RESEARCH ARTICLE

Identifying Pleiotropic Genes in Genome-Wide Association Studies for Multivariate Phenotypes with Mixed Measurement Scales

James J. Yang[1]*, **L. Keoki Williams**[2,3], **Anne Buu**[4]

1 School of Nursing, University of Michigan, Ann Arbor, Michigan, United States of America, **2** Department of Internal Medicine, Henry Ford Health System, Detroit, Michigan, United States of America, **3** The Center for Health Policy and Health Services Research, Henry Ford Health System, Detroit, Michigan, United States of America, **4** Department of Health Behavior and Biological Sciences, University of Michigan, Ann Arbor, Michigan, United States of America

* jjyang@umich.edu

Citation: Yang JJ, Williams LK, Buu A (2017) Identifying Pleiotropic Genes in Genome-Wide Association Studies for Multivariate Phenotypes with Mixed Measurement Scales. PLoS ONE 12(1): e0169893. doi:10.1371/journal.pone.0169893

Editor: Qizhai Li, University of the Chinese Academy of Sciences, CHINA

Received: August 12, 2016

Accepted: December 22, 2016

Published: January 12, 2017

Data Availability Statement: All data are available from deGap at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1.

Funding: 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 funding agencies had no role in study design, analysis, interpretation of results, decision to publish, or preparation of the manuscript. The content is solely

Abstract

We propose a multivariate genome-wide association test for mixed continuous, binary, and ordinal phenotypes. A latent response model is used to estimate the correlation between phenotypes with different measurement scales so that the empirical distribution of the Fisher's combination statistic under the null hypothesis is estimated efficiently. The simulation study shows that our proposed correlation estimation methods have high levels of accuracy. More importantly, our approach conservatively estimates the variance of the test statistic so that the type I error rate is controlled. The simulation also shows that the proposed test maintains the power at the level very close to that of the ideal analysis based on known latent phenotypes while controlling the type I error. In contrast, conventional approaches–dichotomizing all observed phenotypes or treating them as continuous variables–could either reduce the power or employ a linear regression model unfit for the data. Furthermore, the statistical analysis on the database of the Study of Addiction: Genetics and Environment (SAGE) demonstrates that conducting a multivariate test on multiple phenotypes can increase the power of identifying markers that may not be, otherwise, chosen using marginal tests. The proposed method also offers a new approach to analyzing the Fagerström Test for Nicotine Dependence as multivariate phenotypes in genome-wide association studies.

Introduction

Since the first genome-wide association study (GWAS) [1], more than 2,000 loci have been identified to be significantly associated with one or more complex traits [2]. In the early days, researchers have focused on genes associated with well defined functions or specific traits. A systematic review showed that many loci are actually associated with multiple traits [3]. This motivates researchers to study pleiotropy which is a condition in which a single gene affects multiple traits. A well-known example published on Nature is a GWAS involving 107 phenotypes that identified multiple pleiotropy genes [4]. From a statistical point of view, for complex

the responsibility of the authors and does not necessarily represent the official views of the NIH.

Competing Interests: The authors have declared that no competing interests exist.

diseases such as substance use disorders, a gene usually affects multiple traits and yet the effect size on each trait is very small. A GWAS using marginal association tests tends to have low power to detect these small effects. However, if a test can model the association between this gene and multivariate phenotypes *simutaneously*, the statistical power would be greatly increased [5].

When all the multivariate phenotypes are continuous, our team [6] recently conducted a comprehensive review of relevant statistical methods commonly used in the field including the principal component analysis (PCA), the multivariate analysis of variance (MANOVA), the generalizing estimating equations (GEE), the trait-based association test involving the extended Simes procedure (TATES), and the classical Fisher combination test. In the same study, we proposed a new method that relaxes the unrealistic independence assumption of the classical Fisher combination test and is computationally efficient. Our simulations also showed that the proposed method has higher power than existing methods while controlling the type I error rate.

Most of the existing methods that we previously reviewed and compared were designed for continuous multivariate phenotypes. However, it is pretty common in practice that multivariate phenotypes are measured in different scales (i.e. non-commensurate). For example, in substance abuse research, the early onset use of a substance and the lifetime exposure to a substance are both important traits and yet, the former is usually measured as a binary outcome whereas the latter tends to be a continuous or ordinal outcome [7]. The methodological challenge of modeling the association between a gene and non-commensurate phenotypes is that there does not exist a multivariate distribution for mixed data types.

There are two approaches that can handle bivariate phenotypes with one continuous variable and one binary variable [8]. The first approach is to model the bivariate phenotypes by factoring the joint distribution into the product of conditional and marginal distributions [9]. The complexity of this approach, however, increases exponentially when the number of phenotypes increases. The other approach is based on a latent class model with the assumption that conditioning on the latent variable, the bivariate phenotypes are independent. Hence, one can write the joint distribution as a product of the two conditional distributions. Nevertheless, one critical issue with this approach is that the parameters in the latent class model are not identifiable without some constraints as demonstrated in the example in Teixeira-Pinto and Normand (2009) [8].

In comparison to the methods reviewed above, meta-analysis is a more flexible alternative for handling different types of phenotype data [10]. The first step of this approach is to carry out separate analyses for different types of data (e.g. generalized linear models for continuous, binary, or ordinal phenotypes). The p-values from these analyses are later aggregated into a summary statistic. If this summary statistic is more extreme than the critical value, the significance of association between a SNP and multivariate phenotypes is declared. However, the key restriction of this approach is that the sample used to derived the p-value for one phenotype cannot be used to derive the p-value for another phenotype. If we partition the entire sample to multiple subsets for the purpose of meta-analysis, the statistical power would be greatly reduced.

In this study, we extend the Fisher combination function method originally designed for continuous multivariate phenotypes [6] to handle mixed continuous, binary, or ordinal multivariate phenotypes. This new method is also applicable to any number of phenotypes while controlling the type I error rate. Furthermore, the majority of computation time for the proposed method is used to calculate the marginal p-values, whereas the rest of computation time for the Fisher combination function is minimal regardless of the number of phenotypes involved. Therefore, this method is highly effective for exploring multiple combinations of multivariate phenotypes with minimal extra computation time.

This paper is organized as follows. In the next section, we review our previous work on continuous multivariate phenotypes and propose an extension of the previous method to handle multivariate phenotypes with mixed measurement scales. We also show that our proposed method controls the type I error rate. Because the proposed method requires estimation of the correlations for various combinations of phenotypes, we propose the estimation methods in the Estimation of the Correlation between Mixed Phenotypes section. In the Simulation Studies section, we present the results of simulation studies that evaluate the proposed method in terms of the accuracy of correlation and variance estimation, the type I error rate and statistical power. The Real Data Analysis section 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. Discussions and concluding remarks are presented in the Discussions section.

Methods

Previous Work on Continuous Multivariate Phenotypes

In this section, we review our previous work on continuous multivariate phenotypes [6] so that the readers have sufficient background information to understand the proposed method in the next section. For each individual i (= 1, . . ., N), let Z_{ig} (= 0, 1, 2) be the number of reference alleles for SNP g (= 1, . . ., G), and R_{ij}, (j = 1, . . ., M) be the jth phenotypes. To simplify notation, we define $\boldsymbol{Z}^{(g)} = (Z_{1g}, \ldots, Z_{Ng})$ as the gth genotypes; and $\boldsymbol{R}^{(j)} = (R_{1j}, \ldots, R_{Nj})$ as the jth phenotype. Let p_{gj} be the p-value from the marginal test of the association between $\boldsymbol{Z}^{(g)}$ and $\boldsymbol{R}^{(j)}$. Thus, for the gth genotype, we have a collection of p-values $\{p_{g1}, \ldots, p_{gM}\}$ for the M phenotypes. The purpose of this article is to construct an efficient and powerful method for testing the association between the gth SNP and the multivariate phenotypes $\{\boldsymbol{R}^{(1)}, \ldots, \boldsymbol{R}^{(M)}\}$ using these marginal p-values $\{p_{g1}, \ldots, p_{gM}\}$.

Given $\{p_{g1}, \ldots, p_{gM}\}$, the Fisher combination statistic is defined as

$$S^{(g)} = \sum_{j=1}^{M} -2\log(p_{gj}).$$

There are other choices of combination functions but the Fisher combination is chosen because of its asymptotic optimality [11, 12]. Based on the Fisher combination statistic, we may conduct a permutation test to examine the association between the gth SNP and the multivariate phenotypes $\{\boldsymbol{R}^{(1)}, \ldots, \boldsymbol{R}^{(M)}\}$. Although the permutation test is unbiased and asymptotically equivalent to the best parametric tests [13], it is extremely time consuming and thus not feasible for carrying out a whole genome association test that would require performing more than 10^6 permutations.

When the phenotypes $\{\boldsymbol{R}^{(1)}, \ldots, \boldsymbol{R}^{(M)}\}$ follow a multivariate normal distribution, the test statistic $S^{(g)}$ is a sum of chi-squared statistics under the null hypothesis of no association between the genotype and phenotypes. Since multivariate phenotypes are correlated (i.e. the p-values in $S^{(g)}$ are correlated), the null distribution of $S^{(g)}$ follows a gamma distribution with the shape parameter κ and the scale parameter ν [14, 15]. That is,

$$E[S^{(g)}] = \kappa\nu,$$
$$Var[S^{(g)}] = \kappa\nu^2.$$

If we can estimate κ and ν, we can calculate the p-value of $S^{(g)}$ using the gamma distribution rather than the permutation method. This will greatly improve the computation efficiency.

