Published in final edited form as:

Stat Biosci. 2018 April ; 10(1): 117–138. doi:10.1007/s12561-017-9189-9.

A Powerful Test for SNP Effects on Multivariate Binary Outcomes using Kernel Machine Regression

Clemontina A. Davenport,

Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC 27707, USA

Arnab Maity,

Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA

Patrick F. Sullivan, and

Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Jung-Ying Tzeng

Department of Statistics, Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695, USA. Department of Statistics, National Cheng-Kung University, Tainan, Taiwan Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan

Abstract

Evaluating multiple binary outcomes is common in genetic studies of complex diseases. These outcomes are often correlated because they are collected from the same individual and they may share common marker effects. In this paper, we propose a procedure to test for effect of a SNP-set on multiple, possibly correlated, binary responses. We develop a score-based test using a nonparametric modeling framework that jointly models the global effect of the marker set. We account for the nonlinear effects and potentially complicated interaction between markers using reproducing kernels. Our testing procedure only requires estimation under the null hypothesis and we use multivariate generalized estimating equations (GEEs) to estimate the model components to account for the correlation among the outcomes. We evaluate finite sample performance of our test via simulation study and demonstrated our methods using the CATIE antibody study data and the CoLaus Study data.

Keywords

correlated binary responses; generalized estimating equations; IBS kernel; kernel machine; nonparametric regression

1 Introduction

Analyzing multiple binary outcomes collectively is a common problem in studies of complex diseases. For example, in cardiac studies, the joint assessment of abnormal blood pressure, reduced pumping ability, and cardiomyopathy together can give a more complete picture of cardiac function than analyzing each of these factors individually [22]. In the Maternal Life Study [4], multiple binary central and autonomic nervous system signs were measured on infants exposed to cocaine prenatally. Together, these signs are important in determining if the newborn displayed narcotic withdrawal syndrome [8]. Likewise, in genetic studies, often multiple binary phenotypes from each individual are collected and these phenotypes combined are relevant to the disease of interest. We consider the situation where genetic information on a particular set of single nucleotide polymorphisms (SNPs) of interest is also collected for each individual and the primary interest lies in determining whether the marker set has any association with the observed multiple binary outcomes. In this article, our primary goal is to develop a global test for association between the marker set of interest and multiple, possibly correlated, binary outcomes. Our motivating data come from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) antibody study and the CoLaus Study (discussed in detail in Sect. 4). In the CATIE study, the presence or absence of antibodies to three neurotrophic herpesviruses were measured in individuals with schizophrenia in addition to genotype data for 492K SNPs [33]. In the CoLaus study, triglycerides, non-high-density lipoprotein (HDL) cholesterol, and glucose were dichotomized into clinically relevant groups and genotyping was performed to study genetic cardiovascular risk factors [10]. Primary challenges in developing association tests for these kinds of data are to incorporate possible correlation between the multiple binary outcomes and to account for complex and possibly nonlinear interactions among the markers in question. We will develop a testing procedure that accounts for both these issues.

Because genes are the basic units in the biological mechanism and SNPs within a gene tend to work concordantly, it is usually better to study the effects of gene or the whole marker set rather than each individual marker. There are several major issues with individual-marker analysis: 1) the results are often not reproducible because the selection of significant markers is arbitrary leading to different biological conclusions [30], 2) single marker analysis often results in reduced power due to multiple comparisons, and 3) since the markers are analyzed individually, potential interactions between them cannot be captured [28]. Many marker set analysis methods have been recently developed (see e.g., [28,15,41,38,46]) to model the joint effects across multiple markers. To be specific, nonparametric regression methods based on kernel machine regression (KMR) have become very popular in association testing problems because of their ability to incorporate complex interactions [24,15,23,43].

Linear theory and methods are well established in machine learning and statistics, but real-world problems often need nonlinear methods for analysis and to draw conclusions. The kernel machine is a useful tool in these problems because it allows for studying complex effects with linear estimation. Specifically, the kernel function projects the (genotype) data from the original space to a new space where the genetic effects can be modeled linearly [14]. Under the KMR framework, one can use a positive definite kernel function to fit the

joint effect of the marker set. This kernel can be thought of as a similarity measure between pairwise individuals and attempts to associate them to the outcome of interest. Because no parametric assumptions are made on the joint marker set effect, it is advantageous over the typical approach of fitting a parametric generalized linear model (GLM), where the group effect is modeled as a linear combination of the individual SNPs. Parametric methods are easier to fit but are typically restrictive because most of the time there is complex interaction between the SNPs that a linear model cannot account for. Also, violating the linearity assumption may result in loss of power. Thus, kernel based methods offer an attractive and powerful alternative to develop association tests. Support vector machines [37] and their least squares extension [34] are common examples of kernel methods. We direct readers to [14] for an overview of the developments of this method in machine learning.

The elegance of kernel methods lies in the fact that using a positive definite kernel allows hypothesis testing in a GLM framework while still accounting for nonlinear dependence and a large number of SNPs. However, most of these methods are developed for the case when one has a single response variable. For example, [24], [15], and [23] use KMR to model the gene pathway effects on a single, phenotypic, response while accounting for other covariates, and [43] extend KMR to single binary responses using a generalized linear mixed model approach. These papers do not address the common case when multiple phenotypes for each individual are observed, such as in the CATIE data. Recently, [25] used KMR for multiple phenotypes and develop a score-based test statistic focusing solely on continuous responses. [38] and [46] give tests for multiple traits based on GEEs.

One of the main challenges of analyzing multivariate binary responses together in a GLM framework is that the correlation structure is usually unknown and the joint distribution of the outcomes is difficult to construct. Thus a direct development of score-based testing procedures as in KMR with continuous outcomes is not straightforward. We will develop our testing procedure by synergistic use of nonparametric KMR and the generalized estimating equation (GEE) frameworks. When studying multiple phenotypes, correlation can arise for two reasons: 1) the responses are collected on the same individual, and 2) the same set of markers can have effect on more than one response variable. A typical (but naive) approach to analyzing such a data set is to consider one outcome at a time, use an association testing procedure of choice (see e.g., [23]), then perform a Bonferroni correction on the resulting p-values. However this procedure completely ignores any correlation structure between the responses, and may suffer from loss of power. Addressing both sources of correlation could result in increased power to detect genetic effects when there is moderate to high correlation among the outcomes, as we will demonstrate in our numerical studies. We address this problem by using generalized estimating equations (GEEs) to estimate the parameters of the model with a user-specified correlation structure. Even if the correlation structure is misspecified, which is likely in practical situations, GEE estimators are still robust and consistent, and the joint distributions do not need to be specified [19]. We develop a score-type test that is advantageous because estimation only needs to be done under the null model. We show in simulation that our proposed method is comparable in performance to naively testing the outcomes individually when there is low correlation among the outcomes, but is more powerful when the correlation is moderate or high.

The rest of this paper is organized as follows. Sect. 2 provides a review of kernel machine regression for the univariate response model and then describes our proposed testing procedure. In Sect. 3, we evaluate the performance of our proposed procedure via simulation study. We find that as both sources of correlation increase, our method is much more powerful than naively evaluating the outcomes separately, while maintaining proper Type I error rate. We then apply our test statistic to the CATIE and CoLaus data in Sect. 4. Finally we provide some discussion and concluding remarks in Sect. 5.

2 Methods

We first review the KMR testing methodology when there is only one binary outcome and its connection to penalized quasi-likelihood mixed modeling framework. Then we will develop our testing procedure when there are multiple binary outcomes observed for each sampling unit.

2.1 Kernel Machine Association Testing for Univariate Binary Outcome

Assume that for each of n subjects we have a univariate binary outcome, demographic covariates such as age and sex, and genetic information such as the number of minor alleles at each SNP on a chromosome. Let y_i be the binary response for i th individual, $i = 1, \dots, n$, with corresponding $q \times 1$ covariate vector $\mathbf{x}_i$ and $p \times 1$ vector $\mathbf{z}_i$ of SNP information. Define $\mathbf{y} = (y_1, \dots, y_n)^T$ and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$. The primary goal is to study the effect of $\mathbf{z}$ on $\mathbf{y}$ after adjusting for $\mathbf{X}$. Because SNPs may interact in an unknown and complicated way, [23] proposed the model

$$\text{logit}(p_i) = \mathbf{x}_i^T \beta + h(\mathbf{z}_i), \quad (1)$$

where $p_i = P(Y_i = 1 | \mathbf{x}_i, \mathbf{z}_i)$, β is a vector of regression coefficients, and $h(\cdot)$ is an unknown scalar function. By modeling the covariates parametrically and the gene effect nonparametrically, this model is flexible in the genetic effects and allows for nonlinearity and complicated interaction of SNPs in the same pathway.

