald statistic associated with the candidate splitting rule Γ can be expressed as

$$\chi^2(A, \Gamma) = (C\hat{\eta})' \left(C \sum_A C' \right)^{-1} C\hat{\eta}. \quad (3)$$

A large Wald statistic would suggest that the two offspring nodes probably have different effects on the outcome. For a given node A , among all its possible splitting rules, we choose the one that yields the maximum Wald statistic, as given by (3), and use it to grow the tree to the next level. We define the “goodness-of-split” for node A as

$$S(A) = \max_{\Gamma} \chi^2(A, \Gamma), \quad (4)$$

where the maximum is chosen over all permissible splitting rules.

Based on this splitting algorithm, we can grow the tree recursively until the number of samples left in a node is insufficient for further splitting. We recommend a relatively large threshold for the sample size at a node, for example 50, so that a sufficient number of subjects are available to fit using the GEE method. The above fitting algorithm requires that the tree in all iterations have the same size. Thus, after the tree $\mathbf{T}$ is constructed at step 2.1, we prune it back to a predetermined number of terminal nodes, F (say, 10). This can be done in either a stepwise backward or forward fashion. To do it in a forward fashion, we start from the subtree with the root node 1 and its two offspring nodes 2 and 3. Then among nodes 2 and 3, we choose the one with a larger “goodness of split”, as defined by (4), and expand the subtree by including the offspring nodes of the chosen node. Suppose we have obtained a subtree with f terminal nodes, $f < F$. We then can choose the terminal node with the largest “goodness of split” and expand the current subtree by including two offspring nodes of the chosen node to obtain a subtree with $f+1$ terminal nodes. The process can continue until a subtree with F terminal nodes is reached. Similarly, we can reverse the process by moving backward. We can start from a current tree and reduce its size by pruning away the pair of terminal nodes whose parent node has the smallest “goodness of split”. The tree $\mathbf{T}^{(k)}$ in Step 2.1 always has the same size using either strategy. Based on our experience, there is no obvious advantage of one strategy over the other. So, we choose to use the stepwise forward approach to identify the subtree with the given size F .

3.2. Details on Step II and III

The tree $\mathbf{T}$ at the end of Step I has a predetermined size F , which is usually oversized. We aim to identify an optimal subtree of $\mathbf{T}$, so that it presents the joint effect of G parsimoniously, and to evaluate its significance level.

First, we identify a sequence of subtrees of $\mathbf{T}$ as candidates for the optimal subtree. The stepwise forward algorithm that is used in Step I (described in Section 3.1) for selecting a subtree with a given size can again be applied to identify the sequence of candidate trees with various sizes.

Let $\mathbf{T}_f, f = 2, \dots, F$ be the set of nested candidate trees with $\mathbf{T}_f$ having f terminal nodes. For each $\mathbf{T}_f$, we fit the following model using the GEE method,

$h(\mu_m) = \beta'_m \mathbf{x} + \gamma'_m \mathbf{z}(\mathbf{T}_f), 1 \leq m \leq M$. Then we use the Wald statistic for testing whether or not $\gamma = (\gamma'_1, \dots, \gamma'_M)' = \mathbf{0}$, to assess the effect of the $\mathbf{z}(\mathbf{T}_f)$ while adjusting for the effect of $\mathbf{x}$. The Wald statistic can be written as

$$W(\mathbf{T}_f) = \hat{\gamma}' \left(\widehat{\sum}_{\gamma, \gamma} \right)^{-1} \hat{\gamma}, \quad (5)$$

where $\hat{\gamma}$ is the GEE estimation for γ , and where $\widehat{\sum}_{\gamma, \gamma}$ is the principal submatrix corresponding to γ taken from $\widehat{\Sigma}$, which is the (sandwich-type) estimated covariance matrix for all coefficient estimates.

