Published in final edited form as:
Biometrics. 2017 December ; 73(4): 1210–1220. doi:10.1111/biom.12680.

Weighted Pseudolikelihood for SNP set Analysis with Multiple Secondary Outcomes in Case-Control Genetic Association Studies

Tamar Sofer[1,*], **Elizabeth D. Schifano**[2,**], **David C. Christiani**[3,***], and **Xihong Lin**[4,****]

[1]Department of Biostatistics, University of Washington, Seattle, WA 98105; authors contributed equally

[2]Department of Statistics, University of Connecticut, Storrs, CT 06269; authors contributed equally

[3]Department of Environmental Health, Harvard School of Public Health, Boston, MA 02115

[4]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115

Summary

We propose a weighted pseudolikelihood method for analyzing the association of a SNP set, e.g., SNPs in a gene or a genetic pathway or network, with multiple secondary phenotypes in case-control genetic association studies. To boost analysis power, we assume that the SNP-specific effects are shared across all secondary phenotypes using a scaled mean model. We estimate regression parameters using Inverse Probability Weighted (IPW) estimating equations obtained from the weighted pseudolikelihood, which accounts for case-control sampling to prevent potential ascertainment bias. To test the effect of a SNP set, we propose a weighted variance component pseudo-score test. We also propose a penalized IPW pseudolikelihood method for selecting a subset of SNPs that are associated with the multiple secondary phenotypes. We show that the proposed variable selection procedure has the oracle properties and is robust to misspecification of the correlation structure among secondary phenotypes. We select the tuning parameter using a weighted Bayesian Information-like Criterion (wBIC). We evaluate the finite sample performance of the proposed methods via simulations, and illustrate the methods by the analysis of the multiple secondary smoking behavior outcomes in a lung cancer case-control genetic association study.

[*]tsofer@uw.edu
[**]elizabeth.schifano@uconn.edu
[***]dchris@hsph.harvard.edu
[****]xlin@hsph.harvard.edu

Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 4, 5, 6, 7, and 8 are available with this paper at the Biometrics website on Wiley Online Library. Code applying the proposed test and variable selection procedure, as well as examples for applying the procedures, are provided on Github at https://github.com/tamartsi/Multiple_correlated_secondary_outcomes_SNP_set_test.

Keywords

Biased sampling; High-dimensional data; SNP set analysis; Sparsity; Variable selection; Variance component test; Weighted BIC

1. Introduction

Genome-wide association studies (GWAS) measure common genetic variants in a study sample, with the goal of identifying single nucleotide polymorphisms (SNPs) that are associated with a disease or a quantitative trait. GWAS often sample subjects using a case-control design, where disease cases and disease-free controls are genotyped to identify SNPs that are associated with disease susceptibility (e.g., Hunter et al., 2007). There is substantial interest to take advantage of these existing large case-control GWAS to identify common genetic variants that are associated with the *secondary* traits that are often collected in these studies, in addition to the primary disease status. For example, in the motivating lung cancer case-control GWAS conducted in Massachusetts General Hospital (MGH), we are interested in identifying SNPs that are associated with smoking behavior, measured by a few continuous variables. In this paper, we propose statistical methods to study the association of a potentially high-dimensional set of SNPs, e.g., SNPs in a gene, genetic pathway, or network, with *multiple* secondary phenotypes that measure the same underlying trait, e.g., multiple measures of smoking behavior, in case-control GWAS.

Since subjects from a case-control study are often sampled based on a primary disease status, cases are over-represented compared to the underlying population. Thus, careful attention is warranted for inference regarding the secondary phenotypes based on case-control samples. Similar caution is needed in secondary phenotype analysis with trait-dependent sampling of continuous primary outcomes (Lin et al., 2013). Analysis methods that ignore or improperly account for the biased sampling mechanism can lead to biased estimates of the population effects. Selection bias is particularly problematic when both the SNPs and the secondary phenotypes are associated with the primary disease (Monsees et al., 2009). Such associations are likely to exist in the lung cancer GWAS, as it is expected that both SNPs and smoking behavior will affect lung cancer risk. Our task is thus to (1) study the association of interest, while accounting for the case-control sampling scheme. To increase statistical power, we additionally wish to (2) utilize the fact that the multiple phenotypes represent the same underlying trait, and (3) utilize SNP set association tests to borrow strength and information from SNPs within a set. There are no existing methods that simultaneously address these three challenges; this paper aims to fill this gap.

While several approaches have been proposed for modeling a single secondary phenotype (e.g., Lin and Zeng, 2009; Li et al., 2010; Hernán et al., 2004; Monsees et al., 2009), only a few authors have considered multiple secondary phenotypes. He et al. (2012) proposed a retrospective likelihood approach in which they use Gaussian copulas to model the dependence between disease status and the secondary phenotypes. Schifano et al. (2013) proposed a scaled linear regression method to study the effect of a single SNP on multiple continuous secondary phenotypes by accounting for case-control ascertainment through

Inverse Probability Weighting (IPW). Other methods for multivariate outcomes (e.g., Ferreira and Purcell, 2009; Zhou and Stephens, 2014) do not take the sampling mechanism into account. Numerous approaches have also been proposed for SNP *set* analysis of a single phenotype in the absence of ascertainment bias (e.g., Gauderman et al., 2007; Zhou et al., 2013; Han and Pan, 2010; Wu et al., 2010, 2011a).

We propose to first test the association of a SNP set with the secondary outcomes, and then select the most highly associated SNPs using penalized estimating equations. First, we define an IPW pseudolikelihood based on the scaled linear model proposed by Schifano et al. (2013) that assumes the SNP-specific effects are shared across all scaled outcomes. Then, we adapt popular statistical methods for this model, and study their properties. Specifically, following Wu et al. (2011a) we develop a variance component score test for the SNP set, and following Fu (2003); Johnson et al. (2008); Sofer et al. (2014), we develop penalized estimating equations based on the proposed IPW pseudolikelihood, together with tuning parameter selection procedures. A challenging aspect of studying properties of target parameters and equations under biased sampling is that the observations are not identically distributed. The novelty of our approach lies in developing an IPW pseudolikelihood based on the scaled linear model, and in formulating appropriate IPW pseudolikelihood-based testing and selection procedures.

The paper is organized as follows. In Section 2, we describe the lung cancer genetic epidemiological study of multiple secondary smoking phenotypes. We present the model in Section 3, and propose the weighted variance component pseudo-score test in Section 4. In Section 5, we provide the derived penalized estimating equations for variable selection and study their asymptotic properties. We evaluate the performance of the proposed methods using simulation studies in Section 6. In Section 7, we provide the data analysis results from >16K available genes from the motivating study. We conclude with a discussion.

2. Motivating Lung Cancer Case-Control Study

In a lung cancer case-control GWAS conducted at Massachusetts General Hospital (MGH) (Schifano et al., 2013, and references therein), four continuous measures of smoking behavior (age of smoking initiation, years of smoking duration, average number of cigarettes smoked daily, and years of smoking cessation) were collected for both lung cancer cases and controls. Demographic and smoking characteristics of the ever-smoker study population of self-reported Caucasians are provided in Table 1. Genotyping was performed using the Illumina Human610-Quad BeadChip. There were 543,697 SNPs remaining after the quality control process, mapping to 16,270 genes (Kent et al., 2002). A total of $n_0 = 716$ control and $n_1 = 673$ case ever-smoker subjects have genotypic, covariate and smoking information.

Preliminary data analysis indicates that three of the four secondary smoking behavior outcomes (all except years of smoking cessation) are highly associated with the primary lung cancer outcome ($p < 10^{-8}$). Since we do not know *a priori* which SNPs in the GWAS are associated with both the secondary smoking outcomes of interest and the primary lung cancer outcome, we are concerned about ascertainment bias in estimating and testing variants associated with the secondary smoking outcomes.

