8-30-2010

On two-stage hypothesis testing procedures via asymptotically independent statistics

James Y. Dai
Fred Hutchinson Cancer Research Center, jdai@fhcrc.org

Charles Kooperberg
Fred Hutchinson Cancer Research Center

Michael LeBlanc
Fred Hutchinson Cancer Research Center

Ross L. Prentice
Fred Hutchinson Cancer Research Center

On two-stage hypothesis testing procedures via asymptotically independent statistics

James Y. Dai

1100 Fairview Ave N, M2-C200, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109
jdai@fhcrc.org

Charles Kooperberg

Fred Hutchinson Cancer Research Center, Seattle, WA, 98109

Michael LeBlanc

Fred Hutchinson Cancer Research Center, Seattle, WA, 98109

Ross L. Prentice

Fred Hutchinson Cancer Research Center, Seattle, WA, 98109

Summary. Kooperberg and LeBlanc (2008) proposed a two-stage testing procedure to screen for significant interactions in genome-wide association (GWA) studies by a soft threshold on marginal associations (MA), though its theoretical properties and generalization have not been elaborated. In this article, we discuss conditions that are required to achieve strong control of the Family-Wise Error Rate (FWER) by such procedures for low or high-dimensional hypothesis testing. We provide proof of asymptotic independence of marginal association statistics and interaction statistics in linear regression, logistic regression, and Cox proportional hazard models in a randomized clinical trial (RCT) with a rare event. In case-control studies nested within a RCT, a complementary criterion, namely deviation from baseline independence (DBI) in the case-control sample, is advocated as a screening tool for discovering significant interactions or main effects. Simulations and an application to a GWA study in Women's Health Initiative (WHI) are presented to show utilities of the proposed two-stage testing procedures in pharmacogenetic studies.

Keywords: interactions; marginal effects; filtering; pharmacogenetics; randomization; case-only estimator.

1. Introduction

With the advent of high-throughput biotechnologies, e.g., microarray, Single Nucleotide Polymorphism (SNP) chips, and whole-genome sequencing, high-dimensional hypothesis testing has become a routine practice in exploratory biological or epidemiological studies. Statistical power in this setting has been limited by stringent significance rules, e.g., Bonferroni correction of multiple tests, that are required to guard against false positives arising from thousands or millions of tests. While investigators strive to ascertain a large number of biological samples, this is often constrained by cost and sample availability. Strategies on efficient design and analysis are therefore of critical importance.

In genome-wide association (GWA) studies, multi-stage designs have been employed in which all SNPs are screened first for suggestive evidence in a proportion of samples, and the most promising SNPs were tested in the next stage(s) (Satagopan et al., 2004; Prentice and Qi, 2006). Genotyping cost can be substantially reduced, yet with possibly little loss of power as compared to the

one-stage design, whether analysis is based on the replication data only (Satagopan et al., 2004), or combines data from multiple stages (Skol et al., 2006). Efficient analysis of genetic association data has also been extensively studied. Notable strategies encompass addressing local dependence of human genome by haplotype analysis (Lin and Zeng, 2006; Dai et al., 2009a), imputing unmeasured SNPs so that data from different platforms can be combined in meta-analysis (Li and Abecasis, 2006; Browning and Browning, 2007), and exploiting gene-environment independence assumption to gain efficiency (Chatterjee and Carroll, 2005; Dai et al., 2009b). Aside from the sheer number of tests, features of genetic inheritance of complex diseases pose additional threat to adequate power, e.g., multiple risk alleles, each having marginally weak associations, possibly interacting with other alleles or environmental attributes (Kraft, 2009).

When gene-gene or gene-environment interactions are targets of inference, there are ideas scattering in literature to filter out the majority of irrelevant SNPs upfront (Millstein et al., 2006; Kooperberg and LeBlanc, 2008; Murcray et al., 2008). The intuition is that, as long as the statistic used in the filtering stage is independent of the statistic in the testing stage, we only need to correct for the number of the tests actually passing the filtering, thus preserving power on features that are most promising. The filtering criterion was largely formulated by biological premises, e.g., SNPs with interactions are likely to have marginal effects (Kooperberg and LeBlanc, 2008), or gene-gene, gene-environment independence in a case-control sample should be expected if there is no interaction (Millstein et al., 2006; Murcray et al., 2008). Theoretical justification of these procedures, however, has not been elaborated.

In this article, we give a formal treatment of two-stage hypothesis testing procedures via independent statistics. We discuss in Section 2 the conditions that are required for such procedures to maintain strong control of family-wise error rate (FWER) in both high-dimensional and low-dimensional testing. Through use of estimating equation theory, we present in Section 3 a unified approach to prove the asymptotic independence of various statistics previously suggested. We discuss the utility of such procedures through examples in Section 3, that have a broader scope of hypothesis testing than just GWA studies. In Sections 4 and 5, we present simulations and data application to show the benefit of two-stage testing procedures.

2. A class of two-stage procedures and their strong control of FWER

Consider data from n subjects drawn from a cohort based on a prespecified sampling plan. Let Y_i denote the outcome variable, and let $\mathbf{X}_i = (X_{i1}, ..., X_{im})$ denote a collection of m features measured for i^{th} subject. Occasionally, there is a low-dimensional variable of key interest, e.g., a randomized intervention, denoted by Z_i. For different subjects, the random variables $(Y_i, \mathbf{X}_i, Z_i)$ are independent and identically distributed. Let θ_j, $j = 1, ..., m$, denote the parameter of interest. The goal is to test m null hypotheses, $H_{0j} : \theta_j = 0$ versus $H_{1j} : \theta_j \neq 0$.

The test statistic for H_j is often formulated by *asymptotically linear estimators* (ALE) (Newey and Powell, 1990; Robins et al., 1994), scaled by its estimated standard error. An estimator $\hat{\theta}_j$ of θ_j is asymptotically linear if $\sqrt{n}(\hat{\theta}_j - \theta_j) = 1/\sqrt{n}\sum_{i=1}^{n} B_{ij} + o_p(1)$, $\mathrm{E}(B_j) = 0$, $\mathrm{E}(B'_{ij}B_{ij}) < \infty$. The function B_j is referred to as the *influence function* of $\hat{\theta}_j$ in the sense of Casella and Berger (2002). By the Central Limit Theorem and Slutsky's theorem $\sqrt{n}(\hat{\theta}_j - \theta_j)$ is asymptotically normal with mean 0 and variance $\mathrm{E}(B'_{ij}B_{ij})$. Define a Wald test statistic $T_j = \hat{\theta}_j / \sqrt{\widehat{Var}(\hat{\theta}_j)}$.

Now consider a different set hypothesis tests: $K_{0j} : \vartheta_j = 0$ versus $K_{1j} : \vartheta_j \neq 0$. Let $\hat{\vartheta}_j$ denote an asymptotically linear estimator of ϑ_j as defined above. A Wald test statistic is formulated similarly, $T_j^0 = \hat{\vartheta}_j / \sqrt{\widehat{Var}(\hat{\vartheta}_j)}$.

The following two-stage testing procedure is considered: denote by α_0 a prespecified screening factor in the first stage, $0 < \alpha_0 < 1$. The corresponding first-stage critical region is $\Gamma_j^0 = \{T_j^0 : |T_j^0| > C_{1-\alpha_0/2}\}$, where $C_{1-\alpha_0/2}$ is the $1 - \alpha_0/2$ quantile of the standard normal distribution. Suppose there are m_0 features falling in the critical region. Let $0 < \alpha < 1$, and define the second-stage rejection region $\Gamma_j = \{T_j : |T_j| > C_{1-\alpha/2m_0}\}$, where $C_{1-\alpha/2m_0}$ is the $1-\alpha/2m_0$ quantile of the standard normal distribution. We declare a test statistically significant if $T_j^0 \in \Gamma_j^0$ and $T_j \in \Gamma_j$.

We show in the following theorems that with proper conditions, the two-stage testing procedure will control the FWER in the strong sense, though Bonferroni correction is only applied to the second-stage testing. Strong control of FWER means that for any set of null hypotheses, the probability of having at least one false positive test is less than or equal to the prespecified level α (Holm, 1979). The proof is given in the Appendix.

THEOREM 1. *If the asymptotic distribution of $\hat{\theta}_j$ and $\hat{\vartheta}_k$ are multivariate Gaussian, and they are uncorrelated, i.e.,*

$$Cov\left(\sqrt{n}(\hat{\theta}_j - \theta_j), \sqrt{n}(\hat{\vartheta}_k - \vartheta_k)\right) \to_p 0 \qquad \forall \quad j,k \in \{1,...,m\}$$

the proposed two-step procedure preserves FWER at the level α asymptotically in the strong sense, i.e., for any non-empty index set $\mathbf{J} \subseteq \{1,2,...,m\}$

$$lim_{n\to\infty} Pr\left\{\cup_{j\in\mathbf{J}}(T_j^0 \in \Gamma_j^0 \cap T_j \in \Gamma_j)|H_{0j}, K_{0j}\right\} \leq \alpha.$$

Theorem 1 requires that the set of estimators in the first stage and the set of estimators in the second stage are jointly asymptotically independent, under the joint null hypothesis H_{0j}, K_{0j}. This is a rather restrictive condition. For mutually independent features X_{ij}, this condition reduces to uncorrelated $\hat{\vartheta}_j$ and $\hat{\theta}_j$. When X_{ij} are correlated, we show in Section 3 that there are situations where two sets of estimators are jointly asymptotically independent, e.g., H_j is on testing interactions of X_{ij} with a randomized treatment assignment Z_i, and K_j is on the marginal effect of X_{ij}. **Theorem 1** applies regardless of the scale of hypothesis testing. When testing is high dimensional, the conditions can be relaxed as long as there is weak dependence among $\mathbf{X}_i$.

THEOREM 2. *In high-dimensional hypothesis testing, if the asymptotic distribution of $\hat{\vartheta}_j$ and $\hat{\theta}_j$ are multivariate Gaussian and they are uncorrelated, i.e.,*

$$Cov\left(\sqrt{n}(\hat{\vartheta}_j - \vartheta_j), \sqrt{n}(\hat{\theta}_j - \theta_j)\right) \to_p 0 \qquad \forall \quad j \in \{1,...,m\},$$

and $\frac{m_0}{m} \to_p \alpha_0$, the proposed two-step procedure preserves FWER at the level α for large m and n in the strong sense, i.e., for any non-empty index set $\mathbf{J} \subseteq \{1,2,...,m\}$

$$lim_{m\to\infty} lim_{n\to\infty} Pr\left\{\cup_{j\in\mathbf{J}}(T_j^0 \in \Gamma_j^0 \cap T_j \in \Gamma_j)|H_{0j}, K_{0j}\right\} \leq \alpha.$$

Theorem 2 requires marginal asymptotic independence of T_j^0 and T_j under the joint null hypothesis, which is satisfied for numerous examples in Section 3. The conditions required to obtain $\frac{m_0}{m} \to_p \alpha_0$ are those required by the Law of Large Numbers (LLN) for correlated data. For instance, if

$$\text{Cov}\left[I(T_j^0 \in \Gamma_j^0|H_j^0), I(T_k^0 \in \Gamma_k^0|H_k^0)\right] \to 0,$$

when $|j-k|$ gets large, then the LLN for a sequence of $I(T_j^0 \in \Gamma_j^0|K_j^0)$ holds as $m \to \infty$ (White, 2001). This type of serial correlation is exactly the linkage disequilibrium pattern observed in human genome (The International HapMap Consortium, 2005).

