Published in final edited form as:

Stat Methods Med Res. 2018 March ; 27(3): 905–919. doi:10.1177/0962280216647422.

An efficient genome-wide association test for mixed binary and continuous phenotypes with applications to substance abuse research

Anne Buu¹, L Keoki Williams^{2,3}, James J Yang⁴

¹Department of Health Behavior and Biological Sciences, University of Michigan, USA

²Department of Internal Medicine, Henry Ford Health System, USA

³The Center for Health Policy and Health Services Research, Henry Ford Health System, USA

⁴School of Nursing, University of Michigan, USA

Abstract

We propose a new genome-wide association test for mixed binary and continuous phenotypes that uses an efficient numerical method to estimate the empirical distribution of the Fisher's combination statistic under the null hypothesis. Our simulation study shows that the proposed method controls the type I error rate and also maintains its power at the level of the permutation method. More importantly, the computational efficiency of the proposed method is much higher than the one of the permutation method. The simulation results also indicate that the power of the test increases when the genetic effect increases, the minor allele frequency increases, and the correlation between responses decreases. The statistical analysis on the database of the Study of Addiction: Genetics and Environment demonstrates that the proposed method combining multiple phenotypes can increase the power of identifying markers that may not be, otherwise, chosen using marginal tests.

Keywords

Genome-wide association study; Fisher combination function; multivariate permutation; substance abuse

1 Introduction

For complex diseases such as substance use disorders, a gene is usually expressed in multiple traits that tend to be correlated with each other. In most situations, the effect of the gene on each marginal trait is small. If we evaluate the association between the gene and each marginal trait separately, we may not have enough power to detect the association.

Corresponding author: James J Yang, School of Nursing, University of Michigan, 400 North Ingalls St., Ann Arbor, Michigan 48109, USA. jjyang@umich.edu.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Thus, statistical methods that can be used to analyze the gene effect on multiple traits *jointly* are highly desirable.

In studies involving multivariate outcomes (e.g. multiple traits), the outcomes may be non-commensurate. That is, they are measured in different scales. For example, the early onset use of a substance (binary outcome) and the lifetime exposure to a substance (continuous outcome) are both important traits in substance abuse research.¹ The challenge of testing the association between a gene and non-commensurate outcomes is the nonexistence of obvious multivariate distribution to model the mixed outcomes. In this study, we focus on mixed binary and continuous phenotypic traits because they are the most common measurement scales in genetic association studies. Although many diseases are classified into binary categories for clinical purposes (e.g. normal vs. abnormal), a lot of biological and psychological measurements collected in clinical or research settings are continuous variables, which may contain more detailed information about the disease severity or progression. If we could analyze these two types of variables together as multivariate traits, we could greatly increase the information in the outcome and hence heighten the chance of finding the genes associated with the disease.

Teixeira-Pinto and Normand² conducted a comprehensive review of various methods for mixed binary and continuous outcomes and categorized them into two approaches. The first approach is to factor the joint distribution into the product of conditional and marginal distributions.³ The advantage of this approach is that the conditional and marginal distributions can be modeled using appropriate distributions. However, it is difficult to extend it to a more general setting that involves more than two outcomes. The other approach involves a latent variable and assumes that the observed outcomes conditioned on the latent variable are independent. Hence, one can write the joint distribution as the product of the two conditional distributions. Yet, one critical issue of the latent variable approach is that the parameters are not identifiable as demonstrated in the example in Teixeira-Pinto and Normand.² Thus, some constraints in the model are warranted.

Zheng et al.⁴ developed an alternative method for testing the association between a gene and mixed binary and continuous outcomes using the Fisher's combination function. Specifically, they proposed to use the trend test statistic for the association between the gene and the binary outcome and the F -test statistic for the association between the gene and the continuous outcome. The p -values from these two test statistics are later combined into a new statistic, say J , using the Fisher's combination function. If the two test statistics are independent, then J follows χ^2 -distribution with four degrees of freedom. However, in this case, the two test statistics are dependent, because the two outcomes tend to be dependent and the same gene is used for the testing. Thus, the gamma (scaled χ^2)-distribution is more appropriate for inferring the null distribution of J . They conducted a simulation study to estimate the scale and shape parameters of the gamma distribution using the Bootstrap method. Their simulation study showed that the gamma distribution with the estimates of 3.1 for the scale parameter and 1.3 for the shape parameter is a good approximation for the null distribution of J for various values of minor allele frequencies and correlation coefficients between outcomes (Table II of Zheng et al.⁴). Yet, this approach has two disadvantages. First, their estimated null distribution is only applicable for the specific setting involving the

trend test statistic and the F -test statistic. Second, based on our simulation study (see the details in Section 3.2), their association test based on the specific gamma distribution tends to have low power.

In this study, we propose a new method to estimate the empirical distribution of the Fisher's combination statistic under the null hypothesis. Instead of using a computationally intensive method (e.g. the permutation or bootstrap method) to estimate the null distribution of the test statistic, we show that the distribution is actually a function of the correlation between the continuous and binary outcomes. By estimating the correlation using the observed outcomes, we can greatly improve the efficiency of estimating the null distribution. The proposed method has two advantages. First, unlike the methods reviewed in Teixeira-Pinto and Normand,² the proposed method has a great potential to be extended to more than two outcomes with mixed measurement scales. Second, the proposed method has the same level of computational efficiency as the method of Zheng et al.⁴ and yet is more powerful.