Because $-2\log(p_{gj})$ follows a chi-squared distribution with 2 degrees of freedom, we have

$$\kappa v = E[S^{(g)}] = 2M, \tag{1}$$

$$\kappa v^2 = Var[S^{(g)}] = 4M + \sum_{j \neq j'} cov(-2\log(p_{gj}), -2\log(p_{gj'})). \tag{2}$$

Here, the covariance between the p-value of the jth phenotype and the j'th phenotype, $cov(-2\log p_{gj}, -2\log p_{gj'})$, is a function of the correlation between $\boldsymbol{R}^{(j)}$ and $\boldsymbol{R}^{(j')}$ [6, 15]. Define $\rho_{jj'}$ to be the correlation between $\boldsymbol{R}^{(j)}$ and $\boldsymbol{R}^{(j')}$. Our previous work showed that $cov(-2\log(p_{gj}), -2\log(p_{gj'}))$ can be accurately estimated as

$$cov(-2\log(p_{gj}), -2\log(p_{gj'})) \approx \sum_{l=1}^{5} c_l \rho_{jj'}^{2l} - \frac{c_1}{N}(1-\rho_{jj'}^2)^2, \tag{3}$$

where $c_1 = 3.9081$, $c_2 = 0.0313$, $c_3 = 0.1022$, $c_4 = -0.1378$ and $c_5 = 0.0941$. Note that this approximation is very accurate as the maximum difference is less than 0.001. Thus, we can efficiently estimate κ and v using Eqs (1) and (2) with the $cov(\cdot)$ in Eq (2) substituted by the right-hand side of Eq (3).

The Proposed Method for Multivariate Phenotypes with Mixed Measurement Scales

The method reviewed in the previous section is based on the strong assumption that the multivariate phenotypes follow a multivariate normal distribution. Although we have demonstrated its robustness against a long-tail multivariate distribution in a simulation study, it may not be applicable to multivariate phenotypes with mixed measurement scales [6]. In this section, we extend the method to handle mixed continuous, binary, or ordinal phenotypes.

The proposed method is a two-phase approach: the first phase conducts a marginal test for each phenotype; and the second phase uses the Fisher combination function to combine the p-values from the first phase and conducts a multivariate test. The prerequisite for the multivariate test to be valid is that the marginal tests generating p_{gj} are unbiased [16, 17]. In order to meet this criterion, we propose to conduct the marginal tests based on the measurement scales of the phenotypes: using the linear regression for continuous phenotype; the logistic regression for binary phenotypes; and the cumulative logit model for ordinal phenotypes [18]. We propose to use those regression models in Phase 1 not only because they are unbiased but also because we can add covariates in the models to increase the accuracy of testing as well as principal components to correct for population stratification [19].

Once we obtain p-values $\{p_{1g}, \ldots, p_{Mg}\}$ in Phase 1, the second phase is to calculate the correlation $\rho_{jj'}$ between the phenotypes $\boldsymbol{R}^{(j)}$ and $\boldsymbol{R}^{(j')}$, so that we can estimate $cov(-2\log(p_{gj}), -2\log(p_{gj'}))$ using Eq (3). There are two issues that we need to address. First, we need to find appropriate methods for estimating $\rho_{jj'}$ when one or both phenotypes are binary or ordinal. This issue is dealt with in detail in the next section where various estimation methods of correlation are presented for all possible combinations of measurement scales. The second issue is to find the relationship between $cov(-2\log(p_{gj}), -2\log(p_{gj'}))$ and $\sum_{l=1}^{5} c_l \rho_{jj'}^{2l} - \frac{c_1}{N}(1-\rho_{jj'}^2)^2$ when the correlation $\rho_{jj'}$ estimated from non-continuous data is used in Eq (3). We propose a latent response model to address this issue.

Assume that the response $\boldsymbol{R}^{(j)}$ is viewed as a partial or full observation of a continuous *latent response* $\boldsymbol{R}^{*(j)}$. When the phenotype is continuous, $\boldsymbol{R}^{*(j)}$ is fully observed and equal to $\boldsymbol{R}^{(j)}$. However, when the phenotype is binary or ordinal, a certain value of $\boldsymbol{R}^{(j)}$ is observed when

$\boldsymbol{R}^{*(j)}$ falls within an unknown fixed threshold. The binary phenotype is treated as a special case when there is only one threshold. We further assume that $(\boldsymbol{R}^{*(1)}, \ldots, \boldsymbol{R}^{*(M)})$ follows a multivariate normal distribution and the correlation between $\boldsymbol{R}^{*(j)}$ and $\boldsymbol{R}^{*(j')}$ is $\rho_{jj'}$. Let $p^{\dagger}_{gj}$ and $p^{\dagger}_{gj'}$ be the p-values derived from observed binary or ordinal phenotypes. We propose to plug $\hat{\rho}_{jj'}$ in Eq (3) to estimate the true covariance $cov(-2\log(p^{\dagger}_{gj}), -2\log(p^{\dagger}_{gj'}))$. However, under the latent response model, this approach is actually estimating $cov(-2\log(p^{*}_{gj}), -2\log(p^{*}_{gj'}))$, where p^{*}_{gj} and $p^{*}_{gj'}$ are the p-values calculated from the latent variables $\boldsymbol{R}^{*(j)}$ and $\boldsymbol{R}^{*(j')}$. Since the binary or ordinal $\boldsymbol{R}^{(j)}$ and $\boldsymbol{R}^{(j')}$ are derived from the continuous $\boldsymbol{R}^{*(j)}$ and $\boldsymbol{R}^{*(j')}$, the covariance between $\boldsymbol{R}^{(j)}$ and $\boldsymbol{R}^{(j')}$ is smaller:

$$cov(-2\log(p^{\dagger}_{gj}), -2\log^{\dagger}(p_{gj'})) \leq cov(-2\log(p^{*}_{gj}), -2\log(p^{*}_{gj'})).$$

Thus, this approach will over-estimate the covariance of the observed test statistic. In other words, we conservatively estimate the variance of $S^{(g)}$ so that the type I error is controlled. In the Simulation Studies section, we conduct a simulation study to examine the difference between the true covariance and our estimates.

Estimation of the Correlation between Mixed Phenotypes

In this section, we specify various estimation methods of correlation for all possible combinations of measurement scales. Table 1 summarizes the classification of correlation coefficients based on the variable types. We define the following simplified notations for ease of interpretation. Suppose $(U_i, V_i)'$, $i = 1, \ldots, n$, are independent and identical bivariate normal random variables and the correlation between U_i and V_i is ρ. We would like to estimate ρ but either U_i, V_i or both are latent variables. The observed data may be binary (coded 0 or 1) or ordinal (coded as positive integers) depending on the practical situation. In the following sections, we describe different approaches to estimate ρ depending on the types of observed variables. Subscripts or superscripts may be omitted for convenience.

Kendall Correlation: Continuous-Continuous

Suppose we observe X^c_i and Y^c_i where $X^c_i = U_i$ and $Y^c_i = V_i$. To estimate the correlation coefficient ρ, the natural estimator is Pearson's sample correlation r_p. Although Pearson's sample correlation is an asymptotically unbiased estimator of ρ and the variance of r_p reaches the Cramer-Rao lower bound as the sample size increases, it tends to over or underestimate ρ when the sample distribution of $(X^c_i, Y^c_i)'$ deviates from the bivariate normal distribution or the regression of Y^c_i on X^c_i (or vice versa) is nonlinear [20]. Moreover, it cannot handle incomplete data. Our previous work demonstrated that Kendall τ is robust against these problems and thus is chosen to estimate the correlation between continuous variables [6]. Kendall τ is defined as

$$\tau = \frac{K_c - K_d}{n(n-1)/2},$$

Table 1. Different types of correlation coefficients when the variables *X* and *Y* are continuous, binary, or ordinal.

		X		
		Continuous	Binary	Ordinal
Y	Continuous	Kendall	Biserial	Polyserial
	Binary		Tetrachoric	Polychoric
	Ordinal			Polychoric

doi:10.1371/journal.pone.0169893.t001

where K_c is the number of concordant pairs (X_i^c and Y_j^c), and K_d is the number of discordant pairs. To use kendall's τ to estimate ρ, we can use this transformation [21, 22]:

$$r_k = \sin\left(\frac{\pi\tau}{2}\right).$$

Thus, we adopt r_k when both phenotypes are continuous.

Biserial Correlation: Continuous-Binary

Suppose we observed $Y_i^c = V_i$ and $X_i^b = I_{[U_i \geq C]}$, where I is an indicator function and C is a fixed unknown threshold. Pearson proposed the sample *biserial* correlation to estimate ρ [23]. However, its absolute value was shown to exceed 1 when $|\rho| > 0.798$ [24]. Brogden considered the situation when $\rho > 0$ and proposed a better biserial estimator [25]:

$$r_{Brogden} = \frac{\sum_i Y_i^c X_i^b - n\bar{X}^b \bar{Y}^c}{\sum_{i=1}^{\sum X_i} Y^c_{(n-i+1)} - n\bar{X}^b \bar{Y}^c},$$

where $Y^c_{(1)} \leq Y^c_{(2)} \leq \ldots \leq Y^c_{(n)}$. Note that $r_{Brogden} \leq 1$ but $r_{Brogden}$ may be less than -1 when $\rho < 0$. Lord further modified Borgden's estimator as Lord's estimator [26]:

$$r_L = \begin{cases} r_{Brogden} & \text{if } \; r_{Brogden} \geq 0 \\ r^{\dagger}_{Brogden} & \text{if } \; r_{Brogden} < 0, \end{cases}$$

where $r^{\dagger}_{Brogden} = -r_{Brogden}(X_i^b, -Y_i^c)$. The Lord's biserial estimator ensures that r_L is always between -1 and 1 and were shown by simulations to be more efficient in comparison to other estimators [27]. Thus, in this study we adopt r_L to estimate ρ for continuous and binary variables.