Note that both theorems indicate strong control of FWER under the joint null hypotheses, H_{0j} and K_{0j}. The reason is that under the joint null, the test statistics T_j^0 and T_j are the centered z-scores $\sqrt{n}(\hat{\theta}_j-\theta_j)/\sqrt{\hat{V}_2}$, $\sqrt{n}(\hat{\vartheta}_k-\vartheta_k)/\sqrt{\hat{V}_1}$ which are proved to be independent in the Appendix. Under the alternative hypothesis, our simulations suggest that T_j^0 and T_j are approximately independent as well (results not shown).

With minor modification, we can show that when a fixed top m_0 features, rather than a fixed rejection region, are selected for the second-stage testing, we still have strong control of FWER. Details are omitted. Certainly for such two-stage procedures to be useful, they should have more power than a Bonferroni correction for all features. One necessary requirement is that alternative hypothesis H_1 should imply K_1, so that a true alternative should pass the first-stage filtering if the sample size is sufficiently large. For example, a non-zero interaction would suggest a non-zero main effect unless the subgroup effects exactly cancel out. Moreover, the screening statistic T_j^0 should ensure that there is high power for K_1 to pass the filtering. We will discuss this point further in simulation studies.

3. Asymptotically independent statistics

We now discuss asymptotically independent statistics and review a number of examples in two-stage hypothesis testing. To establish the asymptotic *joint* distribution of $\hat{\vartheta}$ and $\hat{\theta}$, it is necessary to study their behaviour under (potentially) misspecified models, since the model indexed by $\hat{\vartheta}$ may disagree with the model indexed by $\hat{\theta}$, e.g., a model is misspecified if it only includes marginal association parameters when actually there are interactions. Maximum likelihood estimation under misspecified models was discussed in White (1982). Let $\boldsymbol{\theta}$ be the set of parameters in the model indexed by θ, and let $\boldsymbol{\vartheta}$ denote the set of parameters in the model indexed by ϑ. Let $\sum_{i=1}^n \boldsymbol{U}_{1i}=0$ be the set of estimating equations solved for $\hat{\boldsymbol{\theta}}$, and let $\sum_{i=1}^n \boldsymbol{U}_{2i}=0$ be the set of estimating equations to be solved for $\hat{\boldsymbol{\vartheta}}$. Suppose $\boldsymbol{\theta}$ is the unique solution to the estimating equations $\mathrm{E}[\boldsymbol{U}_{1i}]=0$, where E denotes expectation under the true distribution. Similarly, $\boldsymbol{\vartheta}$ is the unique solution to the estimating equations $\mathrm{E}[\boldsymbol{U}_{2i}]=0$. Then $\hat{\boldsymbol{\vartheta}} \to_{a.s.} \boldsymbol{\vartheta}$ and $\hat{\boldsymbol{\theta}} \to_{a.s.} \boldsymbol{\theta}$.

Let $A_1=\mathrm{E}[\partial \boldsymbol{U}_{1i}/\partial \boldsymbol{\theta}]$, $A_2=\mathrm{E}[\partial \boldsymbol{U}_{2i}/\partial \boldsymbol{\vartheta}]$, and $B_{kk'}=\mathrm{E}[\boldsymbol{U}_{ki}\boldsymbol{U}_{k'i}]$, $k,k'=1,2$. With suitable regularity conditions (White, 1982) it can be shown that $\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ and $\sqrt{n}(\widehat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta})$ are asymptotically equivalent to $A_k^{-1}n^{-1/2}\sum_{i=1}^n U_{ki}$, $k=1,2$. For each k, the random vector $\boldsymbol{U}_{ki}$ is i.i.d. with zero mean, but for the same i, $\boldsymbol{U}_{1i}$ and $\boldsymbol{U}_{2i}$ are possibly correlated. The *joint* distribution of $\hat{\boldsymbol{\vartheta}}$ and $\hat{\boldsymbol{\theta}}$ is established by the Cramer-Wold device. Let t be a vector of 2 scalars, t_1 and t_2.

$$
\begin{aligned}
t^T\begin{pmatrix}\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\\ \sqrt{n}(\widehat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta})\end{pmatrix} &= t^T\begin{pmatrix}A_1^{-1}n^{-1/2}\sum_{i=1}^n U_{1i}\\ A_2^{-1}n^{-1/2}\sum_{i=1}^n U_{2i}\end{pmatrix}+o_p(1)\\
&= t_1A_1^{-1}n^{-1/2}\sum_{i=1}^n U_{1i}+t_2A_2^{-1}n^{-1/2}\sum_{i=1}^n U_{2i}+o_p(1)\\
&\to_d \mathcal{N}(0,t_1^2A_1^{-1}B_{11}A_1^{-1}+t_2^2A_2^{-1}B_{22}A_2^{-1}+2t_1t_2A_1^{-1}B_{12}A_2^{-1})
\end{aligned}
$$

This leads to the conclusion that the limiting distribution of $\left(\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}),\sqrt{n}(\widehat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta})\right)$ is multivariate Gaussian with zero means and covariance matrix

$$
\begin{bmatrix}A_1^{-1}B_{11}A_1^{-1} & A_1^{-1}B_{12}A_2^{-1}\\ A_2^{-1}B_{21}A_1^{-1} & A_2^{-1}B_{22}A_2^{-1}\end{bmatrix}. \tag{1}
$$

To assess asymptotic independence of $\hat{\vartheta}$ and $\hat{\theta}$, we evaluate the off-diagonal element of their covariance matrix, $A_1^{-1}B_{12}A_2^{-1}$. This provides a unified approach to evaluate asymptotic independence among examples as followed. We first consider marginal independence between two estimators.

3.1. Marginal independence

Example 1: Consider a simple random sample with a continuous Y, e.g., a quantitative trait such as blood pressure, and a collection of high-dimensional features $\mathbf{X}$. The interest is to identify pairwise interactions between two features, say X_1 and X_2, on Y in an ordinary least square regression,

$$\mathrm{E}[Y|X_1, X_2] = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_1 X_2. \tag{2}$$

There are $\binom{m}{2}$ pairwise interactions, which can be computationally infeasible to assess when m is large. Kooperberg and LeBlanc (2008) suggests filtering out features without evidence of marginal association (MA) in a univariate regression,

$$\mathrm{E}[Y|X_1] = \beta_0 + \beta_1 X_1. \tag{3}$$

Let $\boldsymbol{X}_1$ denote the design matrix of (2) with $n \times 4$ dimension, $(1, X_1, X_2, X_1X_2)$. Let $\boldsymbol{X}_2$ denote the design matrix of (3) with 2 columns, $(1, X_1)$. Let $\boldsymbol{Y}$ denote the vector of the outcome variable, and $\boldsymbol{\beta} = (\beta_0, \beta_1)$, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)$. Since the OLS estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ have closed form,

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}, \\ \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T\boldsymbol{Y}, \end{aligned}$$

we can directly compute their covariance

$$\widehat{\mathrm{Cov}}(\hat{\gamma}, \hat{\beta}) = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}\hat{\sigma}^2,$$

where $\hat{\sigma}^2$ is the estimated residual variance under (2). Note that

$$(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

because $(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_1 = \mathbf{I}$ and $\boldsymbol{X}_2$ are contained in $\boldsymbol{X}_1$. Hence

$$(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1} = \begin{bmatrix} .. & .. \\ .. & .. \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

This implies that $\hat{\beta}_1$ and $\hat{\gamma}_3$ are uncorrelated, and hence independent for a normally distributed $\boldsymbol{Y}$ at any sample size.

Example 2a: Consider a case-control study for a binary outcome Y. A collection of features $\mathbf{X}$ were sampled retrospectively conditional on Y. The interest is to identify features that have pairwise interactions, similar to **Example 1**. The standard approach is to fit a logistic regression with interactions,

$$\mathrm{logit}\{\mathrm{E}[Y|X_1, X_2]\} = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_1 X_2. \tag{4}$$

Similarly, Kooperberg and LeBlanc (2008) proposed to filter out features without much marginal association in a univariate logistic regression,

$$\text{logit}\{\text{E}[Y|X_1]\} \quad = \quad \beta_0 + \beta_1 X_1, \tag{5}$$

though the proof was not shown explicitly.

Let $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ be defined as in **Example 1**. Denote $\boldsymbol{U}_{1i}$ the score functions for (4) and $\boldsymbol{U}_{2i}$ the score functions for (5). Note that

$$\begin{aligned} \boldsymbol{U}_{1i} &= \boldsymbol{X}_{1i}(Y_i - E[Y_i|X_{1i}, X_{2i}]), \\ \boldsymbol{U}_{2i} &= \boldsymbol{X}_{2i}(Y_i - E[Y_i|X_{1i}]). \end{aligned}$$

In case-control sampling, the likelihood is the retrospective distributions of covariates conditional on disease status. Remarkably, if a standard logistic regression is fitted to case-control data, $\hat{\beta}_1$ and $\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3$ are the semiparametric maximum likelihood estimators even though biased sampling is ignored (Prentice and Pyke, 1979). The score functions for regression coefficients from the profile likelihood are the same as $\boldsymbol{U}_{1i}$ and $\boldsymbol{U}_{2i}$ except for the intercepts.

We evaluate the terms in the covariance matrix (1). Observe that

$$\begin{aligned} A_1 &= \text{E}\left\{(\boldsymbol{X}_{1i}^T\boldsymbol{X}_{1i})\text{E}[Y_i|\boldsymbol{X}_{1i}](1-\text{E}[Y_i|\boldsymbol{X}_{1i}])\right\}, \\ B_{21} &= \text{E}\left\{(\boldsymbol{X}_{1i}\boldsymbol{X}_{2i}^T)(Y_i-\text{E}[Y_i|\boldsymbol{X}_{1i}])(Y_i-\text{E}[Y_i|\boldsymbol{X}_{2i}])\right\}, \\ A_2 &= \text{E}\left\{(\boldsymbol{X}_{2i}^T\boldsymbol{X}_{2i})\text{E}[Y_i|\boldsymbol{X}_{2i}](1-\text{E}[Y_i|\boldsymbol{X}_{2i}])\right\}. \end{aligned}$$

Thus by the Central Limit Theorem and Slutsky Theorem,

$$\begin{aligned} A_1^{-1}B_{12}A_2^{-1} \quad = \quad & n(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}(\boldsymbol{X}_1^T\boldsymbol{X}_2)(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1} \\ & \left(\frac{1}{n}\sum_{i=1}^n \widehat{\text{E}}[Y_i|\boldsymbol{X}_{1i}](1-\widehat{\text{E}}[Y_i|\boldsymbol{X}_{1i}])]\right) \\ & \left(\frac{1}{n}\sum_{i=1}^n (Y_i-\widehat{\text{E}}[Y_i|\boldsymbol{X}_{1i}])(Y_i-\widehat{\text{E}}[Y_i|\boldsymbol{X}_{2i}])\right) \\ & \left(\frac{1}{n}\sum_{i=1}^n \widehat{\text{E}}[Y_i|\boldsymbol{X}_{2i}](1-\widehat{\text{E}}[Y_i|\boldsymbol{X}_{2i}])]\right) + o_p(1). \end{aligned}$$

We have thus shown that

$$(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1} = \begin{bmatrix} .. & .. \\ .. & .. \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Therefore the lower half submatrix of $A_1^{-1}B_{12}A_2^{-1}$ is 0. This implies that $\hat{\beta}_1$ and $\hat{\gamma}_3$ are asymptotically uncorrelated, and thus asymptotically independent.