3. The Model

3.1 The Scaled Marginal Model

Suppose m positively correlated continuous outcomes $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})^T$ that measure the same underlying trait, a vector of SNPs $\boldsymbol{g}_i$ ($d \times 1$), and a vector of covariates $\boldsymbol{x}_i$ ($p \times 1$), are observed for the i^{th} of n subjects. For simplicity, we assume an additive genetic model with g_{ik} representing the number of copies (or dosages for imputed data) of the minor allele for the k^{th} SNP ($k = 1, \ldots, d$). Let

$$\frac{E(y_{ij}|\boldsymbol{x}_i,\boldsymbol{g}_i)}{\sigma_j}=\boldsymbol{x}_i^T\boldsymbol{\beta}_j+\boldsymbol{g}_i^T\boldsymbol{\alpha},\quad j=1,\ldots,m, \qquad (1)$$

where var $(y_{ij}|\boldsymbol{x}_i,\boldsymbol{g}_i)=\sigma_j^2$ is the outcome specific scale parameter, $\boldsymbol{\beta}_j$ are the covariate effects on scaled outcome j, and $\boldsymbol{\alpha}_k$ is the common effect of the kth SNP on all scaled outcomes $(y_{i1}, \cdots, y_{im})$. Note that these effects may vary across SNPs, but are assumed common across outcomes. This assumption is plausible, since the outcomes are scaled by $\boldsymbol{\sigma}_j$ and are assumed to measure the same underlying trait. The model proposed in Schifano et al. (2013) is a special case of (1) when $d = 1$.

3.2 Weighted pseudolikelihood approach

Pseudolikelihood (PL) provides an attractive and simple framework that does not require a full specification of the likelihood when it is too complex to specify (Gourieroux et al., 1984). The PL function specifies only the first and second moments of outcomes, and allows the second moment (covariance/correlation structure) to be misspecified. A pseudologlikelihood function for $\boldsymbol{\gamma}=(\boldsymbol{\beta}_1^T,\ldots,\boldsymbol{\beta}_m^T,\boldsymbol{\alpha}^T)$ corresponding to model (1), assuming a working correlation matrix $\mathbf{R}_m$ ($m \times m$) for multiple outcomes, is given by

$$p\ell(\boldsymbol{\gamma},\mathbf{R}_m,\boldsymbol{\Psi})=-\frac{1}{2}\sum_{i=1}^{n}(y_i^*-\boldsymbol{Z}_i\gamma)^T\mathbf{R}_m^{-1}(\boldsymbol{y}_i^*-\boldsymbol{Z}_i\gamma), \qquad (2)$$

where $\boldsymbol{\sigma}^2=(\sigma_1^2,\ldots,\sigma_m^2)$, $\boldsymbol{\Psi}$ = *diag*($\boldsymbol{\sigma}^2$), $\boldsymbol{y}_i^*=\boldsymbol{\Psi}^{-1/2}\boldsymbol{y}_i$, $\boldsymbol{Z}_i=(\mathbf{I}_m\otimes\boldsymbol{x}_i^T,\mathbf{1_m}\otimes\mathbf{g_i^T})$, $\mathbf{I}_m$ is an $m \times m$ identity matrix, and $\mathbf{1_m}$ is a $m \times 1$ vector of ones.

If participants were randomly sampled from the population, estimation of the SNP and covariate effects via maximization of (2) would lead to unbiased estimators of the population regression coefficients $\boldsymbol{\gamma}$ provided $E(\boldsymbol{y}_i^*|\boldsymbol{Z}_i)=\boldsymbol{Z}_i\gamma$ is correctly specified. However, under case-control sampling, such an estimator may be biased. We correct this selection bias using the inverse-probability weighted pseudologlikelihood:

$$wp\ell(\boldsymbol{\gamma},\mathbf{R}_m,\boldsymbol{\Psi})=-\frac{1}{2}\sum_{i=1}^{n}w_i(\boldsymbol{y}_i^*-\boldsymbol{Z}_i\gamma)^T\mathbf{R}_m^{-1}(\boldsymbol{y}_i^*-\boldsymbol{Z}_i\gamma). \qquad (3)$$

where $w_i = \pi/p_1$ if $D_i = 1$ and $w_i = (1 - \pi)/(1 - p_1)$ if $D_i = 0$, π is the population disease prevalence, D_i is a case/control indicator (1/0), and $p_1 = n_1/n$ is the proportion of cases in the case-control sample. The weight w_i is proportional to the inverse probability of subject i being sampled in the case-control study dataset (Schifano et al., 2013).

3.3 Estimating Equations

To derive the estimating function for the mean parameter $\boldsymbol{\gamma}$, we differentiate $wp\ell(\boldsymbol{\gamma}, \mathbf{R}_m, \Psi)$ in (3) with respect to $\boldsymbol{\gamma}$:

$$\mathbf{U}_\gamma = \sum_{i=1}^{n} w_i \boldsymbol{Z}_i^T \mathbf{R}_m^{-1} (\boldsymbol{y}_i^* - \boldsymbol{Z}_i \boldsymbol{\gamma}) = 0. \quad (4)$$

Equation (4) is unbiased even when $\mathbf{R}_m$ is misspecified. It is similar to a generalized estimating equation with an identity link, with two notable differences: the weighting, and the use of the scaled outcomes $\boldsymbol{y}^*$ rather than the actual outcomes $\boldsymbol{y}$, and as a result, only a correlation matrix is specified in the equation without scaling parameters. We estimate an unstructured correlation matrix $\mathbf{R}_m$ using the weighted method of moments by solving

$$\mathbf{U}_{R_m} = \sum_{i=1}^{n} w_i \left[(\boldsymbol{y}_i^* - \boldsymbol{Z}_i \boldsymbol{\gamma})(\boldsymbol{y}_i^* - \boldsymbol{Z}_i \boldsymbol{\gamma})^T - \mathbf{R}_m \right] = 0. \quad (5)$$

Other types of method of moments estimators can be used to estimate a structured working correlation matrix $\mathbf{R}_m$. In the simplest case of working independence, $\mathbf{R}_m$ is set to be $\mathbf{I}_m$. To estimate σ_j^2, $j = 1, \ldots, m$, we use the following estimating equations which are free of $\mathbf{R}_m$, following Roy et al. (2003):

$$\mathbf{U}_{\sigma^2} = \sum_{i=1}^{n} w_i \left\{ \frac{y_{ij}}{\sigma_j} \left(\frac{y_{ij}}{\sigma_j} - \boldsymbol{x}_i^T \boldsymbol{\beta}_j - \boldsymbol{g}_i^T \boldsymbol{\alpha} \right) - 1 \right\}. \quad (6)$$

The unpenalized estimators of $\boldsymbol{\gamma}$, $\boldsymbol{\sigma}^2$, and $\mathbf{R}_m$ are obtained by jointly solving (4), (5), and (6). The resulting estimators of $(\boldsymbol{\gamma}, \boldsymbol{\sigma}^2)$ are consistent even when $\mathbf{R}_m$ is misspecified.

4. Testing for SNP set Effect

Under model (1), testing for a SNP set effect on the outcomes corresponds to H_0: $\boldsymbol{\alpha} = 0$. We develop a variance component pseudo-score test under a mixed model formulation of this problem. Specifically, we assume the SNP effects α_k, $k = 1, \ldots, d$, arise from an arbitrary distribution with mean zero and variance τ. We also assume that the third and higher moments of the distribution of α_k, $k = 1 \ldots, d$, are of order $o(\tau)$. Under these assumptions, testing for SNP set effect reduces to testing the variance component with H_0: $\tau = 0$.