Tetrachoric Correlation: Binary-Binary

Suppose we observe $X_i^b = I_{[U_i > C_1]}$ and $Y_i^b = I_{[V_i > C_2]}$ for unknown thresholds C_1 and C_2. Let the proportions in 2 × 2 contingency tables be $p_{11}, p_{12}, p_{21}, p_{22}$. Define the marginal proportions as $p_x = p_{11} + p_{12}$ and $p_y = p_{11} + p_{21}$. Since the underlying variables follow a bivariate normal distribution, Pearson proposed the following likelihood function to find the *tetrachoric* correlation for ρ [28]:

$$L(h, k, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_k^{\infty} \int_h^{\infty} exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right) dxdy$$

where $h = \Phi^{-1}(p_x)$, $k = \Phi^{-1}(p_y)$, and Φ^{-1} is the inverse of the standard normal distribution function. The maximum likelihood estimate (MLE) for ρ is derived by solving

$$L(h, k, \rho) = p_{11}.$$

In this study, we adopt the computational algorithm developed by Good (2006) [29] to calculate this MLE.

Polyserial Correlation: Continuous-Ordinal

Let $Y_i^c = V_i$ and $X_i^o = t$ if $\zeta_{t-1} \leq U_i < \zeta_t$, for a positive integer $t\,(\geq 2)$, where $-\infty = \zeta_0 < \zeta_1 < \ldots < \zeta_{t-1} < \zeta_t = \infty$. Cox derived the likelihood function of observations $(X_i^o, Y_i^c)'$ using the following factorization [30]:

$$\Pi_{i=1}^{n} f(X_i^o = x_i, Y_i^c = y_i) = \Pi_{i=1}^{n} f(y_i) Pr[x_i|y_i] \tag{4}$$

where Y_i^c is a normal random variable and the conditional probability of $X_i^o|Y_i^c$ is

$$Pr[x_i|y_i] = \Phi(\theta_j) - \Phi(\theta_{j-1}),$$

where Φ is the standard normal distribution function and $\theta_j = \frac{\zeta_j - \rho(y_i - \mu)/\sigma}{(1-\rho^2)^{1/2}}$. The MLE is obtained by maximizing the likelihood function in Eq (4). In this study, we adopt the computational algorithm developed by Olsson et al. (1982) [31] to calculate the MLE.

Polychoric Correlation: Binary-Ordinal or Ordinal-Ordinal

Suppose that the variables we observe are both ordinal. That is, the relation between X_i and U_i is

$$\begin{aligned}
X_i &= 1 \text{ if } -\infty = \zeta_0 \leq U_i < \zeta_1 \\
X_i &= 2 \text{ if } \zeta_1 \leq U_i < \zeta_2 \\
&\vdots \\
X_i &= R \text{ if } \zeta_{R-1} \leq U_i < \zeta_R = \infty.
\end{aligned}$$

The relation between Y_i and V_i is similar to this. Pearson and Pearson (1922) [32] proposed the polychoric correlation which can be applied to handle ordinal-ordinal and binary-ordinal (special case) variables. If we arrange the data as a two-way contingency table with observed frequencies n_{ij}, $i = 1, \ldots, R$ and $j = 1, \ldots, C$. Define π_{ij} as the probability of observing n_{ij}. Then the likelihood function is proportion to

$$L(\rho) \propto \Pi_{i=1}^{R} \Pi_{j=1}^{C} \pi_{ij}^{n_{ij}},$$

where $\pi_{ij} = \Phi(\zeta_i, \xi_j) - \Phi(\zeta_{i-1}, \xi_j) - \Phi(\zeta_i, \xi_{j-1}) + \Phi(\zeta_{i-1}, \xi_{j-1})$. Note that $\Phi(\cdot, \cdot)$ is the standard bivariate normal distribution function which is also a function of ρ. Olsson developed an algorithm to derive MLE which was shown by a simulation study to have a small bias with the variance being close to the theoretical value [33]. We adopt this algorithm to calculate the polychoric correlation.

Results

Three simulation studies were conducted to evaluate (1) the accuracy of the proposed estimation methods for correlations between phenotypes; (2) the accuracy of the proposed estimation method for the variance of test statistic; and (3) the type I error rate and statistical power of the proposed multivariate test in comparison to competing methods. A real data analysis was used to identify pleiotropic genes for the risk of nicotine dependence.

Accuracy of the Estimation of Correlation between Phenotypes

We conducted a simulation study to evaluate the accuracy of different correlation estimation methods for mixed continuous, binary, and ordinal data described in the previous section. Because most genome-wide association studies contain more than 1,000 subjects, we simulated 1,000 individuals. For each individual, we simulated a pair of continuous phenotypes from bivariate normal random variables. The correlation ρ for the bivariate normal distribution ranges from −0.9 to 0.9. The binary variables were derived from the continuous variables by

Table 2. Simulation results for the correlation estimation based on Kendall's τ, biserial, polyserial, tetrachoric, or polychoric correlation. The choice of correlation methods depends on the measurement scale. The values of ρ ranges from −0.9 to 0.9. The correlation estimates and standard deviations for various methods are calculated based on 10,000 replications.

ρ	Continuous-Continuous	Continuous-Binary	Continuous-Ordinal	Binary-Binary	Binary-Ordinal	Ordinal-Ordinal
−0.9	−0.8999 (0.0066)	−0.9003 (0.0112)	−0.9009 (0.0082)	−0.8999 (0.0159)	−0.9007 (0.0149)	−0.9004 (0.0120)
−0.8	−0.7997 (0.0124)	−0.8002 (0.0185)	−0.8006 (0.0136)	−0.8000 (0.0249)	−0.8007 (0.0213)	−0.8008 (0.0160)
−0.7	−0.6999 (0.0172)	−0.7004 (0.0238)	−0.7007 (0.0183)	−0.7000 (0.0315)	−0.7009 (0.0269)	−0.7008 (0.0206)
−0.6	−0.6000 (0.0218)	−0.6004 (0.0287)	−0.6004 (0.0225)	−0.6002 (0.0374)	−0.6009 (0.0312)	−0.6008 (0.0252)
−0.5	−0.4999 (0.0253)	−0.5001 (0.0324)	−0.5005 (0.0263)	−0.5000 (0.0418)	−0.5009 (0.0350)	−0.5006 (0.0287)
−0.4	−0.4000 (0.0281)	−0.4005 (0.0351)	−0.4005 (0.0289)	−0.3998 (0.0453)	−0.4005 (0.0380)	−0.4005 (0.0311)
−0.3	−0.3001 (0.0302)	−0.3004 (0.0371)	−0.3004 (0.0310)	−0.3003 (0.0477)	−0.3005 (0.0407)	−0.3005 (0.0333)
−0.2	−0.2004 (0.0317)	−0.2010 (0.0388)	−0.2005 (0.0327)	−0.2010 (0.0493)	−0.2004 (0.0418)	−0.2009 (0.0352)
−0.1	−0.1006 (0.0330)	−0.1008 (0.0398)	−0.1008 (0.0337)	−0.1008 (0.0509)	−0.1010 (0.0428)	−0.1011 (0.0359)
0.0	−0.0000 (0.0331)	0.0002 (0.0397)	0.0001 (0.0339)	−0.0001 (0.0506)	−0.0004 (0.0428)	0.0002 (0.0364)
0.1	0.0998 (0.0324)	0.0999 (0.0392)	0.0998 (0.0331)	0.1000 (0.0499)	0.0999 (0.0420)	0.1000 (0.0354)
0.2	0.1996 (0.0317)	0.2003 (0.0390)	0.1998 (0.0323)	0.1997 (0.0494)	0.1995 (0.0414)	0.2000 (0.0348)
0.3	0.2999 (0.0298)	0.3002 (0.0370)	0.3003 (0.0304)	0.3001 (0.0477)	0.3004 (0.0398)	0.3006 (0.0326)
0.4	0.3993 (0.0279)	0.3996 (0.0352)	0.4000 (0.0287)	0.3994 (0.0449)	0.4003 (0.0372)	0.4003 (0.0311)
0.5	0.4997 (0.0249)	0.5004 (0.0319)	0.5001 (0.0259)	0.5001 (0.0416)	0.5005 (0.0347)	0.5006 (0.0284)
0.6	0.6000 (0.0217)	0.6006 (0.0285)	0.6004 (0.0226)	0.6000 (0.0377)	0.6003 (0.0314)	0.6009 (0.0249)
0.7	0.6999 (0.0172)	0.7004 (0.0237)	0.7006 (0.0184)	0.6999 (0.0317)	0.7005 (0.0269)	0.7008 (0.0206)
0.8	0.8000 (0.0124)	0.8004 (0.0184)	0.8008 (0.0137)	0.8003 (0.0250)	0.8006 (0.0215)	0.8011 (0.0162)
0.9	0.8997 (0.0066)	0.9001 (0.0113)	0.9007 (0.0082)	0.8996 (0.0160)	0.9005 (0.0152)	0.8997 (0.0112)

doi:10.1371/journal.pone.0169893.t002

dividing the observed values into two parts. Similarly, the ordinal variables were derived from the continuous variables by dividing their values into five parts. We later created six different combinations among continuous, binary or ordinal variables. For each simulated pair of phenotypes, we estimated its correlation using the corresponding method described in Estimation of the Correlation between Mixed Phenotypes section. We repeated the process 10,000 times to calculate the mean and standard deviation for each configuration. The simulation results in Table 2 show that all the proposed methods estimate the true ρ well. The confidence intervals cover the true ρ in all situations. The standard deviations are small. Even the largest standard deviation, which occurred with both phenotypes being binary and the correlation being around zero, is about 0.05. Therefore, the accuracy level is high for all of the proposed correlation estimation methods.