This article is organized as follows. In Section 2, we define the statistical model of mixed outcomes and describe the proposed method in detail. We also show that our proposed method controls the type I error rate. In Section 3, we conduct a simulation study to evaluate the performance of the proposed method in comparison to alternative methods under various genetic models. Section 4 presents the results of statistical analysis on the Study of Addiction: Genetics and Environment (SAGE) data to demonstrate the applications in the substance abuse field. Discussion and concluding remarks are presented in Section 5.

2 Methods

We first define the notations used in this article. For the candidate SNP, let $Z_i (= 0, 1, 2)$ be the number of reference alleles for individual i , ($i = 1, \dots, n$) and let $(X_i, W_i)'$ be the bivariate continuous response variables. Suppose Y_i is a dichotomous variable derived from W_i through a cut-off value C :

$$Y_i = \begin{cases} 1 & \text{if } W_i \geq C, \\ 0 & \text{if } W_i < C \end{cases}$$

(1)

Without loss of generality, we assume $C = 0$ in this study. When there is no ambiguity, we drop the subscript i to simplify the notation.

Our purpose is to test the association between the candidate SNP and the bivariate responses. In Section 2.1, we review the test statistic for bivariate outcomes X and W using the Fisher combination function. In Section 2.2, we derive the test statistic for the mixed outcome X and Y . Our method is based on the Fisher combination function because of its flexibility in designing test statistic and asymptotic optimality.^{5,6}

2.1 Bivariate normal responses

The Fisher combination method is based on two components: marginal tests and a combination function. The marginal test needs to be unbiased and the combination function is a continuous, nonincreasing real function. To apply the Fisher combination method for the GWAS study on bivariate normal outcomes $(X, W)'$, a simple regression is used to test the association between the SNP Z and each marginal outcome. Let p_x and p_w be the corresponding marginal p -values. The Fisher combination function is, thus, written as

$$T_1 = -2 \log(p_x) - 2 \log(p_w)$$

Under the null hypothesis of no association between the SNP and the bivariate outcomes, the distribution of T_1 follows a gamma (scaled chi-squared) distribution.^{7,8} Note that for the gamma distribution with the shape parameter k and the scale parameter s , the mean is ks and the variance is ks^2 . In practice, the values of k and s are unknown. One way to avoid estimating k and s is using the permutation method to approximate the null distribution of T_1 statistic. Yet, the computational time of the permutation method increases exponentially as the sample size increases. Since GWAS usually involves a million of association tests, the permutation test is not practically feasible.

An alternative approach is to estimate the shape parameter k and the scale parameter s of the gamma distribution, using the method of moments. We need the first two moments to estimate the two unknown parameters. Under the null hypothesis, the p -values, p_x and p_w , both follow an uniform distribution. Hence, T_1 is the sum of two chi-squared (two degrees of freedom) random variables. Therefore, the mean of T_1 is

$$\mu = E[T_1] = 4$$

and the variance of T_1 is

$$\sigma^2 = \text{Var}[T_1] = 8 + 2\text{cov}[-2 \log(p_x), -2 \log(p_w)]$$

Yang et al.⁹ showed that $\text{cov}[-2 \log(p_x), -2 \log(p_w)]$ can be accurately estimated using the correlation between X and W , ρ_{xw} , as follows:

$$\text{cov}[-2 \log(p_x), -2 \log(p_w)] \approx b_1 \rho_{xw}^2 + b_2 \rho_{xw}^4 + b_3 \rho_{xw}^6 + b_4 \rho_{xw}^8 + b_5 \rho_{xw}^{10} \quad (2)$$

where $b_1 = 3.9081$, $b_2 = 0.0313$, $b_3 = 0.1022$, $b_4 = -0.1378$, and $b_5 = 0.0941$ (see Appendix 1.1 for the detailed derivation). When the true value of ρ_{xw} is known, this estimation is very close to the true covariance. However, in real life situations, the value of ρ_{xw} is unknown and can be estimated using the sample correlation coefficient of X and W : $\hat{\rho}_{xw}$. Let the right hand

side of equation (2) be $\delta(\rho_{xw})$. Under the null hypothesis, the mean of T_1 is 4 and the variance of T_1 can be efficiently calculated as $8 + 2\delta(\hat{\rho}_{xw})$. Thus, using the method of moments, we can estimate the shape parameter k and the scale parameter s using the following two equations:

$$\begin{aligned} ks &= 4, \\ ks^2 &= 8 + 2\delta(\hat{\rho}_{xw}) \end{aligned}$$

That is, the null distribution of T_1 is the gamma distribution with the shape parameter $k = 8 / (4 + \delta(\hat{\rho}_{xw}))$ and the scale parameter $s = 2 + \delta(\hat{\rho}_{xw}) / 2$.