Example 2b: A variation of **Example 2a** is to replace the standard estimators of interactions in (4) by the so-called "case-only" estimators (Piegorsch et al., 1994; Umbach and Weinberg, 1997),

when the disease is rare and the two covariates, such as a gene and an environmental variable, are independent in the population. Despite the substantial efficiency gain, case-only estimators are generally sensitive to departures from gene-environment independence assumption (Albert et al., 2001). In genetic studies within randomized clinical trials (RCT), however, there exists indisputable independence between the treatment assignment and baseline covariates, including genetic variants. Henceforth, we focus on treatment-genotype interactions in a RCT.

Consider a randomized clinical trial with N subjects. Let Y_i denote the indicator variable of whether the i^{th} subject acquires a disease, $i = 1, ..., N$. Assume that $\Pr(Y_i = 1) \approx 0$. Let Z_i denote the indicator variable of whether the treatment or the control assignment is received. Let $\mathbf{X} = (X_1, ..., X_m)$ denote a collection of high-dimensional features, $m \gg N$, e.g., SNPs in a whole-genome association study. If $\mathbf{X}$ were to be measured for every participant, the joint density $(Y_i, Z_i, \mathbf{X}_i)$ is from i.i.d. random variables. However only a proportion of cases and a proportions of controls are sampled to collect $\mathbf{X}$. We consider two logistic regressions: one regression consisting of Y on X_1 in cases and controls as in (5), the other consisting of Z on X_1 in cases only:

$$\text{logit}\{\mathrm{E}[Z|X_1, Y = 1]\} \quad = \quad \delta_0 + \delta_1 X_1. \tag{6}$$

Following the same notations, for the i^{th} subject, the score functions can be expressed as followed:

$$\begin{aligned} \boldsymbol{U}_{i1} &= \boldsymbol{X}_{1i}(Y_i - E[Y_i|\boldsymbol{X}_{1i}]), \\ \boldsymbol{U}_{i2} &= \boldsymbol{X}_{1i}(Z_i - E[Z_i|\boldsymbol{X}_{1i}, Y_i = 1])1_{[Y_i=1]}, \end{aligned}$$

where $\boldsymbol{X}_{1i}$ is the design vector $(1, x_{1i})$. Note that $\boldsymbol{U}_{i2} = 0$ if $Y_i = 0$. The covariance of $\hat{\beta}_1$ and $\hat{\delta}_1$ can again be derived using estimating equation theory. Let $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\delta_0, \delta_1)$. Note that

$$\begin{aligned} B_{21} &= \mathrm{E}\left\{\boldsymbol{X}_{1i}\boldsymbol{X}_{1i}^T(Y_i - \mathrm{E}[Y_i|\boldsymbol{X}_{1i}])(Z_i - \mathrm{E}[Z_i|\boldsymbol{X}_{1i}, Y_i = 1])1_{[Y_i=1]}\right\} \\ &= \Pr(Y_i = 1)\mathrm{E}\left\{\boldsymbol{X}_{1i}\boldsymbol{X}_{1i}^T(1 - \Pr[Y_i = 1|\boldsymbol{X}_{1i}])(Z_i - \mathrm{E}[Z_i|\boldsymbol{X}_{1i}, Y_i = 1])\right\} \\ &= \Pr(Y_i = 1)\mathrm{E}_{\boldsymbol{X}_{1|\mathbf{Y}=1}}\left\{\boldsymbol{X}_{1i}\boldsymbol{X}_{1i}^T(1 - \Pr[Y_i = 1|\boldsymbol{X}_{1i}])\mathrm{E}_{\boldsymbol{Z}|\boldsymbol{X}_{1,}\mathbf{Y}=1}(Z_i - \mathrm{E}[Z_i|\boldsymbol{X}_{1i}, Y_i = 1])\right\} \\ &= 0 \end{aligned}$$

The second to the last equation use the law of iterated expections. Hence the off-diagonal of the covariance matrix is zero and the proof is complete.

Example 3: So far we assume that features that have interactions should also manifest some main effects. Counterexamples can be constructed, e.g., two environmental groups with opposite signs of genetic effects. Another set of screening statistics can be developed to avoid this problem. In the rare-disease scenario considered in **Example 2b**, the association between Z and a feature, X_1 in the combined case-control sample may reveal some clues on interactions. Let

$$\text{logit}\{\mathrm{E}[Z|X_1]\} \quad = \quad \tau_0 + \tau_1 X_1. \tag{7}$$

$$\text{logit}\{\mathrm{E}[Y|X_1, Z]\} \quad = \quad \gamma_0 + \gamma_1 X_1 + \gamma_2 Z + \gamma_3 X_1 Z. \tag{8}$$

The rationale is as followed: when the disease is rare, we expect to have $Z \perp X_1$ in the controls. If there is interaction between Z and X_1, Z and X_1 are dependent in the cases. Due to oversampling of cases, Z and X_1 are dependent in the combined case-control sample. Thus we can select the top

features from the regression $Z \sim X_1$ for further interaction testing. This is somewhat similar to two-step procedures previously proposed for gene-gene (Millstein et al., 2006) and gene-environment interactions (Murcray et al., 2008), though confounding of the gene-gene or gene-environment independence is always an issue in observational studies. Intuitively, because we have not used the disease information Y to guide the screening, we do not have to spend type I error in screening. The formal proof can be pursued similarly by estimating equation theory.

Following the notations in **Example 2b**, two sets of estimating equations are

$$\begin{aligned} \boldsymbol{U}_{i1} &= \boldsymbol{X}_{1i}(Z_i - E[Z_i|\boldsymbol{X}_{1i}]), \\ \boldsymbol{U}_{i2} &= \boldsymbol{X}_{1i}(Y_i - E[Y_i|\boldsymbol{X}_{1i}, Z_i]), \end{aligned}$$

where $\boldsymbol{U}_{i1}$ is the score function of (7) and $\boldsymbol{U}_{i2}$ is the score function of (8). So

$$\begin{aligned} B_{21} &= \mathrm{E}\left\{\boldsymbol{X}_{1i}\boldsymbol{X}_{1i}^T(Z_i - \mathrm{E}[Z_i|\boldsymbol{X}_{1i}])(Y_i - \mathrm{E}[Y_i|\boldsymbol{X}_{1i}, Z_i])\right\} \\ &= \mathrm{E}_{\boldsymbol{Z},\boldsymbol{X}}\left\{\boldsymbol{X}_{1i}\boldsymbol{X}_{1i}^T(Z_i - \mathrm{E}[Z_i|\boldsymbol{X}_{1i}])\mathrm{E}_{\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{X}}(Y_i - \mathrm{E}[Y_i|\boldsymbol{X}_{1i}, Z_i])\right\} \\ &= 0. \end{aligned}$$

The derivation uses the law of iterated expectations, similar to that in **Example 2b**. Hence the off-diagonal of the covariance matrix is zero and the proof is complete. Interestingly, asymptotic independence does not hold when we use the case-only estimators for interactions in the second stage. The reason might be that the information of rare diseases and the independence has been used in formulating the first-stage estimator (7), therefore it cannot be used again in forming case-only estimators.

Note that asymptotic independence also holds between $\hat{\tau}_1$ and $\hat{\gamma}_1$. So it is possible to test for main SNP effects in the control arm. Moreover, the same proof applies if (8) is replaced by any regression model with X_1 and Z as covariates, including

$$\mathrm{logit}\{\mathrm{E}[Y|X_1, Z]\} = \eta_0 + \eta_1 X_1 + \eta_2 Z. \tag{9}$$

so that the adjusted effect of X_1 could be the test of interest after being filtered by $\hat{\tau}_1$ in (7). Since Z is the randomized treatment assignment, the adjusted effect γ_1 approximates the marginal effect β_1 in (5). As $\hat{\tau}_1$ essentially assesses deviation from baseline independence (DBI) between Z and X_1 in the case-control sample, we call it the "DBI" criterion hereafter. These results suggest that in a RCT with a rare outcome, we can use the criterion $\tau_1 \neq 0$ in (7) to screen for SNPs with (adjusted) main effects in (9) or (8), and SNPs with interactions with the randomized treatment in (8). In Section 5, we show in a data example in which using DBI in screening led to some interesting discoveries in adjusted SNP effects.

The utility of DBI can be extended to scenarios where the disease is not rare and there is a known treatment effect. We state this in the following Lemma.

LEMMA 1. *Suppose a case-control sample was drawn from a randomized clinical trial with a binary disease outcome Y and a binary treatment assignment Z. Suppose either one of the following two conditions holds: (a) the disease is rare; (b) the disease is common and $Pr(Y|Z, X) \neq Pr(Y|X)$, i.e., there is a treatment effect conditional on X. Denote by R the indicator of being selected into the case-control sample. For a baseline predictor X, if we observe that*

$$Pr(Z|X, R = 1) \neq Pr(Z|R = 1),$$

then

$$Pr(Y|X, Z) \neq Pr(Y|Z).$$

The proof is straightforward and left to Appendix. In Section 4, we compare the powers of two-stage procedures using MA and DBI under both rare-disease and common-disease scenarios.

Example 4: In RCTs, study endpoint is often time-to-event and primary inference is often based on the Cox proportional hazard model by partial likelihood. The score functions for partial likelihood take a specialized form and the arguments used to show independence in the previous examples do not apply. When the endpoint in a RCT is rare, however, we show below that the estimator for MA in a Cox model is asymptotically independent of a case-only estimator for interaction.

