Biometrika (2012), **99**, 4, *pp.* 929–944
© 2012 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/ass044
Advance Access publication 25 September 2012

Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction

By JAMES Y. DAI, CHARLES KOOPERBERG, MICHAEL LEBLANC AND ROSS L. PRENTICE

Public Health Science Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, U.S.A.

jdai@fhcrc.org clk@fhcrc.org mleblanc@fhcrc.org rprentic@WHI.org

Summary

Several two-stage multiple testing procedures have been proposed to detect gene-environment interaction in genome-wide association studies. In this article, we elucidate general conditions that are required for validity and power of these procedures, and we propose extensions of two-stage procedures using the case-only estimator of gene-treatment interaction in randomized clinical trials. We develop a unified estimating equation approach to proving asymptotic independence between a filtering statistic and an interaction test statistic in a range of situations, including marginal association and interaction in a generalized linear model with a canonical link. We assess the performance of various two-stage procedures in simulations and in genetic studies from Women's Health Initiative clinical trials.

Some key words: Case-only estimator; Filtering; Gene-treatment interaction; Multiple testing; Pharmacogenetics; Randomization.

1. Introduction

Gene-environment interaction is increasingly of interest as it informs disease aetiology, treatment and prevention (Hunter, 2005); yet there are few successes in detecting gene-environment interaction in genome-wide association studies (Rothman et al., 2010). Several aspects of genome-wide gene-environment interaction contribute to the lack of success. Environmental exposures are more difficult to characterize, as individual exposure may be highly variable over time and possibly subject to measurement error. The standard variable-by-variable testing strategy typically has low power due to stringent significance rules that are needed to guard against false positives arising from millions of single-nucleotide polymorphisms. This power shortage is further exacerbated because detecting gene-environment interaction requires a much larger sample size than is needed for detecting marginal association.

At the single genetic variant level, efficient case-only estimators have been proposed for gene-environment interaction and generalized to case-control sampling or two-phase stratified sampling with common diseases (Piegorsch et al., 1994; Umbach & Weinberg, 1997; Chatterjee & Carroll, 2005; Dai et al., 2009); however, they are subject to bias due to deviation from gene-environment independence in observational studies (Albert et al., 2001). Empirical Bayes estimators have been proposed to combine case-only and case-control estimators (Mukherjee & Chatterjee, 2008), which involve a trade-off between efficiency and bias. These estimators improve efficiency in estimating interaction, although the power gain may not surmount Bonferroni correction for millions of tests in genome-wide association studies.

To this end, two-stage multiple testing procedures with independent filtering have been proposed for detecting gene-environment interactions (Kooperberg & LeBlanc 2008; Murcray et al., 2009). Rather than increasing the estimation efficiency of interaction, the idea is to filter out the majority of irrelevant genetic variants initially and only test for interactions among the promising ones. Two types of filtering statistics have been considered: marginal association of a genetic variant (Kooperberg & LeBlanc, 2008) and gene-environment association in the combined case-control sample (Murcray et al., 2009). The latter procedure has been compared with approaches that exploit gene-environment independence (Mukherjee et al., 2012; Cornelis et al., 2012) and extended to studies with case-parent trios (Gauderman et al., 2010). In all these procedures, the statistic used in the filtering stage is suggested to be independent of the statistic in the testing stage under the null hypothesis, so one may only need multiple testing correction for the tests that actually pass the filtering, thereby potentially improving power. Theoretical justification and generalization of these two-stage testing procedures, however, have not been elaborated.

Similar ideas have been considered recently in family-based genetic association studies (Van Steen et al., 2005; Ionita-Laza et al., 2007). Genetic variants are ranked by association evidence from the population-level data and then tested using the family-level data, either in a subset of top-ranking variants (Van Steen et al., 2005) or for all variants by weighted p-values (Ionita-Laza et al., 2007). In the context of microarray gene expression data, two-stage procedures using independent, unsupervised filtering statistics, such as overall variance, have been discussed in Bourgon et al. (2010). While conceptually connected, the two-stage procedures for detecting gene-environment interaction differ from these procedures in that the filtering criteria exploit outcome-dependent information in the same overall dataset, so that the independence of some filtering statistics, for instance filtering marginal genetic association for interaction, requires theoretical development.

In this article, we discuss general properties of two-stage procedures for detecting gene-environment interaction in genome-wide association studies. We prove the asymptotic independence of various filtering and testing statistics previously proposed or newly developed. Our work is motivated by genetic studies within the Women's Health Initiative clinical trials. These studies build on the extensive and unique resource of randomized, placebo-controlled trials in postmenopausal women and aim to assess genetic and environmental influence on a number of clinical endpoints, e.g., cardiovascular events and incident diabetes. The exposure, postmenopausal hormone therapy or dietary intervention, was randomized, so these studies offer gold-standard data for studying gene-environment interaction. While all procedures developed for gene-environment interaction can be used, gene-treatment independence dictated by randomization leads to important extensions of the two-stage procedures, such as use of case-only estimators in the testing stage.