In the standard KMR for binary outcomes, we assume that h lies in a function space $\mathcal{H}_k$. This space is uniquely specified by a symmetric, positive definite kernel function, $K(\cdot, \cdot)$, and is spanned by a set of orthogonal basis functions of the form $\Phi(\mathbf{z}) = \{\phi_j(\mathbf{z})\}_{j=1}^J$ [24].

Typically $\mathcal{H}_k$ is infinite dimensional and the actual basis functions are unknown, thus it is difficult to work with Φ directly. Alternatively, we set the kernel function equal to the inner product defined on $\mathcal{H}_k$, that is, $K(\mathbf{z}_i, \mathbf{z}_j) = \langle \Phi(\mathbf{z}_i), \Phi(\mathbf{z}_j) \rangle_{\mathcal{H}_k}$. This way, instead of representing all $h(\mathbf{z}) \in \mathcal{H}_k$ using the primal representation $h(\mathbf{z}) = \sum_{j=1}^J \omega_j \phi_j(\mathbf{z}) = \Phi(\mathbf{z})^T \omega$, where ω is a vector of coefficients, we can use the dual representation $h(\mathbf{z}) = \sum_{l=1}^n \alpha_l K(\mathbf{z}_l, \mathbf{z}; \rho)$ for some constants α_l . This representation is more convenient to work with because the explicit basis functions do not need to be specified [6,24].

The choice of $K(\cdot, \cdot)$ determines which function space is used to approximate the unknown h . For instance, a d th order polynomial kernel $K(\mathbf{z}_1, \mathbf{z}_2) = (\mathbf{z}_1^T \mathbf{z}_2 + 1)^d$ generates the function space spanned by the basis functions $\phi_j(\mathbf{z}) = \{z_k^a z_{k'}^b : a, b \in \mathbb{N}, a + b \leq d\}$ ($k, k' = 1, \dots, p$) and thus, $d = 1$ and $d = 2$ means we assume $h(\cdot)$ is linear and quadratic in the z 's, respectively. The function space generated by the Gaussian kernel $K(\mathbf{z}_1, \mathbf{z}_2) = \exp\{-\|\mathbf{z}_1 - \mathbf{z}_2\|^2/\rho\}$ is spanned by radial basis functions [5], where ρ is a tuning parameter. There are many kernel options to choose from depending on the problem at hand [14] and in our framework, we can view $K(\cdot, \cdot)$ as a measure of how similar two individuals are in terms of their genotype information. In marker set analysis, a popular kernel for SNP data that allows for SNP interaction (epistasis) is the IBS kernel defined as

$$K(z_i, z_j) = \sum_{s=1}^p \text{IBS}(z_{i,s}, z_{j,s}) / 2p, \quad (2)$$

where $\text{IBS}(z_{i,s}, z_{j,s})$ is the number of alleles shared identical by state (IBS) by individual i and j at SNP s , $s = 1, \dots, p$ [15]. [41] explains that because the number of alleles shared IBS does not depend on different genotype encodings, using the IBS kernel removes the assumption of additivity found in many genetic models. They also suggest using this kernel when the amount of interaction is modest.

Once the kernel is chosen, the parameters β and h in (1) can be estimated by maximizing the penalized log-likelihood using a Fisher scoring or a Newton-Raphson algorithm and [23] construct a score test statistic for the overall genetic effect $H_0: h(\mathbf{z})=0$ (please refer to the Appendix for details). Two advantages of their test statistic are 1) the basis functions for $h(\cdot)$ do not need to be specified, which is often a difficult task, and 2) their test has more power than tests based on a parametric assumption. [23] compared their test to a global test based on the linearity assumption and showed in simulations that their test was as powerful as the global test and suggest using their test as a universal test for both linear and nonlinear pathway effect.

2.2 Testing with Multivariate Binary Outcomes

In this section, we are interested in situations where multiple binary outcomes are measured on each subject. Let y_{ij} be the j th observation for the i th individual, $i = 1, \dots, n$ and $j = 1, \dots, t$ so that the response for a given individual i is $(y_{i1}, \dots, y_{it})$. Define the covariates and SNP information corresponding to y_{ij} as $\mathbf{x}_j$ and $\mathbf{z}_j$, respectively. Multiple outcomes are observed on each subject, but the covariates (e.g., age, sex, etc.) and SNPs do not change over j . Denote $\mathbf{y}_j = (y_{1j}, y_{2j}, \dots, y_{nj})^T$ as the $n \times 1$ vector of all of the j th observations. Because we are grouping the outcomes by observation and not by individual, all of the elements in $\mathbf{y}_j$ are independent of each other but the vectors themselves are not. We combine the covariates in a similar way by defining $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^T$ as the $n \times q$ matrix of covariates for the j th observations. We assume the random variable y_{ij} has a Bernoulli distribution with mean $p_{ij} =$

$E(y_{ij}|\mathbf{x}_i, \mathbf{z}_i)$ and variance $\text{Var}(y_{ij}|\mathbf{x}_i, \mathbf{z}_i) = p_{ij}(1 - p_{ij})$. Marginally, for each $j = 1, \dots, t$, we can write the model

$$\text{logit}(p_{ij}) = \mathbf{x}_i^T \beta_j + h_j(\mathbf{z}_i). \quad (3)$$

Here β_j and h_j indicate that the covariates and SNPs do not necessarily have the same effect on the different responses of each individual. We would like to determine the global effect of the SNPs by testing $H_0: \mathbf{h}(\cdot) = \{h_1(\cdot), \dots, h_t(\cdot)\}^T = \mathbf{0}$.

2.2.1 Test statistic—A naive approach to our testing problem would be to fit the marginal model (3) to each $\mathbf{y}_j$ separately and then test for the significant genetic effect using (9) and adjust for multiple testing, but two issues arise from this naive approach. The first is that the multiple outcomes on a particular individual are correlated and this correlation is not being accounted for. The second is that performing multiple comparison adjustments can result in reduced power. Alternatively, one could use a combined test statistic

$$Q_{\text{naive}}(\hat{\beta}) = (\mathbf{y} - \hat{\mathbf{p}})^T \mathbf{K}^* (\mathbf{y} - \hat{\mathbf{p}}), \quad (4)$$

where $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_t^T)^T$, $\hat{\mathbf{p}} = (\hat{\mathbf{p}}_1^T, \dots, \hat{\mathbf{p}}_t^T)^T$ with $\text{logit}(\hat{\mathbf{p}}_j) = \mathbf{X}\hat{\beta}_j$ and $\hat{\beta} = (\hat{\beta}_1^T, \dots, \hat{\beta}_t^T)^T$ is the vector of estimated regression parameters, and $\mathbf{K}^*$ is an $nt \times nt$ block diagonal matrix with blocks $\mathbf{K}$. The test statistic in (4) combines all the residuals $\mathbf{y}_j - \hat{\mathbf{p}}_j$ and attempts to build a single test to bypass the issue of multiple testing correction. However, Q_{naive} also ignores any correlation structure present among the multiple outcomes and while this strategy works well when the outcomes are truly independent, testing using Q_{naive} results in inflated Type I errors when there is correlation among the outcomes. In a simulation study with $n = 200$ subjects, when the $t = 3$ outcomes per subject were generated to have a compound symmetric correlation structure with correlation 0.5, the Type I error for testing with Q_{naive} was 0.060, and when the correlation was 0.7, the Type I error was 0.079. The more correlated the multiple outcomes are, the more inflated the Type I error becomes and thus Q_{naive} is not a reliable choice for a test statistic. We develop a method for jointly analyzing the multiple binary responses while taking into account the correlation between them but still keeping the flexible modeling of the SNPs. To do this, we utilize the connection of KMR to the mixed model framework but extend it to the case of multiple responses.