We switch to survival analysis notations for this example. Under the proportional hazard model, the hazard function for the failure time Y associated with covariates (Z, X, ZX) is

$$\lambda_(y; Z, X) = \lambda_0(y) \exp(\beta_0 X + \beta_1 Z + \beta_2 ZX),$$

where β_0 is the main effect of genotype X, β_1 is the main effect of treatment Z, β_2 is the interaction, and λ_0 is an unspecified baseline hazard function. When Y is subject to independent right-censorship, we observe $T = \min(Y, C)$ and $\Delta = I(Y \leq C)$, where C is the censoring time. Let $(T_i, \Delta_i, Z_i, X_i)$, $i = 1, ..., n$ be n independent replicates. Then the partial likelihood function for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_1)$ is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\beta}' \boldsymbol{X}_i)}{\sum_{j=1}^{n} R_j(T_i) \exp(\boldsymbol{\beta}' \boldsymbol{X}_i)} \right\}^{\Delta_i},$$

where $\boldsymbol{X}_i$ is the 3×1 design vector, $R_j(t)$ is the at-risk indicator $I(T_j \geq t)$. The corresponding score function equals

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Delta_i \left\{ \boldsymbol{X}_i - \frac{S^{(1)}(\boldsymbol{\beta}, \boldsymbol{X}_i)}{S^{(0)}(\boldsymbol{\beta}, \boldsymbol{X}_i)} \right\},$$

where $S^{(0)}(\boldsymbol{\beta}, t) = \sum_{j=1}^{n} R_j(t) \exp(\boldsymbol{\beta}' \boldsymbol{X}_j)$ and $S^{(1)}(\boldsymbol{\beta}, t) = \sum_{j=1}^{n} R_j(t) \exp(\boldsymbol{\beta}' \boldsymbol{X}_j) \boldsymbol{X}_j$. If we fit a Cox model with only the marginal effect of X, the model may be misspecified. Let α denote the MA parameter in the Cox model. Under misspecified Cox models, the robust variance-covariance estimator for $\hat{\alpha}$ is $A^{-1}(\hat{\alpha}) B(\hat{\alpha}) A^{-1}(\hat{\alpha})$ (Lin and Wei, 1989), where $B(\alpha) = \sum_{i=1}^{n} W_i(\alpha) W_i(\alpha)'$,

$$W_i(\alpha) = \Delta_i \left\{ X_i - \frac{S^{(1)}(\alpha, X_i)}{S^{(0)}(\alpha, X_i)} \right\} - \sum_{j=1}^{n} \frac{\Delta_j R_i(T_j) \exp(\alpha' X_j)}{n S^{(0)}(\alpha, X_j)} \left\{ X_j - \frac{S^{(1)}(\alpha, X_j)}{S^{(0)}(\alpha, X_j)} \right\},$$

and

$$A(\alpha) = \sum_{i=1}^{n} \Delta_i \left\{ \frac{S^{(2)}(\alpha, X_i)}{S^{(0)}(\alpha, X_i)} - \frac{S^{(1)}(\alpha, X_i) S^{(1)}(\alpha, X_i)'}{S^{(0)}(\alpha, X_i)^2} \right\}.$$

Let $s^{(r)}(\alpha, t) = E[S^{(r)}(\alpha, t)]$, $r = 0, 1$, $N_i(t) = I\{T_i \leq t, \Delta_i = 1\}$, $\bar{N}(t) = \sum N_i(t)$. Lin and Wei (1989) showed that $n^{-1/2} \sum_{i=1}^{n} W_i(\alpha)$ is asymptotically equivalent to $n^{-1/2} \sum_{i=1}^{n} w_i(\alpha)$,

$$w_i(\alpha) = \int_0^{\infty} \left\{ X_i - \frac{s^{(1)}(\alpha, X_i)}{s^{(0)}(\alpha, X_i)} \right\} dN_i(t) - \int_0^{\infty} \frac{R_i(t) \exp(\alpha' X_i)}{s^{(0)}(\alpha, t)} \left\{ X_i - \frac{s^{(1)}(\alpha, t)}{s^{(0)}(\alpha, t)} \right\} d\tilde{F}(t),$$

where $\tilde{F}_n(t) = \bar{N}(t)/n$ and $\tilde{F}(t) = E[\tilde{F}_n(t)]$.

Now suppose one wishes to assess and compare the treatment hazard ratios stratified by the genotype $X = x$ valued at $0, 1, 2$. The hazard rate at time $T = t$ from randomization, may be specified as

$$\lambda(t; Z, X) = \lambda_{0x} \exp(\beta_1 Z + \beta_2 ZX),$$

so that $\exp(\beta_1)$ is the treatment hazard ratio for subjects with $x = 0$, and $\exp(\beta_2)$ indexes an additive interaction for subjects with $x = 1, 2$.

Observe that

$$\begin{aligned}\frac{\Pr(Z=1|T=t,X)}{\Pr(Z=0|T=t,X)} &= \frac{\Pr(T=t|Z=1,X)\Pr(Z=1|x)}{\Pr(T=t|Z=0,X)\Pr(Z=0|X)} \\ &= \frac{\Pr(T\geq t|Z=1,X)\Pr(Z=1|x)}{\Pr(T\geq t|Z=0,X)\Pr(Z=0|X)}\exp(\beta_1 Z+\beta_2 ZX) \\ &= \frac{\Pr(Z=1|T\geq t,X)}{\Pr(Z=0|T\geq t,X)}\exp(\beta_1 Z+\beta_2 ZX).\end{aligned}$$

Z and X are independent by design, the event is rare and censoring rates are equal in two arms, it will follow that

$$\frac{\Pr(Z=1|T\geq t,X)}{\Pr(Z=0|T\geq t,X)} = \frac{q}{1-q}$$

to an excellent approximation, where q is the fraction of the trial cohort assigned to the treatment. It follows that the β_1 and β_2 can be estimated using logistic regression of Z on X with $\log(q/(1-q))$ as an "offset". Though estimated by a logistic regression, these estimators have a hazard ratio interpretation in this context. Note that this version of case-only estimator allows for estimation of treatment hazard ratio in each subgroup, not just the interaction parameter in "standard" case-only estimators in logistic regression (Piegorsch et al., 1994). See Vittinghoff and Bauer (2006) for related work.

Let $u_i(\boldsymbol{\beta})$ denote the estimating functions for this case-only estimator. Following the same argument we used in **Example 2b**, $E[u_i(\boldsymbol{\beta})' w_i(\alpha)] = 0$. Briefly, observe that $u_i(\boldsymbol{\beta})$ is based on the distribution of $Z|X, \Delta = 1$ when $\Delta = 1$, and is zero otherwise. On the other hand, w_i is on the distribution of $X|\Delta$. This leads to zero asymptotic covariance for $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$. We thus show that in RCT with a rare endpoint, the estimator of the marginal association in a Cox model is independent of the case-only estimator of the interaction. This result can be extended to Cox-model marginal association analyses based on such cohort sampling techniques as nested case-control and case-cohort sampling.

3.2. Joint independence

For two sets of asymptotically linear estimators, for instances $\{\hat{\beta}_{1j}, j = 1, ..., m\}$ and $\{\hat{\gamma}_{3k}, k = 1, ..., m\}$, joint independence implies that $\hat{\beta}_{1j}$ and $\hat{\gamma}_{3k}$ are uncorrelated $\forall j, k$, since their joint distribution is multivariate Gaussian. This is a much stronger condition than no marginal correlation between $\hat{\beta}_{1j}$ and $\hat{\gamma}_{3j}$ for the same j. In fact, pathological examples can be constructed to show that merely marginal independence in the two set of estimators could yield inflated FWER when m is small. In randomized clinical trials, however, joint independence can be achieved for estimators in **Example 2a** and **3** when the randomized assignment Z is involved, whether data are from the full cohort or a case-control sample. Next we show the proof for joint independence of $\{\hat{\beta}_{1j}, j = 1, ..., m\}$ in (5) and $\{\hat{\gamma}_{3k}, k = 1, ..., m\}$ in (8).

Let $\bar{Z}$ denote the mean of Z_i, and let $\bar{X}_j$ denote the mean of X_{ij}. We consider logistic regressions with centered versions of X_{ij} and Z_i:

$$\begin{aligned}\text{logit}\{\text{E}[Y_i|X_{ik}]\} &= \beta_{0k} + \beta_{1k}(X_{ik} - \bar{X}_k), \\ \text{logit}\{\text{E}[Y_i|X_{ij}, Z_i]\} &= \gamma_{0j} + \gamma_{1j}(X_{ij} - \bar{X}_j) + \gamma_{2j}(Z_i - \bar{Z}) + \gamma_{3j}(X_{ij} - \bar{X}_j)(Z_i - \bar{Z}).\end{aligned}$$

Again let $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k})$ and $\boldsymbol{\gamma}_j = (\gamma_{0j}, \gamma_{1j}, \gamma_{2j}, \gamma_{3j})$. Following the same notations in **Example 2a**, where U_{i1} denote the score vector for $\boldsymbol{\gamma}$, U_{i2} denote the score vector for $\boldsymbol{\beta}$. Denote by V the covariance matrix of $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_j$, $V = A_1^{-1} B_{12} A_2^{-1}$. For $j \neq k$, not every element of B_{21} is 0. However, it is easy to see that the lower half of this 4×2 matrix is 0. Denote by $B_{12,ll'}$ the element of B_{12} in the l^{th} row and the l'^{th} column. Then

$$\begin{aligned}
B_{12,31} &= \mathrm{E}\left\{(Z_i - \bar{Z})(Y_i - \mathrm{E}[Y_i|X_{ij}, Z_i])(Y_i - \mathrm{E}[Y_i|X_{ik}])\right\} \\
&= \mathrm{E}[(Z_i - \bar{Z})]\mathrm{E}\left\{(Y_i - \mathrm{E}[Y_i|X_{ij}, Z_i])(Y_i - \mathrm{E}[Y_i|X_{ik}])\right\} = 0, \\
B_{12,41} &= \mathrm{E}\left\{(Z_i - \bar{Z})(X_{ij} - \bar{X}_j)(Y_i - \mathrm{E}[Y_i|X_{ij}, Z_i])(Y_i - \mathrm{E}[Y_i|X_{ik}])\right\} \\
&= \mathrm{E}[Z_i - \bar{Z}]\mathrm{E}\left\{(X_{ij} - \bar{X}_j)(Y_i - \mathrm{E}[Y_i|X_{ij}, Z_i])(Y_i - \mathrm{E}[Y_i|X_{ik}])\right\} = 0, \\
B_{12,32} &= \mathrm{E}\left\{(Z_i - \bar{Z})(X_{ik} - \bar{X}_k)(Y_i - \mathrm{E}[Y_i|X_{ij}, Z_i])(Y_i - \mathrm{E}[Y_i|X_{ik}])\right\} = 0, \\
B_{12,42} &= \mathrm{E}\left\{(Z_i - \bar{Z})(X_{ik} - \bar{X}_k)(X_{ij} - \bar{X}_j)(Y_i - \mathrm{E}[Y_i|X_{ij}, Z_i])(Y_i - \mathrm{E}[Y_i|X_{ik}])\right\} = 0.
\end{aligned}$$

Similarly, we can show that the off-diagonal 2×2 submatrices of A_1 are 0. A little matrix algebra yields $V_{23} = V_{24} = 0$. Thus we proved the asymptotic independence of $\hat{\beta}_{1k}$ and $\hat{\gamma}_{3j}$, $\forall j \neq k$. We have shown previously the asymptotic independence of $\hat{\beta}_{1j}$ and $\hat{\gamma}_{3j}$. This leads to strong independence of $\{\hat{\beta}_{1j}, j = 1, ..., m\}$ and $\{\hat{\gamma}_{3k}, k = 1, ..., m\}$.

Note the same proof applies to the estimators in **Example 3**. As **Theorem 1** stated, joint independence establishes strong control of FWER by a two-stage procedure, regardless the number of tests. As such, in a RCT where there are a number of baseline covariates, we could use both MA and DBI to screen for interactions between treatment- and baseline covariates.