2. Control of the familywise error rate by two-stage testing procedures

2·1. *A class of two-stage multiple testing procedures*

Consider a genetic study with n independent subjects drawn from a cohort based on a prespecified sampling plan. Let Y_i denote the outcome variable, and let $X_i = (X_{i1}, \ldots, X_{im})$ denote a collection of m genetic variants measured for the ith subject. There is a key environmental variable or a randomized intervention, denoted by Z_i, along with a list of known predictors or potential confounders W_i. For different subjects, the random variables (Y_i, X_i, Z_i, W_i) are independent and identically distributed. Let θ_j denote the gene-environment interaction between Z and X_j in a regression model. The goal is to test m null hypotheses $H_{0j} : \theta_j = 0$ against alternative hypotheses $H_{1j} : \theta_j \neq 0$.

The test statistic for $H_{0j}: \theta_j = 0$ is often formulated by asymptotically linear estimators (Newey & Powell, 1990; Robins et al., 1994) and scaled by its estimated standard error. An estimator $\hat{\theta}_j$ is asymptotically linear if $n^{1/2}(\hat{\theta}_j - \theta_j) = n^{-1/2}\sum_{i=1}^{n} B_{ij} + o_p(1)$ where $E(B_{ij}) = 0$ and $E(B_{ij}^{\mathrm{T}} B_{ij}) < \infty$. The function B_{ij} is referred to as the influence function of $\hat{\theta}_j$, in the sense of Casella & Berger (2002). By the central limit theorem and Slutsky's theorem, $n^{1/2}(\hat{\theta}_j - \theta_j)$ is asymptotically normal with mean zero and variance $E(B_{ij}^{\mathrm{T}} B_{ij})$. We define a Wald statistic $T_j = \hat{\theta}_j / \hat{\mathrm{var}}(\hat{\theta}_j)^{1/2}$ for testing H_{0j}.

Now consider a different set of hypothesis tests for the filtering step, $K_{0j}: \zeta_j = 0$ versus $K_{1j}: \zeta_j \neq 0$. Let $\hat{\zeta}_j$ denote an asymptotically linear estimator of ζ_j. A Wald statistic is formulated similarly as $T_j^0 = \hat{\zeta}_j / \hat{\mathrm{var}}(\hat{\zeta}_j)^{1/2}$. We call T_j^0 the filtering statistic, since it is potentially informative as a filter for testing H_{0j}.

We discuss the following two-stage testing procedure. Denote by α_0 a prespecified tuning parameter, with $0 < \alpha_0 < 1$, that defines the first-stage rejection region $\Gamma_j^0 = \{T_j^0 : |T_j^0| > \Phi^{-1}(1 - \alpha_0/2)\}$, where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution. Suppose that m_0 genetic variants pass the filter. Denote by α (with $0 < \alpha < 1$) the targeted familywise error rate that defines the second-stage rejection region $\Gamma_j = \{T_j : |T_j| > \Phi^{-1}\{1 - \alpha/(2m_0)\}$. We declare the jth test to be statistically significant if $T_j^0 \in \Gamma_j^0$ and $T_j \in \Gamma_j$.

2·2. *Pairwise asymptotic independence*

When hypothesis testing is high-dimensional and adjacent genetic variants are correlated, which are features of genome-wide association studies, we prove that asymptotic independence between the pairs of estimators $\hat{\zeta}_j, \hat{\theta}_j$ and weak dependence among genetic variants are sufficient to control the familywise error rate in the strong sense, even though Bonferroni correction is applied only to the second-stage testing. Strong control of the familywise error rate 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).

THEOREM 1. *If the asymptotic joint distribution of $\hat{\zeta}_j$ and $\hat{\theta}_j$ is multivariate Gaussian with zero asymptotic covariance, i.e.* $\mathrm{cov}\{n^{1/2}(\hat{\zeta}_j - \zeta_j), n^{1/2}(\hat{\theta}_j - \theta_j)\} \to 0$ *in probability for all $j \in \{1, \ldots, m\}$, and m_0/m converges to a constant α_0' in probability, then the proposed two-step procedure preserves the familywise error rate at the level α for large m and n in the strong sense; that is, for any nonempty index set $\mathcal{J} \subseteq \{1, \ldots, m\}$ and under the null hypotheses H_{0j} for all j,* $\lim_{m\to\infty} \lim_{n\to\infty} \mathrm{pr}\{\bigcup_{j\in\mathcal{J}} (T_j^0 \in \Gamma_j^0) \cap (T_j \in \Gamma_j)\} \leqslant \alpha$.

The proof is given in the Appendix. The conditions for obtaining $m_0/m \to \alpha_0'$ in probability are those required by the law of large numbers for correlated data. For instance, if

$$\mathrm{cov}\{I(T_j^0 \in \Gamma_j^0), I(T_k^0 \in \Gamma_k^0)\} \to 0$$

as $|j - k|$ gets large, then the law of large numbers for a sequence of $I(T_j^0 \in \Gamma_j^0)$ holds as $m \to \infty$ (White, 2000). This type of serial correlation is exactly the linkage disequilibrium pattern observed in the human genome (International HapMap Consortium, 2005).