Note that fitting the KMR model for each outcome individually is equivalent to iteratively fitting the working models $\tilde{\mathbf{y}}_j = \mathbf{X}\beta_j + \mathbf{h}_j + \epsilon_j$, $j = 1, \dots, t$, where $\tilde{\mathbf{y}}_j = \mathbf{X}\beta_j + \mathbf{h}_j + \mathbf{D}_j^{-1}(\mathbf{y}_j - \hat{\mathbf{p}}_j)$, $\mathbf{h}_j = (h_{1j}, \dots, h_{nj})^T$, $\epsilon_j = (\epsilon_{1j}, \dots, \epsilon_{nj})^T$ and $\mathbf{D}_j = \text{diag}\{p_{1j}(1 - p_{1j}), \dots, p_{nj}(1 - p_{nj})\}$. Univariate test statistics for each response are derived from the working models using the assumptions that $\epsilon_j \sim N(\mathbf{0}, \mathbf{D}_j^{-1})$ and $\mathbf{h}_j \sim N(\mathbf{0}, \tau_j \mathbf{K})$. Thus the naive test statistic $Q_{\text{naive}}(\hat{\beta})$ can be seen as being derived from the joint working model

$$\begin{bmatrix} \tilde{\mathbf{y}}_1 \\ \vdots \\ \tilde{\mathbf{y}}_t \end{bmatrix} = \begin{bmatrix} \mathbf{X}\beta_1 \\ \vdots \\ \mathbf{X}\beta_t \end{bmatrix} + \begin{bmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_t \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_t \end{bmatrix}, \quad (5)$$

with the assumptions that $\mathbf{h} = (\mathbf{h}_1^T, \dots, \mathbf{h}_t^T)^T \sim \text{MVN}(\mathbf{0}, \mathbf{V}_h)$ and $\varepsilon = (\varepsilon_1^T, \dots, \varepsilon_t^T)^T \sim \text{MVN}(\mathbf{0}, \mathbf{D}^{-1})$, where $\mathbf{V}_h$ is a block diagonal matrix with $\tau_j \mathbf{K}$ as the blocks and $\mathbf{D}$ is the block diagonal matrix with blocks $\mathbf{D}_1, \dots, \mathbf{D}_t$. This, however does not take into account any correlation between the multiple outcomes.

In practice, the true correlation structure among the outcomes is unknown, so we posit a working correlation structure for $(\varepsilon_{1l}, \dots, \varepsilon_{it})$ in the form of $\mathbf{G}(\theta)$ where $\mathbf{G}$ is a matrix that is known up to a parameter θ . By constructing the vector of errors as $\varepsilon = (\varepsilon_1^T, \dots, \varepsilon_t^T)^T$, the working correlation of all the errors has the form $\mathbf{S} = \mathbf{G}(\theta) \otimes \mathbf{I}_n$, where $\otimes$ denotes the Kronecker product and $\mathbf{I}_n$ is an $n \times n$ identity matrix. Thus, if we were to use a working unstructured correlation matrix, then $\mathbf{G}(\theta)$ has 1's along the diagonal and $\mathbf{G}(\theta)_{jk} = \theta_{jk}$ when $j \neq k$. Note that the marginal variances of each component of ε remain the same, that is, $\text{Cov}(\varepsilon_j) = \mathbf{D}_j^{-1}$. Combined with the fact that $\text{Cor}(\varepsilon) = \mathbf{S}$, we can take the correlation into account by modifying the working model in (5) so that $\varepsilon \sim \text{MVN}(\mathbf{0}, \mathbf{D}^{-1/2} \mathbf{S} \mathbf{D}^{-1/2})$. The parameters β_j and $\mathbf{h}_j$ can now be estimated using BLUE and BLUP respectively, and the variance components τ_j can be estimated by maximizing the restricted quasi-likelihood criterion.

Our main goal is to test for genetic pathway effects on the multiple binary outcomes $H_0: \mathbf{h}(\cdot) = \mathbf{0}$. To do this, we propose a score-type test statistic based on the derivative of the quasi-likelihood and adopt a similar technique to that of [25] to obtain our final test statistic as

$$Q(\hat{\beta}, \hat{\theta}) = (\mathbf{y} - \hat{\mathbf{p}})^T \left(\hat{\mathbf{S}} \mathbf{D}^{1/2} \right)^{-1} \hat{\mathbf{D}}^{1/2} \mathbf{K}^* \hat{\mathbf{D}}^{1/2} \left(\hat{\mathbf{D}}^{1/2} \hat{\mathbf{S}} \right)^{-1} (\mathbf{y} - \hat{\mathbf{p}}), \quad (6)$$

where $\mathbf{K}_j^*$ is a block diagonal matrix with the j th block equal to $\mathbf{K}$ and all other blocks $\mathbf{0}$ and the hats on $\hat{\mathbf{p}}, \hat{\mathbf{S}}$, and $\hat{\mathbf{D}}$ indicate the evaluation of these matrices at $\hat{\beta}$ and $\hat{\theta}$ (see Appendix). Recall that $\mathbf{S} = \mathbf{G}(\theta) \otimes \mathbf{I}_n$ and hence if no correlation is present, then $\mathbf{S}$ becomes an $nt \times nt$ identity matrix and our statistic reduces to Q_{naive} in (4).

We chose a logistic model for the conditional probability but a probit or some other model could be substituted. Our test statistic still has the advantages of that in [23] but also accounts for correlation in the multiple responses.

2.2.2 Estimation under null hypothesis and null distribution of Q—In order to evaluate Q , β and θ must be estimated under H_0 . To do this, we use the generalized estimating equation (GEE) framework [19] which is a multivariate extension of the quasi-

likelihood approach discussed in detail by [1]. GEEs are advantageous over likelihood based methods because the joint likelihood for the multiple binary outcomes is difficult to construct and in the GEE framework, this joint likelihood does not need to be specified. Also, in practice, the true structure of $\mathbf{G}(\theta)$ is unknown and a working or posited structure is used instead. Contrary to likelihood based methods, if the correlation structure is misspecified, the GEE estimators will still be unbiased and robust. GEE estimators are asymptotically normal and consistent, where consistency depends only on a correctly specified model and not the correlation structure $\mathbf{G}(\theta)$ [19]. In our procedure, we used a completely unstructured correlation matrix for $\mathbf{G}(\theta)$.

We then use a simulation-based technique to get the p-values of our test statistic in (6). We rewrite the test statistic as $Q = \hat{\mathbf{R}}^T \Lambda \hat{\mathbf{R}}$, where Λ is a diagonal matrix of eigenvalues, $\hat{\mathbf{R}}$ is an appropriate vector (see Appendix). Similar to the univariate testing procedure, one could use a chi-squares approximation (see e.g., [23], [40], [7]). Instead we adopt a simulation based procedure to compute p-values. We generate random vectors $\mathbf{R}_b^* \sim \text{MVN}(\mathbf{0}, \mathbf{I})$, $b = 1, \dots, B$, and compute $Q_b^* = \mathbf{R}_b^{*T} \Lambda \mathbf{R}_b^*$. The p-value is defined as $(1/B) \sum_{b=1}^B I[Q_b^* > Q]$ where I denotes the indicator function.

There are two things to note about our proposed test. 1) Our statistic is different than in [21] in that we are only using the quadratic portion of the score. The bias correction for binary responses presented by [21] is not needed here because the affected trace portion of the test statistic is omitted. 2) We approximate Q using a mixture of chi-squared distributions, even though the responses are non-Gaussian. This approach is typical in the literature (see eg. [45,11,47]) and [45] (as well as our simulation studies below) have shown that it performs well in terms of Type I error and power.

3 Simulations and Results

We conducted a simulation study to evaluate the performance of our test statistic. To generate the observations we first generated threshold variables from the model

$$v_{ij} = \mathbf{x}_i^T \beta + a h_j(\mathbf{z}_i) + \varepsilon_{ij}, \quad (7)$$

where the covariates were $\mathbf{x}_i = \{1, x_{i1} \sim N(0,1), x_{i2} \sim N(0,1)\}^T$ and $\beta = (1, 0.5, 0.5)^T$. We simulated $t = 3$ repeated observations for each subject $i = 1, \dots, n$ with multivariate normal errors, where $(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})^T$ had zero mean, unit variance, and a compound symmetric correlation structure with correlation parameter θ . Then the correlated binary response was generated as $y_{ij} = I[v_{ij} > 0]$. The variable a in (7) was used to study the size and power of our test with $a=0$ corresponding to size and $a=0.25, 0.5, 0.75, 1$ corresponding to increasing power. We chose the sample size n to be 200 or 400. To simulate the SNPs, we identified the unique SNP patterns and frequency of occurrence from the *SLC17A1* gene in the CATIE data (see Table 1). We then sampled from these unique patterns with replacement weighted by the frequency which allowed us to preserve the LD structure between the SNPs. We

varied the number of SNPs by using only the 9 relevant SNPs and by adding 21 extraneous SNPs to the 9 relevant ones by sampling from $\{0,1,2\}$ with probability $(0.35,0.2,0.45)$. For the choice of h_j we used two cases: (1)

$$h_1(\mathbf{z}_i) = \frac{1}{4}\{2z_{i9} - z_{i8} + z_{i2}^2 - (z_{i1} - z_{i4})^2 + z_{i5}z_{i6} - (z_{i3} + z_{i6} + z_{i7})\} \text{ and } h_2(\mathbf{z}_i) = h_3(\mathbf{z}_i) = 0, \text{ and (2)}$$

$$h_1(\mathbf{z}_i) = \frac{1}{4}\{2z_{i9} - z_{i8} + z_{i2}^2 - (z_{i1} - z_{i4})^2 + z_{i5}z_{i6} - (z_{i3} + z_{i6} + z_{i7})\} + 0.5z_{i3} \text{ and } h_2(\mathbf{z}_i) = h_3(\mathbf{z}_i) = 0.5z_{i3}. \text{ Note that in Case 1, the only source of correlation is given by the parameter } \theta, \text{ and in Case 2, the two sources of correlation are from } \theta \text{ and from } h_2 \text{ and } h_3. \text{ We used the IBS kernel in (2) to model } h_j. \text{ Lastly, we chose the correlation parameter of the errors to be } \theta = 0, 0.3, 0.5, \text{ and } 0.7.$$