Second, we choose as the “optimal” tree, from the list of candidate trees $\mathbf{T}_f, f = 2, \dots, F$, the one with the smallest “empirical” p-value associated with the Wald statistic defined by (5). Then we evaluate the statistical significance of the chosen tree model. That is, we find the adjusted p-value for the observed minimum empirical p-value. Because of the selection process involved in identifying each candidate tree $\mathbf{T}_f, f = 2, \dots, F$, under the null hypothesis $W(\mathbf{T}_f)$ no longer follows the Chi-squared distribution, a permutation procedure is needed to obtain the empirical p-value P_f for each candidate tree $\mathbf{T}_f$. To evaluate the significance level of the observed minimum p-value $\min_{2 \leq f \leq F} P_f$, one generally needs a two-layer permutation procedure (Westfall and Young, 1993), which is very computationally intensive. Instead, we adopt the computationally more efficient *minP* algorithm by Ge et al. (2003) to estimate $P_f, f = 2, \dots, F$, as well as the (adjusted) p-value for $\min_{2 \leq f \leq F} P_f$ through a single-layer permutation procedure. A detailed description of the permutation algorithm is given in the Web Appendix A.

3.3. Univariate analysis with multiple comparison adjustment

For comparison with the multivariate approach described above, we consider a univariate approach (called the univariate-MinP) that analyzes each individual outcome using the univariate PLTR and then adjusts for multiple comparisons. A more detailed description of the univariate-MinP approach is given in the Web Appendix B.

4. Simulation

For our simulation, we studied a candidate gene region with 11 SNPs. Their haplotype frequencies are given in Figure 3 of Yeager et al. (2007). There were a total of 25 haplotypes in the study population. To assign genotypes at the 11 SNPs for a subject in the study population, we independently selected pairs of haplotypes according to the haplotype frequencies. To evaluate the power of the proposed multivariate tree model and of the univariate-MinP approach, we designated the 5th SNP in the candidate region as the single functional locus (one-locus risk model) that influences the quantitative traits under study. We assumed that this functional SNP was not typed and that genotypes at the other 10 SNPs were available for the association study. The power was calculated as the probability that the joint effect on the outcome from the genotypes at the 10 measured SNPs is significant under a given significance level (according to the adjusted p-value from the multivariate tree model or the univariate-MinP method). In addition to the genotypes, we considered a non-genetic binary covariate X that had to be adjusted for in the analysis. We let $\Pr(X = 1) = 0.4$ and $\Pr(X = 0) = 0.6$, and we further assumed that X and the genotypes were independent in the study population. Let g_5 be the coded value for the unobserved genotype at the functional locus, with values of

0, 1, and 2 representing the copy number of the minor allele “1”. Conditioned on X and g_5 , we generated a multivariate quantitative trait $Y = (y_1, \dots, y_M)'$ from the multivariate normal distribution $MVN(\mu, \mathbf{V})$, where $\mu = \beta X + \gamma g_5$, with $\beta = (\beta_1, \dots, \beta_M)'$ and $\gamma = (\gamma_1, \dots, \gamma_M)'$ being the given coefficients and $\mathbf{V}$ being the covariance matrix. In our simulation, we considered a multivariate trait consisting of 2 or 3 phenotypes (i.e., $M = 2$, or 3). For simplicity, we assumed a covariate effect $\beta_m = 0.2$, $m = 1, \dots, M$, for all conducted simulations. Besides the multivariate quantitative traits, we considered multivariate binary traits by dichotomizing the $Y = (y_1, \dots, y_M)'$ generated from $MVN(\mu, \mathbf{V})$ using the expected median value of y_m , $m = 1, \dots, M$, as the cut-point.

To evaluate the type I error rate, we considered a bivariate trait (quantitative or its dichotomized version) unrelated to the genotype, i.e., $\gamma_m = 0$, $m = 1, 2$. We assumed the covariance matrix $\mathbf{V} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, with $\rho = 0, 0.1$, or 0.2 . For each value of ρ , we generated 2000 datasets, each consisting of 800 subjects with their bivariate trait, genotypes at 10 SNPs and covariate X simulated according to the procedure described above.

To evaluate power, we simulated a bivariate trait (quantitative or its dichotomized version) by

assuming $\gamma_m = 0.18$ for both phenotypes, and the covariance matrix $\mathbf{V} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, with ρ being chosen among $(-0.3, -0.2, -0.1, 0.0, 0.1, 0.2, 0.3)$. We also simulated a trivariate trait with the

associated variance-covariance matrix $\mathbf{V} = \begin{pmatrix} 1.00 & 0.20 & -0.20 \\ 0.20 & 1.00 & -0.30 \\ -0.20 & -0.30 & 1.00 \end{pmatrix}$ under the following three scenarios, Scenario I: $\gamma_1 = 0.18$, and $\gamma_2 = \gamma_3 = 0$; Scenario II: $\gamma_1 = \gamma_2 = 0.18$, and $\gamma_3 = 0$; and Scenario III: $\gamma_1 = \gamma_2 = \gamma_3 = 0.18$. This covariance matrix was similar to the one observed in the application described later. Under each considered case, we simulated 500 datasets, each consisting of 800 subjects for the power evaluation.