We present a series of examples that use different independent filtering statistics to test for gene-environment interaction in genome-wide association studies. To establish the asymptotic joint distribution of $\hat{\zeta}_j$ and $\hat{\theta}_j$, it is necessary to study their behaviour under potentially misspecified models, since the model indexed by ζ_j may disagree with the model indexed by θ_j; for example, a model is misspecified if it only includes marginal association parameters when actually there are interactions. The true disease model for complex diseases may include multiple loci, multiple environmental exposures, and possibly multiple gene-gene and gene-environment interactions.

Maximum likelihood estimation under misspecified models was discussed in White (1982). Let θ be the vector of parameters in the model indexed by θ_j, and let ζ denote the vector of parameters in the model indexed by ζ_j. Let $\sum_{i=1}^n U_{1i} = 0$ be the set of estimating equations solved for $\hat\theta$, and let $\sum_{i=1}^n U_{2i} = 0$ be the set of estimating equations to be solved for $\hat\zeta$. Suppose that θ^* is the unique solution to the estimating equations $E(U_{1i}) = 0$, where E denotes expectation under the true distribution. Similarly, suppose that ζ^* is the unique solution to the estimating equations $E(U_{2i}) = 0$. Then $\hat\zeta \to \zeta^*$ almost surely and $\hat\theta \to \theta^*$ almost surely.

Let $A_1 = E(\partial U_{1i}/\partial\theta)$, $A_2 = E(\partial U_{2i}/\partial\zeta)$ and $B_{kk'} = E(U_{ki}U_{k'i})$ $(k, k' = 1, 2)$. Under suitable regularity conditions (White, 1982), it can be shown that $n^{1/2}(\hat\theta - \theta^*)$ and $n^{1/2}(\hat\zeta - \zeta^*)$ 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 U_{ki} is independent and identically distributed with zero mean, but for the same i, U_{1i} and U_{2i} are possibly correlated. The joint distribution of $\hat\zeta$ and $\hat\theta$ is established by the Cramer–Wold device. The limiting distribution of $\{n^{1/2}(\hat\theta - \theta^*),\ n^{1/2}(\hat\zeta - \zeta^*)\}$ is multivariate Gaussian with zero means and covariance matrix

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

To assess asymptotic independence of $\hat\zeta$ and $\hat\theta$, we need to evaluate the off-diagonal element of the covariance matrix, $A_1^{-1}B_{12}A_2^{-1}$. This provides a unified approach to proving asymptotic independence, as illustrated in the examples below.

2·3. *Filtering by marginal association*

In this section, we provide formal justification of the independence of marginal association and interaction in generalized linear models with a canonical link. This result holds quite generally for the parameter estimates in two nested generalized linear models with a canonical link function. To make the result more visible, we state the theorem in the general setting of two nested models, and leave the result for gene-environment interaction as a corollary.

THEOREM 2. *Let $(Y_i, V_{i1}, \ldots, V_{ip})$ $(i = 1, \ldots, n)$ denote independent and identically distributed random variables sampled from a joint probability function $\mathcal{P}$, where Y is an outcome variable in a generalized linear model with a canonical link function g, and $(V_{i1}, \ldots, V_{ip})$ are p covariates. Let $(V_{i1}, \ldots, V_{iq})$, with $q < p$, be the first q covariates in the set $(V_{i1}, \ldots, V_{ip})$. Consider two nested generalized linear models*

$$g\{E(Y \mid V_1, \ldots, V_q)\} = \beta_0 + \sum_{j=1}^{q} \beta_j V_j, \tag{2}$$

$$g\{E(Y \mid V_1, \ldots, V_p)\} = \gamma_0 + \sum_{j=1}^{p} \gamma_j V_j. \tag{3}$$

Under regularity conditions for maximum likelihood estimation under misspecified models, the maximum likelihood estimators $(\hat\beta_0, \ldots, \hat\beta_q)$ and $(\hat\gamma_{q+1}, \ldots, \hat\gamma_p)$ are asymptotically independent.

COROLLARY 1. *Let Y be an outcome variable in a generalized linear model with a canonical link function g, and let X be the genetic variable, Z the environmental variable and W the additional covariates. Consider two nested generalized linear models*

$$\begin{aligned} g\{E(Y \mid X, W)\} &= \beta_0 + \beta_1 X + \beta_2 W, \\ g\{E(Y \mid X, Z, W)\} &= \gamma_0 + \gamma_1 X + \gamma_2 Z + \gamma_3 XZ + \gamma_4 W. \end{aligned} \tag{4}$$

Then the maximum likelihood estimators $\hat{\beta}_1$ and $\hat{\gamma}_3$ are asymptotically independent.