Accuracy of the Estimation of Variance for Test Statistic

We conducted a simulation study to evaluate the relative accuracy of the proposed variance estimation method when the phenotypes are a mixture of continuous, binary or ordinal variables. The simulation was based on 1,000 simulated individuals. Given the correlation ρ ranging from −0.9 to 0.9, we simulated bivariate normal random variables $(U, V)'$ for each individual. The binary and ordinal variables were generated from U or V following the same procedure as the simulation study described in the previous section. We also independently simulated the genotypes Z of all individuals with the minor allele frequency 0.5. We used the linear regression to test the association between a genotype and a continuous phenotype; the logistic regression for a binary phenotype; and the cumulative logit model for an ordinal phenotype. The process was repeated 10,000 times so that we have 10,000 pairs of p-values for each of the six types of combination (i.e. continuous-continuous, continuous-ordinal, ordinal-

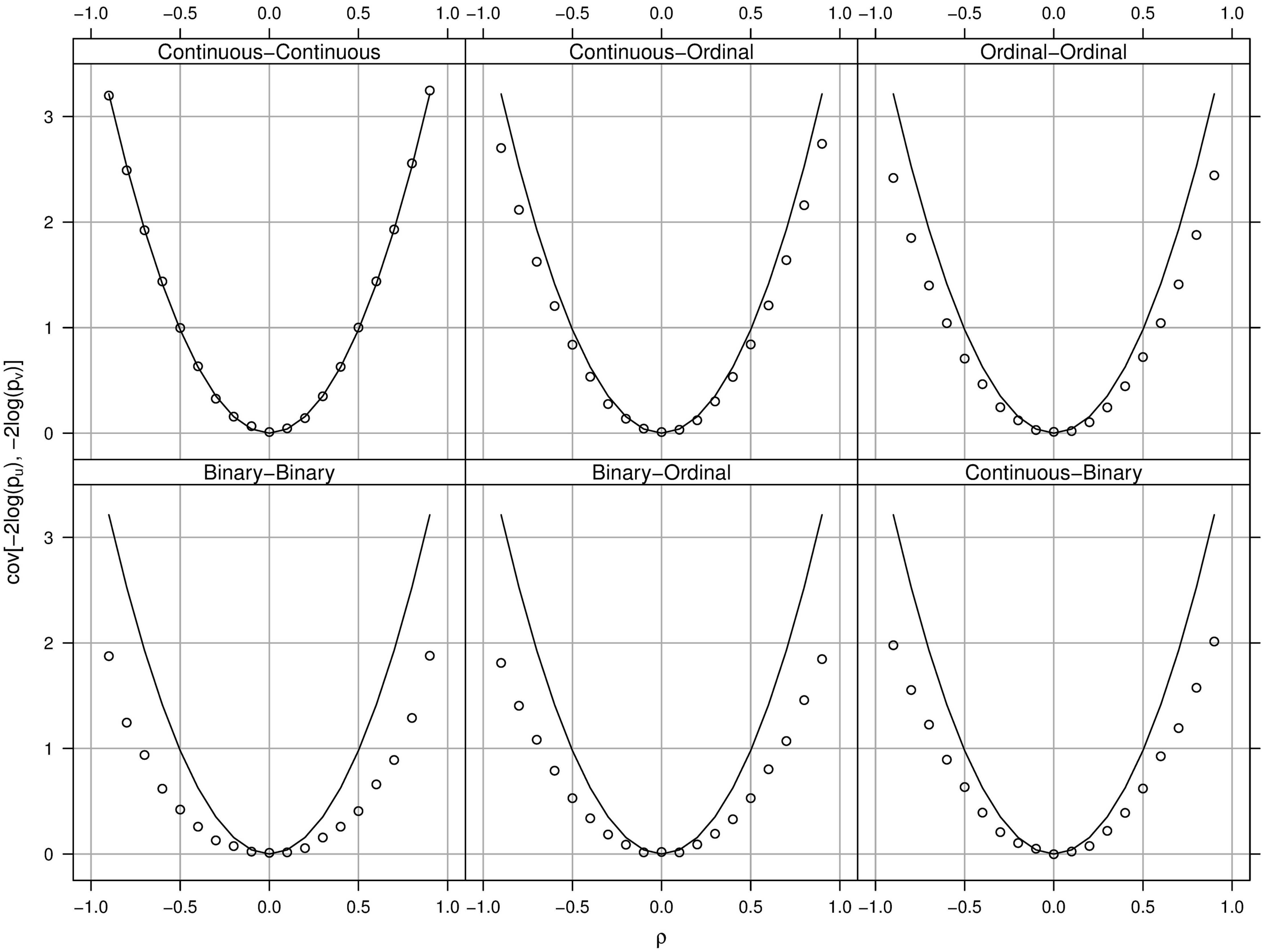

Fig 1. The relationship between the covariance *cov*[−2*log*(p_u), −2*log*(p_v)] and the correlation *ρ*. The title in each panel indicates the types of data simulated. The solid curve in each panel corresponds to our covariance estimates using Eq (3). The dotted curves are the true covariances calculated from the simulated data.

doi:10.1371/journal.pone.0169893.g001

ordinal, binary-binary, binary-ordinal, and continuous-binary). Based on these simulated *p*-values, we can calculate their covariance which is considered the *true* covariance for the purpose of comparison. Since the values of ρ are known in the experiment, they were plugged in Eq (3) to produce the *estimate* of the covariance based on the proposed method. Fig 1 summarizes the simulation results with the solid curves being the estimates and the dotted curves being the true covariances. The findings from this simulation study are summarized as follows:

1. The covariances depend on the values of ρ. When $\rho = 0$, the covariance is zero. When $|\rho|$ increases, the covariance increases.
2. The true covariance is always less than or equal to the covariance estimate using the right-hand side of Eq (3).
3. When both phenotypes are continuous (the top-left panel), the covariance estimates match the true covariances.

4. When one or both phenotypes are not continuous, the covariance estimates tend to over-estimate the true covariances.
5. The difference between the true covariances and the estimates varies across different data types. For example, the difference for continuous-ordinal data is relatively small. On the other hand, the difference for binary-binary data is the largest among all six combinations.

In summary, the results indicate that our proposed estimation method tends to slightly over-estimate the true covariance. For example, when $\rho = 0.5$ and both the observed phenotypes are binary, our estimate of $Var[S^{(g)}]$ is 9.9956 which over-estimates the target value of 8.813 by 13%. Nevertheless, when one or both phenotypes are continuous or ordinal, the difference is much smaller. Furthermore, because our approach conservatively estimates the variance of the test statistic $S^{(g)}$, the resulting type I error rate is controlled.

Type I Error Rate and Statistical Power of the Proposed Multivariate Test

A simulation study was conducted to evaluate the performance of the proposed method in terms of the type I error rate and statistical power. We considered a pleiotropic gene model in which multivariate phenotypes were modeled as a function of the candidate gene with varied effect sizes. For each individual, we simulated the genotype $Z = (0, 1, 2)$ based on the minor allele frequency (MAF) which is uniformly distributed on [0.1, 0.5]. Therefore, Z represents the number of reference alleles for a SNP. A total of 100 individuals were generated. The latent phenotypes were simulated from multivariate normal (MVN) random variables. We considered $(V_1, V_2, \ldots, V_6)' \sim MVN((\mu_1, \mu_2, \ldots, \mu_6)', \Sigma)$ where the diagonal elements of Σ are 1 and the off diagonal elements of Σ are ρ. The value of $\mu_i(i = 1, \ldots, 6)$ was defined as

$$\mu_i = \begin{cases} -e_i & \text{if } Z = 0 \\ 0 & \text{if } Z = 1 \\ e_i & \text{if } Z = 2, \end{cases}$$

where e_i is the genetic effect size.

The observed phenotypes $(U_1, U_2, \ldots, U_6)'$ with mixed measurement scales were derived from the simulated latent variables $(V_1, V_2, \ldots, V_6)'$. Let $U_i = V_i(i = 1, 2)$ represent the continuous measurements. By dividing $V_i(i = 3, 4)$ into two intervals with the cut-off value C, we derived binary measurements: $U_i = 1$ if $V_i > C$ or $U_i = 0$ if $V_i \leq C(i = 3, 4)$. For the ordinal scale, we divided $V_i(i = 5, 6)$ into five intervals using 4 cut-off points and assigned the values of 1 to 5 to $U_i(i = 5, 6)$ accordingly.