The weighted pseudo-score test statistic for H_0: $\tau = 0$ is given by

$$U_{\tau,0}(\hat{\boldsymbol{\beta}}_0)=\frac{1}{2}\left(\boldsymbol{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_0\right)^T \mathbf{W}\mathbf{R}_{mn}^{-1}\boldsymbol{G}\boldsymbol{G}^T\mathbf{R}_{mn}^{-1}\mathbf{W}\left(\boldsymbol{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_0\right),$$

where $\mathbf{X}=(\mathbf{I}_m \otimes \boldsymbol{x}_1^T, \ldots, \mathbf{I}_m \otimes \boldsymbol{x}_n^T)^T$, $\boldsymbol{G}=((\mathbf{1}_m \otimes \boldsymbol{g}_1^T)^T, \ldots, (\mathbf{1}_m \otimes \boldsymbol{g}_n^T)^T)^T$, $\mathbf{R}_{mn}^{-1}=\mathbf{I}_n \otimes \mathbf{R}_m^{-1}$, $\boldsymbol{y}^*=(\boldsymbol{y}_i^{*T}, \ldots, \boldsymbol{y}_n^{*T})$ are the vectorized data, and $\mathbf{W}$ = diag $(w_1\mathbf{1}_m, \ldots, w_n\mathbf{1}_m)$ is the $nm \times nm$ matrix with the weights for each subject (repeated m times for m outcomes) as the diagonal elements. The vector $\hat{\beta}_0$ consists of the estimated regression parameters under the null model, which is given in equation (1) but with only the covariates $\boldsymbol{x}_i$. Note that the outcome-specific variances in $\hat{\boldsymbol{\Psi}}$, suppressed in the notation as $\boldsymbol{y}_i^*=\hat{\boldsymbol{\Psi}}^{-1/2}\boldsymbol{y}_i$, are also estimated under the null model where $\boldsymbol{\alpha}$ = 0. In principle, different assumptions on the SNP effects could be incorporated into the matrix $\boldsymbol{G}\boldsymbol{G}^T$. For instance, specifying $\boldsymbol{G}=((\mathbf{I}_m \otimes \boldsymbol{g}_1^T)^T, \ldots, (\mathbf{I}_m \otimes \boldsymbol{g}_n^T)^T)^T$ allows different SNP effects across scaled outcomes. We emphasize that even if the common effect assumption does not hold, this test is still valid as it controls type I error, since under the null all SNP effects are identical to 0.

Under the null hypothesis, $U_{\tau,0}$ follows a mixture of independent $\chi^2_{(1)}$ distributions of the form $\sum_l \lambda_l \chi^2_{l,(1)}$, where λ_l are calculated from the data. In Supplementary Materials Sections S2.1–S2.3, we provide the full derivation of the variance component pseudo-score test for multiple (and a single) secondary outcomes and the details regarding the computation of the null distribution of the test statistic. The distribution of this mixture is calculated using Davies method (Davies, 1980), implemented in standard software.

5. SNP Selection via Penalized Estimation

To investigate which subset of SNPs are associated with the secondary phenotype(s), we provide a weighted penalized estimating equations approach for SNP selection.

5.1 The Penalized weighted pseudologlikelihood

Let $P_\lambda(\alpha_k) \geqslant 0$, $k = 1, \ldots, d$ denote a penalty function with a tuning parameter λ, where $P_\lambda(\cdot)$ is from a family of penalty functions that are singular at the origin and induce sparsity in the estimates; further required conditions on $P_\lambda(\cdot)$ will be specified later. Such penalties include the LASSO (Tibshirani, 1996), and "oracle"-type penalties such as SCAD (Fan and Li, 2001) and MCP (Zhang, 2010) penalties that have attractive asymptotic properties of consistency and uniform sparsity under weak conditions. The penalized negative weighted pseudologlikelihood, in which the penalty is applied on the sub-vector $\boldsymbol{\alpha}$ of $\boldsymbol{\gamma}$, can be written in the general form: $-wp\ell(\boldsymbol{\gamma}, \mathbf{R}_m, \boldsymbol{\Psi})+\sum_{k=1}^{d} P_\lambda(\alpha_k)$.

5.2 The Penalized Weighted Estimating Equations

To estimate $\boldsymbol{\gamma}$ with a sparse estimator of $\boldsymbol{\alpha}$, we incorporate the sparse penalty into the estimating function for $\boldsymbol{\gamma}$ following the approach of Johnson et al. (2008). Thus, the estimating equation for $\boldsymbol{\gamma}$ for a case-control study of n individuals is given by

$$\mathbf{U}^P_\gamma=\sum_{i=1}^{n} w_i \boldsymbol{Z}_i^T \mathbf{R}_m^{-1}(\boldsymbol{y}_i^* - \boldsymbol{Z}_i\boldsymbol{\gamma}) - n\boldsymbol{q}_\lambda(|\boldsymbol{\gamma}|)\circ \mathrm{sgn}(\boldsymbol{\gamma})=0 \qquad (7)$$

where $\boldsymbol{q}_\lambda(|\boldsymbol{\gamma}|)=(0_{mp}^T, q_\lambda(|\alpha_1|), \ldots, q_\lambda(|\alpha_d|))^T$, and $q_\lambda(\cdot)=P_\lambda'(\cdot)$ is the sub-gradient of sparsity-inducing penalty function $P_\lambda(\cdot)$, and ○ denotes a component-wise product. In our application, we penalize only the SNP coefficients $\boldsymbol{\alpha}$.

5.3 Asymptotic properties of the estimators

Under the regularity conditions provided in the Supplementary Materials (Section S3.1), the parameters estimated using the IPW pseudologlikelihood converge to the true parameters as the sample size n increases, even if the number of SNPs d diverges to infinity. For the following result, we assume that $d/n \to 0$, but the results could easily be extended to $\log(d)/n \to 0$ if the outcomes are normally distributed, in a similar manner to Sofer et al. (2014). Theorem 1 assumes an oracle penalty function, and that the weights w_i, $i = 1, \ldots, n$ are bounded, i.e., that the disease prevalence is not 0 or 1. Denote by $\| \cdot \|_F$ the Frobenius norm of a matrix. The proof of Theorem 1 is provided in the Supplementary Materials (Section S3.2).

Theorem 1—Let n, d → ∞ such that d/n → 0, and let m, p be fixed. Let A denote the set of indices of SNP coefficient vector $\boldsymbol{\alpha}$ that are not zero, i.e., $A = \{k: \alpha_k \neq 0\}$ and $|A| = s < \infty$. Let $Z_{i,A}$ be the i^th design matrix with the subset of columns corresponding to the true model, and similarly γ_A the subvector of γ corresponding to the true model. Let the regularity conditions in Supplementary Materials Section S3.1 hold. Let $\mathbf{R}_T$ be the true, unknown correlation matrix. The parameters $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\Psi}}, \hat{\mathbf{R}}_m)$ estimated by solving $\mathbf{U}^p_\gamma$, $\mathbf{U}_{\sigma^2}$, and $\mathbf{U}_{R_m}$ satisfy:

1. The estimator $\hat{\gamma}$ is

- **a.** Consistent for $\boldsymbol{\gamma}$, i.e. $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|=O_p(\sqrt{d/n})$
- **b.** Uniformly sparse, i.e. $P(\hat{\alpha}_k \neq 0 \text{ for all } k \in A) \to 1$
- **c.** Asymptotically normal in the true model: Let

$$\boldsymbol{\Upsilon}=\left(\sum_{i=1}^{n} w_i \boldsymbol{Z}_{i,A}^T \hat{\mathbf{R}}_m^{-1} \boldsymbol{Z}_{i,A}\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{w}_i \boldsymbol{Z}_{i,A}^T \hat{\mathbf{R}}_m^{-1} \mathbf{R}_T \hat{\mathbf{R}}_m^{-1} \boldsymbol{Z}_{i,A}\right) \left(\sum_{i=1}^{n} w_i \boldsymbol{Z}_{i,A}^T \hat{\mathbf{R}}_m^{-1} \boldsymbol{Z}_{i,A}\right)^{-1}.$$

then

$$\sqrt{n/d}\,\boldsymbol{\Upsilon}^{-1/2}(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_A) \xrightarrow{\mathscr{L}} \mathscr{N}(0, \mathbf{I}).$$

2. The estimator $\hat{\boldsymbol{\Psi}}$ is consistent $\|\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}\|_F=O_P(\sqrt{d/n})$

3. If an unstructured working correlation matrix is specified, then $\|\hat{\mathbf{R}}_m - \mathbf{R}_T\|_F = O_P(\sqrt{d/n})$ If $\|\hat{\mathbf{R}}_m - \mathbf{R}_T\|_F = O_P(\sqrt{d/n})$, then the covariance estimator of $\hat{\boldsymbol{\gamma}}$ is efficient with $\boldsymbol{\Upsilon} = \left(\sum_{i=1}^{n} w_i \boldsymbol{Z}_{i,A}^T \mathbf{R}_T^{-1} Z_{i,A}\right)^{-1}$ asymptotically.