The proof of Theorem 2 is given in the Appendix. Corollary 1 follows from Theorem 2 immediately. To our knowledge, this independence of two estimators in two nested generalized linear models is novel to the statistical literature. It holds under both the null and the alternative hypotheses, and thus is more general than the independence result suggested in Hjort & Claeskens (2003, Lemma 3.2). It covers both binary traits and quantitative traits often investigated in genetic studies. For linear models, the independence holds exactly for any sample size. In the logistic regression model, our simulations yield nearly zero empirical correlation when the sample size is a few hundred. For case-control sampling, the likelihood comprises the retrospective distributions of covariates conditional on disease status. It is well known that if a standard logistic regression is fitted to case-control data, then $\hat{\beta}_1$ and $\hat{\gamma}_3$ are the semiparametric maximum likelihood estimators even when biased sampling is ignored (Prentice & Pyke, 1979); therefore the proof of Theorem 2 still applies.

Remark 1. Using marginal association as a filter for gene-environment interaction implicitly assumes that a gene which has an interaction with the environmental variable is likely to also display evidence of marginal association with the phenotype. This is plausible when the interaction effect is in the same direction as the main genetic effect. For qualitative interaction, where the subgroup effects may cancel out when averaged, using marginal association as a filter for interaction will yield limited power. On the other hand, marginal genetic association itself is often the primary goal of genetic association studies, so it is convenient to look for gene-environment interaction among genetic variants that demonstrate marginal association. Furthermore, the use of marginal genetic association as a filter for interaction is applicable to both qualitative and quantitative traits. This is in contrast to gene-environment association, the other filtering criterion we discuss later, which can only be used for case-control studies.

Exploiting the case-only estimator in the second-stage test, we next propose two extensions of the two-stage multiple testing procedure that use marginal genetic association as a filter. The first extension is to use the case-only estimator in case-control genetic studies within randomized clinical trials. When gene-environment independence holds in the study population and the disease is rare, the standard estimator of interaction in case-control sampling is not as efficient as the case-only estimator (Piegorsch et al., 1994; Umbach & Weinberg, 1997). Despite exhibiting a substantial efficiency gain, the case-only estimator is generally sensitive to departures from the gene-environment independence assumption (Albert et al., 2001). In genetic studies within randomized clinical trials, however, independence between the treatment assignment and genetic variants is dictated by the study design. In such settings, if we use two-stage procedures to test for gene-treatment interaction, it is appealing to replace the standard case-control interaction estimator by the case-only estimator, even though the independence of marginal association and the case-only estimator has not yet been established. We now provide this result.

Proposition 1. *Consider a case-control genetic association study in a randomized clinical trial. The data are composed of n observations, each of which has a binary disease outcome Y, a binary treatment variable Z, a genetic variant X and a vector of covariates W. Two logistic regression models are considered,*

$$\operatorname{logit}\{E(Y \mid X, W)\} = \beta_0 + \beta_1 X + \beta_2 W, \tag{5}$$

$$\operatorname{logit}\{E(Z \mid X, Y = 1)\} = \gamma_0 + \gamma_{\text{co}} X, \tag{6}$$

where β_1 is the marginal genetic effect and γ_{co} is the case-only interaction between Z and X. If Z is independent of X and the disease outcome $Y = 1$ is rare, then the parameter γ_{co} is equivalent to γ_3 in (4). Furthermore, the maximum likelihood estimators for β_1 and γ_{co} are asymptotically independent.

The proof of Proposition 1, given in the Appendix, is quite different from that of Theorem 2, since we do not have nested models. If gene-environment independence holds, use of the case-only estimator for testing gene-treatment interaction in the second stage would substantially improve the power, as we show in simulations.

The second extension is to consider two-stage procedures for survival outcomes. In randomized clinical trials, study endpoints are often time-to-event outcomes and analyses are often based on the Cox proportional hazard model. One may wonder whether the independence of marginal association and interaction carries over to the estimators in the Cox proportional hazard model. The key to the proof of independence in Theorem 1 is that the estimating equations can be expressed as the sum of independent and identically distributed terms in the form of $X\{Y - E(Y \mid X)\}$. The score functions for partial likelihood take a specialized form, and the arguments used to show independence in the previous examples do not apply in general. However, in the special case where the endpoint is rare, we show that the estimator for marginal association in a Cox model is asymptotically independent of the case-only estimator of the interaction.

Proposition 2. *Consider a randomized clinical trial with survival time Y, subject to independent right censorship. We observe $T = \min(Y, C)$ and $\Delta = I(Y \leqslant C)$, where C is the censoring time. Denote by Z the randomized treatment, and denote by X a genetic variant. Let $(T_i, \Delta_i, Z_i, X_i)$ $(i = 1, \ldots, n)$, be independent replicates. Consider the Cox models*

$$\lambda(t; Z, X, W) = \lambda_{0x}(t) \exp(\gamma_2 Z + \gamma_3 Z X + \gamma_4 W), \tag{7}$$

$$\lambda(t; X, W) = \lambda_0(t) \exp(\beta_1 X + \beta_2 W), \tag{8}$$

where (7) is a stratified Cox proportional hazard model to assess the treatment hazard ratio and (8) is a Cox proportional hazard model to assess the main effect of X. If the disease is rare and the censoring rate is equal in two randomized arms, the typical case-only interaction in the model $\operatorname{logit}\{E(Z \mid X, \Delta = 1)\} = \gamma_0 + \gamma_{\text{co}} X$ *is equivalent to the interaction γ_3 in (7). Furthermore, the estimators $\hat{\beta}_1$ and $\hat{\gamma}_{\text{co}}$ are asymptotically independent.*

The proof is given in the Supplementary Material. This result generalizes the use of the case-only estimator in two-stage procedures to randomized clinical trials with rare time-to-event outcomes. It can be further extended to Cox-model marginal association analyses based on such cohort sampling techniques as nested case-control and case-cohort sampling, as long as the event in the trial is rare.