3.3. A counterexample

We review a counterexample in which the independence of two statistics does not hold, so that an adaptive two-stage procedure would fail to preserve the type I error. This example pertains to choice of the estimator for a gene-environment interaction in an observational study. The case-only estimator is efficient, yet the required gene-environment independence is often subject to confounding. A naive adaptive procedure is to first test the independence of gene and environment factors in the controls, then use the case-only estimator or the standard interaction estimator when the first hypothesis is not rejected (Albert et al., 2001). Following the notations in **Example 2b**, let Z denote the environmental factor and X denote the genetic factor. We note that in the following three regression models,

$$\begin{aligned}
\mathrm{logit}\{\mathrm{E}[Z|X, Y = 0]\} &= \nu_0 + \nu_1 X_1, \\
\mathrm{logit}\{\mathrm{E}[Y|X, Z]\} &= \gamma_0 + \gamma_1 X + \gamma_2 Z + \gamma_3 XZ, \\
\mathrm{logit}\{\mathrm{E}[Z|X, Y = 1]\} &= \delta_0 + \delta_1 X.
\end{aligned}$$

$\hat{\nu}_1$ is independent of δ_1 since they use different data, yet $\hat{\nu}_1$ is not independent of γ_3. So using $\hat{\nu}_1$ to decide whether to use $\hat{\gamma}_3$ or $\hat{\delta}_1$ will incur an inflated type I error (Albert et al., 2001).

4. Simulations

We first examine empirical correlations of various pairs of statistics in small samples by simulation. We generated 20,000 simulated datasets from the logistic regression model (4), in which parameters are $\boldsymbol{\gamma} = (-1, 0, 0, \gamma_3)$ with varying levels of interactions $\gamma_3 = 0, 0.5$, or 1. The sample sizes of the simulated datasets were 200, 1000, and 5000. X_1 and X_2 were generated as bivariate normal

Table 1. The empirical correlations for the marginal effect $\hat{\beta}_1$ and the interaction $\hat{\gamma}_3$ in **Example 2a** with simple random sampling among 20,000 simulated data. X_1 and X_2 have a standard bivariate normal distribution with varying degree of correlation. The parameters are from the following models: β_1: logit$\{\mathsf{E}[Y|X]\} = \beta_0 + \beta_1 X$; γ_3: logit$\{\mathsf{E}[Y|X,Z]\} = \gamma_0 + \gamma_1 X + \gamma_2 Z + \gamma_3 XZ$.

		$n = 200$		$n = 1000$		$n = 5000$	
		Estimator	Z-score	Estimator	Z-score	Estimator	Z-score
$\gamma_3 = 0$	cor(X_1, X_2)=0.0	-0.014	-0.014	-0.007	-0.008	-0.003	-0.004
	cor(X_1, X_2)=0.4	-0.019	-0.017	-0.006	-0.006	-0.001	-0.001
	cor(X_1, X_2)=0.8	-0.018	-0.015	-0.005	-0.006	0.001	0.001
$\gamma_3 = 0.5$	cor(X_1, X_2)=0.0	-0.007	-0.009	-0.007	-0.007	-0.007	-0.007
	cor(X_1, X_2)=0.4	0.008	0.007	0.002	0.003	-0.010	-0.009
	cor(X_1, X_2)=0.8	0.004	0.004	-0.011	-0.010	0.004	0.005
$\gamma_3 = 1$	cor(X_1, X_2)=0.0	-0.001	-0.002	0.005	0.006	0.005	0.005
	cor(X_1, X_2)=0.4	0.008	0.008	-0.011	-0.011	-0.018	-0.018
	cor(X_1, X_2)=0.8	0.003	0.005	-0.007	-0.007	-0.019	-0.019

distribution with means 0 and variances 1. We introduced a range of correlation between X_1 and X_2 from 0 to 0.8. Table 1 shows the empirical correlation of $\hat{\beta}_1$ and $\hat{\gamma}_3$, and empirical correlation of their corresponding centered z-scores in various parameter settings. The centered z-score is the centered estimator (estimator substracted by the mean), scaled by the sandwich variance estimator. Clearly even with a sample size of 200, the asymptotic independence holds fairly well across different levels of interactions and different levels of covariate correlations.

We also simulated case-control samples that are nested within a randomized clinical trial (**Example 2b** and **3**). A binary randomized treatment variable was generated from $Ber(0.5)$, the SNP has minor allele frequency 0.2 and diploids were formed assuming the Hardy-Weinberg equilibrium. The logistic regression model (4) with $\boldsymbol{\gamma} = (-4.5, 0, 0, \gamma_3)$ was used to generate the disease data for $N = 2 \times 10^4, 10^5, 5 \times 10^5$. The marginal disease probability is around 0.01. Again we varied the size of interaction at $0, 0.5$ and 1. The sizes of the second-stage case-control samples are approximately 400, 2000, 10000, with a 1:1 case-control ratio. Table 2 shows the empirical correlations of various statistics in 20,000 simulated datasets. With sample size 400 and beyond, the correlation of various pairs of statistics is fairly close to zero as theory predicts. The pair of statistics in the counterexample display a consistent correlation around -0.7, confirming that using such statistics in an adaptive testing procedure would yield an inflated type I error.

In Table 3, we assess the empirical FWER over 1000 simulated datasets, each containing 10,000 SNPs for 1000 subjects. The minor allele frequencies were randomly generated from a uniform distribution from 0.1 to 0.5. The SNPs are either independent ($\rho = 0.5$) or have a serial correlation ($\rho = 0.5$). A binary treatment assignment was randomly generated by $Ber(0.5)$. The binary disease status was generated by the logistic model (8) with $\boldsymbol{\gamma} = (-4, 0, 0, 0)$ or $(-4, 0, \log(1.5), 0)$, the latter assumes a mild treatment effect. Various two-stage procedures in **Example 2a, 2b** and **3** were applied to screen for main effects or interactions with α_0= $0.001, 0.01,$ or 0.1, and $\alpha = 0.05$. For each pair of statistics, we alternated the screening statistics and the testing statistics. For example, we could use interactions as screening criteria for testing main effects. Across all settings, FWERs are controlled at the level of 0.05; as expected, none of them falls outside of 95% confidence intervals of 0.05.

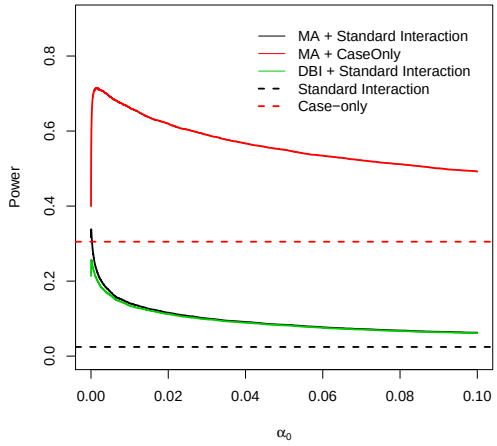
We study in Figure 1 and Figure 2 the power performance for the procedures in simulations with half a million independent SNPs, one of which is true alternative hypothesis and the rest are all nulls. The SNPs were generated based on Hardy-Weinberg equilibrium. The simulation was designed with a case-control study nested in a randomized clinical trial (**Example 2b** and **3**). The

Table 2. The empirical correlations of various statistics in **Example 2a-3** with case-control sampling among 20,000 simulated case-control data.The parameters are from the following models: β_1: logit$\{\mathsf{E}[Y|X]\} = \beta_0 + \beta_1 X$; γ_3: logit$\{\mathsf{E}[Y|X,Z]\} = \gamma_0 + \gamma_1 X + \gamma_2 Z + \gamma_3 XZ$; δ_1: logit$\{\mathsf{E}[Z|X,Y=1]\} = \delta_0 + \delta_1 X$; τ_1: logit$\{\mathsf{E}[Z|X]\} = \tau_0 + \tau_1 X$; ν_1: logit$\{\mathsf{E}[Z|X,Y=0]\} = \nu_0 + \nu_1 X$.

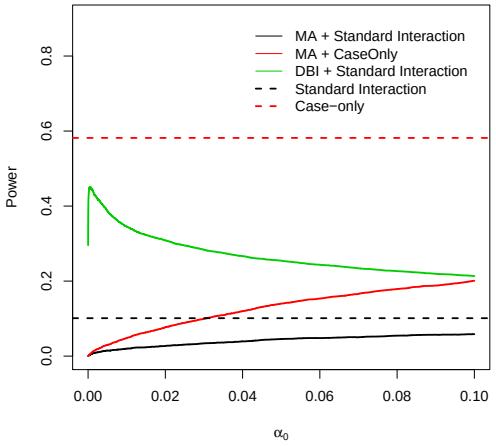
		$n = 400$		$n = 2000$		$n = 10000$	
		Estimator	Z-score	Estimator	Z-score	Estimator	Z-score
$\gamma_3 = 0$	$\beta_1 : \gamma_3$	-0.004	-0.003	0.002	0.003	0.003	0.003
	$\beta_1 : \delta_1$	-0.004	0.002	0.001	0.001	-0.001	-0.001
	τ_1: γ_3	$-$2e-4	-0.001	0.003	0.003	0.001	0.001
	τ_1: γ_1	0.001	0.002	-0.004	-0.004	-0.003	-0.004
	ν_1: γ_3	-0.702	-0.703	-0.706	-0.706	-0.709	-0.709
$\gamma_3 = 0.5$	$\beta_1 : \gamma_3$	-0.008	-0.006	0.005	0.005	0.006	0.006
	$\beta_1 : \delta_1$	-0.004	-0.004	0.002	0.003	0.008	0.008
	τ_1: γ_3	-0.018	-0.019	2e-4	0.001	-0.004	-0.004
	τ_1: γ_1	0.016	0.019	-0.001	-0.001	0.007	0.007
	ν_1: γ_3	-0.720	-0.722	-0.719	-0.726	-0.719	-0.719
$\gamma_3 = 1$	$\beta_1 : \gamma_3$	-0.010	-0.005	0.005	0.006	-0.016	-0.016
	$\beta_1 : \delta_1$	-0.009	0.004	0.013	0.014	-0.012	-0.011
	τ_1: γ_3	-0.012	-0.011	0.005	0.005	-0.004	-0.004
	τ_1: γ_1	0.006	0.010	0.004	0.005	0.002	0.002
	ν_1: γ_3	-0.701	-0.705	-0.705	-0.706	-0.701	-0.701

Table 3. The empirical FWER in 1000 simulations. 10,000 SNPs are simulated for 1000 cases and 1000 controls by case-control sampling. SNPs are generated either independently or with serial correlation 0.5. The disease status is generated by the logistic model (8). The parameters are from the following models: β_1, logit$\{\mathsf{E}[Y|X]\} = \beta_0 + \beta_1 X$; γ_3, logit$\{\mathsf{E}[Y|X,Z]\} = \gamma_0 + \gamma_1 X + \gamma_2 Z + \gamma_3 XZ$; δ_1, logit$\{\mathsf{E}[Z|X,Y=1]\} = \delta_0 + \delta_1 X$; τ_1, logit$\{\mathsf{E}[Z|X]\} = \tau_0 + \tau_1 X$. $\beta_1 \to \gamma_3$ indicates using β_1 for screening, and γ_3 for testing.