Note that condition (c) applies when the working correlation matrix $\mathbf{R}_m$ is misspecified. Also note that for a set of non-zero estimated effects $\hat{A}$, we have that the sandwich covariance matrix is obtained by subsetting the matrices $\mathbf{Z}_i$ according to the set of non-zero SNP effects to obtain $\mathbf{Z}_{i,\hat{A}}$, $i = 1, \ldots, n$. Then the sandwich covariance matrix is the corresponding $\boldsymbol{\Upsilon}_{\hat{A}}$. Asymptotic normality is only guaranteed for the true, unknown, model, as effect estimates of truly zero effects are usually estimated as zero, by asymptotic sparsity.

5.4 Tuning Parameter Selection with wBIC

We propose a weighted Bayesian Information-like Criterion (wBIC) that inverse-probability weights each observation's contribution to the pseudologlikelihood when selecting the tuning parameter λ. Let $\hat{A}$ denote a set of indices indicating the non-zero coefficients of $\hat{\boldsymbol{\gamma}}$. Let $\hat{\boldsymbol{\gamma}}_A$ denote the $|\hat{A}|$-dimensional vector of nonzero regression coefficients, and let $\mathbf{Z}_{\hat{A}}$ and $\mathbf{Z}_{i,\hat{A}}$ be the corresponding design matrices for all observations and a single observation, respectively. Let $\hat{\boldsymbol{\gamma}}_{\hat{A}}$, $\hat{\mathbf{R}}_{m,\hat{A}}$ and $\hat{\boldsymbol{\Psi}}_{\hat{A}}$ be the estimates associated with the model indexed by the coefficients in $\hat{A}$. For the model indexed by the coefficients in $\hat{A}$, define the wBIC as

$$\begin{aligned} \mathrm{wBIC}_{\hat{A}} &= -2wp\ell(\hat{\boldsymbol{\gamma}}_A, \hat{\mathbf{R}}_{m,\hat{A}}, \hat{\boldsymbol{\Psi}}_{\hat{A}}) + d^*_{\hat{A}} \log n \\ &\propto n^{-1}\sum_{i=1}^{n} w_i(\boldsymbol{y}_i^* - \boldsymbol{Z}_{i,\hat{A}}\hat{\boldsymbol{\gamma}}_{\hat{A}})'\hat{\mathbf{R}}_{m,\hat{A}}^{-1}(\boldsymbol{y}_i^* - \boldsymbol{Z}_{i,\hat{A}}\hat{\boldsymbol{\gamma}}_{\hat{A}}) + n^{-1}d^*_{\hat{A}}\log n \end{aligned} \quad (8)$$

with $d^*_{\hat{A}} = tr(\mathbf{H}_{\hat{A}}^{-1}\mathbf{V}_{\hat{A}})$ as a measure of model complexity (e.g., Gao and Song, 2010) where

$$\mathbf{H}_{\hat{A}} = E\left\{-\frac{\partial^2 \log PL}{\partial\boldsymbol{\gamma}_{\hat{A}}\partial\boldsymbol{\gamma}'_{\hat{A}}}\right\} \text{ and } \mathbf{V}_{\hat{A}} = \mathrm{var}\left\{\frac{\partial \log PL}{\partial\boldsymbol{\gamma}_{\hat{A}}}\right\}.$$

In practice, we use the estimate $\hat{d}^*_{\hat{A}} = tr(\hat{\mathbf{H}}_{\hat{A}}^{-1}\hat{\mathbf{V}}_{\hat{A}})$ where

$$\hat{\mathbf{H}}_{\hat{A}} = \sum_{i=1}^{n} w_i \boldsymbol{Z}_{i,\hat{A}}^T \hat{\mathbf{R}}_{m,\hat{A}}^{-1} \boldsymbol{Z}_{i,\hat{A}} \text{ and} \quad (9)$$

$$\hat{\mathbf{V}}_{\hat{A}} = \sum_{t=1}^{n}\left\{w_i \boldsymbol{Z}_{i,\hat{A}}^T \hat{\mathbf{R}}_{m,\hat{A}}^{-1}(\boldsymbol{y}_i^* - \boldsymbol{Z}_{i,\hat{A}}\hat{\boldsymbol{\gamma}}_{\hat{A}})\right\}\left\{w_i \boldsymbol{Z}_{i,\hat{A}}^T \hat{\mathbf{R}}_{m,\hat{A}}^{-1}(y_i^* - \boldsymbol{Z}_{i,\hat{A}}\hat{\boldsymbol{\gamma}}_{\hat{A}})\right\}^T. \quad (10)$$

5.5 Computation of the penalized estimators

5.5.1 Computation of estimators for a fixed tuning parameter—We use an MM algorithm following Johnson et al. (2008). We begin by fitting the unpenalized model to estimate $\hat{\gamma}^{(0)}$, $\boldsymbol{\sigma}^{2(0)}$ and $\mathbf{R}_m^{(0)}$. Note that if $mp + d > n$, then other starting values are recommended, as discussed in the Supplementary Materials, Section S3.3.4. We then fix $\boldsymbol{\sigma}^{2(0)}$ and $\mathbf{R}_m^{(0)}$ for the penalized estimation of $\boldsymbol{\gamma}$, where the latter is done by using a local quadratic approximation (Fan and Li, 2001), so that

$$\hat{\boldsymbol{\gamma}}^{(k+1)}=\hat{\boldsymbol{\gamma}}^{(k)}+\left\{\mathbf{A}(\hat{\boldsymbol{\gamma}}^{(k)})+n\sum\nolimits_{\lambda}(\hat{\boldsymbol{\gamma}}^{(k)})\right\}^{-1}\mathbf{U}_{\gamma}^{P}(\hat{\boldsymbol{\gamma}}^{(k)}), \quad (11)$$

where $\mathbf{A}(\hat{\gamma})=\sum\nolimits_{i=1}^{n} w_i \boldsymbol{Z}_i^T[\mathbf{R}_m^{(0)}]^{-1}\boldsymbol{Z}_i$, $\Sigma_\lambda(\boldsymbol{\gamma}) = diag\{0_{mp}, q\lambda(/\alpha_1/)/(\varepsilon + /\alpha_1/), \ldots, q_\lambda(/\alpha_d/)/(\varepsilon + /\alpha_d/)\}$, and $\mathbf{R}_m=\mathbf{R}_m^{(0)}$ and $\boldsymbol{y}^* = \boldsymbol{\Psi}^{-1/2,(0)}\boldsymbol{y}$ in $\mathbf{U}_\gamma^P$. The mp zeros along the $\Sigma_\lambda(\boldsymbol{\gamma})$ diagonal correspond to the unpenalized covariate effects. The updates are repeated until convergence.

5.5.2 Using the wBIC for tuning parameter selection—For small scaled coefficients common in GWAS, we recommend comparing wBIC on "re-evaluated data" for different choices of λ. By this, we mean (a.) set a λ sequence; (b.) for each λ in the sequence, perform the weighted penalized estimation described in Section 5.5.1 to identify the set A of nonzero regression coefficients in $\boldsymbol{\gamma}$; and (c.) fit weighted, but *unpenalized* models containing only those variables in A from (b.), and select appropriate λ by minimizing the criteria using this "re-evaluated" data. This procedure had better finite-sample performance than using the penalized version of the effect estimates in calculating wBIC, likely due to the bias caused by penalization, which could be high when both effect and sample sizes are small.

6. Simulation Studies

We simulated data in which the SNP(s) affect both the primary disease outcome (D), and the secondary outcomes of interest, which in turn also affect D. Various scenarios were investigated that are reported here and in the Supplementary Materials. The Supplementary Materials (Sections S2.4 and S3.3) include results from additional simulations to study the power gain resulting from the assumption of SNP-specific common effects across outcomes, the effect of different working correlation structures, the effect of using incorrect weights, the loss in power caused by using outcomes that are not associated with the SNP set, and the effect of a large number of (null) SNPs relative to the sample size n. In Sections S2.4.7 and S3.3.3 we also demonstrate scenarios which induce a selection bias towards the null, thus reducing the power of the unweighted procedures.