The two-stage procedures discussed in this section essentially use two independent pieces of information in the data, one for filtering and the other for testing. Since both are of interest in genetic studies, we have constructed in a companion paper (Dai et al., 2012) a joint test for marginal effect and gene-environment interaction simultaneously, although its purpose is quite different from that of the two-stage procedures described here.

2·4. *Filtering by gene-environment association*

If the disease outcome Y is binary and rare, a different filtering criterion to detect gene-environment interaction, namely gene-environment association in the combined case-control samples, was proposed by Murcray et al. (2009). The rationale is that when the disease is rare, we expect to have Z independent of X in the controls if Z and X are independent in the cohort. If there is an interaction between Z and X, then they will be dependent in the cases, as suggested by the case-only estimator. Owing to oversampling of cases, Z and X should also be dependent in the combined case-control sample.

The filtering statistic appears to be unsupervised, since gene-environment association is assessed in the combined case-control sample irrespective of the disease status Y, so the filtering statistic should be independent of the test statistic. On the other hand, the information used for filtering comes from oversampling of cases, so the filtering statistic does contain outcome information. The formal proof of independence uses estimating equation theory. We state this result in the context of testing for gene-environment interaction.

PROPOSITION 3. *Consider a case-control genetic association study. The data are composed of n subjects, each with (Y, Z, X, W) as previously defined. Let R denote the indicator of being selected into the case-control sample. Consider the logistic regression models*

$$\text{logit}\{E(Z \mid X, W, R = 1)\} = \tau_0 + \tau_1 X + \tau_2 W, \tag{9}$$

$$\text{logit}\{E(Y \mid X, Z, W)\} = \gamma_0 + \gamma_1 X + \gamma_2 Z + \gamma_{\text{cc}} XZ + \gamma_4 W, \tag{10}$$

where τ_1 is the gene-environment association in the combined case-control sample and γ_{cc} is the standard interaction in case-control studies. Then the maximum likelihood estimators for τ_1 and γ_{cc} are asymptotically independent.

The proof is given in the Appendix. Murcray et al. (2009) provided a proof of the independence between $\hat{\tau}_1$ and $\hat{\gamma}_{\text{cc}}$. Our proof is more general and accommodates confounders in both models. The proof of independence between $\hat{\tau}_1$ and $\hat{\gamma}_{\text{cc}}$ does not require the independence of Z and X in the population, but this assumption is useful for the filter to have power to screen for gene-environment interaction. In observational studies, one cannot be sure that the gene-environment independence assumption holds for every genetic variant, although a large deviation from gene-environment independence can be detected by comparing the proportion of variants passing the filter to the threshold α_0.

Remark 2. The gene-environment association in the combined case-control sample, namely τ_1 in (9), is a diluted version of the case-only interaction γ_{co} when the disease is rare. We make three cautionary comments on its use. First, when the disease is common, it is less clear whether the gene-environment association is useful for detecting gene-treatment interaction, as gene-treatment independence no longer holds in the controls. Second, the case-control sampling ratio could be critical to the power of the gene-environment association as a screening tool. More controls would make τ_1 less informative about the interaction γ_{cc}. Third, asymptotic independence does not hold between the estimated gene-environment association and the case-only estimator, thus one cannot use gene-environment association as a filter when the case-only estimator is being employed for testing interaction.

On the other hand, because of the unsupervised nature of τ_1, its estimator is independent of estimators of any parameters in a regression of Y on Z and X. The use of τ_1 as a filter can be expanded to the adjusted marginal effect in the model

$$\text{logit}\{E(Y \mid X, Z, W)\} = \beta_0 + \beta_{\text{adj}} X + \beta_2 Z + \beta_3 W.$$

By the same proof as for Proposition 3, $\hat{\beta}_{\text{adj}}$ is independent of $\hat{\tau}_1$. When the treatment Z influences Y, the adjusted effect β_{adj} could be of interest. One could use $\hat{\tau}_1$ as a filter for $\hat{\beta}_{\text{adj}}$, because a nonzero τ_1 implies a nonzero interaction γ_{cc}, which may also imply a nonzero adjusted effect β_{adj}.

The derivation of our results on asymptotic independence also applies to the likelihood ratio test and the score test, since the asymptotic covariance matrix for these test statistics can be similarly expressed in the form (1). See, for example, the derivation of the asymptotic distribution of the likelihood ratio test in van der Vaart (1998, Ch. 16).