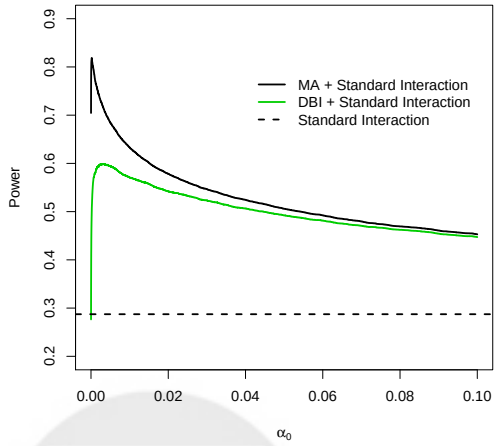
		$\boldsymbol{\gamma} = (-4, 0, 0, 0)$			$\boldsymbol{\gamma} = (-4, 0, \log(1.5), 0)$		
		$\alpha_0 = 0.001$	$\alpha_0 = 0.01$	$\alpha_0 = 0.1$	$\alpha_0 = 0.001$	$\alpha_0 = 0.01$	$\alpha_0 = 0.1$
$\rho = 0$	$\beta_1 \to \gamma_3$	0.037	0.045	0.050	0.050	0.046	0.043
	$\gamma_3 \to \beta_1$	0.043	0.036	0.042	0.048	0.045	0.054
	$\beta_1 \to \delta_1$	0.042	0.051	0.043	0.041	0.030	0.037
	$\delta_1 \to \beta_1$	0.044	0.049	0.034	0.038	0.043	0.060
	$\alpha_1 \to \gamma_3$	0.042	0.047	0.032	0.042	0.035	0.034
	$\alpha_1 \to \gamma_1$	0.036	0.033	0.047	0.036	0.036	0.026
$\rho = 0.5$	$\beta_1 \to \gamma_3$	0.044	0.046	0.046	0.053	0.040	0.039
	$\gamma_3 \to \beta_1$	0.046	0.048	0.050	0.043	0.040	0.047
	$\beta_1 \to \delta_1$	0.052	0.050	0.043	0.051	0.051	0.044
	$\delta_1 \to \beta_1$	0.057	0.044	0.051	0.043	0.051	0.045
	$\alpha_1 \to \gamma_3$	0.041	0.034	0.039	0.046	0.045	0.046
	$\alpha_1 \to \gamma_1$	0.047	0.037	0.048	0.045	0.043	0.040



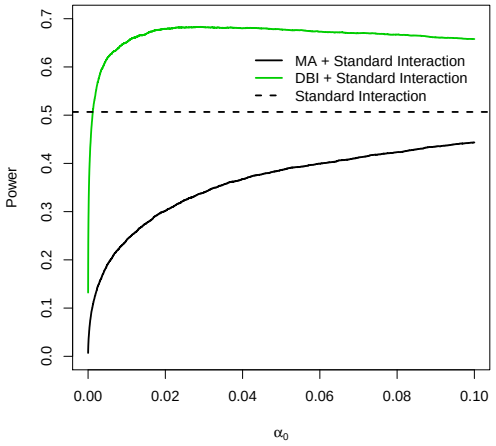
(a) $\boldsymbol{\gamma} = (-4, 0, 0, \log 2)$



(b) $\boldsymbol{\gamma} = (-4, -0.5 \log 2, 0, \log 2)$

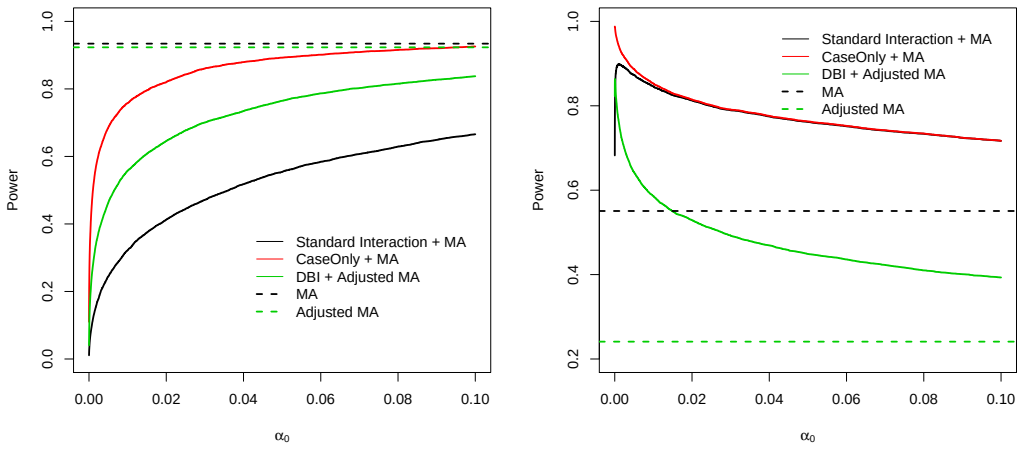


(c) $\boldsymbol{\gamma} = (-2, 0, -\log 1.5, \log 2)$



(d) $\boldsymbol{\gamma} = (-2, -\log 1.8, \log 1.5, \log 2)$

Fig. 1. Power to detect the treatment-SNP interaction using two-stage procedures in simulations. Assuming one of 1 million SNPs carries disease risk, the risk model takes form of logit$\{\mathsf{E}[Y|X,Z]\} = \gamma_0 + \gamma_1 X + \gamma_2 Z + \gamma_3 XZ$. The parameters are from the following models: MA - β_1, logit$\{\mathsf{E}[Y|X]\} = \beta_0 + \beta_1 X$; Standard Interaction - γ_3, logit$\{\mathsf{E}[Y|X,Z]\} = \gamma_0 + \gamma_1 X + \gamma_2 Z + \gamma_3 XZ$; CaseOnly - δ_1, logit$\{\mathsf{E}[Z|X, Y=1]\} = \delta_0 + \delta_1 X$; DBI - τ_1, logit$\{\mathsf{E}[Z|X]\} = \tau_0 + \tau_1 X$.



(a) $\boldsymbol{\gamma} = (-4, \log 1.5, 0, \log 1.5)$ (b) $\boldsymbol{\gamma} = (-4, \log 1.8, \log 0.8, -\log 2)$

Fig. 2. Power to detect the SNP main effect using two-stage procedures in simulations. Assuming one of 1 million SNPs carries disease risk, the risk model takes form of logit$\{\mathsf{E}[Y|X,Z]\} = \gamma_0 + \gamma_1 X + \gamma_2 Z + \gamma_3 XZ$. The parameters are from the following models: MA - β_1, logit$\{\mathsf{E}[Y|X]\} = \beta_0 + \beta_1 X$; γ_3, Standard Interaction - logit$\{\mathsf{E}[Y|X,Z]\} = \gamma_0 + \gamma_1 X + \gamma_2 Z + \gamma_3 XZ$; Case-Only - δ_1, logit$\{\mathsf{E}[Z|X,Y=1]\} = \delta_0 + \delta_1 X$; DBI - τ_1, logit$\{\mathsf{E}[Z|X]\} = \tau_0 + \tau_1 X$; Adjusted MA - η_1, logit$\{\mathsf{E}[Y|X,Z]\} = \eta_0 + \eta_1 X + \eta_2 Z$

randomization ratio to the treatment arm and the control arm is 1:1, and the case-control sampling ratio is also 1:1. The power was computed as the percentage of simulations where the SNP with the signal was declared to be significant in 10,000 simulations. With almost half a million null tests, the number of tests passing the first stage criterion would vary little from $m\alpha_0$, so we used the second stage cut-off $C_{1-\alpha/(2m\alpha_0)}$ as if it was fixed in every simulation. We devised various parameter settings and a range of α_0 to study the full spectrum of operating characteristics.

Figure 1 shows the power to detect an interaction between the signal SNP (X) and the treatment Z. Denote by N the total sample size, n the sample size for the case-control sample, p is the minor allele frequency of the target SNP, $\boldsymbol{\gamma}$ is the vector of parameters in the model (8) that generates the data. Across 4 graphs, the red curve is the power of screening main effects by $\hat{\beta}_1$ before testing interactions using the case-only estimator ($\hat{\delta}_1$), for α_0 valued from 0.00001 to 0.1; the black curve is the power of screening by main effects $\hat{\beta}_1$ before testing interactions using the standard estimator ($\hat{\gamma}_3$); the green line is the power of screening by DBI $\hat{\tau}_1$ before testing interaction ($\hat{\gamma}_3$). The red horizontal dotted line is the power of testing for case-only interactions by Bonferroni correction for half an million test; the black horizontal dotted line is the power of testing for standard interactions by Bonferroni correction for half a million tests. These two dotted lines provide benchmarks for a comparison of power.

In Figure 1(a), $N = 50,000, n \approx 1000, p = 0.1, \boldsymbol{\gamma} = (-4, 0, 0, \log 2)$. The marginal disease probability is around 0.02. Clearly, the case-only estimator provides a substantial lift of power on top of the standard interaction estimator. The two-stage procedure with MA and case-only estimator yields substantial power improvement over other procedures. The two-stage procedure with DBI and standard interaction estimators is slightly outperformed by first screening main effects. In Figure 1(b), $N = 50,000, n \approx 1000, p = 0.2, \boldsymbol{\gamma} = (-4, -0.5\log 2, 0, \log 2)$, so that the interaction between the treatment and the SNP is qualitative, i.e., the sign of the SNP effect differs in different treatment groups. The main effect of the SNP is negligible, which lead to poor power performance of the two-stage procedures using main effects for an initial screen. The two-stage procedure using DBI avoids the cancellation of opposite SNP effects, thus yields a noticeable power gain over the two-stage procedures using main effects. The best procedure in this scenario is, however, the case-only estimator with Bonferroni correction (the dotted red line), suggesting that in rare disease settings, case-only estimators are preferable wherever possible. We show in **Example 3** that DBI can be extended to settings with common diseases, as long as there is a treatment effect. In Figure 1(c), $N = 50,000, n \approx 1000, p = 0.2, \boldsymbol{\gamma} = (-2, 0, -\log 1.5, \log 2)$. The disease probability is roughly 10% and there is a mild treatment effect and a strong interaction effect. Case-only estimators are no longer eligible, yet the two-stage procedure using MA to filter still performs better than the two-stage procedure using DBI, both improving upon the standard interactions. In Figure 1(d), $N = 50,000, n \approx 1000, p = 0.3, \boldsymbol{\gamma} = (-2, -\log 1.8, \log 1.5, \log 2)$. There is a qualitative interaction and thus a small marginal SNP effect, hence the two-stage procedure using MA to screen has little power to detect the true interaction, while screening by DBI in this scenario yields much better power.

Figure 2 shows the power to detect the SNP effect using two-stage procedures either using interaction or DBI as the screening criterion. The labels for various procedures are similar to those in Figure 1, except that the SNP effect, either marginal or adjusted for treatment, is of interest after screening. In Figure 2(a), $N = 50,000, n \approx 1000, p = 0.1, \boldsymbol{\gamma} = (-4, \log 1.5, 0, \log 1.5)$. Clearly in this setting, the best procedure is to test the marginal effect directly, two-stage procedures using either interaction or DBI in screening does not perform comparably. The reason is that testing for interaction, even with case-only estimator, is more costly to sample size than testing for main effect, therefore screening by interactions for marginal effects does not help when there is a moderate size of interaction. In Figure 2(b), $N = 50,000, n \approx 1000, p = 0.1, \boldsymbol{\gamma} = (-4, -\log 1.8, \log 0.8, \log 2)$.