2.2 Mixed continuous and binary responses

When the responses are continuous and binary (i.e. X and Y), the Fisher combination function can still be applicable to test the association between the SNP (i.e. Z) and the responses. Let p_x be the p -value for the test statistic of the association between Z and X ; and p_y be the p -value corresponding to the association between Z and Y . Using the Fisher combination function, we define the test statistic for the association between Z and (X, Y) as

$$T_2 = -2 \log(p_x) - 2 \log(p_y)$$

(3)

Since T_2 is the sum of two chi-squared statistics, it follows a gamma distribution. Similar to T_1 in the previous section, the mean of T_2 is

$$E[T_2] = 4$$

and the variance of T_2 is

$$\text{Var}[T_2] = 8 + 2\text{cov}[-2 \log(p_x), -2 \log(p_y)]$$

Unlike the case of bivariate normal responses, the covariance here is, however, not tractable. When the responses are independent, $\text{cov}[-2 \log(p_x), -2 \log(p_y)] = 0$, and therefore the null distribution of T_2 follows a chi-squared distribution with four degrees of freedom and the variance of T_2 being 8. Yet, when the responses are correlated, Yang⁸ has showed that this approach is too liberal and thus tends to inflate the type I error rate. On the other hand, using the Cauchy-Schwarz inequality, we can derive the following result:

$$\text{cov}[-2 \log(p_x), -2 \log(p_y)] \leq \sqrt{\text{var}(-2 \log(p_x))\text{var}(-2 \log(p_y))} = 4$$

Hence, the maximum of the variance of T_2 is 16. If we adopt this value, the test would be too conservative and thus tend to have low power. Therefore, our goal is to find a good trade-off between these two extreme values of $\text{Var}[T_2]$ (i.e. 8 and 16).

According to Section 2.1, if we observe bivariate continuous responses X and W , we would use the sample correlation coefficient $\hat{\rho}_{xw}$ to estimate the variance of the test statistic through equation (2). However, when we only observe mixed continuous and binary responses X and Y , we may assume that Y is derived from a latent variable W through equation (1). Pearson¹⁰ and Tate^{11,12} proposed the sample biserial correlation, which is a consistent estimate of ρ_{xw} . It is defined as

$$\hat{\rho}_{xw} = \frac{\hat{r}_{xy} S_y}{\phi(\Phi^{-1}(1 - \bar{y}))}$$

where ϕ and Φ are the standard Gaussian density and distribution functions, respectively;

$$\bar{x} = \sum_{i=1}^n x_i / n; \bar{y} = \sum_{i=1}^n y_i / n; S_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n; S_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / n;$$

$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / n$; and $\hat{r}_{xy} = S_{xy} / (S_x S_y)$ is the sample correlation coefficient between X and Y (see Appendix 1.2 for the detailed derivation of the biserial correlation).

We propose to plug $\hat{\rho}_{xy}$ (which is a consistent estimate of ρ_{xw}) in equation (2) and use the resulting estimate of $\text{cov}[-2 \log(p_x), -2 \log(p_w)]$ to estimate $\text{cov}[-2 \log(p_x), -2 \log(p_y)]$. Although this approach tends to slightly overestimate $\text{Var}[T_2]$ (see Section 3.1), it is a better estimate than the lower bound (8) or the upper bound (16).

3 Simulation studies

3.1 Accuracy of variance estimation

To evaluate the relative accuracy of the proposed variance estimation method in Section 2.2, we conducted a simulation study. The simulation was based on 1000 simulated individuals. Given the correlation ρ , we simulated bivariate random variables of $(X, W)'$ for each individual. The random variable Y was derived from W using equation (1) with $C=0$. We also independently simulated genotypes Z of all individuals with the minor allele frequency 0.5. We used a simple linear regression to test the association between Z and X and derived the p -value: p_x ; in a similar way, we derived the p -value for the association between Z and W : p_w . For the binary response Y , we used the Cochran–Armitage trend statistic to test the association between Z and Y and derived the corresponding p -value: p_y . The process was repeated 100,000 times, so that we have 100,000 pairs of (p_x, p_w) and (p_x, p_y) . Based on these simulated p -values, we can calculate $\text{cov}[-2 \log(p_x), -2 \log(p_w)]$ and $\text{cov}[-2 \log(p_x), -2 \log(p_y)]$. By varying the value of ρ , we can investigate the relationship between ρ and the covariance.

The simulation results are shown in Figure 1. The solid curve corresponds to the estimates of $\text{cov}[-2 \log(p_x), -2 \log(p_w)]$ based on the simulated bivariate normal responses. The estimates of $\text{cov}[-2 \log(p_x), -2 \log(p_y)]$ based on the simulated mixed continuous and

binary responses are, on the other hand, delineated by the dash curve. The differences between these two curves show that: (a) the difference in the two covariance estimates reaches 0 at $\rho = 0$; (b) the difference increases when the absolute value of ρ increases; and (c) the covariance corresponding to mixed responses is smaller than the covariance corresponding to bivariate continuous responses for every value of ρ . The finding in (c) indicates that our proposed method, which uses an estimate of $\text{cov}[-2 \log(p_x), -2 \log(p_w)]$ tends to slightly over-estimate $\text{cov}[-2 \log(p_x), -2 \log(p_y)]$. For example, when $\rho = 0.5$, our estimate of $\text{Var}[T_2]$ is 9.96, which over-estimates the target value (9.22) by 8%. Nevertheless, this approach conservatively estimates the variance of T_2 so that the resulting type I error rate is controlled.

3.2 Performance of the proposed method

3.2.1 Simulation design.—We conducted a Monte Carlo simulation study to compare the proposed method with the method of Zheng et al.⁴ and the permutation method, in terms of the type I error rate and statistical power. For each subject, the genotype $Z = (0, 1, 2)$ representing the total number of reference allele was generated based on the minor allele frequency (MAF) ζ . That is, the genotype coded as the number of reference allele was simulated from the multinomial distribution with probabilities $((1 - \zeta)^2, 2\zeta(1 - \zeta), \zeta^2)$. Given the genotype, we generated the responses for various values of ρ using the following model:

$$\begin{pmatrix} X \\ W \end{pmatrix} \sim \text{BVN} \left(\begin{pmatrix} \mu_x \\ \mu_w \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \quad (4)$$

Note that the values of μ_x and μ_w represent the genetic effect, which was defined as follows:

$$\mu_l = \begin{cases} -e_l, & \text{if } Z = 0 \\ e_l(2\theta - 1), & \text{if } Z = 1 \\ e_l, & \text{if } Z = 2 \end{cases} \quad (5)$$

where $l = x$ or w . The parameter θ controlled whether the genetic effect was recessive ($\theta = 0$), additive ($\theta = 1/2$), or dominant ($\theta = 1$).¹³ The value of Y was then derived using equation (1).