3. POWER CONSIDERATIONS AND THE TUNING PARAMETER α_0

We have shown that the proposed two-stage procedures control the familywise error rate. For such two-stage procedures to be useful, they should also have higher power than the benchmark one-stage Bonferroni correction for all genetic variants. Intuitively, the power advantage of the two-stage procedures, if any, comes from the less stringent significance rule due to a much smaller number of genetic variants passing the filter. Moreover, a true alternative should have a high probability of passing the filter. This implies two conditions.

First, the alternative hypothesis H_{1j} should imply the alternative hypothesis in the filtering stage K_{1j}; otherwise, filtering would not enrich for true alternatives H_{1j}. This is equivalent to the condition discussed for two-stage procedures used in microarray experiments (Bourgon et al., 2010), that the filtering statistics and test statistics need to be correlated under alternatives. This condition is true for the gene-environment association filtering when gene-environment independence, the rare-disease assumption and oversampling of cases are all met. It is not always true for filtering by marginal association, since genetic variants with gene-environment interaction could have zero marginal effect. In this sense, marginal association is a leaky filter that could lose some true alternatives in interaction.

Second, even if H_{1j} implies K_{1j}, the probability of passing the filter for a true alternative H_{1j} should be sufficiently high. A weak filter may offset the benefit of fewer tests in the second stage. To see this, it is necessary to approximate the power for the two-stage procedure and for the one-stage Bonferroni test. For a hypothesized disease risk model H_{1j} and a sampling scheme, the power to detect a genetic variant for a large sample size n and a large number of genetic variants m is approximately

$$\left[1 - \Phi\left\{\Phi^{-1}\left(1 - \frac{\alpha_0}{2}\right) - \frac{|\mu^0|}{\sigma^0/n^{1/2}}\right\}\right]\left[1 - \Phi\left\{\Phi^{-1}\left(1 - \frac{\alpha}{2m\alpha_0}\right) - \frac{|\mu|}{\sigma/n^{1/2}}\right\}\right], \quad (11)$$

where μ^0 and σ^0 are the mean and asymptotic standard deviation of the first-stage filtering estimator, and μ and σ are the mean and asymptotic standard deviation of the second-stage testing estimator. We have assumed in (11) that the genetic variants are independent, so that $m_0 = m\alpha_0$ for a large m. Clearly, α_0 plays an important role in power performance. An unduly small α_0 would allow too few genetic variants to pass the filter, thus adversely affecting the power. As α_0 increases, the probability of passing the filter will be greater, but the multiple testing penalty for

Table 1. *Empirical familywise error rate* (%) *for various two-stage procedures in* 1000 *simulations. A total of* 10 000 *genetic variants were simulated for* 1000 *cases and* 1000 *controls by case-control sampling. Genetic variants were generated either independently or with serial correlation* 0·5. *The disease status was generated by the logistic model* (10)

		$\boldsymbol{\gamma}=(-4, 0, 0, 0)$			$\boldsymbol{\gamma}=(-4, 0, \log 1{\cdot}5, 0)$		
		$\alpha_0=0{\cdot}001$	$\alpha_0=0{\cdot}01$	$\alpha_0=0{\cdot}1$	$\alpha_0=0{\cdot}001$	$\alpha_0=0{\cdot}01$	$\alpha_0=0{\cdot}1$
$\rho=0$	$\beta_1\rightarrow\gamma_{\rm cc}$	3·7	4·5	5·0	5·0	4·6	4·3
	$\beta_1\rightarrow\gamma_{\rm co}$	4·2	5·1	4·3	4·1	3·0	3·7
	$\tau_1\rightarrow\gamma_{\rm cc}$	4·2	4·7	3·2	4·2	3·5	3·4
	$\tau_1\rightarrow\beta_{\rm adj}$	3·6	3·3	4·7	3·6	3·6	2·6
$\rho=0{\cdot}5$	$\beta_1\rightarrow\gamma_{\rm cc}$	4·4	4·6	4·6	5·3	4·0	3·9
	$\beta_1\rightarrow\gamma_{\rm co}$	5·2	5·0	4·3	5·1	5·1	4·4
	$\tau_1\rightarrow\gamma_{\rm cc}$	4·1	3·4	3·9	4·6	4·5	4·6
	$\tau_1\rightarrow\beta_{\rm adj}$	4·7	3·7	4·8	4·5	4·3	4·0

The parameters in the second column are from the following models: $\text{logit}\{E(Y\mid X)\}=\beta_0+\beta_1X$; $\text{logit}\{E(Y\mid X, Z)\}=\gamma_0+\gamma_1X+\gamma_2Z+\gamma_{\rm cc}XZ$; $\text{logit}\{E(Z\mid X, Y=1)\}=\gamma_0+\gamma_{\rm co}X$; $\text{logit}\{E(Z\mid X)\}=\tau_0+\tau_1X$; $\text{logit}\{E(Y\mid X, Z)\}=\beta_0+\beta_{\rm adj}X+\beta_2Z$.

genetic variants passing the filter will also increase. With prior assumptions on the true model, such as (μ, σ) and (μ^0, σ^0), an optimal α_0 can be computed based on (11).