6.1 Data generation

We generated genotypes of a general population of 100,000 people using Hapgen2 based on the CEU population for 87 SNPs near or within gene *CDH18*, and then sample n_1 cases and n_0 controls from the population. We selected "causal" SNPs, and specified their effect sizes

$\boldsymbol{\alpha}$ on the scaled secondary outcomes $\left(y^*_{ij}, i=1,\ldots,n, j=1,\ldots,m=4\right)$ under the common effect assumption. Thus, we had $y^*_{ij}=\boldsymbol{\alpha}^T\boldsymbol{g}^c_i+\varepsilon_{ij}$ and $\varepsilon_i\sim\mathcal{N}\left(\mathbf{0},\mathbf{R}\right)$, where $\boldsymbol{g}^c_i$ is the vector (possibly of length 1) of causal genotypes of individual i, **0** is the vector with m zero entries, and **R** is the correlation matrix between m scaled secondary outcomes. The correlation matrix **R** matched the observed (unstructured) correlations in the lung cancer data set. The secondary outcomes were then scaled so that $\boldsymbol{y}$ = diag(ψ)$\boldsymbol{y}$*, where ψ is the vector of outcome standard deviations estimated from the data.

The disease probability for each individual was generated from the logistic model $\text{logit}\left[p\left(D_i=1\right)\right]=\beta_0+\boldsymbol{\beta}^T_{\tilde{g}}\tilde{\boldsymbol{g}}^c_i+\boldsymbol{\beta}^T_y\boldsymbol{y}$. Note that either $\tilde{\boldsymbol{g}}^c=\boldsymbol{g}^c$or $\boldsymbol{g}^c\subset\tilde{\boldsymbol{g}}^c$, depending on the simulation setting. Disease status for each individual was sampled from a binary random variable using the calculated disease probabilities. The parameter β_0 was set so that the desired population prevalence was achieved (rare: $\boldsymbol{\pi}\approx 0.01$ or common: $\boldsymbol{\pi}\approx 0.08$). The lung cancer data most closely resemble the rare disease setting, but for more general applicability, we investigate both the rare and common disease settings.

6.2 Testing

We compared the type I error and power of several tests under the rare and common disease settings. The simulations had n_0 = n_1 = 500 cases and controls. We selected a single causal SNP, rs6869352 (minor allele frequency (MAF) = 0.37 in the CEU population; $|r^2| > 0.5$ with 9 other SNPs), from the gene *CDH18* affecting both the primary and secondary outcomes (i.e., $\tilde{g}^c=g^c$, both of length 1). Following Monsees et al. (2009) and consistent with the effect sizes observed in the lung cancer data, we set the primary disease model parameters to be β_g= log(1.7)/2, log(1.7), and $\beta_{yj}\in$ {log(2)/2, log(2)}, j = 1,…,m. The compared tests for type I error estimation included the 'unweighted' pseudo-score test that assumed the weights for all cases and controls to be 1 (i.e., ignored disease status), the pseudo-score test using controls only ('controls'), the proposed IPW pseudo-score test ('IPW'), as well as MANOVA, and the method of Conneely and Boehnke (2007) 'CB' that regresses each outcome against each variant and reports the minimum p-value adjusted for multiplicity of both traits and SNPs. Both MANOVA and CB are unweighted. For power estimation, we also compare the 'minP-oracle' method, which first tests all SNPs in the set individually using SMAT (Schifano et al., 2013), and then applies a Bonferroni correction to the smallest p-value based on the (oracle) effective number of tests n_e, where n_e was the number such that under the null, min{p_k}/n_e = 0.001, k = 1,…, 87, (since the p-value threshold for power was 0.001). Note that this test cannot be implemented in practice.

Table 2 provides the estimated type I error for p-values thresholds 0.01 and 0.001 over 10^5 simulations under the null hypothesis of no SNP set effect on the multiple secondary outcomes, and Figure 1 provides power curves of the tests that controlled the type I error for β_g = log(1.7) when varying the SNP effect on the scaled outcomes α, over 5,000 simulations; power curves for all investigated tests under additional settings are provided in Supplementary Materials Section S2.4.3. As expected, 'IPW' and 'controls' always have estimated type I error rates close the nominal level, with 'IPW' being somewhat protective. When a disease is rare, they have almost identical power curves, but when the disease is more prevalent 'IPW' becomes more powerful, since it takes advantage of the information from the cases that now have a higher representation in the population. Here, the minP-oracle test is always less powerful than IPW. However, this may not be always true, e.g., in scenarios with less LD between variants and a stronger effect of the causal SNP, we expect the minP-oracle to be more powerful.

In general, all unweighted methods have higher bias when the disease is rare (i.e., when the case-control sample is less representative of the population) as compared to common. Bias for the unweighted methods also increases when β_g and β_y increase. In the specific settings of common disease and β_g = log(1.7),β_{yj}= log(2)/2, MANOVA controlled the type I error. However, it also had a very low power. In the Supplementary Materials (Section S2.4.2), we also provide the estimates of the size of the 'IPW' method for lower type I error levels.

6.3 Variable selection

Data were simulated in the same manner as in the testing simulations except that we removed the SNPs with absolute pairwise correlations greater than 0.75 in order to more reliably evaluate the performance of the proposed variable selection procedure. Our method allows for SNPs that are in LD and hence stringent LD pruning is not needed.

Of the remaining 39 SNPs, we set three SNPs as causal for the secondary outcomes: rs4242066 (G1), rs17222312 (G2), and rs12655266 (G3) (MAF between 0.08–0.12), i.e., $\boldsymbol{g}^c$ = {G1, G2, G3}. These three SNPs were weakly correlated with each other, with absolute correlation ranging between 0.051–0.115. In both Scenarios 1 and 2 depicted in Figure 2, G1–G3 are associated with the secondary outcomes Y1–Y4, as well as D, and D affects the Sampling (S) in the case-control dataset. G1–G3 are moderately correlated with the remaining SNPs in the set, G4–G39 (absolute correlation between G1–G3 and G4–G39: 0.040 – 0.554). In Scenario 2, G4–G6 (rs347743, rs4866042, rs6869352) are associated with D, but not Y1–Y4 conditional on G1–G3; thus, $g^c \neq \tilde{g}^c = \{\mathrm{G1},\ \mathrm{G2},\ \mathrm{G3},\ \mathrm{G4},\ \mathrm{G5},\ \mathrm{G6}\}$. In both scenarios we set $\alpha_k = -c\log_{10}(MAF_k)$ where c = 0.25 for each causal SNP k, k = 1,…, 3, leading to common effect sizes between 0.23–0.28. Following Monsees et al. (2009), the disease model parameters were set to $\boldsymbol{\beta}_g$ = (log(1.7),…, log(1.7))^T and $\boldsymbol{\beta}_y$ = (log(2),…, log(2)). Each simulated dataset has $n_0 = n_1$ = 1000. Note that for variable selection, we wish to identify SNPs G1–G3 as nonzero; nonzero effects for SNPs G4–G39 would represent false positives in the presence of causal SNPs G1–G3.

Across 500 simulations, we compare the performance of the models selected based on wBIC with two other criteria: an unweighted BIC (uwBIC) and a control-only BIC (cBIC). For uwBIC, the inverse probability weights are not used. For cBIC, we analyze controls-only data. Precise definitions of these criteria are provided in the Supplementary Materials (Section S3.3). Otherwise, we always used the procedure described in Section 5.5.2 with the MCP penalty with fixed parameter a = 3.7. The λ sequence contained 100 values between 10^{-5} and 15, equally spaced on the log scale. We used estimates of σ_j^2 and $\mathbf{R}_m$ from the full unpenalized model containing all covariates.

We evaluated the performance of the proposed variable selection procedure based on the following criteria: #Nonzero: the number of α_k estimated as non-zero (averaged over simulations; true=3); T: Proportion of simulations selecting the true model; FP: average number of zero α_k's that are incorrectly estimated as non-zero (false positives) across simulations (max=36); FN: average number of non-zero α_k's that are incorrectly estimated as zero (false negatives) across simulations (max=3); and Prediction Error $(\text{PE}){:}\frac{1}{n_{\text{new}}}(\boldsymbol{y}^*_{\text{new}} - \boldsymbol{Z}_{\text{new}}\hat{\boldsymbol{\gamma}})^T(\boldsymbol{y}^*_{\text{new}} - \boldsymbol{Z}_{\text{new}}\hat{\boldsymbol{\gamma}})$, where $\mathbf{Z}_{new}$, $\boldsymbol{y}_{new}$ designate a dataset with $n_{new} = 500$ observations generated under the same model as the training data used to estimate the model parameters, $\hat{\gamma}$ and $\hat{\boldsymbol{\Psi}}=\text{diag}(\widehat{\boldsymbol{\sigma}^2})$ are estimates of the parameters from the training data, and $\boldsymbol{y}^*_{\text{new}}=\text{blockdiag}(\hat{\boldsymbol{\Psi}})^{-1/2}\boldsymbol{y}_{\text{new}}$.