In addition to the parameter θ , we manipulated the following three parameters to evaluate the performance of competing methods under different situations. The minor allele frequencies $\zeta = 0.1, 0.3, 0.5$ were chosen. The correlation between responses ranged from

zero ($\rho = 0$), moderate ($\rho = 0.4$), to high ($\rho = 0.8$). Moreover, the values of e_x and e_w were set at 0 or 0.4 to represent four different genetic effect sizes. Note that the setting of $e_x = e_w = 0$ corresponds to the null hypothesis of no genetic effect. For each parameter configuration, we simulated 100 subjects.

We compared the proposed method with three competing methods: (a) the method proposed by Zheng et al.⁴ that employs the Gamma distribution with the scale parameter = 3.1 and the shape parameter = 1.3 as the reference null distribution; (b) the permutation method for testing the association between Z and $(X, Y)'$; and (c) the permutation method for testing the association between Z and $(X, W)'$. Since the values of the latent variable W were observed in the simulation, the third method was carried out to serve as the upper bound among all the competing methods.

For Methods 2 and 3, 1000 permutations were conducted to calculate the p -values of the test statistic. Under the null hypothesis ($e_x = e_w = 0$), we repeated the simulation 10,000 times to have enough precision for comparing type I errors. Under the alternative hypotheses, we repeated the simulation 2500 times so that the simulation standard error (10^{-2}) is sufficient for comparing powers. Based on the data of 10,000 and 2500 replications, the empirical type I error rate and power were calculated as the proportion of p -values less than 0.05, respectively.

3.2.2 Simulation results.—The empirical Type I error rates at 0.05 level for the four methods are listed in Table 1. For 10,000 replications, the standard deviation of the empirical Type I error rate is 0.005. As demonstrated in the table, all the four methods control the type I error for various values of MAF and correlations between responses.

The empirical powers at the significance level of 0.05 for the recessive, additive, and dominant genetic effect models are listed in Tables 2 to 4, respectively. For 2500 replications, the standard deviation of the simulated power is 0.01. The findings are summarized as follows:

1. The comparison of test performance across the three genetic effect models shows that the tests have the highest power under the dominant model, whereas the power under the recessive model is the lowest.
2. Among the four competing methods, the Zheng's method has the lowest power.
3. The power of all tests increases as the genetic effects, e_x or e_w , increase.
4. The power of all tests increases as MAF increases for the recessive and additive genetic models, while the statistical power under the dominant effect model does not increase with MAF.
5. The power of all tests decreases as the correlation between responses, ρ , increases. This is expected since a large ρ value decreases the information contained in bivariate responses.
6. The power of the proposed method is very close to the permutation method for mixed responses. Yet, the computational efficiency of the proposed method is

much higher. For example, in our simulation study, the proposed method is 1000 times more efficient than the permutation method. More importantly, when the sample size is large, the computational time for exact permutation becomes intractable, whereas the proposed method is not affected in terms of computational efficiency. In fact, the accuracy of correlation estimation for the proposed method improves as the sample size increases.

7. If we can observe the latent variable W , the power of the test using X and W is generally higher than the one using the mixed responses X and Y .
8. When both e_x and e_w are nonzero, the power of the tests is higher than the power in the setting with only e_x or e_w being nonzero. Furthermore, when one of e_x and e_w is zero (i.e. has no effect), the reduction in power is more severe in the setting of $e_x = 0$, because the continuous response X contains more information than the binary response Y . For example, in the setting of $\text{MAF} = 0.3$ and $\rho = 0.4$, the proposed test has the power of 0.776 when $e_x = e_w = 0.4$. However, when $e_w = 0$ (i.e. the binary response Y is not associated with the genotype), the power of the proposed test is reduced to 0.549; when $e_x = 0$ (i.e. the continuous response X is not associated with the genotype), the power of the proposed test is further reduced to 0.361.

4 Real data analysis

We conducted real data analysis using the database from the SAGE, http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1. The SAGE is a case-control study that gathered data from three large scale studies in the substance abuse field: the Collaborative Study on the Genetics of Alcoholism (COGA), the Family Study of Cocaine Dependence (FSCD), and the Collaborative Genetic Study of Nicotine Dependence (COGEND). The total number of individuals with individual level data available is 4121. Each individual was genotyped using the Illumina Human 1 M-Duo beadchip, which contains over one million SNP markers. We chose unrelated individuals with SNPs that were restricted to 22 autosomes and passed the quality control measures according to the procedures in Anderson et al.¹⁴ In addition, individuals were removed if their phenotypes were missing. The final number of unrelated individuals included in the analysis was 3409 (1653 male and 1756 female). The total number of SNP markers that passed the quality control was 726,247.