Since the marginal SNP effect is small and the interaction is fairly big, using interaction or DBI to find marginal effect yields much better power than one-stage Bonferroni correction for MA or adjusted MA.

Taking collectively, each of the proposed two-stage procedures has unique niche in power performance. There are situations where none of them improves power upon the one-stage all-SNP Bonferroni test, see Figure 1(b) and Figure 2(a). It is useful to screening by MA for interactions, since testing main effect is generally more powerful than testing interaction, so that a SNP having interaction is likely to have high probability to pass the filter. However when there are qualitative interactions, which might be not common, the two-stage procedure using MA does not perform well since the marginal effect is small. The DBI criterion offers an alternative screening procedure which can improve power in this setting, see Figure 1d. It appears less useful to screen for marginal effects by interactions or DBI, unless there are qualitative interactions (Figure 2(b)). In any case the marginal effect is usually of primary interest, so we may well have good power to test all SNPs.

4.1. optimal α_0

Clearly in Figure 1 and 2, an optimal α can be achieved, since more tests passing the first stage will incur more penalty by Bonferroni correction. For simulations with half a million SNPs in Figure 1 and 2, the optimal α_0 for power performance is fairly small, in the range of $0.0001 \sim 0.001$. Analytically, it is possible to find the optimal α_0 for a hypothesized disease risk model and a sampling plan, since power to detect a feature can be written out approximately,

$$\Pr(T_j^0 \in \Gamma_j^0, T_j \in \Gamma_j | H_{1j}) \approx \Pr(T_j^0 > C_{1-\alpha_0/2} | H_{1j}) \Pr(T_j > C_{1-\alpha/(2m\alpha_0)} | H_{1j}).$$

With prior assumptions on an alternative hypothesis, the asymptotic distribution of T_j^0 and T_j under H_{1j} can be derived for a fixed sample size, so that an optimal α_0 can be computed. Alternatively, power simulations can be conducted using our R-package `powerGWASinteraction` from CRAN to obtain the optimal α_0. In any case, it is important to note that we need to set α_0 fixed before performing hypothesis testing. Data-adaptive selection of α_0 will undermine error control.

5. Data application

The Women's Health Initiative (WHI), one of the largest studies of postmenopausal women's health in the U.S., is composed of four randomized clinical trials (CT) and an observational study (OS). An elevated invasive breast cancer risk was found among women assigned to estrogen plus progestin, with suggestive evidence of risk reduction among women assigned either to estrogen-alone or to a low-fat dietary pattern. To discover the genetic variants that may influence the risk, perhaps jointly with the interventions, WHI launched a genome-wide association study with a three stage design (Prentice and Qi, 2006). In the third stage, a total of 9039 SNPs were selected from previous stages or other studies, and were genotyped among 2,166 invasive breast cancer cases in the CT and 1:1 matched controls. Primary analyses have been presented recently (Prentice et al., 2009, 2010). Seven SNPs in the fibroblast growth factor receptor 2 (*FGFR2*) met criteria for genome-wide significance. Recognizing limited power in detecting interactions, the investigators focused the search for treatment-genotype interactions to the top seven SNPs ranked by MA (Prentice et al., 2009, 2010), as well as a number of SNPs that have been reported in the literature to be associated with breast cancer (Huang et al., 2010). Since invasive breast cancer is a rare event in the study, the investigators used the case-only estimators described in **Example 4**. A number of SNPs showed suggestive evidence of interactions with one or more interventions. The analyses presented here are exploratory and supplementary to the findings from these primary analyses.

Table 4. The results of two-stage procedures applied to the WHI GWAS, using DBI as a screening criterion. Four SNPs out of the top 50 SNPs ranked by the DBI criterion reach statistical significance in testing the adjusted marginal SNP effect.

		rs7705343		rs13159598		rs9790879		rs4415084	
		OR	p-value	OR	p-value	OR	p-value	OR	p-value
E-alone	$\hat{\alpha}_1$	1.5823	0.0006	1.5217	0.0014	1.4701	0.0035	1.4269	0.0063
All trials	$\hat{\eta}_1$	1.1672	0.0006	1.1653	0.0007	1.1649	0.0007	1.1695	0.0005
E-alone	$\hat{\delta}_1$	1.5231	0.0298	1.4657	0.0444	1.3842	0.0907	1.3832	0.0936

Our results justified the focused search for interactions in a subset of SNPs ranked by top marginal association. In addition, we also explored the two-stage procedures using the DBI criterion to look for significant marginal effects and interactions (**Example 3**). This is done separately for each of the 4 randomized trials. In the first stage, we ranked SNPs by p-values for DBI. We tested for main effects and interactions for the top 50 SNPs ranked by DBI. Among the top 50 SNPs ranked by DBI in the E-alone trial, there are four SNPs that pass the Bonferroni correction for 50 SNPs in testing for adjusted SNP effect. Table 3 shows the parameter estimates of these 4 SNPs. The adjusted additive SNP effect $\hat{\eta}_1$ was estimated from case-control data for all 4 trials, adjusted for matching variables, important baseline predictors and randomization indicators. The effect size (odds ratio) is fairly modest around 1.16. In the E-alone trial, there seems to be a weak interaction between the SNPs and the treatment. The effect sizes of the interactions (case-only estimator $\hat{\delta}_1$) are around $1.4 \sim 1.5$.

Interestingly, these four SNPs are all located in the *MRPS30* gene which have been shown to have suggestive evidence of interaction with multiple clinical interventions (Huang et al., 2010), though the findings there are guided by prior studies. None of them would reach the genome-wide significance level for either marginal effect or interaction, yet they reach the FWER level of 0.05 for marginal association by our two-stage procedure. The reason might be that these four SNPs have weak main effects and weak interactions with the E-alone intervention. The DBI criteria seem to synthesize these weak effects and prioritizes them for further testing. This data example suggests that two-stage procedures can be used as data-adaptive tool, as opposed to candidate genes from prior studies, for discovering novel genes affecting disease risk. Certainly this search strategy only serves as a supplement to the standard one-stage Bonferroni test, since it missed the seven SNPs in the *FGFR2* gene.

6. Discussion

We studied conditions that are required to maintain strong control of FWER for a class of two-stage hypothesis testing procedures previously proposed. We provided a unified approach to prove asymptotic independence by estimating equation theory. Two types of screening statistics are discussed, one is based on marginal association (MA) and the other is based on deviation from baseline independence (DBI). In the majority of simulation settings, one or both of the proposed procedures outperform the standard one-stage testing with Bonferroni correction.

The impact of these results is profound to discovering baseline features that influence treatment effect in a RCT, whether they are low or high-dimensional. In randomized clinical trials, subgroup analyses are heavily criticized as having low power to detect interactions in general, which could be further exacerbated by data-adaptive procedures. Our results suggest that one could select predictors with evidence of MA or with evidence of DBI to test for interactions, without spending type I error in finding these candidate predictors. We expect that the two-stage procedures will be most

useful to discovering interaction in pharmacogenetic studies, where it is almost certain that these studies have low power to detect interactions. In fact, we are involved in a GWA study within the WHI clinical trials as part of the Genomics and Randomized Trials Network (GARNET) program (`www.garnetstudy.org`) in which we plan to use this strategy. Using interactions or DBI to screen for marginal effect appears less powerful in general, at most serving as exploratory supplement to primary analyses. In any case, it is worthwhile to attempt these two-stage procedures and the standard 1-stage Bonferroni test, while splitting FWER among them.

The asymptotic independence of the two-stage test statistics can be extended beyond the examples presented. Indeed, for any estimator of parameters in a generalized linear model (GLM), the score function can be written in a form of $\mathbf{X}(\mathbf{Y} - \mathrm{E}[\mathbf{Y}|\mathbf{X}]) = 0$. The proof in Section 3 applies immediately to the independence of estimators of MA and interaction. Thus the proposed two-stage procedure can be applied to any GLM outcome, e.g., Poisson counts. We show in **Example 4** that approximation can be made for survival data in RCT with a rare event, so that the independence can be carried over to Cox partial likelihood. For survival data with a common event, however, the proposed two-stage procedures do not work in general. In future study, we intend to investigate the possibility of asymptotic independence using parametric survival analysis.

7. Acknowledgment

The authors of this manuscript were supported in part by U01 AI068615, U01 AI068617, U01 HG005152, R01 CA90998, P01 CA53996. The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts N01WH22110, 24152, 32100-2, 32105-6, 32108-9, 32111-13, 32115, 32118-32119, 32122, 42107-26, 42129-32, and 44221.

8. Appendix

8.1. Proof of Theorem 1

PROOF. With mild regularity conditions, standard estimating equation theory imply that

$$\begin{aligned}\sqrt{n}(\hat{\vartheta}_k - \vartheta_k) &\to_d \mathcal{N}(0, V_1),\\ \sqrt{n}(\hat{\theta}_j - \theta_j) &\to_d \mathcal{N}(0, V_2),\end{aligned}$$

where V_1 and V_2 are asymptotic variances which can be estimated by their respective empirical averages, $\hat{V}_1$ and $\hat{V}_2$. By the Law of Large Number, $\hat{V}_1$ and $\hat{V}_2$ are consistent estimators.

Since

$$\mathrm{Cov}\left(\sqrt{n}(\hat{\vartheta}_k - \vartheta_k), \sqrt{n}(\hat{\theta}_j - \theta_j)\right) \to 0 \qquad \text{as } n \to \infty,$$

and

$$\begin{aligned}\frac{\sqrt{n}(\hat{\vartheta}_k - \vartheta_k)}{\sqrt{V_1}} - \frac{\sqrt{n}(\hat{\vartheta}_k - \vartheta_k)}{\sqrt{\hat{V}_1}} &= o_p(1),\\ \frac{\sqrt{n}(\hat{\theta}_j - \theta_j)}{\sqrt{V_2}} - \frac{\sqrt{n}(\hat{\theta}_j - \theta_j)}{\sqrt{\hat{V}_2}} &= o_p(1),\end{aligned}$$

we derive that

$$\begin{pmatrix} \frac{\sqrt{n}(\hat{\vartheta}_k - \vartheta_k)}{\sqrt{\hat{V}_1}} \\ \frac{\sqrt{n}(\hat{\theta}_j - \theta_j)}{\sqrt{\hat{V}_2}} \end{pmatrix} \to_d \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right).$$

Since $T_k^0 = \hat{\vartheta}_k / \sqrt{\hat{V}_1/n}$ and $T_j = \hat{\theta}_j / \sqrt{\hat{V}_2/n}$, this implies T_k^0 and T_j are asymptotically independent $\forall j, k$ under the global null hypothesis.