Similarly, the power of the standard one-stage Bonferroni correction can be approximated as

$$1-\Phi[\Phi^{-1}\{1-\alpha/(2m)\}-n^{1/2}|\mu|/\sigma]. \tag{12}$$

For the power of the two-stage procedure (11) to exceed the power of the one-stage procedure (12), the probability of passing the filter for a true alternative should be larger than the ratio of the second components in (11) and (12). For a well-powered one-stage Bonferroni procedure, one may not be able to find a filter that further improves power.

4. Simulations

We first examine the correlation of various estimators in small samples, one for linear regression and another for logistic regression. The results are satisfactory: for a sample size of a few hundred, the correlation is nearly zero in all simulation settings. These results are included in the Supplementary Material.

We next assess the empirical familywise error rate over 1000 simulated datasets, each containing 10 000 genetic variants for 1000 subjects. The minor allele frequencies were randomly generated from a uniform distribution from 0·1 to 0·5. The diploids were formed assuming the Hardy–Weinberg equilibrium. The genetic variants are either independent or have a serial correlation 0·5. A binary treatment assignment was randomly generated as a Ber(0·5) distribution. The binary disease status was generated by the logistic model $\text{logit}\{E(Y=1\mid X, Z)\}=\gamma_0+\gamma_1X+\gamma_2Z+\gamma_3XZ$ with parameters $(-4, 0, 0, 0)$ or $(-4, 0, \log 1{\cdot}5, 0)$; the latter assumes a mild treatment effect. The two-stage procedures were applied to screen for interactions with $\alpha_0=0{\cdot}001$, 0·01 or 0·1 and $\alpha=0{\cdot}05$. Table 1 shows that the familywise error rate is controlled at the level 0·05 as expected. Control of the familywise error rate was also observed with other parameter settings not reported here.

We investigate the power of the two-stage procedures. We assume a case-control genetic study within a randomized clinical trial with half a million independent genetic variants, of which

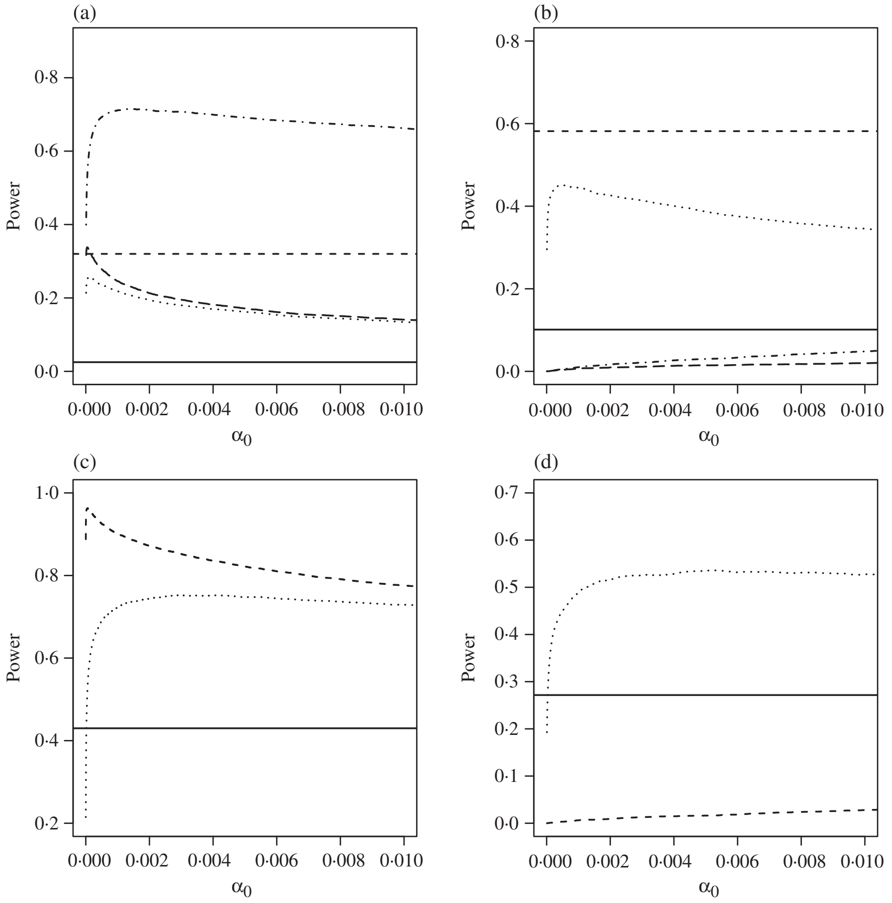

Fig. 1. Power comparison of the two-stage procedures for detecting gene-treatment interaction. The four panels represent the scenarios (a) rare disease and quantitative interaction, (b) rare disease and qualitative interaction, (c) common disease and quantitative interaction, and (d) common disease and qualitative interaction. Five testing methods are plotted: the one-stage case-control interaction (solid), the one-stage case-only interaction (short dashes), the two-stage procedure with marginal association filtering and standard interaction testing (long dashes), the two-stage procedure with marginal association filtering and case-only interaction testing (dot-dash), and the two-stage procedure with gene-environment association filtering and standard interaction testing (dots).