Although the *number of alcohol symptoms* has been commonly adopted as a measure of the phenotype of alcohol use disorder (AUD), *early onset of alcohol use* has been found to be associated with risky trajectories of progression to AUD.¹⁵ Thus, to identify the gene that contributes to susceptibility to alcohol addiction, we may need to consider both of these phenotypes together. Following the alcohol research literature,¹⁵ we defined early onset of regular drinking (i.e. drinking once a month for 6 months or more) as the onset age of 14 or younger (coded as 1; as 2 otherwise). Figure 2 shows the distributions of the two phenotype variables used in the analysis. The two variables, number of alcohol symptoms and early onset of regular drink, are correlated with the biserial correlation coefficient being -0.4761 .

The negative correlation indicates that early onset drinkers tend to have more alcohol symptoms.

Since the study samples are from different ethnic groups, the differences in allele frequency across ethnic groups are likely to produce false positive association.¹⁶ To evaluate and adjust for population stratification, we conducted principal components analysis.¹⁷ The results showed that the first three principal components are sufficient to adjust for population stratification.

The proposed combination method is a two-step procedure. In the first step, we conducted marginal genomewide association tests for each phenotypic outcome variable. The simple linear regression was used for the number of alcohol symptoms (continuous variable), whereas the logistic regression was employed for the early onset status of regular drinking (binary variable). The first three principal components were added as covariates to adjust for population stratification. The second step involved using the biserial correlation between the mixed phenotypes as well as the marginal p -values derived from Step 1 to calculate the p -values of the genome-wide association tests on the mixed continuous and binary outcomes.

Figure 3 displays the QQ-plot of the p -values for the association between the early onset of regular drink and SNPs. The QQ-plot of the number of alcohol symptoms and SNPs are shown in Figure 4. The fact that the p -values lie on the diagonal lines in both figures indicates that the model assumptions of our marginal analysis are adequate. The QQ-plot of the p -values for the association between the mixed phenotypes and SNPs is depicted in Figure 5. In this figure, the majority of p -values are on the diagonal line but, unlike Figures 3 and 4, about 10 smallest p -values are above the diagonal line. This difference indicates that our proposed method identifies more significant SNPs than the marginal tests.

To evaluate the accuracy of the proposed method, we employed the permutation method based on the Fisher combination function. We conducted 1000 permutations so the standard deviation of the Monte Carlo p -value is 0.0158. The procedure of the permutation method is as follows. We first calculated the observed statistic, T_0 , using equation (3). The labels of bivariate phenotype variables were then randomly permuted to form the permuted data sets. The vector of the bivariate phenotypes instead of the individual phenotype was permuted to preserve the correlation between the two phenotype variables. For the i^{th} permuted data set, we calculated T_i using equation (3), ($i = 1, \dots, 1,000$). The p -value of each SNP was, thus, computed as the proportion of T_i being greater than or equal to T_0 .

The differences in p -values between the proposed method and the permutation method are plotted in Figure 6. The figure shows that the differences between these two methods are less than 0.1 for all the 726,247 SNPs and they are within the variation among the resulting estimates based on 1000 permutations. On the other hand, the computational time is dramatically different between the two methods. The Plink software¹⁸ was adopted to calculate the marginal p -values because of its well-known efficiency. Using our workstation with 2.53 GHz Intel(R) Xeon(R) CPU and 64 GB RAM, it took only 5 min to calculate all statistics using the proposed method. However, it took more than 64 h to conduct

simulations for the permutation method as well as more than 500 GB of storage to deposit the resulting 1000 sets of statistics.

To identify the SNPs associated with susceptibility to alcohol addiction, we set the reduced type I error rate at 10^{-6} . Based on the marginal p -values, we identified one SNP (rs8019635, $p = 9.03 \times 10^{-7}$) to be associated with the number of alcohol symptoms and one SNP (rs6800626, $p = 3.71 \times 10^{-7}$) with the early onset of regular drinking. On the other hand, using the proposed method, we identified two SNPs (rs6800626, $p = 9.75 \times 10^{-7}$ and rs17427389, $p = 7.60 \times 10^{-7}$) to be associated with both phenotype variables. Although rs6800626 was also identified using the marginal test based on the early onset of regular drinking, rs17427389 was not detected by either marginal tests. Yet, the literature has found rs17427389 in the PLEKHG1 region to be associated with alcohol dependence.^{19,20} This demonstrates that combining multiple phenotypes can increase the power of identifying markers that may not be, otherwise, chosen using marginal tests. In addition, marginal tests may identify those SNPs that only contribute to a particular phenotype. Therefore, if our goal is to identify the genes that contribute to the common liability to alcohol addiction, the proposed method is a better approach than marginal tests.

5 Discussions

In this study, we propose a new genome-wide association test for mixed binary and continuous phenotypes that uses an efficient numerical method to estimate the empirical distribution of the Fisher's combination statistic under the null hypothesis. The simulation shows that the proposed method controls the type I error rate and also maintains its power at the level of the permutation method. More importantly, the computational efficiency of the proposed method is much higher than the one of the permutation method. The comparison of test performance across three genetic effect models (recessive, additive, and dominant) shows that the tests have the highest power under the dominant model and the lowest power under the recessive model. Furthermore, the power of all competing methods based on the Fisher's combination statistic increases when the genetic effects increase, the minor allele frequency increases, and the correlation between responses decreases.

Our real data analysis using the SAGE database shows that the p -values estimated by the proposed method are very close to the ones by the permutation method. More importantly, the computational efficiency of the proposed method is much higher than the permutation method. The results of our analysis also demonstrate that combining multiple phenotypes can increase the power of identifying markers that may not be, otherwise, chosen using marginal tests. Thus, the proposed method is a better approach than marginal tests to identify the genes that contribute to the common liability to complex diseases such as alcohol use disorder.