The values of ρ were set at 0, 0.35 and 0.75 to represent independent, moderate dependent, or highly dependent multivariate phenotypes. We also manipulated the values of $e_1, \ldots, e_6$ to be 0, 0.5, 0.7, or 0.9 to represent different genetic effect sizes. Note that the configuration of all $e_1, \ldots, e_6$ being equal to zero represents the null condition of no genetic effect.

We compared the proposed method (labelled as *Mixed*), with three alternative approaches. When there was no method available for handling mixed phenotypes, people tended to analyze them as phenotypes in the same measurement scale. One commonly adopted approach is to dichotomize each of $(U_1, U_2, \ldots, U_6)'$ and carry out the analysis with the marginal p-values derived from a logistic regression model (labelled as *Dichotomous*). Another naive approach is to treat each of $(U_1, U_2, \ldots, U_6)'$ as continuous measurements and carry out the analysis with the marginal p-values derived from a linear regression model (labelled as *Continuous*). We also compared our method with the ideal situation when the analysis is conducted on the latent

Table 3. Simulation results for the empirical power with varied correlations ρ and genetic effect sizes (e_1, . . ., e_6). The *Latent* column is the power with the proposed method applied to 6 latent phenotypes. The *Mixed* column is the power with the proposed method applied to 6 observed phenotypes. The *Dichotomous* column is the power when the observed phenotypes are dichotomized. The *Continuous* column is the power when the phenotypes in mixed measurements are treated as continuous variables. The number of iterations is 10^6 when all genetic effect sizes are zero and 10^4 for other situations.

ρ	e_1	e_2	e_3	e_4	e_5	e_6	Latent	Mixed	Dichotomous	Continuous
0	0	0	0	0	0	0	0.00009	0.00008	0.00004	0.00009
0.35	0	0	0	0	0	0	0.00057	0.00024	0.00005	0.00025
0.75	0	0	0	0	0	0	0.00027	0.00007	0.00002	0.00012
0	0.5	0.5	0.5	0.5	0.5	0.5	0.96	0.91	0.77	0.91
0.35	0.5	0.5	0.5	0.5	0.5	0.5	0.84	0.73	0.49	0.73
0.75	0.5	0.5	0.5	0.5	0.5	0.5	0.51	0.36	0.19	0.39
0	0	0.7	0	0.7	0	0.7	0.94	0.88	0.68	0.88
0.35	0	0.7	0	0.7	0	0.7	0.82	0.68	0.42	0.70
0.75	0	0.7	0	0.7	0	0.7	0.36	0.20	0.08	0.25
0	0.9	0.9	0	0	0	0	0.94	0.94	0.67	0.94
0.35	0.9	0.9	0	0	0	0	0.84	0.84	0.42	0.83
0.75	0.9	0.9	0	0	0	0	0.37	0.37	0.05	0.38
0	0	0	0.9	0.9	0	0	0.95	0.71	0.68	0.74
0.35	0	0	0.9	0.9	0	0	0.85	0.43	0.41	0.49
0.75	0	0	0.9	0.9	0	0	0.37	0.05	0.05	0.09
0	0	0	0	0	0.9	0.9	0.95	0.90	0.69	0.92
0.35	0	0	0	0	0.9	0.9	0.85	0.72	0.42	0.78
0.75	0	0	0	0	0.9	0.9	0.38	0.16	0.05	0.30

doi:10.1371/journal.pone.0169893.t003

phenotypes $(V_1, V_2, . . ., V_6)'$. This approach is labelled as *Latent* and serves as our gold standard.

The empirical type I error rates were set at 10^{-4}. In order to evaluate the performance of competing methods in terms of controlling the type I error, we carried out 10^6 replications under the null conditions. For the conditions with non-zero effect sizes, 10^4 replications are sufficient to show the differences in power. The simulation results are shown in Table 3. The findings are summarized as follows:

1. Based on 10^6 iterations, a half of the width of 95% confidence interval is 0.00139. Therefore, under the null conditions when all the effect sizes are zero ($e_i = 0$, $i = 1, . . ., 6$), all the four methods control the type I error.
2. When some or all effect sizes (e_i) are nonzero, the power of all methods decreases as the correlation ρ increases. The decrease in power is expected because highly correlated phenotypes contain less information than phenotypes with low correlations.
3. Among all four methods, the *Latent* has the highest power and the *Dichotomous* has the lowest power. When mixed measurement scales are observed, dichotomizing all observed measurements could reduce the power by half.
4. The power of the proposed method (*Mixed*) is very close to that of the *Latent* which is the gold standard. Because the true values of latent phenotypes are unknown in real situations, this result demonstrates that our proposed method can provide an efficient and powerful way to conduct multivariate testing with phenotypes in mixed measurements in practice.
5. If we treat all mixed measurement scales as continuous variables and apply the proposed method to it (i.e. the *Continuous*), the power is close to that of the proposed method. In

some situations, it had even higher power than the proposed method. However, modeling binary or ordinal responses as continuous variables is both mathematically and practically questionable. For instance, in the binary case, the mean response value is within 0 and 1 but the predicted values from a linear regression model would cover the entire real line [18]. Furthermore, Guisan and Harrell (2000) [34] provided four reasons why applying a linear regression model is statistically incorrect when the outcome is ordinal. Therefore, we do not recommend the use of the *Continuous* approach.

6. In comparison to the situation when all phenotypes are associated with the pleiotropic gene, the power of all methods tends to be reduced when a half of the phenotypes are not associated with the gene ($e_1 = e_3 = e_5 = 0$). Even when the nonzero effect sizes (e_2, e_4, e_6) increase from 0.5 to 0.7, the power is still lower than that in the situation when all genetic effects are at the 0.5 level. This implies that selecting relevant phenotypes is a prerequisite for maintaining the power level of a multivariate test.

7. We also investigated whether the power varies with different measurement scales, and found that the power for continuous phenotypes ($e_1 = e_2 = 0.9$, $e_j = 0$, $j = 3, 4, 5, 6$) is the highest; the power for ordinal phenotypes ($e_5 = e_6 = 0.9$, $e_j = 0$, $j = 1, 2, 3, 4$) is the next highest; and that for binary phenotypes ($e_3 = e_4 = 0.9$, $e_j = 0$, $j = 1, 2, 5, 6$) is the lowest. Such reduction in power from continuous to ordinal is relative small in comparison to the reduction from ordinal to binary. Thus, collecting phenotype data in continuous or ordinal measurement scales has the advantage of increasing statistical power.

Real Data Analysis

We conducted real data analysis using the database from the Study of Addiction: Genetics and Environment (SAGE). 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 4,121. Each individual was genotyped using the Illumina Human 1M-Duo beadchip which contains over 1 million SNP markers.

The Fagerström Test for Nicotine Dependence (FTND) is a commonly adopted instrument for assessing the intensity of physical addiction to nicotine [35]. It consists of six items of which some are ordinal and the others are binary:

1. `ftnd_1`: How soon after you wake up do you smoke your first cigarette? (3 = within 5 minutes; 2 = 6–30 minutes; 1 = 31–60 minutes; 0 = after 60 minutes)

2. `ftnd_2`: Do you find it difficult to refrain from smoking in places where it is forbidden? (1 = yes; 0 = no)

3. `ftnd_3`: Which cigarette would you hate most to give up? (1 = the first one in the morning; 0 = all others)

4. `ftnd_4`: How many cigarettes per day do you smoke? (0 = 10 or less; 1 = 11–20; 2 = 21–30; 3 = 31 or more)

5. `ftnd_5`: Do you smoke more frequently during the first hours after waking than during the rest of the day? (1 = yes; 0 = no)

6. `ftnd_6`: Do you smoke when you are so ill that you are in bed most of the day? (1 = yes; 0 = no)

The items are summed to yield a total score of 0–10. The higher the total score, the more intense is the patient's physical dependence on nicotine. Clinically, the score of 6 or higher indicates high nicotine dependency and represents individuals who would be particularly likely to benefit from tapering and/or the prescription of nicotine replacement therapy as an adjunct to standard counseling. The score of 5 or less, on the other hand, suggests low to moderate nicotine dependency and represents individuals for whom standard counseling is most appropriate. The total score, `ftnd_total`, or the binary variable, `ftnd_total (binary)`, coded as 0 or 1 depending on whether `ftnd_total`< 6 would be popular choices for the phenotype measure. However, they have some important issues including (1) throwing away potentially important information; (2) assigning different weights to the items; and (3) choosing an arbitrary cutoff. Our proposed method, therefore, provides a new approach to conduct a multivariate test based on the original six items in mixed measurement scales.

From the original 4,121 individuals, we eliminated individuals whose FTND phenotype information was not available. Since a few individuals were related family members, the `KING` program [36] was used to identify and select unrelated individuals. The final number of unrelated individuals included in the analysis was 2,775 (1,288 males, 1,487 females). Our analysis only included 22 autosomes and the 753,238 SNP's that passed the quality control procedures [37]. The phenotype distributions among the 2,775 individuals are presented in Fig 2 using bar-plots. Four of the FTND items (`ftnd_2`, `ftnd_3`, `ftnd_5`, and `ftnd_6`) are binary and the other two FTND items (`ftnd_1` and `ftnd_4`) are ordinal ranging from 0 to 3. The FTND total score ranges from 0 to 9. The sample correlations among the 6 FTND items are shown in Table 4. The correlations range from 0.44 (`ftnd_2` and `ftnd_5`) to 0.77 (`ftnd_1` and `ftnd_6`).