We analyzed each simulated dataset using model (2) allowing coefficients to vary for different outcomes. In the tree-building process, we considered splitting rules based on genotypes (coded as 0, 1, and 2, corresponding to the copy number of a chosen allelic type) on a single SNP, and we treated genotypes as continuous variables, i.e., the legitimate splitting rules are: 0 vs. 1 and 2; and 0 and 1 vs. 2, but not 1 vs. 0 and 2. We used different assumptions about the correlation matrices in the GEE method (e.g., independent and unstructured) and obtained quite similar results based on simulated datasets (see Web Table 1). In the following, we present results based on analyses using “independence” as the working correlation matrix in the GEE method.

Simulation results for type I error evaluation are given in Table 1. It appears that the proposed method has a type I error rate close to the nominal levels ($\alpha = 0.01, 0.05$) under all covariance structures considered for both continuous and binary traits. We also evaluated the type I error rate of the multivariate tree model when the tree size was pre-specified. Results are given in Web Table 2. We can see from the table that the p-value estimated from the permutation procedure for a tree with a given size has the desired property.

Power calculations for the simulated bivariate outcomes are given in Table 2. Under the given covariance matrix, we compared the power of the multivariate analysis (applying the proposed method to study two phenotypes jointly) with that of the univariate-MinP procedure (multiple-comparison adjusted univariate analysis). Results in Table 2 demonstrate the advantage of using the multivariate approach for both the quantitative and binary traits. Interestingly, the power improvement over the univariate-MinP analysis increases as the correlation between the two outcomes changes from positive to negative. The explanation is that the negative correlation between the two outcomes helps to reduce the variance in the coefficients'

estimation when the underlying effects (γ_1 and γ_2) for the two correlated phenotypes point in same direction (both either positive or negative).

Power calculations for the simulation of a trivariate quantitative outcome are given in Table 3. For each simulated dataset, we analyzed various subsets of the multivariate outcome using the joint analysis as well as the univariate-MinP analysis. From Table 3 we can observe the gain in power when individual phenotypes related to the same gene were studied together over the analysis of individual phenotypes. The power to detect the gene reaches its highest level when all relevant phenotypes are studied jointly. But when the multivariate trait under study includes one or more phenotypes not related to the gene under study, the power is diminished to some extent.

We further evaluated the performance of the proposed method by simulating data under a 2-locus risk model, and we found that our method performed very reasonably (See Web Table 3). The simulation design is described in the Web Appendix C.

We evaluated the convergence properties of the iterative algorithm through extensive simulation studies. We found that the algorithm usually converged within 50 iterations in the simulation setup described above. When the sample size was 500, the algorithm converged in 995~1000 out of 1000 replicated datasets under the various considered scenarios. When the sample size increased to 1000, the algorithm converged in 998-1000 out of 1000 replicated datasets. The working matrix did not appear to have an impact on the convergence properties of the algorithm; neither did the type of outcome (continuous or binary). When the algorithm did not converge, it always iterated between two tree models. In these rare situations, we chose the tree model with the larger Wald test statistic.

5. Application

As part of the Cancer Genetic Markers of Susceptibility (CGEMS) project (Yeager et al., 2007), a multi-stage genome-wide association study of prostate cancer has been conducted. The initial scan was performed on a nested case-control sample drawn from the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. To demonstrate the application of the proposed model, we used genotype data generated in this CGEMS prostate cancer project to study the association between the candidate gene CYP2B6 and the multivariate trait, which consists of the following three dichotomized smoking behavior phenotypes relating to smoking intensity, duration of smoking and age at smoking initiation, $Y_{\text{cig-day}}$, $Y_{\text{smk-yr}}$, and $Y_{\text{smk-age}}$, respectively, where $Y = 1$ if the number of cigarettes smoked per day is 4 or more, and 0 otherwise; $Y_{\text{smk-yr}} = 1$ if years of smoking is 29 or more and 0 otherwise; and $Y_{\text{smk-age}} = 1$ if the age at smoking initiation is 18 or below and 0 otherwise. Each dichotomous cut-point was chosen to be close to the corresponding median value. The gene CYP2B6 was previously reported to be associated with nicotine metabolism (Ring et al., 2007) and nicotine dependence (Saccone et al., 2007).