In this study, we only considered mixed binary and continuous phenotypic traits because they are the most common measurement scales in genetic association studies. However, in many fields such as the substance abuse field, ordinal outcomes are also very common.²¹ The proposed method has a great potential to be extended to a more general setting that involves multivariate responses with continuous, binary, and ordinal responses.

Acknowledgements

The funding agencies had no role in study design, analysis, interpretation of results, decision to publish, or preparation of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by National Institutes of Health (NIH) grants, K01 AA016591 & R01 DA035183 to A. Buu, and R01 AI079139 & R01 HL118267 to L. K. Williams. The authors acknowledge the Study of Addiction: Genetics and Environment (SAGE), which is part of the Gene Environment Association Studies (GENEVA) initiative supported by the National Human Genome Research Institute (dbGaP study accession phs000092.v1.p1).

Appendix 1

Appendix 1.1 The derivation of equation (2)

Let (x, y) be the standard bivariate normal random variables with the density function

$$f(x, y | \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right)$$

where ρ is the correlation between x and y .

Let p_x and p_y be the p -values of the statistics x and y , respectively. Then the relation between p_x and x is $p_x = 2\Phi(-|x|)$, where Φ is the standard normal distribution function. Similarly, $p_y = 2\Phi(-|y|)$. We can write their covariance as

$$\begin{aligned} g(\rho) &= \text{cov}(-2\log(p_x), -2\log(p_y)) \\ &= E((-2\log(p_x))(-2\log(p_y))) - E(-2\log(p_x))E(-2\log(p_y)) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 4\log(2\Phi(-|x|))\log(2\Phi(-|y|))f(x, y | \rho)dx dy - 4 \\ &\quad (6) \end{aligned}$$

Therefore, for a given value of ρ , we can calculate $g(\rho) = \text{cov}(-2\log(p_x), -2\log(p_y))$ numerically using equation (6). Let $\rho_1, \dots, \rho_m$ be the ρ values covering $(0, 1)$ and $g(\rho_1), \dots, g(\rho_m)$ be the corresponding values calculated using equation (6). We can find the values of b_1, b_2, b_3, b_4 , and b_5 using the following equation:

$$\arg \min_{b_1, b_2, b_3, b_4, b_5} \sum_{j=1}^m \left(g(\rho_j) - b_1 \rho_j^2 - b_2 \rho_j^4 - b_3 \rho_j^6 - b_4 \rho_j^8 - b_5 \rho_j^{10} \right)^2$$

(7)

The derived values can, thus, be used to calculate $\text{cov}(-2 \log(p_x), -2 \log(p_y))$. That is,

$$\text{cov}(-2 \log(p_x), -2 \log(p_y)) \approx b_1 \rho^2 + b_2 \rho^4 + b_3 \rho^6 + b_4 \rho^8 + b_5 \rho^{10}$$

where $b_1 = 3.9081$, $b_2 = 0.0313$, $b_3 = 0.1022$, $b_4 = -0.1378$, and $b_5 = 0.0941$.

Appendix 1.2 Biserial correlation

Without loss of generality, we assume that (X, W) are the standard bivariate normal random variables where the correlation between X and W is ρ_{xw} . The random variable Y is defined as $Y = I(W \geq C)$. The covariance between X and Y can be written as

$$\text{cov}(X, Y) = \rho_{xy} \sigma_x \sigma_y = \rho_{xy} \sigma_y$$

(8)

It can also be derived as follows:

$$\begin{aligned} \text{cov}(X, Y) &= E(XY) - E(X)E(Y) \\ &= E(XY) \\ &= E(XI(W \geq C)) \\ &= \int_C^\infty \int_{-\infty}^\infty x f(x, w | \rho_{xw}) dx dw \\ &= \rho_{xw} \phi(C) \end{aligned}$$

(9)

where ϕ is the standard normal density function. Using equation (8) and equation (9), we can obtain the following result:

$$\rho_{xw} = \frac{\rho_{xy}\sigma_y}{\phi(C)}$$

(10)

Therefore, to estimate ρ_{xw} from the random samples (x_i, y_i) , $i = 1, \dots, n$, we need to calculate the following sample statistics: $\bar{x} = \sum_{i=1}^n x_i / n$; $\bar{y} = \sum_{i=1}^n y_i / n$; $S_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n$; $S_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / n$; $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / n$; and $\hat{r}_{xy} = S_{xy} / (S_x S_y)$. The cut-off value C is estimated by $\hat{C} = \Phi^{-1}(1 - \bar{y})$. Plugging these statistics into equation (10), we can derive the sample biserial correlation