We now prove the main result. Observe that m_0 is a random variable taking values at $0, 1, ..., m$. Let $\mathbf{I}_k$ denote a set of k distinct integers from $\{1, ..., m\}$, that index the hypotheses passing the filter in the first stage. Let $\mathcal{I}_k$ denote the collection (set) of all possible $\mathbf{I}_k$. Note that

$$
\begin{aligned}
& \lim_{n\to\infty} \Pr\left\{\cup_{j\in\mathbf{J}}(T_j^0 \in \Gamma_j^0 \cap T_j \in \Gamma_j)|H_{0j}, K_{0j}\right\} \\
= \; & \lim_{n\to\infty} \sum_{k=1}^{m} \sum_{\mathbf{I}_k\in\mathcal{I}_k} \Pr\left\{\left[\cap_{j\in(\mathbf{J}\cap\mathbf{I}_k)}(T_j^0 \in \Gamma_j^0|K_{0j})\right] \cap \left[\cup_{j\in(\mathbf{J}\cap\mathbf{I}_k)}(T_j \in \Gamma_j|H_{0j})\right]\right\} \\
= \; & \lim_{n\to\infty} \sum_{k=1}^{m} \sum_{\mathbf{I}_k\in\mathcal{I}_k} \Pr\left\{\cap_{j\in(\mathbf{J}\cap\mathbf{I}_k)}(T_j^0 \in \Gamma_j^0|K_{0j})\right]\right\} \Pr\left\{\cup_{j\in(\mathbf{J}\cap\mathbf{I}_k)}(T_j \in \Gamma_j|H_{0j})\right\} \qquad (10) \\
\leq \; & \lim_{n\to\infty} \sum_{k=1}^{m} \sum_{\mathbf{I}_k\in\mathcal{I}_k} \Pr\left\{\cap_{j\in(\mathbf{J}\cap\mathbf{I}_k)}(T_j^0 \in \Gamma_j^0|K_{0j})\right]\right\} \left\{\sum_{j\in(\mathbf{J}\cap\mathbf{I}_k)} \Pr\left(|T_j| \geq C_{1-\alpha/2k}|H_{0j}\right)\right\} \qquad (11) \\
\leq \; & \lim_{n\to\infty} \sum_{k=1}^{m} \sum_{\mathbf{I}_k\in\mathcal{I}_k} \Pr\left\{\cap_{j\in(\mathbf{J}\cap\mathbf{I}_k)}(T_j^0 \in \Gamma_j^0|K_{0j})\right]\right\} \alpha \qquad (12) \\
\leq \; & \alpha \qquad (13)
\end{aligned}
$$

The equality (10) holds by the asymptotic independence, the inequality (11) uses the Bonferroni inequality, (12) uses the fact that the size of the set $\{j : j \in (\mathbf{J} \cap \mathbf{I}_k)\}$ is less than or equal to k, and (13) holds because

$$\sum_{k=0}^{m} \sum_{\mathbf{I}_k\in\mathcal{I}_k} \Pr\left\{\cap_{j\in(\mathbf{J}\cap\mathbf{I}_k)}(T_j^0 \in \Gamma_j^0|K_{0j})\right]\right\} = 1$$

if we denote by $\mathbf{I}_0$ the empty set. □

8.2. Proof of Theorem 2

PROOF. The proof of asymptotic independence of T_j^0 and T_j under the global null hypothesis is same as that in **Theorem 1**. Observe that

$$m_0 = \sum_{j=1}^{m} I(T_j^0 \in \Gamma_j^0|K_j^0).$$

Unless T_j^0s are independent, $\mathrm{E}[\frac{m0}{m}]$ is generally not equal to α_0. However if $\frac{m_0}{m} \to_p \alpha_0'$, we can prove the main result as followed:

$$
\begin{aligned}
& \lim_{m\to\infty}\lim_{n\to\infty} \Pr\left\{\cup_{j\in\mathbf{J}}(T_j^0 \in \Gamma_j^0 \cap T_j \in \Gamma_j)|H_{0j}, K_{0j}\right\} \\
\leq \; & \lim_{m\to\infty}\lim_{n\to\infty} \sum_{j=1}^{J} \Pr\left\{(T_j^0 \in \Gamma_j^0 \cap T_j \in \Gamma_j)|H_{0j}, K_{0j}\right\} \qquad (14) \\
= \; & \lim_{m\to\infty}\lim_{n\to\infty} \sum_{j=1}^{J} \Pr\left(T_j^0 \in \Gamma_j^0|K_{0j}\right) \Pr\left(T_j \in \Gamma_j|H_{0j}\right) \qquad (15)
\end{aligned}
$$

$$\begin{aligned} &= \lim_{m\to\infty}\lim_{n\to\infty}\left\{\frac{1}{m}\sum_{j=1}^{J}\Pr\left(T_j^0\in\Gamma_j^0|K_{0j}\right)\right\}\frac{m\alpha}{m_0} \\ &\leq \alpha_0\frac{\alpha}{\alpha_0} \qquad (16) \\ &= \alpha. \end{aligned}$$

The inequality (14) uses Bonferroni inequality, (15) uses the asymptotic independence of T_j^0 and T_j, (16) holds because

$$\frac{1}{m}\sum_{j=1}^{J}\Pr\left\{T_j^0\in\Gamma_j^0|H_{0j}\right\}\leq\alpha_0,$$

and by Slutsky's Theorem

$$\frac{m\alpha}{m_0}\to_p\frac{\alpha}{\alpha_0}.$$

□

8.3. Proof of Lemma 1

PROOF. (a) If the disease is rare, $\Pr(Z|X,R=1)\neq\Pr(Z|R=1)$ implies $\Pr(Z|X,Y=1)\neq\Pr(Z|Y=1)$, hence $\Pr(Y|X,Z)\neq\Pr(Y|Z)$.

(b) If the disease is common and $\Pr(Y|Z,X)=\Pr(Y|Z)$,

$$\Pr(X|Y,Z)=\frac{\Pr(Y|X,Z)\Pr(X|Z)}{\Pr(Y|Z)}=\Pr(X|Z)=\Pr(X).$$

This implies $\Pr(X|Y)=\Pr(X)$. Now

$$\begin{aligned} \Pr(Z,X|Y) &= \frac{\Pr(Y|Z,X)\Pr(Z)\Pr(X)}{\Pr(Y)} \\ &= \frac{\Pr(Y|Z)\Pr(Z)\Pr(X)}{\Pr(Y)} \\ &= \Pr(Z|Y)\Pr(X) \\ &= \Pr(Z|Y)\Pr(X|Y). \end{aligned}$$

Hence $Z\perp X|Y$. Since the sampling depends on Y soly, this implies that Z and X are independent in the selected case-control samples, i.e.,

$$\Pr(Z|X,R=1)=\Pr(Z|R=1).$$

Note that if $\Pr(Y|X,Z)=\Pr(Y|X)$, the same argument leads to the conclusion that Z and X are independent in the selected case-control sample. So we need the condition $\Pr(Y|X,Z)\neq\Pr(Y|X)$ to ensure that it is possible to observe $\Pr(Z|X,R=1)\neq\Pr(Z|R=1)$. □

References

Albert, P. S., D. Ratnasinghe, J. Tangrea, and S. Wacholder (2001). Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology 154*, 587–693.

Browning, S. R. and B. L. Browning (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics 81*, 1084–1097.

Casella, G. and R. L. Berger (2002). *Statistical Inference*. Pacific Grove, CA: Duxbury.

Chatterjee, N. and R. J. Carroll (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika 92*, 399–418.

Dai, J. Y., M. LeBlanc, N. L. Smith, B. Psaty, and C. Kooperberg (2009a). SHARE: an adaptive algorithm to select the most informative set of snps for candidate genetic association. *Biostatistics 10*, 680–693.

Dai, J. Y., M. LeBlanc, and C. Kooperberg (2009b). Semiparametric estimation exploiting covariate independence in two-phase randomized trials. *Biometrics 65*, 178–187.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand J Statist 6*, 65–70.

Huang, Y., D. G. Ballinger, U. Peters, D. A. Hinds, D. R. Cox, E. Beilarz, R. T. Chlebowski, J. E. Rossouw, A. McTienan, T. Rohan, and R. L. Prentice (2010). Variation in the MRPS30 region and the effects of hormonal and nutritional interventions on postmenopausal breast cancer incidence. pp. Submitted.

Kooperberg, C. and M. LeBlanc (2008). Increasing the power of identifying gene-gene interactions in genome-wide association studies. *Genetic Epidemiology 32*, 255–263.

Kraft, P. (2009). Genetic risk prediction - are we there yet. *New England Journal of Medicine 360*, 1701–1703.

Li, Y. and G. R. Abecasis (2006). Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *American Journal of Human Genetics S79*, 2290.

Lin, D. Y. and L. J. Wei (1989). The robust inference for the Cox proportional hazard models. *Journal of American Statistical Association 84*, 1074–1078.

Lin, D. Y. and D. Zeng (2006). Likelihood-based inference on haplotype effects in genetic association studies. *Journal of American Statistical Association 101*, 89–104.

Millstein, J., D. V. Conti, F. D. Gilliland, and J. W. Gauderman (2006). A testing framework for identifying susceptibility genes in the presence of epistasis. *American Journal of Human Genetics 78*, 15–27.

Murcray, C. E., J. P. Lewinger, and J. W. Gauderman (2008). Gene-environment interaction in genome-wide association studies. *American Journal of Epidemiology 169*, 219–226.

Newey, W. K. and J. Powell (1990). Efficient estimation of linear and type i censored regression models under conditional quantile restrictions. *Econometric Theory 6*, 295–317.

Piegorsch, W. W., C. R. Weinberg, and J. A. Taylor (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. *Statistics in Medicine 13*, 153–162.

Prentice, R. L., Y. Huang, D. A. Hinds, U. Peters, D. R. Cox, E. Beilharz, R. T. Chlebowski, J. E. Rossouw, B. Caan, and D. Ballinger (2010). Variation in the FGFR2 gene and the effect of low-fat dietary pattern on invasive breast cancer. *Cancer Epidemiology, Biomarkers & Prevention 19*, 74–79.

Prentice, R. L., Y. Huang, D. A. Hinds, U. Peters, M. Pettinger, D. R. Cox, E. Beilharz, R. T. Chlebowski, J. E. Rossouw, B. Caan, and D. Ballinger (2009). Variation in the FGFR2 gene and the effects of postmenopausal hormone therapy on invasive breast cancer. *Cancer Epidemiology, Biomarkers & Prevention 18*, 3079–3085.

Prentice, R. L. and R. Pyke (1979). Logistic disease incidence models and case-control studies. *Biometrika 66*, 403–411.

Prentice, R. L. and L. Qi (2006). Aspects of the design and analysis of high-dimensional snp studies for disease risk estimation. *Biostatistics 7*, 339–354.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association 89*, 846–866.

Satagopan, J. M., E. S. Venkatraman, and C. B. Begg (2004). Two-stage designs for gene-disease studies with sample size constraints. *Biometrics 60*, 589–597.

Skol, A. D., L. J. Scott, G. R. Abecasis, and M. Boehnke (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet 38*, 209–213.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature 437*, 1299–1320.

Umbach, D. M. and C. R. Weinberg (1997). Designing and analyzing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine 16*, 1731–1743.

Vittinghoff, E. and D. C. Bauer (2006). Case-only analysis of treatment-covariate interactions in clinical trials. *Biometrics 62*, 769–776.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1–25.

White, H. (2001). *Asymptotic theory for econometricsians*, Chapter 3. Howard House, Wagon Lane, Bingley BD16 1WA, UK: Emerald Group Publishing Limited.