The variable selection results are provided in Table 3. Note first that, particularly under a low disease prevalence, the wBIC and cBIC perform similarly, but the wBIC performs better because it incorporates information from both the controls and the cases (appropriately downweighted), and becomes more beneficial as the disease prevalence increases. Consider now uwBIC. Under Scenario 1, it outperforms wBIC and cBIC in terms of T, FP, and FN, but wBIC performs best in terms of PE. However, the effect estimates are biased with uwBIC when we do not account for case-control ascertainment (see Supplementary Materials Section S3.3.1, Table S3, and related discussion). Under Scenario 2, uwBIC selection performs worse than wBIC and cBIC, particularly using the measures T and FP, and more so when the disease is rare. This is because G4–G6 are often selected by uwBIC due to the uncontrolled selection bias, and are then false positives. SNP G5 was the most common offender, likely due to its somewhat high absolute correlation of 0.55 with G1. Note that uwBIC can also cause an increase in FN in settings where the selection bias results in weakened estimated effects of the causal SNPs; see Supplementary Materials Section S3.3.3.

In Table S3 in the Supplemental Materials, we also report the estimation performance in terms of bias, standard deviation (SD), and root-mean-squared error (RMSE) for $\alpha_1 - \alpha_3$ corresponding to G1–G3. These metrics are reported as averages across all simulations, and as averages across the simulations in which the corresponding effect was estimated as non-zero. In all settings, the magnitude of the bias is largest for the uwBIC approach, consistent with expectations. The SD for estimates of wBIC are often larger than those of uwBIC. When computed conditionally on the SNP being selected, the RMSEs of the wBIC estimates are all smaller than their uwBIC counterparts except for G2 under common disease.

7. Analysis of the Lung Cancer Data

We performed a genome-wide SNP set analysis of the 16,270 genes in the MGH smoking behavior data set described in Section 2 using the weighted pseudo-score test, assuming a disease prevalence of 7.35×10^{-3}. The (secondary) outcomes were $\left(-\sqrt{\text{DURATION}}, \sqrt{\text{INITIATION}}, -\sqrt{\text{CPD}}, \sqrt{\text{CESSATION}}\right)$. We adjusted for age, gender, college education, and 4 principal components to correct for population substructure. The estimated effects (in absolute value) of the outcomes on the primary lung cancer outcome were (0.72, 0.66, 0.39, and 0.05), some of which are larger than the effects investigated in the simulations ($\beta_{yj} \in \{\log(2), \log(2)/2\}$). We first fit a “null model” (without genotypes); this step took less than 2 minutes. We then scanned the genome and tested the effects of each of ~16K available genes. This step took a little under 7 hours when using a single processor of an Intel(R) Xeon(R) CPU E5-2620 @ 2.00GHz compute node.

We considered only the 11,890 genes with at least 3 variants. Therefore, the type I error threshold for declaring the genome-wide significance of the associations between genes and smoking behavior is $0.05/11{,}890 = 4.2 \times 10^{-6}$. None of the genes passed the significance threshold, but we present the top associated genes (as determined by their *p*-values) in Table 4. For each gene, we report the number of SNPs it contained, its *p*-value, and a comparison to SMAT in terms of computing time and in terms of *p*-value. For SMAT we report the minimum *p*-value in the SNP set, both with and without a multiple testing adjustment for the number of SNPs.

One can see that the SNP set test is faster by an order of magnitude for various sizes of SNP sets, and that the *p*-value obtained using the SNP set test for these genes is often comparable to the unadjusted SMAT *p*-value, and usually smaller than the adjusted SMAT p-value. The *CDH18* gene reported in Schifano et al. (2013) (but not a top gene in the present analysis) had *p*-value 0.005 when using the proposed SNP set test, while its minimum SMAT p-value was 9.5×10^{-8} (1.4×10^{-5} after applying Bonferroni adjustment testing 150 variants) suggesting that in the presence of an extremely strong effect of a few SNPs (stronger than in the simulated data) that are in low LD with the remaining SNPs, the variance component test may be less powerful than a single-SNP based test.

Next, we performed variable selection for the SNPs in the top associated genes using the MCP penalty with the wBIC. The complete list of nonzero SNPs from the top genes, the SNPs in LD with these SNPs with $|r| > 0.75$ (Yi et al., 2015), and also timing results, are provided in Supplementary Materials Section S3.4.

8. Discussion

We presented the use of pseudolikelihood methods to test for the effect of a SNP set on multiple secondary traits and to perform variable selection in case-control genetic epidemiological studies. Specifically, we proposed a variance component test for the SNP set effect on multiple secondary outcomes, and a penalized estimation procedure to simultaneously estimate the SNP and covariate effects while performing variable selection, both based on the IPW pseudolikelihood. We further proposed a weighted BIC to select the tuning parameter required for model selection. We provide theoretical justifications for the proposed methods. Our simulation study shows that the proposed methods perform well.

We emphasize that our approach does not provide *p*-values for individual SNPs; we only have a *p*-value for the SNP set test. However, the test controls type I error, and therefore, the probability of the event that at least one variant in the associated set is being selected is smaller than the probability of rejecting the null of no association between the SNP set and the outcome, which is bounded by this type I error when the null hypothesis is true. It is a topic of future work to assign a *p*-value for a specific SNP post testing and selection.

Our proposed method assumes the secondary outcomes are positively correlated after a proper transformation and measure the same underlying trait. An alternative approach is to develop a latent variable model for multiple secondary outcomes and regress the latent variable on a SNP-set. Latent variable IPW SNP procedures can be developed and possibly used in conjunction with the pseudo-score test for a single secondary outcome (see Supplementary Materials Section S2.3). Future research along these lines is needed.

In some of our simulations settings (e.g., common disease, small secondary outcome and genotype effects on disease), the naive unweighted test protected the type I error, and was also more powerful than the IPW test. However, as shown in Tchetgen Tchetgen (2014), the association between a genotype and a single secondary outcome is equal to their association in the general population, plus a term that can be decomposed into two variationally independent functions: a function measuring the disease dependence on the genotype, and a so-called selection bias function, describing the difference between the secondary outcome mean in the cases and controls. Both functions depend on the genotype, and therefore, a data-driven method to estimate such bias is not feasible. It is also less clear how to estimate this bias when multiple secondary outcomes are used. Moreover, in some settings the bias terms can also “negate” the effect of the SNP on the disease, thereby reducing the power of the unweighted estimator and also degrading model selection performance (see Supplementary Materials, Sections S2.4.7 and S3.3.3). Therefore, we recommend using the weighted procedures.

In the simulations, we compared our proposed test to the “minP-oracle” that tests each variant in the set using SMAT, and applies a Bonferroni correction using the effective number of tests, as determined by the simulations. Unfortunately, for real data, many existing methods that correct for the effective number of tests are based on the LD structure of the SNP set, assuming that the test statistics share the same correlation structure. This assumption does not hold for IPW tests.

Motivated by the lung cancer GWAS data set, we focused on common variants. However, as with other variance component tests, our proposed test can be useful for testing rare variants as well. Extending the proposed approach for rare variants, and incorporating the burden test via a similar approach as with SKAT-O (Lee et al., 2012) is for future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health [R37 CA076404 to TS, ES007142 to EDS, CA074386, CA092824, and P50CA090578 to DC, R37 CA076404 and P01 CA134294 to XL] and TS was also supported by NHLBI contract HHSN268201300005C to the HCHS/SOL GAC.

References

Conneely KN, Boehnke M. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. The American Journal of Human Genetics. 1980; 81:1158–1168.

Davies R. Algorithm as 155: The distribution of a linear combination of chi-2 random variables. Journal of the Royal Statistical Society C. 1980; 29:323–333.

Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Journal of the American Statistical Association. 2001; 96:1348–1360.