$$\hat{\rho}_{xw} = \frac{\hat{r}_{xy} S_y}{\phi(\Phi^{-1}(1 - \bar{y}))}$$

References

1. Flory JD and Manuck SB. Impulsiveness and cigarette smoking. *Psychosom Med* 2009; 71: 431–437. [PubMed: 19251874]
2. Teixeira-Pinto A and Normand SLT. Correlated bivariate continuous and binary outcomes: Issues and applications. *Stat Med* 2009; 28: 1753–1773. [PubMed: 19358234]
3. Fitzmaurice GM and Laird NM. Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics* 1997; 53: 110–122. [PubMed: 9147588]
4. Zheng G, Wu CO, Kwak M, et al. Joint analysis of binary and quantitative traits with data sharing and outcome-dependent sampling. *Genet Epidemiol* 2012; 36: 263–273. [PubMed: 22460626]
5. Littell RC and Folks JL. Asymptotic optimality of Fisher's method of combining independent tests. *J Am Stat Assoc* 1971; 66: 802–806.
6. Littell RC and Folks JL. Asymptotic optimality of Fisher's method of combining independent tests. *J Am Stat Assoc* 1973; 68: 193–194.
7. Brown MB. Method for combining non-independent, one-sided tests of significance. *Biometrics* 1975; 31: 987–992.
8. Yang JJ. Distribution of Fisher's combination statistic when the tests are dependent. *J Stat Comput Simul* 2010; 80: 1–12.
9. Yang JJ, Li J, Williams LK, et al. An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function. *BMC Bioinform* 2016; 17: 1–11.
10. Pearson K On a new method of determining correlation between a measured character A, and a character B, of which only the percentage of cases wherein B exceeds (or falls short of) a given intensity is recorded for each grade of A. *Biometrika* 1909; 7: 96–105.
11. Tate RF. Applications of correlation models for biserial data. *J Am Stat Assoc* 1955; 50: 1078–1095.
12. Tate RF. The theory of correlation between two continuous variables when one is dichotomized. *Biometrika* 1955; 42: 205–216.
13. Falconer DS and Mackay TFC. *Introduction to quantitative genetics*. 4th ed. Longmans Green, Harlow, Essex, UK, 1996.
14. Anderson CA, Pettersson FH, Clarke GM, et al. Data quality control in genetic case-control association studies. *Nat Protocols* 2010; 5: 1564–1573. [PubMed: 21085122]

15. Buu A, Wang W, Schroder SA, et al. Developmental emergence of alcohol use disorder symptoms and their potential as early indicators for progression to alcohol dependence in a high risk sample: A longitudinal study from childhood to early adulthood. *J Abnorm Psychol* 2012; 121: 897–908. [PubMed: 21842966]
16. Yang JJ, Li J, Buu A, et al. Efficient inference of local ancestry. *Bioinformatics* 2013; 29: 2750–2756. [PubMed: 23958727]
17. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; 38: 904–909. [PubMed: 16862161]
18. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; 81: 559–575. [PubMed: 17701901]
19. Bierut LJ, Agrawal A, Bucholz KK, et al. A genome-wide association study of alcohol dependence. *Proc Natl Acad Sci USA* 2010; 107: 5082–5087. [PubMed: 20202923]
20. Zuo L, Zhang F, Zhang H, et al. Genome-wide search for replicable risk gene regions in alcohol and nicotine co-dependence. *Am J Med Genet Part B-Neuropsychiatr Genet* 2012; 159B: 437–444.
21. Dziak JJ, Li RZ, Zimmerman MA, et al. Time-varying effect models for ordinal responses with applications in substance abuse research. *Stat Med* 2014; 33: 5126–5137. [PubMed: 25209555]

Figure 1.

The relationship between covariance and the correlation ρ . The solid curve corresponds to $\text{cov}[-2 \log(p_x), -2 \log(p_w)]$. The dash curve corresponds to $\text{cov}[-2 \log(p_x), -2 \log(p_y)]$.

Figure 2.
The distributions of phenotypes: the early onset of regular drinking and the number of alcohol symptoms.

Figure 3.

The Q-Q plot of observed p -values versus expected p -values for the early onset of regular drinking.

Figure 4.

The Q-Q plot of observed p -values versus expected p -values for the number of alcohol symptoms endorsed.

Figure 5.

The Q-Q plot of observed p -values versus expected p -values for the bivariate phenotypes of the early onset of regular drinking and the number of alcohol symptoms.

Figure 6.

The difference in p -values calculated from two methods: the proposed method and the permutation method (with 1000 permutations).

Table 1.

Empirical type I error rates at the 0.05 level under the null hypothesis of no genetic effect with varied minor allele frequency (MAF) and correlation between responses (ρ).

MAF	ρ	Proposed method	Zheng's method	Permutation method	Permutation method
		X and Y	X and Y	X and Y	X and W
0.1	0	0.0471	0.0245	0.0481	0.0482
	0.4	0.0464	0.0295	0.0488	0.0481
	0.8	0.0403	0.0421	0.0481	0.0492
0.3	0	0.0489	0.0271	0.0495	0.0486
	0.4	0.0452	0.0306	0.0487	0.0491
	0.8	0.0428	0.0456	0.0497	0.0499
0.5	0	0.0485	0.0245	0.0480	0.0461
	0.4	0.0439	0.0296	0.0471	0.0461
	0.8	0.0427	0.0446	0.0484	0.0461

Table 2.

Empirical power under the recessive model ($\theta = 0$) with varied genetic effect size (e_x, e_w), minor allele frequency (MAF), and correlation between responses (ρ).