17 SNPs were genotyped at and around the CYP2B6 gene (20 kb up and downstream) at chromosome 19q13.2. In the linear part of model (2), we adjusted for prostate cancer status (prostate cancer case or control), and age at diagnosis in five-year intervals (55–59, 60–64, 65–69, 70–74), which were stratification variables in the original data collection. We used the tree structure to summarize the joint effect of the 17 SNPs and tested whether there was a significant joint effect from those 17 SNPs on the study phenotypes. During the tree building process, similar to the simulation study, we considered splitting rules based on genotypes (coded as 0, 1, and 2, corresponding to the copy number of a chosen allelic type) on a single SNP. The legitimate splitting rules included: (a) 0 vs. 1 and 2 and (b) 0 and 1 vs. 2.

We focused on 1,318 ever smokers with complete data for all non-genetic covariates and genotypes for the 17 SNPs. Genotypes frequencies are given in Web Table 4. We used the proposed method to assess the association between the gene and various subgroups of the three phenotypes by finding the optimal tree model to summarize the joint effect of those 17 SNPs. Results are summarized in Table 4. The model-selection adjusted p-value, which was obtained through the permutation procedure described in the Methods section, indicated the significance level of the association. When we analyzed each binary trait separately, significant associations between the gene and $Y_{\text{cig-day}}$ (adjusted p-value = 0.012), and between the gene and $Y_{\text{smk-yr}}$ (adjusted p-value = 0.032) were observed. But there was no evidence for a significant association between the gene and $Y_{\text{smk-age}}$ (adjusted p-value = 0.392). By analyzing traits $Y_{\text{cig-day}}$ (smoking intensity) and $Y_{\text{smk-yr}}$ (smoking duration) jointly, we detected a stronger association (adjusted p-value = 0.009) than those from univariate analyses. When the third trait $Y_{\text{smk-age}}$ was included to form a trivariate outcome, the signal for association became weaker (adjusted p-value = 0.015).

More detailed results for the joint analysis of $Y_{\text{cig-day}}$ and $Y_{\text{smk-yr}}$ are given in Figure 1 and Table 5. The final optimal tree (Figure 1), which summarized the joint effect of 4 SNPs selected from the 17 candidate ones, defined a risk factor with 5 categories. We also used the GEE method to obtain the naïve estimation of OR for each risk category (using the category with the most subjects as the baseline) after adjusting for other covariates (prostate cancer status and age groups). These are summarized in Table 5. Those OR estimates and their associated confidence intervals were biased since the risk factor was defined to fit the outcome. As a comparison, the final tree model for the analysis of all three phenotypes ($Y_{\text{cig-day}}$, $Y_{\text{smk-yr}}$, and $Y_{\text{smk-age}}$) is shown in Web Figure 1.

6. Discussion

Because of the lack of understanding of the underlying etiology and the uncertain nosology of complex traits, usually several phenotypic measures are obtained for the complex trait under study, with no single phenotypic measurement fully quantifying the complicated characteristics of the complex trait. An attractive tactic in these situations is to combine the relevant phenotypes into a single multivariate trait. We propose a partial linear tree-based regression model to study the joint effect of a given set of markers on a multivariate outcome while allowing for adjustment for the effect of covariates. Within the model, the joint effect is approximated by a nonparametric tree structure that is more flexible than the standard linear modeling approach. The new method provides a novel multilocus association test for multivariate outcomes. It can be useful in the setting of candidate gene association studies, where the main goal is to establish the association between the outcome and the candidate gene. It can also be useful in pathway analysis, where one evaluates the association between the outcome and a panel of genes in the pathway.

The proposed model leverages on the well-established GEE framework and thus can be used to study such types of multivariate outcomes as continuous and binary. Unlike the existing tree-based model for multivariate outcome, the proposed method can formally test the statistical significance of the joint effect, captured by the tree model, while adjusting for the covariates' effects linearly. Zhang and Singer (1999) suggested another way of adjusting for the covariates in the tree model framework by using them on top of the tree. It would be interesting to compare different ways of covariate adjustment in the tree-framework.