Ferreira MAR, Purcell SM. A multivariate test of association. Bioinformatics. 2009; 25:132–133. [PubMed: 19019849]

Fu WJ. Penalized estimating equations. Biometrics. 2003; 59:126–132. [PubMed: 12762449]

Gao X, Song PK. Composite likelihood bayesian information criteria for model selection in high-dimensional data. Journal of the American Statistical Association. 2010; 105:1531–1540.

Gauderman W, Murcray C, Gilliland F, Conti D. Testing association between disease and multiple SNPs in a candidate gene. Genetic epidemiology. 2007; 31:383–95. [PubMed: 17410554]

Gourieroux C, Monfort A, Trognon A. Pseudo maximum likelihood methods: Theory. Econometrica. 1984; 19:681–700.

Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. Human heredity. 2010; 70:42–54. [PubMed: 20413981]

He J, Li H, Edmondson AC, Rader DJ, Li M. A gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies. Biostatistics. 2012; 13:497–508. [PubMed: 21933777]

Hernán, Ma, Hernández-Díaz, S., Robins, JM. A Structural Approach to Selection Bias. Epidemiology. 2004; 15:615–625. [PubMed: 15308962]

Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature Genetics. 2007; 39:870–874. [PubMed: 17529973]

Johnson BA, Lin D, Zeng D. Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models. Journal of the American Statistical Association. 2008; 103:672–680. [PubMed: 20376193]

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC http://genome.ucsc.edu/. Genome Research. 2002; 12:996–1006. [PubMed: 12045153]

Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012; 13:762–75. [PubMed: 22699862]

Li H, Gail AMH, Berndt S, Chatterjee N. Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. Genetic Epidemiology. 2010; 34:427–433. [PubMed: 20583284]

Lin DY, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. Genetic Epidemiology. 2009; 33:356–365.

Lin DY, Zeng D, Tang ZZ. Quantitative trait analysis in sequencing studies under trait-dependent sampling. Proceedings of the National Academy of Sciences. 2013; 110:12247–12252.

Monsees GM, Tamimi RM, Kraft P. Genome-wide association scans for secondary traits using case-control samples. Genetic Epidemioligy. 2009; 33:717–728.

Roy J, Lin X, Ryan LM. Scaled marginal models for multiple continuous outcomes. Biostatistics. 2003; 4:371–383. [PubMed: 12925505]

Schifano E, Li L, Christiani D, Lin X. Genome-wide Association Analysis for Multiple Continuous Phenotypes. American Journal of Human Genetics. 2013; 92:744–759. [PubMed: 23643383]

Sofer T, Dicker L, Lin X. Variable selection for high-dimensional multivariate outcomes. Statistica Sinica. 2014; 24:1633–1654. [PubMed: 28642637]

Tchetgen Tchetgen EJ. A general regression framework for a secondary outcome in case-control studies. Biostatistics. 2014; 15:117–128. [PubMed: 24152770]

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society (Series B). 1996; 58:267–288.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-set analysis for case-control genome-wide association studies. American Journal of Human Genetics. 2010; 86:929–42. [PubMed: 20560208]

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data using the sequence kernel association test. American Journal of Human Genetics. 2011a; 89:82–93. [PubMed: 21737059]

Yi H, Breheny P, Imam N, Liu Y, Hoeschele I. Penalized Multimarker vs. Single-Marker Regression Methods for Genome-Wide Association Studies of Quantitative Traits. Genetics. 2015; 199:205–22. [PubMed: 25354699]

Zhang CH. Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics. 2010; 38:894–942.

Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genetics. 2013; 9:e1003264. [PubMed: 23408905]

Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature Methods. 2014; 11:407–409. [PubMed: 24531419]

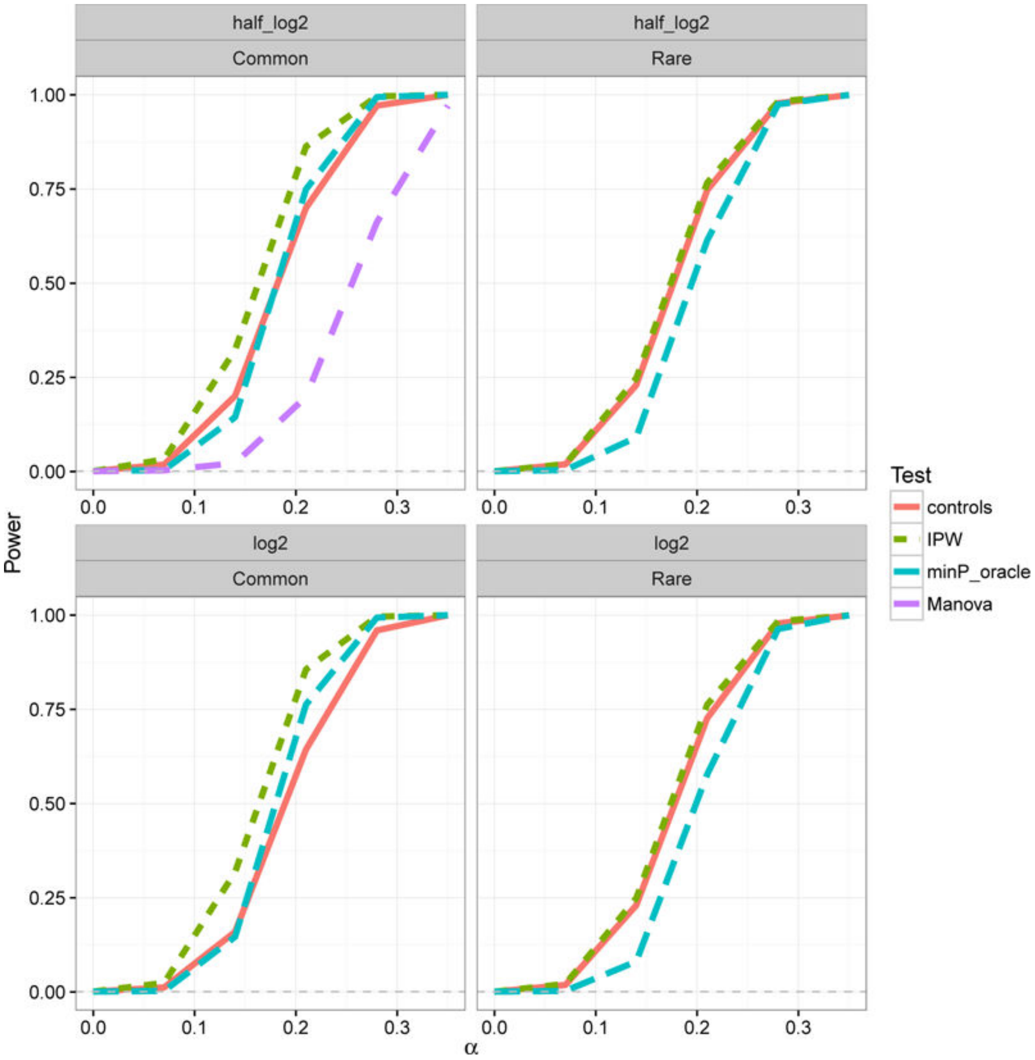

Figure 1.

Power curves of the SNP set tests, for all tests that controlled the type I error: the IPW and controls-only pseudoscore tests, “minP oracle” test that performs association testing for each SNP separately, and takes the minimum p-value with Bonferroni adjustment for the (oracle) effective number of tests, and MANOVA. The figure compares two disease prevalence settings: common (8%) and rare (1%). There is a single causal SNP G, affecting both the disease (D) and the secondary outcomes (Y). The effect of G on D is $\beta_g = \log(1.7)$, and the effect of Y on D (β_{yj}) is either log(2) or log(2)/2. Power was calculated as the proportion of simulations with p-value< 0.001.