We conducted marginal genome-wide association tests on the six FTND items and the two derived FTND scores. We also conducted the multivariate test based on the six FTND items. For the binary variables, the logistic regression was employed. For the ordinal variables, the cumulative logit model was used. Because the `ftnd_total` score was treated as a continuous variable, the linear regression was applied. For all the regression models, in addition to the genotype (coded as 0, 1, or 2), we included each individual's age (from 18 to 74 years old), gender, and race (850 black and 1,925 white) as covariates to eliminate potential confounders. We also carried out the principal component analysis [19] to examine population stratification but did not include any principal components in the model because the first principal component perfectly matches with race (i.e. the multicolinearity issue). The marginal p-values are summarized using QQ-plots in Fig 3; the p-values based on the multivariate test using the proposed method are shown in Fig 4. The fact that more observed p-values are above the diagonal line in Fig 4 (in comparison to Fig 3) indicates that the multivariate test is more powerful and thus may identify more significant SNPs associated with the six FTND items.

To identify the SNPs associated with susceptibility to FTND, we set the reduced type I error rate at 10^{-6}. Based on the marginal p-values, we identify 1 SNP (rs821722, $p = 9.54 \times 10^{-7}$) to be associated with `ftnd_1` and 1 SNP (rs3138134, $p = 7.94 \times 10^{-7}$) with `ftnd_3`. The other four FTND items are not associated with any SNP. The derived phenotype based on `ftnd_total` is also not associated with any SNP. Besides, the SNP that is associated with `ftnd_1` is also associated with `ftnd_total (binary)`. On the other hand, using the proposed multivariate test, we identify 9 SNPs (rs17538699, rs17798885, rs2245261, rs4077464, rs4658846, rs4658847, rs6553017, rs7672047, rs944582) to be associated with the six FTND phenotype variables. 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 risk shared by the six FTND items, the proposed method is a better approach than marginal tests.

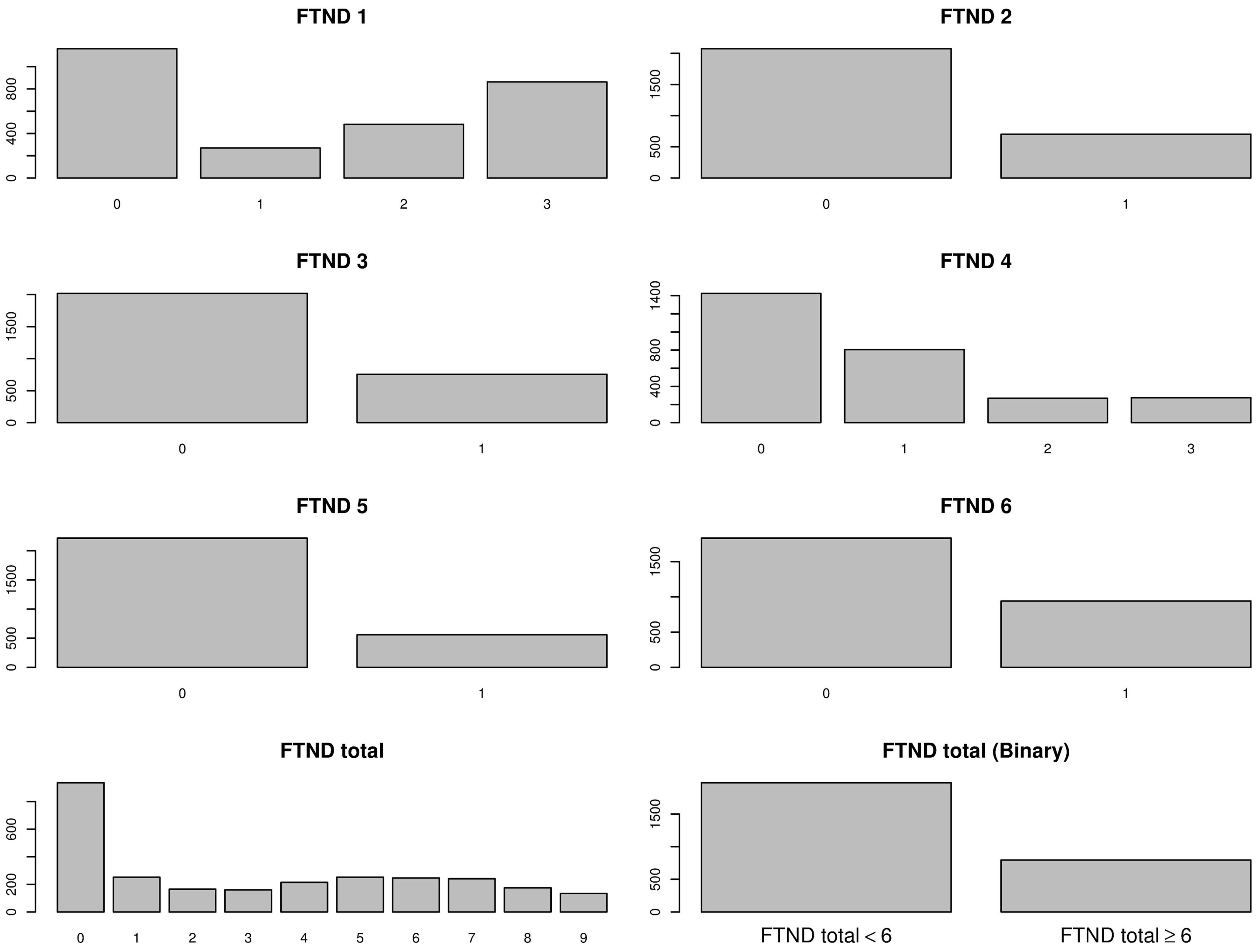

Fig 2. The distributions of phenotypes: FTND 1, FTND 2, . . ., FTND 6, FTND total, and FTND total (Binary). FTND total (Binary) is derived from FTDN total according to whether FTND total score is less than 6 or not.

doi:10.1371/journal.pone.0169893.g002

Discussion

In this study, we propose a new multivariate method for GWAS when the multivariate phenotypes are a mixture of continuous, binary, or ordinal variables. We use a latent response model to unify different data types for estimating correlation between phenotypes. The first phase of

Table 4. The correlations among the 6 FTND items.

correlation	ftnd_2	ftnd_3	ftnd_4	ftnd_5	ftnd_6
ftnd_1	0.6815	0.6758	0.7528	0.6446	0.7770
ftnd_2		0.4579	0.5895	0.4403	0.6838
ftnd_3			0.4822	0.6394	0.5294
ftnd_4				0.4452	0.6702
ftnd_5					0.5110

doi:10.1371/journal.pone.0169893.t004

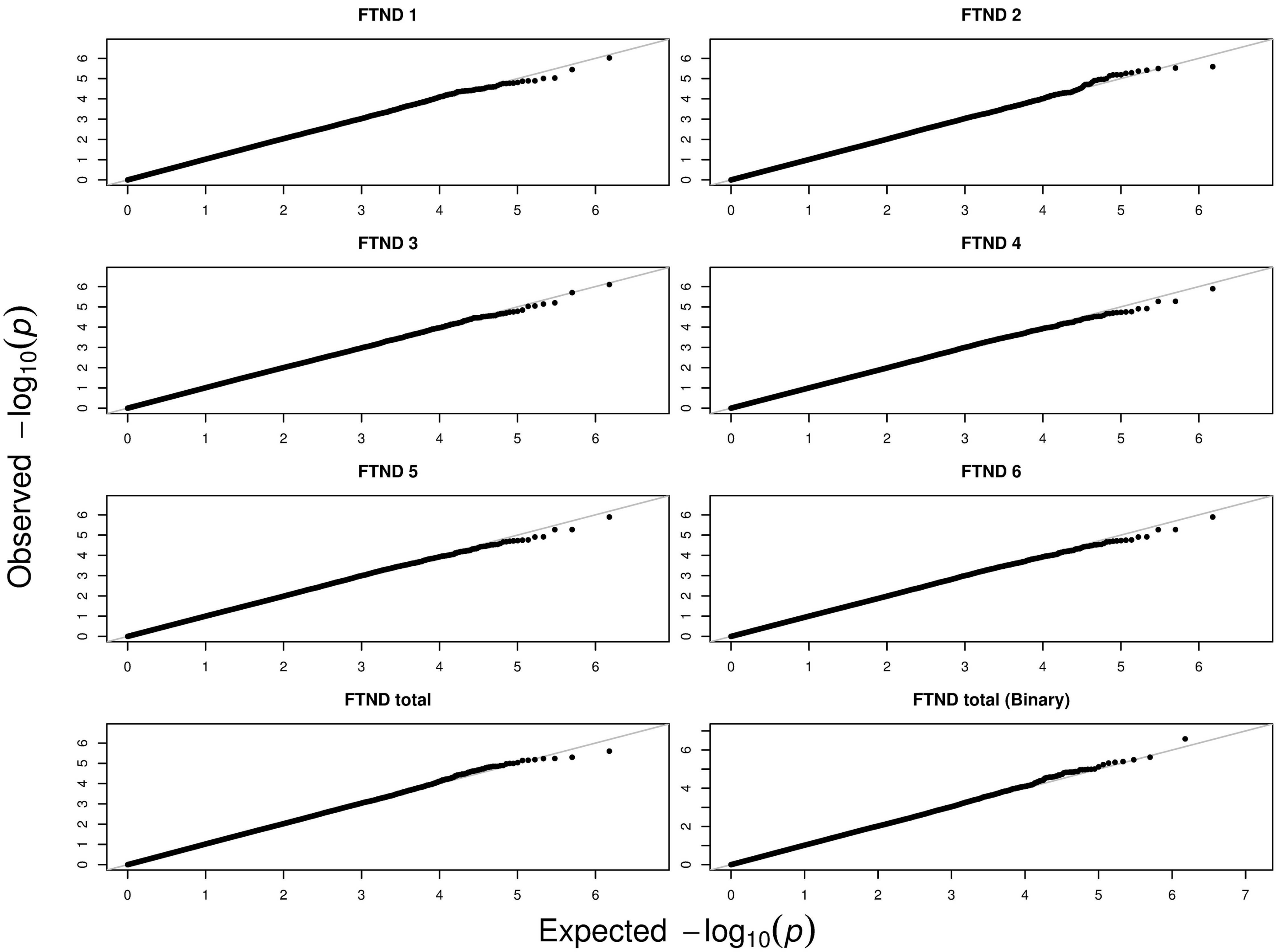

Fig 3. The Q-Q plots of observed *p*-values versus expected *p*-values based on the marginal tests.