Compared with the univariate PLTR, the multivariate extension gains its efficiency by combining information from multiple outcomes under the assumption that there is a common latent risk factor underlying all the considered univariate outcomes (but allowing the effect from a given risk category to vary across different phenotypes). This multivariate approach

becomes less effective if the assumption is violated. For example, in a candidate gene association study of a bivariate trait (Y_1 , Y_2), if only Y_1 is related to the gene under study, but not Y_2 , an analysis of the bivariate outcome will be less powerful than a univariate analysis of Y_1 , since the inclusion of Y_2 only introduces noise and increases the degrees of freedom (if coefficients are allowed to vary across outcomes). Thus caution is needed when selecting multiple outcomes to study jointly. Finally, we would like to point out that the application of our new method is not limited to genetic association studies. The proposed model provides an effective inference procedure for testing the relationship between a multivariate outcome and any set of potential risk factors, and it could be useful in many other biomedical research settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research utilized the high-performance computational capabilities of the Biowulf PC/Linux cluster at the National Institutes of Health, Bethesda, Maryland, USA (<http://biowulf.nih.gov>). The work of K Yu, Q Li, N Caporaso, and N Chatterjee was supported in part by the Intramural Program of the NIH and the National Cancer Institute. The work of Q Li was also partially supported by the National Science Foundation of China, No. 10371126. AW Bergen was supported by U01 DA020830.

References

- Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. Classification and regression trees. Wadsworth Publishing Co., Inc; Belmont, CA: 1984.
- Chen J, Yu K, Hsing A, Therneau TM. A partially linear tree-based regression model for assessing complex joint gene-gene and gene-environment effects. *Genetic Epidemiology* 2007;32:238–251. [PubMed: 17266115]
- Cook NR, Zee RY, Ridker PM. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Statistics in Medicine* 2003;23:1439–1453. [PubMed: 15116352]
- Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. *Test* 2003;18:1–44.
- Halekoh U, Hojsgaard S, Yan J. The R package geepack for generalized estimating equations. *Journal of Statistical Software* 2006;15:1–11.
- Lange C, Silverman EK, Xu X, Weiss ST, Laird NM. A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics* 2003;4:195–206. [PubMed: 12925516]
- Larsen DR, Speckman PL. Multivariate regression trees for analysis of abundance data. *Biometrics* 2004;60:543–549. [PubMed: 15180683]
- LeBlanc M, Crowley J. Survival trees by goodness of split. *Journal of the American Statistical Association* 1993;88:457–467.
- Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986;73:13–22.
- Ring HZ, Valdes AM, Nishita DM, Prasad S, Jacob P, Tyndale RF, Swan GE, Benowitz NL. Gene-gene interactions between CYP2B6 and CYP2A6 in nicotine metabolism. *Pharmacogenetics and Genomics* 2007;17:1007–1015. [PubMed: 18004205]
- Saccone SF, Hinrichs AL, Saccone NL, Chase GA, Konvicka K, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau O, et al. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Human Molecular Genetics* 2007;16:36–49. [PubMed: 17135278]
- Segal MR. Tree-structured methods for longitudinal data. *Journal of the American Statistical Association* 1992;87:407–418.

- Therneau T, Atkinson EG. An introduction to recursive partitioning using the RPART routines. Mayo Foundation technical report. 1997
- Westfall, PH.; Young, BS. Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment. Wiley; New York: 1993.
- Xu X, Tian L, Wei LJ. Combining dependent tests for linkage association across multiple phenotypic traits. *Biostatistics* 2003;4:223–229. [PubMed: 12925518]
- Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genetics* 2007;39:645–649. [PubMed: 17401363]
- Yu K, Xu J, Rao DC, Province M. Using tree-based recursive partitioning methods to group haplotypes for increased power in association studies. *Annals of Human Genetics* 2005;69:577–589. [PubMed: 16138916]
- Yu K, Martin R, Rothman N, Zheng T, Lan Q. Two-sample comparison based on prediction error, with applications to candidate gene association studies. *Annals of Human Genetics* 2007;71:107–118. [PubMed: 17227481]
- Zhang HP. Classification trees for multiple binary responses. *Journal of American Statistical Association* 1998;93:180–193.
- Zhang HP, Bonney G. Use of classification trees for association studies. *Genetic Epidemiology* 2000;19:323–332. [PubMed: 11108642]
- Zhang, HP.; Singer, R. Recursive Partitioning in the Health Sciences. Springer; New York: 1999.

Figure 1.