Under Case 1 when $\theta = 0$, we expect the logistic KMR from [23] to perform optimally because the responses are independent. Under Case 2, correlation arises from the errors and from the SNP effects and even when $\theta = 0$, the responses are not independent. We expect our proposed method to perform better for all of Case 2. We ran 20,000 simulations when $a = 0$ to examine the size of our test at $\alpha = 0.005, 0.01$, and 0.05 , and 1,000 simulations to examine the power at $\alpha = 0.05$. We compared our method to the naive method of testing each of the responses $\mathbf{y}_1, \mathbf{y}_2$, and $\mathbf{y}_3$ individually and using a Bonferroni correction for multiple testing. To perform the testing and get the p-values for the naive method, we used the SKAT package in R described by [41]. We also compared our method to the GEE Wald test. Due to perfect collinearity between the SNPs in some simulations, we could not model them directly, thus we first performed principal component analysis on the SNPs and then retained components with corresponding eigenvalues greater than 1 as proxies. We then used a Wald test to jointly test for the association of the retained components.

The results comparing the size of the three methods are found in Table 2. Overall, the KMR approaches had size close to the nominal values when only the relevant SNPs were included. For $n=200$ and $\theta=0.7$, we found that our method had inflated size for small nominal values but stabled out as the nominal value increased. For larger sample size, this inflation disappeared and our method had size very close to the nominal values. When the irrelevant SNPs were added, our method became more conservative as the sample size and θ increased. This is due to the fact that the uncorrelated irrelevant SNPs have no effect on the response and outnumber the SNPs that do. Due to the nature of the Bonferroni correction, testing the repeated observations individually became very conservative when the irrelevant SNPs were added, especially for the smaller sample size. For the most part, the GEE Wald test had inflated Type I error, especially when the number of SNPs increased, as seen in [46]. For larger sample size, the size of the KMR methods were very similar for all θ .

The results for the power of the test for $n = 200$ are given in Fig. 1. Our proposed method is denoted by “MV” for multivariate KMR, the naive approach is denoted by “UV” for univariate KMR with a Bonferroni correction, and GEE Wald test is denoted by “GEE”. When $h_2 = h_3 = 0$ (Case 1, top panel), the two KMR methods generally had comparable power while the GEE method had less power. For moderate to high correlations in the errors, our method had slightly higher power over the UV method and when the correlation was low or nonexistent, the naive method had slightly higher power over our proposed method. When the extraneous SNPs were included (dashed lines), we found that both KMR methods had an

overall decrease in power while the GEE method showed an increase in power under case 2. For any fixed a , as θ increased, the power of our method increased, while the power of the naive method showed no discernible trend, and the power of the GEE method decreased. When h_2 and h_3 were nonzero (Case 2, bottom panels), we saw higher power in our proposed method for all values of a and all error correlations. This is due to the fact that our method accounts for the two sources of correlation in the responses, one from the errors and one from the SNP effects. Recall that even when $\theta = 0$ yielding uncorrelated errors, the dependence in the SNP effects is still causing the responses to be correlated.

The results for $n = 400$ given in Fig. 2 follow similar trends as those for $n = 200$ in that both KMR methods have comparable power with our method having slightly higher power for larger error correlation and the GEEWald method having uniformly lower power. For Case 2 and when extraneous SNPs were added, we again saw our method perform universally better than the naive method and the GEE method, which had a great deal less power. When $\theta = 0.7$ our method's power was nearly double that of the naive method in Case 2.

In summary, our simulation study results showed that both KMR methods had appropriate size when the relevant SNPs were used and became conservative when the irrelevant SNPs were added, while the GEE method generally had liberal type I errors. We also found that the two methods were comparable when the SNP effect was sparse with our proposed method generally having slightly higher power only when θ was large. When the sample size was large and correlation was high, we did see more gain in using our method over the other methods in terms of power. When dependency was present in the SNP effects, our proposed method had higher power than both alternative methods regardless of the amount of error correlation. Again, our method performs so much better because it accounts for both sources of correlation that may be present in the response. Thus when using our method over the other methods, there is no loss when correlation in the errors is small, but there is much gain if there is even a little correlation in the SNP effect no matter what the correlation in the errors is.

4 Real Data Analyses

4.1 CATIE Data Results

The CATIE antibody study from [44], is based on the original CATIE study described by [20]. Studies have found that variation in the MHC region is a replicable risk factor of schizophrenia. This variation plays a known role in the body's responses to neurotrophic infectious agents. [44] also show that there is an association between exposure to these infectious agents and cognitive deficits in individuals with schizophrenia. Our main goal is to analyze the relationship between the presence or absence of antibodies of the infectious agents and the genes located in the MHC region which are associated with schizophrenia.

The philosophy of the CATIE study was to assess antipsychotic drug treatments on a wide range of individuals with schizophrenia. There were 1460 participants and 51% of them gave DNA samples with genotype data available on 492K SNPs [33]. Using the CATIE samples, [44] measured the presence or absence of IgG class antibodies to three herpesviruses, Herpes Simplex Virus type 1 (HSV-1), Herpes Simplex Virus type 2 (HSV-2)

and Cytomegalovirus (CMV). They found an association between exposure to these viruses and cognitive impairments in individuals with schizophrenia. In this analysis, we study the relationship between genes located in the MHC region that are known to be associated with schizophrenia, and presence or absence of antibodies to the herpesviruses listed above.

We applied our proposed multivariate kernel machine method to the CATIE data and compared it with the univariate method applied to each of the responses individually, and to the GEE Wald testing method. The three outcomes of each individual were the presence or absence of the three herpesviruses, the covariates were sex and age, and the genetic information was the number of minor alleles at each SNP for genes associated with schizophrenia and located around the MHC region. We used the IBS kernel to model the SNP effect. We used $B = 10,000$ resampled statistics in order to compute the p-values for our method. For the univariate case, we reported the minimum of the three p-values after correcting for multiple testing.

The results for our data analysis are presented in Table 3. The multivariate, univariate, and GEE methods are denoted by “MV”, “UV”, and “GEE”, respectively. Our method identified two significant genes, *BTN2A2* (p-value 0.0018) and MHC (p-value 0.0015). The univariate method also identified these two regions as significant along with two additional genes, *BTN2A1* and *POM121L2* (both p-values 0.0012). The GEE Wald method identified four genes in addition to those identified by the MV method, *SLC17A1*, *SLC17A3*, *BTN2A1*, and *NOTCH4* (all p-values < 0.0005). One possible reason our method was unable to detect the two genes identified by the UV method is because there were a small number of SNPs in the sets with little correlation, similar to Case 1 in our simulation where the UV method was more favorable. The estimated correlation matrices of these SNPs are shown in Figure 3 (first and second panels), and we see that *POM121L2* had very little correlation, while *BTN2A1* had more where the test was marginally non-significant. Another reason could be that the relationship among the SNPs is not well captured by the IBS kernel. When the linear and quadratic kernel were used (not presented), our method was able to detect these and other regions as significant. Choosing the correct kernel to use is a recurring problem in the KM literature since in practice, the true interaction between SNPs in a set is unknown. The GEE Wald method’s detection of *SLC17A3* and *NOTCH4* could be due to the increased number of SNPs in the sets (13 and 24, respectively) and the resulting inflated Type I error.