doi:10.1371/journal.pone.0169893.g003

our method uses regression models with different link functions to accommodate different measurement scales of the phenotypes. These regression models not only enable us to evaluate the goodness-of-fit but also provide a way for adding covariates to adjust for potential confounders. The second phase of our method employs continuous latent responses to handle the correlation estimation of mixed data types. The simulation study demonstrates that our proposed correlation estimation methods have high levels of accuracy. The results also show that our approach conservatively estimates the variance of the test statistic so that the type I error rate is controlled.

We conducted a simulation study to evaluate the proposed multivariate test in terms of the type I error rate and statistical power when the observed phenotypes are in mixed measurement scales, in comparison to three competing methods: (1) the ideal analysis when the latent phenotypes are known; (2) a conventional approach that dichotomizes all phenotypes; and (3) a conventional approach treating all phenotypes as continuous. The simulation result shows that the proposed method maintains the power at the level very close to that of the ideal

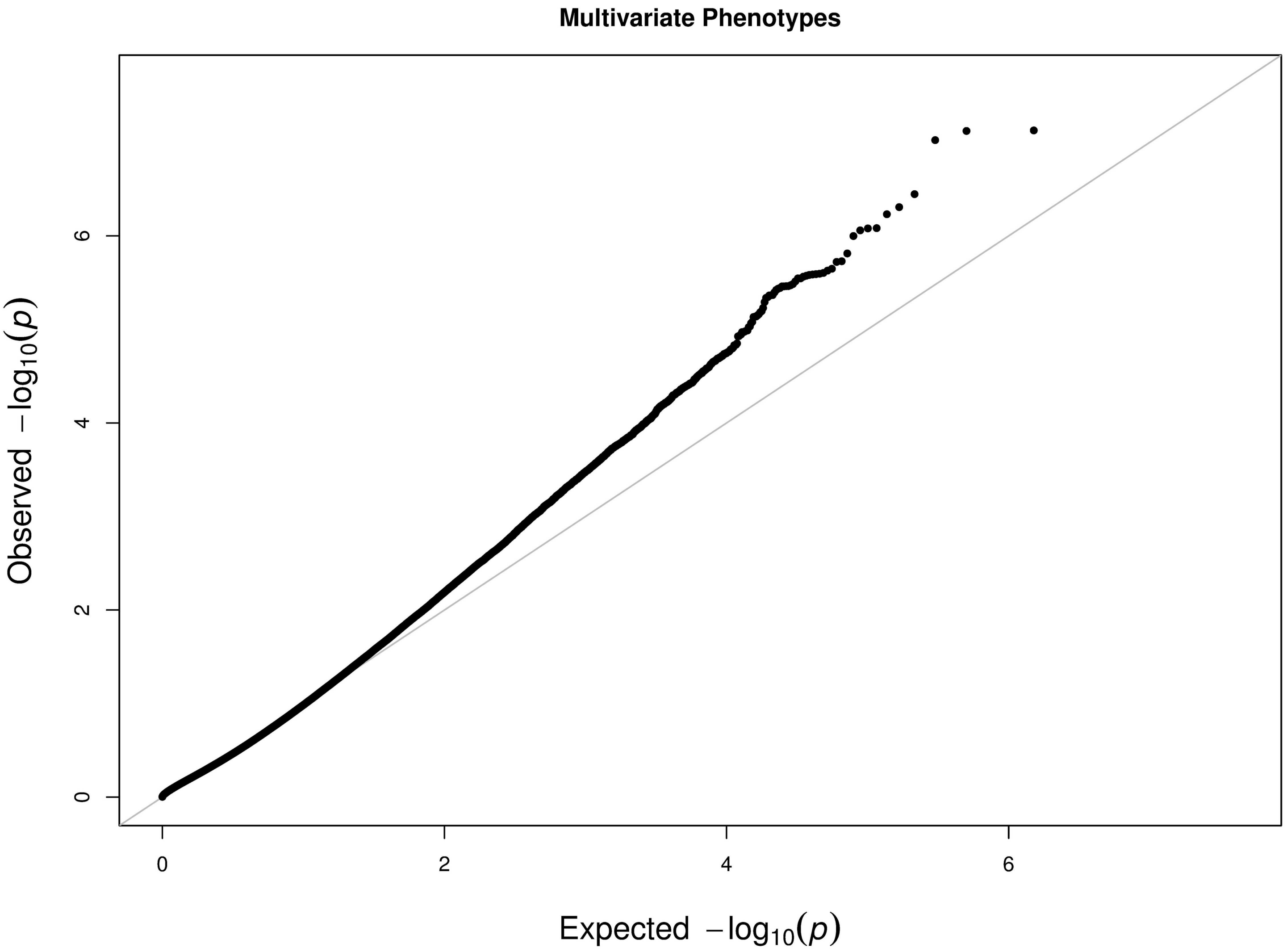

Fig 4. The Q-Q plot of observed *p*-values versus expected *p*-values based on the multivariate test.

doi:10.1371/journal.pone.0169893.g004

analysis while controlling the type I error. Furthermore, when mixed measurement scales are observed, dichotomizing all observed measurements could reduce the power by half. Although the power for treating all mixed measurement scales as continuous variables is close to that for the proposed method, this conventional approach is not recommended because fitting a linear regression model on categorical variables is both mathematically and practically questionable.

Our real data analysis using the well-known database, SAGE, in the addiction field demonstrates that conducting a multivariate test on multiple phenotypes can increase the power of identifying markers that may not be, otherwise, chosen using marginal tests. The proposed method also offers a new approach to analyzing the items rather than the total score of FTND as multivariate phenotypes in GWAS. In summary, the proposed method is a better approach than marginal tests to identify pleiotropic genes that contribute to the common liability to complex diseases such as substance use disorders.

Although the proposed method was designed to handle continuous, binary, and ordinal phenotypes, it can be extended to deal with count data. Under the framework of our two-phase approach, the first phase would employ a Poisson or negative binomial regression model to conduct a marginal test on count data; and the second phase would treat the count data as continuous in calculation of pairwise correlations, because both Poisson and negative binomial distributions can be approximated by a normal distribution based on the large sample theory [38]. The proposed method involving Kendall τ is also robust against deviation from a normal distribution. Nevertheless, future research is needed to further extend the method to handle zero-inflated count data such as the number of alcohol use disorder symptoms [39, 40]. It is also important to extend the proposed method to deal with nominal phenotypes such as disease subtypes. For example, Zucker (1994) [41] proposed a well-known developmental theory that classifies alcoholism into 4 subtypes: antisocial alcoholism, developmentally limited alcoholism, negative affect alcoholism, and the primary alcoholism (isolated, episodic, and developmentally cumulative).

In standard case-control studies, the proportion of cases in the sample may be much higher than that in the population. To deal with this ascertainment bias, many studies employed a liability threshold model [42] assuming an underlying latent random variable, which is normally distributed in the population and has a certain threshold that determines the disease status. Zöllner and Pritchard (2007) [43] proposed another approach to correct the ascertainment bias directly based on population prevalence of the disease phenotype and sampling scheme. They also conducted a simulation study showing that when the association test is powerful or the sample size is in thousands (applicable to our setting), the ascertainment bias is negligible and thus the correction may not be necessary. These existing methods and simulation results are, however, based on GWAS with the case-control design involving a binary phenotype. How to extend these methods to handle GWAS with multivariate phenotypes is, therefore, a very important and yet complex question for future research because the definition of "case" is unclear, especially when the phenotypes are continuous.

The method proposed in this study was designed for GWAS with independent subjects. Due to reduced costs for SNP arrays, in recent years, many family studies have collected GWAS data [44–46] so relatedness has become a new component to account for in modelling. The linear mixed model (LMM) has been used to adjust for correlation between related subjects with a univariate phenotype [47]. When multivariate phenotypes from related subjects are considered, the association test between a SNP and multivariate phenotypes needs to account for an additional level of correlation. Since the computational bottleneck is to estimate genetic correlation matrix, direct implementation of the LMM can only handle a sample size in hundreds. Thus, the computation becomes very expensive when LMM is extended to multivariate phenotypes. Zhou and Stephens (2014) [48] proposed an efficient matrix-variate linear mixed model (mvLMM) to identify pleiotropic genes while controlling for correlation among a large sample of related subjects. Theoretically, their method is applicable to any number of phenotypes. However, the complexity of their method and computational speed increase with the number of phenotypes. Specifically, th number of parameters in the mvLMM and the computational time for the EM algorithm is quadratically proportional to the number of phenotypes. Therefore, mvLMM is only applicable to a modest number of phenotypes (fewer than 10 traits). A future direction of research is to extend the proposed method to handle related subjects and compare it with mvLMM on performance.