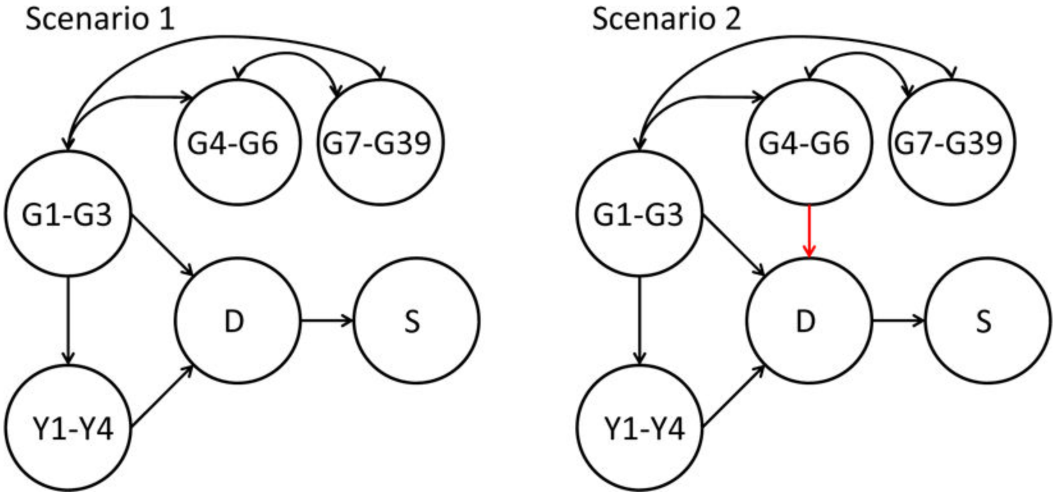

Figure 2.

Variable Selection Simulation Scenarios. In both Scenarios 1 and 2, SNPs G1–G3 are associated with the secondary outcomes Y1–Y4, as well as Disease Status (D), and Disease Status affects the Sampling (S) in the case-control dataset. SNPs G1–G3 are moderately correlated with the remaining SNPs in the set, G6–G39. In Scenario 2, SNPs G4–G6 are associated with Disease Status (D), but not the secondary outcomes Y1–Y4 conditional on G1–G3.

Table 1

Demographic Characteristics of MGH Caucasian, ever-smoker study participants, according to current smoking status. Entries are Mean (S.D.) for continuous variables and Count (%) for binary.

	Control		**Case**	
	Former (N=492)	**Current (N=224)**	**Former (N=393)**	**Current (N=280)**
Age	61.87 (10.51)	53.84 (11.48)	68.38 (9.08)	60.79 (10.17)
Gender(M)	256 (52.03%)	81 (36.16%)	220 (55.98%)	144 (51.43%)
College Grad (Y)	171 (34.76%)	46 (20.54%)	133 (33.84%)	62 (22.14%)
Age of Smoking Initiation	16.97 (3.60)	16.83 (4.57)	17.28 (4.38)	16.48 (3.79)
Smoking Duration	26.40 (14.63)	35.64 (11.93)	39.42 (14.45)	43.58 (10.24)
Average CPD	21.26 (15.02)	20.55 (11.53)	28.97 (14.87)	27.56 (13.29)
Years of Smoking Cessation	20.91 (11.90)	0.05 (0.17)	17.25 (11.98)	0.14 (0.22)

Table 2

Type I error comparisons, reported as ratios between the estimated and desired levels, of the various compared tests based on 10^5 simulations under the null of no SNP set effect on the secondary outcomes: ‘unweighted,’-the pseudo-score test assuming weights equal to 1 for all observations, which is equivalent to a test that ignores disease status; ‘controls only’-the pseudo-score test using only control participants; ‘IPW’-the proposed inverse probability weighted pseudo-score test, MANOVA, and ‘CB’-the method of Conneely and Boehnke (2007) for adjusting the minimum p-value for multiple testing. These simulations set the effect of a single causal SNP G on the disease status (D) as $\beta_g \in \{log(1.7)/2, log(1.7)\}$, and, the effect of the secondary outcomes (Y) as $\beta_g \in \{log(2)/2, log(2)\}$, $j = 1,\ldots, 4$. Scenarios with inflated type I error are marked, with an asterisk. We determined that a test has inflated type I error if the 95% confidence interval around an estimated type I error rate does not include the desired value (which is the p-value threshold used). The confidence intervals used the null standard error, so in practice tests are inflated under the 0.01 p-value threshold if the ratio between the estimated and desired type I error is larger than 1.06, and under the 0.001 p-value threshold if this ratio is larger than 1.19.

***p*-value threshold**	β_y	β_g	**unweighted**	**controls only**	**IPW**	**MANOVA**	**CB**
		Rare disease (1% prevalence)					
0.01	log(2)/2	log(1.7)/2	1.15*	0.95	0.95	0.95	1.02
0.001	log(2)/2	log(1.7)/2	1.24*	0.90	0.89	1.05	1.03
0.01	log(2)	log(1.7)/2	1.60*	0.96	0.94	1.05	1.31*
0.001	log(2)	log(1.7)/2	2.02*	0.91	0.90	1.15	1.43*
0.01	log(2)/2	log(1.7)	1.71*	0.96	0.95	1.07*	1.26*
0.001	log(2)/2	log(1.7)	2.23*	0.91	0.90	1.10	1.35*
0.01	log(2)	log(1.7)	4.06*	0.96	0.95	1.48*	3.14*
0.001	log(2)	log(1.7)	7.23*	0.92	0.89	1.70*	5.79*
		Common disease (8% prevalence)					
0.01	log(2)/2	log(1.7)/2	1.06*	0.96	0.94	1.05	0.98
0.001	log(2)/2	log(1.7)/2	1.09	0.92	0.88	0.80	0.90
0.01	log(2)	log(1.7)/2	1.23*	0.99	0.90	1.07*	1.09*
0.001	log(2)	log(1.7)/2	1.34*	0.95	0.83	1.45	1.10
0.01	log(2)/2	log(1.7)	1.31*	0.99	0.94	1.02	1.08*
0.001	log(2)/2	log(1.7)	1.52*	0.95	0.87	1.00	1.23*
0.01	log(2)	log(1.7)	2.05*	1.05	0.89	1.24*	1.45*
0.001	log(2)	log(1.7)	2.86*	1.10	0.84	1.05	1.77*

Table 3

Variable selection performance using the penalized pseudolikelihood method. The tuning parameter was chosen using: wBIC-weighted BIC; uwBIC-unweighted BIC; cBIC-BIC using controls only.

	$\pi = 0.01$					**$\pi = 0.08$**				
	# Nonzero	**T**	**FP**	**FN**	**PE**	**# Nonzero**	**T**	**FP**	**FN**	**PE**
Scenario 1										
uwBIC	3.13	0.84	0.164	0.030	4.082	3.15	0.84	0.168	0.020	4.229
wBIC	2.93	0.63	0.196	0.270	4.062	3.02	0.68	0.206	0.188	4.055
cBIC	3.02	0.57	0.292	0.276	4.076	3.03	0.55	0.324	0.292	4.096
Scenario 2										
uwBIC	3.70	0.40	0.738	0.042	4.116	3.38	0.66	0.398	0.016	4.379
wBIC	3.00	0.68	0.200	0.202	4.031	3.01	0.65	0.218	0.206	4.168
cBIC	3.06	0.60	0.302	0.240	4.047	3.02	0.52	0.336	0.318	4.197

Table 4

Comparison between the computing times and p-values obtained, by the variance component pseudoscore test and SMAT (obtained on each SNP in the SNP set), for the top genes identified in the gene-based GWAS. Computation time is given in seconds; for SMAT, it is the time of estimating the effect of all SNPs in the set. For SMAT, we also report the minimum p-value in the SNP set obtained by SMAT, with and without a Bonferroni multiple testing adjustment for the number of SNPs in the set.

Gene Symbol	**Chr**	**# SNPs**	**time varComp**	**time SMAT**	***p* varComp**	**min-*p* SMAT (unadjusted)**	**min-*p* SMAT (adjusted)**
MVP	16	4	1.44	4.30	3.43E–05	1.99E–05	7.96E–05
CCDC73	11	23	1.67	26.80	9.23E–05	2.01E–05	4.62E–04
ARHGAP35	19	6	1.64	6.95	4.60E–04	6.21E–04	3.73E–03
NUP93	16	16	1.41	17.10	4.80E–04	6.01E–06	9.62E–05
SPRED1	15	13	1.42	14.52	5.94E–04	2.11E–04	2.74E–03
NBEA	13	105	2.74	124.61	6.08E–04	6.88E–05	7.22E–03