4.2 CoLaus Data Results

High levels of non-HDL cholesterol and triglycerides in the body are associated with heart disease risk [3,2,27] and the National Heart, Lung, and Blood Institute (NHLBI) give the cutoffs of 160 mg/dL and 200 mg/dL, respectively for undesirable levels [9]. Those with diabetes and glucose levels greater than 126 mg/dL are also at increased risk for heart disease [13]. Genetic factors can play a role in these increased levels, and we hope to investigate this relationship further. This will be done using the CoLaus study, a population based study among Caucasians in Switzerland to investigate the prevalence of cardiovascular and metabolic risk factors and to find new genetic determinants associated with cardiovascular disease [10]. Of the 6188 total participants, 1704 subjects with genetic data and relevant covariate information were included in analysis. Genotypes were generated

from a targeted sequencing study in 202 drug target genes [29]. We focus our analysis on those SNPs with minor allele frequencies (MAF) > 0.05 from 178 autosomal genes, for which missing genotypes were imputed using MaCH [18]. To apply our method, we dichotomized the three responses of interest (i.e., non-HDL cholesterol, triglycerides, and blood glucose levels) using the NHLBI clinical cutoffs and adjusted for age, sex, and ethnicity using the top 5 principal components. We compared our method to the univariate version in a similar manner to the CATIE analysis.

The genes corresponding to the 12 lowest MV, UV, or GEE p-values are presented in Table 4. Due to the loss in power for multiple testing, the multivariate method only detected one gene to be significantly associated with the three binary outcomes (*HCRT1*, p-value = 0.0002) while the univariate and GEE Wald methods did not identify any genes. The estimated correlation matrix of the SNPs for gene *HCRT1* is shown in Figure 3 (last panel), showing moderate correlation between the SNPs, comparable to Case 2 in our simulation study, where the multivariate test performed better. *HCRT1* (hypocretin (orexin) receptor 1, alias OX1R) has been shown to be associated with glucose regulation ([32,36,12]), as well as cholesterol and triglycerides ([39,16,31]).

5 Discussion

In this paper, we model and test the joint effect of a marker set on multiple binary responses using kernel machine regression. The kernel method allows for nonlinearity and complicated interactions among markers in the same pathway and avoids the issues of parametric techniques and assumptions. Our procedure adjusts for other covariates and accounts for two sources of correlation that can arise in multiple outcomes, namely the random individual effect, and the shared marker effects. We developed a score-based test using the GEE framework to determine the joint marker effects on all the responses collectively. We conducted a simulation study to assess the performance of our method compared to evaluating the responses individually and using a Bonferroni correction. We found that the two methods were comparable when the responses were independent or had weak correlation solely due to the individual random effect. When correlation from shared markers was introduced, our method was uniformly more powerful power, regardless of individual random effect.

We found similar results when we applied our methods to two real datasets. In the CATIE study, where the source of correlation among the responses were most likely from the subjects, the univariate method was able to detect more genes as significantly associated with the outcomes of interest. However in the CoLaus study, where correlation among the responses came from both the individuals and the SNPs, the multivariate method identified the sole significant gene. Because there is no meaningful loss when correlation is low and considerable gain when correlation is moderate to high, we recommend our method as a general tool for multiple binary phenotype analysis.

In this work, we focus on SNP-set effects using the IBS kernel, but our method is not restricted to just genetic data. Environmental, biological, and any other types of factors of interest that have nonlinear effects or complicated interactions can be modeled in this

framework as long as an appropriate kernel is chosen. For example, many measurements are taken during a standard echocardiogram and a combination of these variables can be used to predict presence of heart failure or different types of cardiomyopathies [26]. One could then develop tests using the methods in this paper to determine if certain combinations of echo parameters have significant effect on the binary outcomes of interest. Predictors that interact with markers can also be modeled using kernel machine regression and tests for significance can be performed. Thus our method can be used as a device in studying phenotypic data with genetic predictors or general correlated binary outcomes with environmental or biological factors.

In general, the choice of kernel plays a key role and an inappropriate kernel can yield poor performance and loss of power [35]. Methods to overcome this rely on “learning the kernel”, where the kernel is adapted to the current problem. A common example is to use cross-validation techniques to select the optimal kernel from a set of predefined kernels [17]. [42] propose using a composite kernel that is some weighted combination of a set of candidate kernels. In this case, the composite kernel will have comparable performance to the unknown optimal kernel, and will perform better than any individual kernel. They also present a perturbation approach to selecting the kernel with the minimum p-value among a set of candidate kernels, and their method allows for correlation among the covariates and SNPs.

Another important issue to be noted is that in the derivation of the null distribution of the test statistic, we implicitly assume that the working responses $\tilde{\mathbf{y}}$ is normal. This may not be a satisfactory assumption of non-Gaussian, especially binary data. As a result, our proposed test is not an exact test. However, such strategies have been successfully applied to univariate tests for binary data, see for example, [23], [40], [7], [45], among others. Also, in our numerical studies we observed that the Type I errors remained close to nominal levels in all the situations we investigated.

Finally, our method can also serve as a preliminary analysis in a genome-wide association study (GWAS). In GWAS, many markers across the entire genome are identified as important to the trait or phenotype of interest. As mentioned in the introduction, testing marker-by-marker has many drawbacks, and testing all markers together will have very little, if any power to detect signals. Thus marker sets can be appropriately established gene-by-gene and our methodology can be used as a screening process to identify important genes by detecting significant global effects of the marker sets. Once these genes are identified, one can conduct further gene-level analysis to interpret the global signal and to determine which SNPs in the set are driving the global effect. We note that our method is scalable to genome wide studies. For $n = 200$ individuals, $t = 3$ responses per individual, $p = 30$ SNPs, $B = 10000$, and using the IBS kernel, on average, our test takes 3.3 seconds to run on an Intel i7 2.60 GHz processor (using one core) with 8 GB of RAM.

Acknowledgments

The authors thank Dr. Robert Yolken at Johns Hopkins University for providing the antibody data. The authors also thank Drs. Peter Vollenweider and Gerard Waeber, PIs of the CoLaus study, and Drs. Meg Ehm and Matthew Nelson, collaborators at GlaxoSmithKline for providing the Co- Laus phenotype and sequence data. This work was

supported by National Institutes of Health grants R00 ES017744 (to A.M.), R01 MH084022 (to J.Y.T. and P.F.S.), and P01 CA142538 (to J.Y.T.).

References

1. Agresti A. Categorical data analysis. 2. Wiley; New York: 2002.
2. Arsenault BJ, Rana JS, Stroes ESG, Desprs J-P, Shah PK, Kastelein JJP, Wareham NJ, Boekholdt SM, Khaw K-T. Beyond low-density lipoprotein cholesterol: respective contributions of nonhigh-density lipoprotein cholesterol levels, triglycerides, and the total cholesterol/high-density lipoprotein cholesterol ratio to coronary heart disease risk in apparently healthy men and women. *J Am Coll Cardiol*. 2010; 55:3541.
3. Austin MA, Hokanson JE, Edwards KL. Hypertriglyceridemia as a cardiovascular risk factor. *Am J Cardiol*. 1998; 81(7):B12B.
4. Bauer CR, Shankaran S, Bada HS, Lester B, Wright LL, Krause-Steinrauf H, Smeriglio VL, Finnegan LP, Maza PL, Verter J. The maternal lifestyle study: drug exposure during pregnancy and short-term maternal outcomes. *Am J Obstet Gynecol*. 2002; 186:487–495. [PubMed: 11904612]
5. Buhmann MD. Radial basis functions: theory and implementations. Cambridge University Press; Cambridge: 2003.
6. Burges CJ. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discov*. 1998; 2:121–167.
7. Chen J, Chen W, Zhao N, Wu MC, Schaid DJ. Small-sample kernel association tests for human genetic and microbiome association studies. *Genetic Epidemiology*. 2016; 40:5–19. [PubMed: 26643881]
8. Das A, Poole WK, Bada HS. A repeated measures approach for simultaneous modeling of multiple neurobehavioral outcomes in newborns exposed to cocaine in utero. *Am J Epidemiol*. 2004; 159:891–899. [PubMed: 15105182]
9. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive summary of the third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). *JAMA*. 2001; 285:2486–2497. [PubMed: 11368702]
10. Firmann M, Mayor V, Vidal PM, Bochud M, Pecoud A, Hayoz D, Paccaud F, Preisig M, Song KS, Yuan X, et al. The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovasc Disord*. 2008; 8:6. [PubMed: 18366642]
11. Freytag S, Bickeböller H, Amos CI, Kneib T, Schlather M. A novel kernel for correcting size bias in the logistic kernel machine test with an application to rheumatoid arthritis. *Hum Hered*. 2012; 74:97–108. [PubMed: 23466369]
12. Girault EM, Foppen E, Ackermans MT, Fliers E, Kalsbeek A. Central administration of an orexin receptor 1 antagonist prevents the stimulatory effect of Olanzapine on endogenous glucose production. *Brain Res*. 2013; 1527:238–245. [PubMed: 23830851]
13. Grundy SM, Cleeman JI, Daniels SR, Donato KA, Eckel RH, Franklin BA, Gordon DJ, Krauss RM, Savage PJ, Smith SC Jr, et al. Diagnosis and management of the metabolic syndrome: an American Heart Association/National Heart, Lung, and Blood Institute scientific statement. *Circulation*. 2005; 112:2735–2752. [PubMed: 16157765]
14. Hofmann T, Schölkopf B, Smola AJ. Kernel methods in machine learning. *Ann Stat*. 2008; 36:1171–1220.
15. Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Gen*. 2008; 82:386–397.
16. Kralisch S, Klein J, Lossner U, Bluher M, Paschke R, Stumvoll M, Fasshauer M. Isoproterenol, TNFalpha, and insulin downregulate adipose triglyceride lipase in 3T3-L1 adipocytes. *Mol Cell Endocrinol*. 2005; 240:43–49. [PubMed: 16009485]
17. Lanckriet GRG, Cristianini N, Bartlett P, El Ghaoui L, Jordan M. Learning the kernel matrix with semidefinite programming. *J Mach Learn Res*. 2004; 5:27–72.

18. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol.* 2010; 34:816–834. [PubMed: 21058334]
19. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika.* 1986; 73:13–22.
20. Lieberman JA, Stroup TS, McEvoy JP, Swartz MS, Rosenheck RA, Perkins DO, Keef RSE, Davis SM, Davis CE, Lebowitz BD, et al. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *N Engl J Med.* 2005; 353:1209–1223. [PubMed: 16172203]
21. Lin X. Variance component testing in generalised linear models with random effects. *Biometrika.* 1997; 84:309–326.
22. Lipsitz SR, Fitzmaurice GM, Ibrahim JG, Sinha D, Parzen M, Lipshultz S. Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: an application to acquired immune deficiency syndrome data. *J R Stat Soc.* 2009; 172:3–20.
23. Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics.* 2008; 9:292. [PubMed: 18577223]
24. Liu D, Lin X, Ghosh D. Semiparametric regression of multi-dimensional genetic pathway data: least squares kernel machines and linear mixed models. *Biometrics.* 2007; 63:1077–1088.
25. Maity A, Sullivan PF, Tzeng JY. Multivariate phenotype association analysis by marker-set kernel machine regressions. *Genet Epidemiol.* 2012; 36:686–695. [PubMed: 22899176]
26. McCartan C, Mason R, Jayasinghe SR, Griffiths LR. Cardiomyopathy classification: ongoing debate in the genomics era. *Biochem Res Int.* 2012; 2012:796926. [PubMed: 22924131]
27. Miller M, Stone NJ, Ballantyne C, Bittner V, Criqui MH, Ginsberg HN, Goldberg AC, Howard WJ, Jacobson MS, Kris-Etherton PM, et al. Triglycerides and cardiovascular disease: a scientific statement from the American Heart Association. *Circulation.* 2011; 123:2292–2333. [PubMed: 21502576]
28. Nam D, Kim SY. Gene-set approach for expression pattern analysis. *Brief Bioinform.* 2008; 9:189–197. [PubMed: 18202032]
29. Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science.* 2012; 337:100–114. [PubMed: 22604722]
30. Pan KH, Lih CJ, Cohen SN. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proc Natl Acad Sci USA.* 2005; 102:8961–8965. [PubMed: 15951424]
31. Shen Y, Zhao Y, Zheng D, Chang X, Ju S, Guo L. Effects of orexin A on GLUT4 expression and lipid content via MAPK signaling in 3T3-L1 adipocytes. *J Steroid Biochem Mol Biol.* 2013; 138:376–383. [PubMed: 23907013]
32. Sikder D, Kodadek T. The neurohormone orexin stimulates hypoxia-inducible factor-1 activity. *Genes Dev.* 2007; 21:2995–3005. [PubMed: 18006690]
33. Sullivan PF, Lin D, Tzeng JY, van den Oord E, Perkins D, Stroup TS, Wagner M, Lee S, Wright FA, Zou F, et al. Genomewide association for schizophrenia in the CATIE study: results of Stage 1. *Mol Psychiatry.* 2008; 13:570–584. [PubMed: 18347602]
34. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Least squares support vector machines. *World Scientific; Singapore:* 2002.
35. Szafranski M, Grandvalet Y, Rakotomamonjy A. Composite kernel learning. *Mach Learn.* 2010; 79:73–103.
36. Tsuneki H, Wada T, Sasaoka T. Role of orexin in the central regulation of glucose and energy homeostasis. *Endocr J.* 2012; 59:365–374. [PubMed: 22293586]
37. Vapnik VN. Statistical learning theory. *Wiley; New York:* 1998.
38. Wang X, Lee S, Zhu X, Redline S, Lin X. GEE-based SNP set association test for continuous and discrete traits in family based association studies. *Genet Epidemiol.* 2013; 37:778786.
39. Wortley KE, Chang GQ, Davydova Z, Leibowitz SF. Peptides that regulate food intake: orexin gene expression is increased during states of hypertriglyceridemia. *Am J Physiol.* 2003; 284:R1454–R1465.

40. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP Set Analysis for Case-Control GenomeWide Association Studies. *American Journal of Human Genetics*. 2010; 86:929–942. [PubMed: 20560208]
41. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). *Am J Hum Genet*. 2011; 89:82–93. [PubMed: 21737059]
42. Wu M, Maity A, Lee S, Simmons EM, Harmon QE, Lin X, Engel S, Molldrem JJ, Armistead PM. Kernel machine SNP-set testing under multiple candidate kernels. *Genet Epidemiol*. 2013; 37:267–275. [PubMed: 23471868]
43. Yan Q, Tiwari HK, Yi N, Gao G, Zhang K, Lin W, Lou XY, Cui X, Liu N. A sequence kernel association test for dichotomous traits in family samples under a generalized linear mixed model. *Hum Hered*. 2015; 79:60–68. [PubMed: 25791389]
44. Yolken RH, Torrey EF, Lieberman JA, Yang S, Dickerson FB. Serological evidence of exposure to herpes simplex virus Type 1 is associated with cognitive deficits in the CATIE schizophrenia sample. *Schizophr Res*. 2011; 128:61–65. [PubMed: 21353483]
45. Zhang D, Lin X. Hypothesis testing in semiparametric additive mixed models. *Biostatistics*. 2003; 4:57–74. [PubMed: 12925330]
46. Zhang Y, Xu z, Shen X, Pan W. the Alzheimer's Disease Neuroimaging Initiative. Testing for association with multiple Traits in generalized estimation equations, with application to neuroimaging data. *Neuroimage*. 2014; 96:309–325. [PubMed: 24704269]
47. Zhao Y, Chen F, Zhai R, Lin X, Diao N. Association test based on SNP set: logistic kernel machine based test vs. principal component analysis. *PLoS ONE*. 2012; 7:e44978. [PubMed: 23028716]

Appendix

From section 2.1, the parameters β and h in (1) can be estimated by maximizing the penalized log-likelihood using a Fisher scoring or a Newton-Raphson algorithm. [23] show that, by treating β as a vector of fixed effects and $\mathbf{h} = (h_1, \dots, h_n)^T$ as a vector of random effects, the logistic KM estimator is the same as the penalized quasi-likelihood estimator from a logistic mixed model $\text{logit}(p_i) = \mathbf{x}_i^T \beta + h_i$, where $\mathbf{h} \sim N(\mathbf{0}, \tau \mathbf{K})$, $\tau = 1/\lambda$, λ is the penalty parameter from the penalized likelihood, and $\mathbf{K}$ is a square matrix whose (i, j) th element is (2). The normal equations given in (5) of [23] coincides with iteratively fitting a working linear mixed model $\tilde{\mathbf{y}} = \mathbf{X}\beta + \mathbf{h} + \mathbf{e}$ until convergence, where β and $\mathbf{h}$ are estimated using BLUE and BLUP respectively and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{D})$ where $\mathbf{D} = \text{diag}\{p_i(1-p_i)\}$. The regularization parameter τ can be estimated by treating it as a variance component and maximizing the REML criterion

$$\ell \approx -\frac{1}{2} \log |\mathbf{V}_u| - \frac{1}{2} \log |\mathbf{X}^T \mathbf{V}_u^{-1} \mathbf{X}| - \frac{1}{2} (\tilde{\mathbf{y}} - \mathbf{X}\hat{\beta})^T \mathbf{V}_u^{-1} (\tilde{\mathbf{y}} - \mathbf{X}\hat{\beta}), \quad (8)$$

where $\mathbf{V}_u = \mathbf{D}^{-1} + \tau \mathbf{K}$ and $\tilde{\mathbf{y}} = \mathbf{X}\beta + \mathbf{h} + \mathbf{D}^{-1}(\mathbf{y} - \mathbf{p})$. We refer to [23] for full details.