The final tree model based on SNPs within the gene CYP2B6 when two dichotomized traits relating to smoking intensity ($Y_{\text{cig-day}}$) and smoking duration ($Y_{\text{smk-yr}}$) were analyzed simultaneously.

Estimated type I errors of the multivariate tree model under the analysis of a bivariate phenotype with the covariance matrix $V = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. For either the continuous outcome or its dichotomized version, the type I error rate was estimated based on 2,000 simulated datasets, with each consisting of 800 subjects.

Table 1

Covariance	Significance level = 0.05		Significance level = 0.01	
	Continuous outcome	Dichotomized outcome	Continuous outcome	Dichotomized outcome
$\rho = 0.0$	0.052	0.046	0.011	0.008
$\rho = 0.1$	0.055	0.051	0.012	0.011
$\rho = 0.2$	0.047	0.048	0.011	0.009

Power comparison at the significance level of 0.05 for the analyses of a bivariate phenotype with the covariance matrix $V = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ simulated under the one locus risk model. For either the continuous outcome or its dichotomized version, the power was estimated based on 500 simulated datasets, each consisting of 800 subjects.

Table 2

Covariance	Continuous outcome		Dichotomized outcome	
	Joint analysis	Univariate-MinP	Joint analysis	Univariate-MinP
$\rho = 0.3$	0.65	0.60	0.47	0.44
$\rho = 0.2$	0.69	0.62	0.51	0.42
$\rho = 0.1$	0.73	0.64	0.53	0.43
$\rho = 0.0$	0.78	0.64	0.57	0.45
$\rho = -0.1$	0.82	0.66	0.59	0.43
$\rho = -0.2$	0.87	0.66	0.65	0.44
$\rho = -0.3$	0.91	0.67	0.70	0.46

Power comparison for the analyses of three phenotypes Y_1 , Y_2 , and Y_3 at the significance level of 0.05. The power was estimated based on 500 simulated datasets, with each consisting of 800 subjects.

Table 3

Phenotypes related to the gene	Selected phenotypes for testing			
	Test	Y_1	Y_1, Y_2	Y_1, Y_2, Y_3
Scenario I:	Joint analysis	0.55	0.44	0.36
	Univariate-MinP	0.55	0.42	0.36
Y_1	Joint analysis	0.55	0.69	0.65
	Univariate-MinP	0.55	0.64	0.57
Scenario II:	Joint analysis	0.55	0.69	0.96
	Univariate-MinP	0.55	0.64	0.74
Y_1, Y_2	Joint analysis	0.55	0.69	0.96
	Univariate-MinP	0.55	0.64	0.74
Scenario III:	Joint analysis	0.55	0.69	0.96
	Univariate-MinP	0.55	0.64	0.74
Y_1, Y_2, Y_3	Joint analysis	0.55	0.69	0.96
	Univariate-MinP	0.55	0.64	0.74

Table 4

Joint association studies of selected smoking behavior related phenotypes.

Testing phenotypes	Optimal tree-size	Adjusted P-value
$Y_{\text{cig-day}}$	4	0.012
$Y_{\text{smk-yr}}$	3	0.032
$Y_{\text{smk-age}}$	3	0.392
$Y_{\text{cig-day}}, Y_{\text{smk-yr}}$	5	0.009
$Y_{\text{cig-day}}, Y_{\text{smk-age}}$	4	0.029
$Y_{\text{smk-yr}}, Y_{\text{smk-age}}$	3	0.144
$Y_{\text{cig-day}}, Y_{\text{smk-yr}}, Y_{\text{smk-age}}$	4	0.015

Naïve OR estimates for the terminal nodes in the final tree model when two smoking behavior phenotypes $Y_{\text{cig-day}}$ and $Y_{\text{smk-yr}}$ were analyzed jointly.

Table 5

Node id	N	$Y_{\text{cig-day}}$		$Y_{\text{smk-yr}}$	
		OR	95% C.I.	OR	95% C.I.
Node 1	61	1.21	0.71-2.06	0.71	0.41-1.23
Node 2	339	0.89	0.68-1.17	1.45	1.10-1.91
Node 3	319	0.56	0.42-0.75	1.34	1.01-1.77
Node 4	538	Ref*	Ref*	Ref*	Ref*
Node 5	61	1.91	1.09-3.34	1.66	0.97-2.84

*The risk category level represented by Node 4 is used as the reference when calculating the OR.