e_x	e_w	MAF	ρ	Proposed method	Zheng's method	Permutation method	Permutation method
				X and Y	X and Y	X and Y	X and W
0.4	0.4	0.1	0	0.0764	0.0440	0.0764	0.0804
			0.4	0.0592	0.0408	0.0636	0.0672
			0.8	0.0532	0.0540	0.0632	0.0672
		0.3	0	0.4124	0.3248	0.4072	0.4724
			0.4	0.3656	0.3184	0.3728	0.4260
			0.8	0.2940	0.3048	0.3208	0.3392
		0.5	0	0.8984	0.8408	0.8972	0.9444
			0.4	0.8352	0.7964	0.8424	0.8904
			0.8	0.7512	0.7572	0.7672	0.8024
0	0.4	0.1	0	0.0600	0.0320	0.0624	0.0672
			0.4	0.0424	0.0256	0.0444	0.0536
			0.8	0.0420	0.0424	0.0492	0.0520
		0.3	0	0.1716	0.1092	0.1676	0.2392
			0.4	0.1620	0.1196	0.1712	0.2228
			0.8	0.1120	0.1136	0.1256	0.1532
		0.5	0	0.4972	0.3988	0.4948	0.6712
			0.4	0.4448	0.3692	0.4556	0.6200
			0.8	0.3676	0.3592	0.3928	0.5124
0.4	0	0.1	0	0.0644	0.0316	0.0624	0.0628
			0.4	0.0488	0.0332	0.0496	0.0540
			0.8	0.0456	0.0484	0.0560	0.0504
		0.3	0	0.2752	0.1968	0.2728	0.2748
			0.4	0.2280	0.1676	0.2316	0.2168
			0.8	0.1540	0.1560	0.1740	0.1444
		0.5	0	0.6668	0.5704	0.6680	0.6608
			0.4	0.6336	0.5624	0.6420	0.6292
			0.8	0.5432	0.5368	0.5708	0.5108

Table 3.

Empirical power under the additive model ($\theta = 1/2$) with varied genetic effect size (e_x , e_w), minor allele frequency (MAF) and correlation between responses (ρ).

e_x	e_w	MAF	ρ	Proposed method	Zheng's method	Permutation method	Permutation method
				X and Y	X and Y	X and Y	X and W
0.4	0.4	0.1	0	0.4860	0.3836	0.4872	0.5500
			0.4	0.4140	0.3624	0.4276	0.4732
			0.8	0.3660	0.3732	0.3900	0.4224
		0.3	0	0.8300	0.7560	0.8308	0.8960
			0.4	0.7932	0.7496	0.8032	0.8556
			0.8	0.6868	0.6912	0.7064	0.7456
		0.5	0	0.9088	0.8504	0.9072	0.9544
			0.4	0.8448	0.8088	0.8488	0.9008
			0.8	0.7620	0.7676	0.7776	0.8196
0	0.4	0.1	0	0.2024	0.1436	0.2016	0.2948
			0.4	0.1768	0.1344	0.1864	0.2492
			0.8	0.1448	0.1464	0.1632	0.1812
		0.3	0	0.3980	0.3020	0.3948	0.5904
			0.4	0.3836	0.3124	0.3880	0.5864
			0.8	0.2704	0.2700	0.2972	0.4136
		0.5	0	0.4812	0.3840	0.4744	0.6840
			0.4	0.4392	0.3628	0.4468	0.6452
			0.8	0.3492	0.3440	0.3772	0.5344
0.4	0	0.1	0	0.2932	0.2156	0.2928	0.2924
			0.4	0.2616	0.1968	0.2704	0.2556
			0.8	0.1964	0.1968	0.2176	0.1836
		0.3	0	0.6128	0.5168	0.6076	0.6144
			0.4	0.5656	0.4816	0.5760	0.5560
			0.8	0.4524	0.4556	0.4840	0.4236
		0.5	0	0.6832	0.5924	0.6812	0.6744
			0.4	0.6532	0.5828	0.6612	0.6452
			0.8	0.5508	0.5544	0.5836	0.5296

Table 4.

Empirical power under the dominant model ($\theta = 1$) with varied genetic effect size (e_x , e_w), minor allele frequency (MAF), and correlation between responses (ρ).

e_x	e_w	MAF	ρ	Proposed method	Zheng's method	Permutation method	Permutation method
				X and Y	X and Y	X and Y	X and W
0.4	0.4	0.1	0	0.9348	0.8968	0.9372	0.9696
			0.4	0.8960	0.8688	0.9028	0.9360
			0.8	0.8336	0.8396	0.8484	0.8820
		0.3	0	0.9884	0.9768	0.9880	0.9984
			0.4	0.9700	0.9616	0.9724	0.9840
			0.8	0.9300	0.9348	0.9396	0.9608
		0.5	0	0.8928	0.8384	0.8920	0.9412
			0.4	0.8380	0.7968	0.8464	0.8892
			0.8	0.7512	0.7580	0.7724	0.8028
0	0.4	0.1	0	0.5592	0.4592	0.5596	0.7492
			0.4	0.5100	0.4368	0.5200	0.7072
			0.8	0.4336	0.4276	0.4704	0.6300
		0.3	0	0.7028	0.6080	0.7032	0.8720
			0.4	0.6848	0.6112	0.6964	0.8768
			0.8	0.5772	0.5760	0.6052	0.8336
		0.5	0	0.4916	0.3808	0.4932	0.6632
			0.4	0.4592	0.3864	0.4660	0.6260
			0.8	0.3644	0.3628	0.3864	0.5204
0.4	0	0.1	0	0.7520	0.6644	0.7544	0.7504
			0.4	0.7312	0.6572	0.7416	0.7224
			0.8	0.6432	0.6452	0.6780	0.6360
		0.3	0	0.8788	0.8204	0.8812	0.8752
			0.4	0.8788	0.8324	0.8812	0.8828
			0.8	0.8240	0.8256	0.8460	0.8288
		0.5	0	0.6684	0.5644	0.6700	0.6556
			0.4	0.6408	0.5588	0.6464	0.6312
			0.8	0.5412	0.5384	0.5664	0.5128