Testing the overall genetic effect $H_0: h(\mathbf{z}) = 0$ for univariate responses is equivalent to testing $H_0: \tau = 0$. [23] propose the following score test statistic based on the derivative of (8) with respect to τ

$$S = \frac{Q(\hat{\beta}_0) - p_Q}{\sigma_Q}, \quad (9)$$

where $Q(\hat{\beta}_0) = (\tilde{\mathbf{y}} - \mathbf{X}\hat{\beta}_0)^T \mathbf{D} \mathbf{K} \mathbf{D} (\tilde{\mathbf{y}} - \mathbf{X}\hat{\beta}_0) = (\mathbf{y} - \hat{\mathbf{p}}_0)^T \mathbf{K} (\mathbf{y} - \hat{\mathbf{p}}_0)$, $\text{logit}(\hat{\mathbf{p}}_0) = \mathbf{X}\hat{\beta}_0$, $\hat{\beta}_0$ is the MLE of β under the null logistic model, $p_Q = \text{tr}\{\mathbf{P}_0 \mathbf{K}\}$, $\sigma_Q = 2\text{tr}\{\mathbf{P}_0 \mathbf{K} \mathbf{P}_0 \mathbf{K}\}$, and $\mathbf{P}_0 = \mathbf{D}_0 - \mathbf{D}_0 \mathbf{X} (\mathbf{X}^T \mathbf{D}_0 \mathbf{X})^{-1} \mathbf{X}^T \mathbf{D}_0$ where $\mathbf{D}_0 = \text{diag}\{\hat{p}_0(1 - \hat{p}_0)\}$.

From section 2.2.1, we modify the working model in (5). Define $\mathbf{D}$ as the block diagonal matrix with blocks $\mathbf{D}_1, \dots, \mathbf{D}_t$. Then the variance-covariance matrix of ε is $\mathbf{D}^{-1} \mathbf{D}^{1/2} \mathbf{S} \mathbf{D}^{1/2} \mathbf{D}^{-1} = \mathbf{D}^{-1/2} \mathbf{S} \mathbf{D}^{-1/2}$ and the modified working model will have the same form as (5) but with $\varepsilon \sim \text{MVN}(\mathbf{0}, \mathbf{D}^{-1/2} \mathbf{S} \mathbf{D}^{-1/2})$. The parameters β_j and $\mathbf{h}_j$ can now be estimated using BLUE and BLUP respectively, and the variance components τ_j can be estimated by maximizing the restricted quasi-likelihood criterion

$$\ell \approx -\frac{1}{2} \log |\mathbf{V}_m| - \frac{1}{2} \log |\mathbf{X}^T \mathbf{V}_m^{-1} \mathbf{X}| - \frac{1}{2} \begin{pmatrix} \tilde{\mathbf{y}}_1 - \mathbf{X}\hat{\beta}_1 \\ \vdots \\ \tilde{\mathbf{y}}_t - \mathbf{X}\hat{\beta}_t \end{pmatrix}^T \mathbf{V}_m^{-1} \begin{pmatrix} \tilde{\mathbf{y}}_1 - \mathbf{X}\hat{\beta}_1 \\ \vdots \\ \tilde{\mathbf{y}}_t - \mathbf{X}\hat{\beta}_t \end{pmatrix}, \quad (10)$$

where $\mathbf{V}_m = \mathbf{D}^{-1/2} \mathbf{S} \mathbf{D}^{-1/2} + \mathbf{V}_h$.

The main goal is to test for genetic pathway effects $H_0: \mathbf{h}(\cdot) = \mathbf{0}$ which is equivalent to testing $H_0: \tau_1 = \dots = \tau_t = 0$. To do this, we propose a score-type test statistic based on the derivative of the quasi-likelihood like that in (10). Taking the derivative of the criterion in (10) with respect to τ_j for $j = 1, \dots, t$ and then setting $\tau_j = 0$, the score function for τ_j is $S_j = Q_j(\beta, \theta) - p_{jQ}$, where

$$Q_j(\beta, \theta) = (\mathbf{y} - \hat{\mathbf{p}})^T (\mathbf{S} \mathbf{D}^{1/2})^{-1} \mathbf{D}^{1/2} \mathbf{K}_j^* \mathbf{D}^{1/2} (\mathbf{D}^{1/2} \mathbf{S})^{-1} (\mathbf{y} - \hat{\mathbf{p}})$$

and $p_{jQ} = \text{tr}\{\mathbf{P} \mathbf{K}_j^*\}$. Because the τ_j 's are considered as variance components and thus are nonnegative, testing $H_0: \tau_1 = \dots = \tau_t = 0$ is equivalent to testing $H_0: \tau_1 + \dots + \tau_t = 0$ and we adopt a similar technique to [25].

From section 2.2.2, in order to evaluate Q in (6), we estimate β and θ under the null using GEEs. We posit the GEEs under H_0

$$\tilde{\mathbf{X}}^T \mathbf{D} \mathbf{V}_{m0}^{-1} (\mathbf{y} - \mathbf{p}) = \mathbf{0}, \quad (11)$$

where $\tilde{\mathbf{X}}$ is an $nt \times nt$ block diagonal matrix with elements $\mathbf{X}$ and $\mathbf{V}_{m0} = \mathbf{D}^{-1/2} \mathbf{S} \mathbf{D}^{-1/2}$. If there is no genetic pathway effect (H_0 : $\mathbf{h}(\cdot) = \mathbf{0}$ is true), then $\mathbf{V}_{m0}$ is the working variance-covariance matrix of $\mathbf{y}$. To solve (11), [19] suggest using a modified Fisher scoring algorithm to find β and a method of moments estimation for θ . The updating equation is

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + (\tilde{\mathbf{X}}^T \hat{\mathbf{D}} \hat{\mathbf{V}}_{m0}^{-1} \hat{\mathbf{D}} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \hat{\mathbf{D}} \hat{\mathbf{V}}_{m0}^{-1} (\mathbf{y} - \hat{\mathbf{p}}).$$

The initial estimates in the first iteration come from fitting a generalized linear model assuming independence.

We then use a simulation-based technique to get the p-values of Q . It can be shown that, under H_0 , $\text{var}(\mathbf{y}_j - \hat{\mathbf{p}}_j)$ can be approximated as $\hat{\mathbf{D}}_j - \hat{\mathbf{D}}_j \mathbf{X} (\mathbf{X}^T \hat{\mathbf{D}}_j \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{D}}_j = \hat{\mathbf{P}}_j$. This follows from the fact that under H_0 , $\mathbf{y}_j - \hat{\mathbf{p}}_j = \mathbf{D}_j (\tilde{\mathbf{y}}_j - \mathbf{X} \hat{\beta}_j)$, and that $\text{Cov}(\tilde{\mathbf{y}}_j - \mathbf{X} \hat{\beta}_j) = \mathbf{D}_j^{-1} - \mathbf{D}_j^{-1} \mathbf{X} (\mathbf{X}^T \mathbf{D}_j \mathbf{X})^{-1} \mathbf{X}^T \mathbf{D}_j^{-1}$ from linear model theory. If the multiple outcomes were independent, then the variance-covariance matrix of $\mathbf{y} - \hat{\mathbf{p}}$ would be $\hat{\mathbf{P}}$, a block diagonal matrix with elements $\hat{\mathbf{P}}_j$, but since the outcomes are correlated, the variance-covariance matrix of $\mathbf{y} - \hat{\mathbf{p}}$ is $\hat{\mathbf{P}}^{1/2} \hat{\mathbf{S}} \hat{\mathbf{P}}^{1/2}$, which is no longer block diagonal. By defining $\hat{\mathbf{M}} = \hat{\mathbf{P}}^{1/2} \hat{\mathbf{S}} \hat{\mathbf{P}}^{1/2}$, (6) can be rewritten as $Q = \{\hat{\mathbf{M}}^{-1/2} (\mathbf{y} - \hat{\mathbf{p}})\}^T \hat{\mathbf{B}} \{\hat{\mathbf{M}}^{-1/2} (\mathbf{y} - \hat{\mathbf{p}})\}$ where $\hat{\mathbf{B}} = \hat{\mathbf{M}}^{1/2} (\hat{\mathbf{S}} \hat{\mathbf{D}}^{1/2})^{-1} \hat{\mathbf{D}}^{1/2} \mathbf{K} \hat{\mathbf{D}}^{1/2} (\hat{\mathbf{D}}^{1/2} \hat{\mathbf{S}})^{-1} \hat{\mathbf{M}}^{1/2}$. Using eigenvalue decomposition, we can write $\hat{\mathbf{B}} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$ where $\mathbf{U}$ is the matrix of orthogonal eigenvectors of $\hat{\mathbf{B}}$ and $\mathbf{\Lambda}$ is a diagonal matrix whose elements are the corresponding eigenvalues. Thus $Q = \hat{\mathbf{R}}^T \mathbf{\Lambda} \hat{\mathbf{R}}$, where $\hat{\mathbf{R}} = \mathbf{U}^T \hat{\mathbf{M}}^{-1/2} (\mathbf{y} - \hat{\mathbf{p}})$.