Acknowledgments

We 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).

Author Contributions

Conceptualization: JJY LKW AB.

Formal analysis: JJY AB.

Funding acquisition: LKW AB.

Investigation: LKW AB.

Methodology: JJY AB.

Software: JJY.

Writing – original draft: JJY.

Writing – review & editing: LKW AB.

References

1. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005 Apr; 308(5720):385–389. doi: 10.1126/science.1109557 PMID: 15761122
2. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research. 2014 Jan; 42(D1):D1001–D1006. doi: 10.1093/nar/gkt1229 PMID: 24316577
3. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, et al. Abundant Pleiotropy in Human Complex Diseases and Traits. American Journal of Human Genetics. 2011 Nov; 89 (5):607–618. doi: 10.1016/j.ajhg.2011.10.004 PMID: 22077970
4. Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010 Jun; 465(7298):627–631. doi: 10.1038/nature08800 PMID: 20336072
5. Allison DB, Thiel B, St Jean P, Elston RC, Infante MC, Schork NJ. Multiple phenotype modeling in gene-mapping studies of quantitative traits: Power advantages. American Journal of Human Genetics. 1998 Oct; 63(4):1190–1201. doi: 10.1086/302038 PMID: 9758596
6. Yang JJ, Li J, Williams LK, Buu A. An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function. BMC Bioinformatics. 2016; 17(1):1–11. doi: 10.1186/s12859-015-0868-6
7. Flory JD, Manuck SB. Impulsiveness and Cigarette Smoking. Psychosomatic Medicine. 2009 May; 71 (4):431–437. doi: 10.1097/PSY.0b013e3181988c2d PMID: 19251874
8. Teixeira-Pinto A, Normand SLT. Correlated bivariate continuous and binary outcomes: Issues and applications. Statistics In Medicine. 2009 Jun; 28(13):1753–1773. doi: 10.1002/sim.3588 PMID: 19358234
9. Fitzmaurice GM, Laird NM. Regression models for mixed discrete and continuous responses with potentially missing values. Biometrics. 1997 Mar; 53(1):110–122. doi: 10.2307/2533101 PMID: 9147588
10. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nature Reviews Genetics. 2013 Jul; 14(7):483–495. doi: 10.1038/nrg3461 PMID: 23752797
11. Littell RC, Folks JL. Asymptotic Optimality of Fisher's Method of Combining Independent Tests. Journal of the American Statistical Association. 1971; 66(336):802–806. doi: 10.1080/01621459.1971.10482347
12. Littell RC, Folks JL. Asymptotic Optimality of Fisher's Method of Combining Independent Tests II. Journal of the American Statistical Association. 1973; 68(341):193–194. doi: 10.1080/01621459.1973.10481362
13. Hoeffding W. The Large-sample Power of Tests Based on Permutation of Observations. Annals of Mathematical Statistics. 1952; 23(2):169–192. doi: 10.1214/aoms/1177729436
14. Brown MB. Method For Combining Non-independent, One-sided Tests of Significance. Biometrics. 1975; 31(4):987–992. doi: 10.2307/2529826
15. Yang JJ. Distribution of Fisher's combination statistic when the tests are dependent. Journal of Statistical Computation and Simulation. 2010 Jan; 80(1–2):1–12. doi: 10.1080/00949650802412607

16. Pesarin F. Multivariate permutation tests with applications in biostatistics. John Wiley & Sons; 2001.

17. Pesarin F, Salmaso L. Permutation tests for complex data. Chichester: John Wiley & Sons; 2010.

18. Agresti A. Categorical Data Analysis. 2nd ed. Wiley Series in Probability and Statistics. Wiley; 2002.

19. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006 Aug; 38 (8):904–909. doi: 10.1038/ng1847 PMID: 16862161

20. Schaeffer MS, Levitt EE. Concerning Kendall Tau, a Nonparametric Correlation-coefficient. Psychological Bulletin. 1956; 53(4):338–346. doi: 10.1037/h0045013 PMID: 13336201

21. Kendall MG. Rank and Product-moment Correlation. Biometrika. 1949; 36(1–2):177–193. doi: 10.1093/biomet/36.1-2.177 PMID: 18132091

22. Kendall M, Gibbons JD. Rank Correlation Methods. 5th ed. London: Oxford; 1990.

23. 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 Jul; 7:96–105. doi: 10.2307/2345365

24. Tate RF. The Theory of Correlation Between Two Continuous Variables When One Is Dichotomized. Biometrika. 1955; 42(1–2):205–216. doi: 10.2307/2333437

25. Brogden HE. A new coefficient; application to biserial correlation and to estimation of selective efficiency. Psychometrika. 1949 Sep; 14(3):169–82. doi: 10.1007/BF02289151 PMID: 24536228

26. Lord FM. Biserial Estimates of Correlation. Psychometrika. 1963; 28(1):81–85. doi: 10.1007/BF02289550

27. Bedrick EJ. A Comparison of Generalized and Modified Sample Biserial Correlation Estimators. Psychometrika. 1992 Jun; 57(2):183–201. doi: 10.1007/BF02294504

28. Pearson K. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences. 1896; 187:253–318. doi: 10.1098/rsta.1896.0007

29. Good IJ. Comments, conjectures and conclusions. Journal of Statistical Computation and Simulation. 2006; 76(8):737–740. doi: 10.1080/10629360500108186

30. Cox NJ. On the Estimation of Spatial Auto-correlation in Geomorphology. Earth Surface Processes and Landforms. 1983; 8(1):89–93. doi: 10.1002/esp.3290080109

31. Olsson U, Drasgow F, Dorans NJ. The Polyserial Correlation-coefficient. Psychometrika. 1982; 47 (3):337–347. doi: 10.1007/BF02294164

32. Pearson K, Pearson ES. On polychoric coefficients of correlation. Biometrika. 1922 Jul; 14:127–156. doi: 10.1093/biomet/14.1-2.127

33. Olsson U. Maximum Likelihood Estimation of the Polychloric Correlation-coefficient. Psychometrika. 1979; 44(4):443–460. doi: 10.1007/BF02296207

34. Guisan A, Harrell FE. Ordinal response regression models in ecology. Journal of Vegetation Science. 2000 Oct; 11(5):617–626. doi: 10.2307/3236568

35. Heatherton TF, Kozlowski LT, Frecker RC, Fagerström KO. The Fagerström Test for Nicotine Dependence—a Revision of the Fagerström Tolerance Questionnaire. British Journal of Addiction. 1991 Sep; 86(9):1119–1127. doi: 10.1111/j.1360-0443.1991.tb01879.x PMID: 1932883

36. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010 Nov; 26(22):2867–2873. doi: 10.1093/bioinformatics/btq559 PMID: 20926424

37. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nature Protocols. 2010 Sep; 5(9):1564–1573. doi: 10.1038/nprot.2010.116 PMID: 21085122

38. Casella G, Berger RL. Statistical Inference. Duxbury advanced series. Brooks/Cole Publishing Company; 1990.

39. Buu A, Johnson NJ, Li R, Tan X. New variable selection methods for zero-inflated count data with applications to the substance abuse field. Statistics in Medicine. 2011 Aug; 30(18):2326–2340. doi: 10.1002/sim.4268 PMID: 21563207

40. Buu A, Li RZ, Tan XM, Zucker RA. Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. Statistics in Medicine. 2012 Dec; 31(29):4074–4086. doi: 10.1002/sim.5510 PMID: 22826194

41. Zucker RA. Pathways to alcohol problems and alcoholism: A developmental account of the evidence for multiple alcoholisms and for contextual contributions to risk. In: Zucker RA, Boyd G, Howard J, editors. Research monograph-26, The development of alcohol problems: exploring the biopsychosocial matrix

of risk. vol. 26. Rockville, MD (6000 Executive Boulevard, Rockville 20892): National Institute on Alcohol Abuse and Alcoholism; 1994. p. 255–289.

42. Falconer DS. Inheritance of Liability to Certain Diseases Estimated from Incidence among Relatives. Annals of Human Genetics. 1965; 29:51–76. doi: 10.1111/j.1469-1809.1965.tb00500.x

43. Zöllner S, Pritchard JK. Overcoming the winner's curse: Estimating penetrance parameters from case-control data. American Journal of Human Genetics. 2007 Apr; 80(4):605–615. doi: 10.1086/512821 PMID: 17357068

44. Laird NM, Lange C. Family-based designs in the age of large-scale gene-association studies. Nature Reviews Genetics. 2006 May; 7(5):385–394. doi: 10.1038/nrg1839 PMID: 16619052

45. Ott J, Kamatani Y, Lathrop M. Family-based designs for genome-wide association studies. Nature Reviews Genetics. 2011 Jul; 12(7):465–474. doi: 10.1038/nrg2989 PMID: 21629274

46. Laird NM, Lange C. The Role of Family-Based Designs in Genome-Wide Association Studies. Statistical Science. 2009 Nov; 24(4):388–397. doi: 10.1214/08-STS280

47. Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics. 2006 Feb; 38 (2):203–208. doi: 10.1038/ng1702 PMID: 16380716

48. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature Methods. 2014 Apr; 11(4):407–409. doi: 10.1038/nmeth.2848 PMID: 24531419