Fig. 1.

Power calculations for multivariate KMR (circle lines), univariate logistic KMR with Bonferroni correction (triangle lines), and PC GEE (plus lines) for sample size $n = 200$, number of SNPs $p = 9$ (solid lines) and $p = 30$ (dashed lines), and different correlations θ . The top panel (a)–(d) refers to Case 1 where the response is only correlated through the subject effects. The bottom panel (e)–(h) refers to Case 2 where the responses have correlation from shared marker effects. The effect of the SNPs is scaled by a .

Fig. 2.

Power calculations for multivariate KMR (circle lines), univariate logistic KMR with Bonferroni correction (triangle lines), and PC GEE (plus lines) for sample size $n = 400$, number of SNPs $p = 9$ (solid lines) and $p = 30$ (dashed lines), and different correlations θ . The top panel (a)–(d) refers to Case 1 where the response is only correlated through the subject effects. The bottom panel (e)–(h) refers to Case 2 where the responses have correlation from shared marker effects. The effect of the SNPs is scaled by a .

Fig. 3. Estimated correlation matrices among SNPs for *BTN2A1* (a) and *POM121L2* (b) from the CATIE study, and *HCRTR1* (c) from the CoLaus study.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Frequency table of genotype of *SLC17A1* gene in CATIE schizophrenia data.

Genotype	Frequency	%
221212211	114	0.165
222202220	103	0.149
112111121	72	0.104
111121112	46	0.067
220222202	41	0.059
112112221	38	0.055
111122212	28	0.041
211122212	26	0.038
111222212	24	0.035
002021122	21	0.030
212112221	15	0.022
002020022	10	0.014
102122222	10	0.014
102022222	9	0.013
220222211	8	0.012
002121122	8	0.012
002122222	8	0.012
102021122	7	0.010
202022222	7	0.010
221212220	6	0.009
111121121	6	0.009
220222212	6	0.009
002022222	6	0.009
221212221	5	0.007
112212221	5	0.007
011121112	5	0.007
111222222	4	0.006
121212211	3	0.004

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Genotype	Frequency	%
1111121202	3	0.004
120222202	3	0.004
211222212	3	0.004
211122222	3	0.004
002222222	3	0.004
112111211	2	0.003
111222202	2	0.003
011122212	2	0.003
002122212	2	0.003
011222212	2	0.003
002222200	2	0.003
112112220	1	0.001
212111211	1	0.001
222222222	1	0.001
212112211	1	0.001
112212211	1	0.001
111122121	1	0.001
212212221	1	0.001
102122221	1	0.001
0020202	1	0.001
10212112	1	0.001
002021212	1	0.001
111121212	1	0.001
211022212	1	0.001
102122212	1	0.001
102222212	1	0.001
102020022	1	0.001
011121122	1	0.001
102121122	1	0.001
012112122	1	0.001
111122222	1	0.001

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Genotype	Frequency	%
1	1	0.001
0	2	
2	2	
2	2	
2	2	
2	2	
2	2	
1		

Table 2

Type I error rates for multivariate KMR (MV) univariate logistic KMR with Bonferroni correction (UV), and PC GEE (GEE) for different nominal values α , sample sizes n , number of SNPs p , and correlations θ .

α	$\theta = 0$			$\theta = 0.3$			$\theta = 0.5$			$\theta = 0.7$		
	.005	.01	.05	.005	.01	.05	.005	.01	.05	.005	.01	.05
$p = 9$	n = 200											
	MV	.0059	.012	.053	.0062	.012	.054	.0062	.011	.051	.0084	.013
	UV	.0056	.011	.049	.0053	.010	.049	.0055	.011	.047	.0051	.011
	GEE	.0084	.016	.067	.0087	.018	.069	.0090	.017	.064	.0084	.015
	n = 300											
	MV	.0050	.010	.052	.0060	.011	.048	.0063	.011	.050	.0090	.016
$p = 30$	UV	.0034	.007	.038	.0034	.007	.041	.0032	.006	.038	.0038	.007
	GEE	.0340	.052	.149	.0323	.052	.154	.0306	.049	.146	.0280	.045
	n = 400											
	MV	.0068	.011	.052	.0058	.011	.049	.0049	.010	.049	.0051	.009
	UV	.0060	.011	.049	.0046	.010	.048	.0042	.010	.049	.0058	.011
	GEE	.0068	.013	.059	.0066	.012	.060	.0058	.012	.058	.0060	.011
$p = 90$	n = 200											
	MV	.0056	.011	.052	.0049	.009	.048	.0054	.009	.042	.0057	.010
	UV	.0052	.010	.046	.0042	.009	.042	.0047	.009	.044	.0050	.009
	GEE	.0168	.027	.099	.0141	.025	.092	.0155	.027	.096	.0131	.023
	n = 300											
	MV	.0056	.011	.052	.0049	.009	.048	.0054	.009	.042	.0057	.010
$p = 90$	UV	.0052	.010	.046	.0042	.009	.042	.0047	.009	.044	.0050	.009
	GEE	.0168	.027	.099	.0141	.025	.092	.0155	.027	.096	.0131	.023

Table 3

Resulting p-values for multivariate (MV), univariate (UV), and GEE Wald testing for the CATIE Antibody study using the IBS kernel. Each gene had a different sample size, n , and different number of SNPs p . The Bonferroni correction is $0.05/12 = 0.0042$ in order to analyze 12 genes and significant results are in bold.

Gene	n	p	MV	UV	GEE
<i>SLC17A1</i>	690	9	0.0074	0.0042	<0.0001
<i>SLC17A3</i>	388	13	0.4452	0.2430	<0.0001
<i>BTN3A2</i>	632	5	0.1070	0.0534	0.0737
<i>BTN2A2</i>	419	5	0.0018	0.0006	<0.0001
<i>BTN2A1</i>	661	4	0.0044	0.0012	<0.0001
<i>HIST1H2AG</i>	419	2	0.5366	0.3387	0.3384
<i>HIST1H2BJ</i>	645	2	0.1180	0.0135	0.0778
<i>PRSS16</i>	408	1	0.2739	0.1257	0.0229
<i>POM121L2</i>	643	4	0.0405	0.0012	0.0606
<i>ZNF184</i>	389	8	0.9296	1	0.9767
<i>NOTCH4</i>	587	24	0.2118	0.1356	0.0004
MHC	490	787	0.0015	<0.0001	<0.0001

Table 4

Resulting 12 lowest p-values for multivariate (MV), univariate (UV), or GEE Wald (GEE) testing for the CoLaus study using the IBS kernel. Each gene had a sample size of $n = 1704$, and a different number of SNPs p . The Bonferroni correction is $0.05/178 = 0.00028$ in order to analyze 178 genes and significant results are in bold.

Gene	p	MV	UV	GEE
<i>HCRTR1</i>	7	0.0002	0.0005	0.0021
<i>CCL11</i>	2	0.0190	0.0005	0.3815
<i>IL6</i>	1	0.0015	0.0007	0.0054
<i>NOS2</i>	8	0.0900	0.0011	0.0055
<i>OSM</i>	1	0.1347	0.0181	0.1922
<i>CD4</i>	9	0.0182	0.0701	0.8607
<i>RORC</i>	4	0.0228	0.0533	0.4667
<i>SLC6A9</i>	2	0.0828	0.0255	0.1727
<i>PYGB</i>	20	0.0362	0.1869	0.7654
<i>MAPK11</i>	2	0.1423	0.0402	0.0955
<i>APCS</i>	2	0.0503	0.1066	0.1826
<i>PTGER1</i>	4	0.2354	0.0528	0.4039