one is a true association and the rest are null associations. The diploids were formed assuming Hardy–Weinberg equilibrium. The randomization ratio to the treatment arm and the control arm is 1:1, and the case-control sampling ratio is also 1:1. With almost half a million independent 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 $\Phi^{-1}\{1-\alpha/(2m\alpha_0)\}$ as if it were fixed in every simulation. The power was computed as the percentage of simulations where the genetic variant with the signal was declared to be significant when the familywise error rate is controlled at 0·05 in 10 000 simulations. We attempted numerous parameter settings and a range of values for α_0 to study operating characteristics. Figure 1 shows several scenarios where the relative power comparison for detecting gene-treatment interaction is representative. Additional simulations using gene-environment association to screen for adjusted genetic effect are left to the Supplementary Material.

Figures 1(a) and 1(b) show the power to detect an interaction between the signal genetic variant X and the treatment Z when the disease is rare, with prevalence 0·02. We display power versus α_0 for each parameter setting in order to examine the effect of α_0. Let N be the total sample size, n the sample size for the case-control sample, f the minor allele frequency of the target genetic variant, and γ the vector of parameters in model (10) that generates the data.

Figure 1(a) shows a setting with a quantitative interaction between the genetic variant and the treatment: $N = 50\,000$, $n \approx 1000$, $f = 0{\cdot}1$ and $\gamma = (-4, 0, 0, \log 2)$. Clearly, the two-stage procedure with filtering by marginal association and testing by the case-only interaction yields substantially improved power over other procedures. The two-stage procedure with filtering by marginal association and testing by the standard case-control interaction provides approximately 30% power, similar to the one-stage case-only procedure with Bonferroni correction. The two-stage procedure with filtering by the gene-environment association and testing by the standard interaction is outperformed by both the one-stage case-only procedure and the other two-stage procedures. The one-stage standard interaction procedure yields almost no power in this scenario. This example demonstrates the advantage of using the case-only estimator in the two-stage procedure with marginal association filtering.

In Fig. 1(b), we simulate a setting with a qualitative interaction between the genetic variant and the treatment, i.e., the sign of the genetic variant effect differs in different treatment groups: $N = 50\,000$, $n \approx 1000$, $f = 0{\cdot}2$ and $\gamma = (-4, -0{\cdot}5 \log 2, 0, \log 2)$. The main effect of the genetic variant in this setting is negligible, which leads to poor performance of the two-stage procedures using marginal association for an initial screen. The two-stage procedure using the gene-environment association avoids the cancellation of opposite genetic variant effects and so yields a noticeable power gain over the two-stage procedures using marginal association. The best procedure in this scenario, however, is the case-only estimator with Bonferroni correction. This example confirms that for qualitative interactions, marginal association is not effective as a filter for gene-treatment interaction.

Figures 1(c) and 1(d) show the power to detect the interaction when the disease prevalence is 10%. In this scenario, the case-only estimator may have bias and an inflated type I error because the rare disease assumption is not met. In Fig. 1(c), a model with a quantitative interaction was generated: $N = 10\,000$, $n \approx 1000$, $f = 0{\cdot}1$ and $\gamma = (-2, 0, 0, \log 2)$. In this setting, the best procedure is the two-stage procedure with filtering by marginal association and testing by standard interaction, while the worst is the one-stage procedure with standard interaction. Because of the 1:1 case-control sampling ratio, the two-stage procedure with the gene-environment association shows improved power over the one-stage procedure. In Fig. 1(d), a model with a qualitative interaction was generated: $N = 10\,000$, $n \approx 1000$, $f = 0{\cdot}1$ and $\gamma = (-2, -0{\cdot}5 \log 2, 0, \log 2)$. The two-stage procedure with marginal association as a filter has no power at all, while the gene-environment association as a filter outperforms the one-stage procedure.

Taken collectively, each of the two-stage procedures has a niche in power performance. There are situations where none of them improves power upon the one-stage Bonferroni procedure. For simulations with half a million genetic variants, the optimal α_0 for power performance is in the range of 0·0001 to 0·001. This is in agreement with previous work (Kooperberg & LeBlanc, 2008; Murcray et al., 2009).

5. Application

The Women's Health Initiative is one of the largest studies of postmenopausal women's health in the United States, and is composed of four randomized clinical trial components and a prospective observational study. An elevated invasive breast cancer risk was found among women

Table 2. *Results of the two-stage procedures applied to the Women's Health Initiative study, using gene-environment association as a screening criterion applied to the estrogen-alone trial. Four genetic variants out of the top* 50 *genetic variants reach statistical significance in testing the adjusted marginal genetic effect*

		rs7705343		rs13159598		rs9790879		rs4415084	
		Odds ratio	p-value	Odds ratio	p-value	Odds ratio	p-value	Odds ratio	p-value
E-alone	$\hat{\tau}_1$	1·5823	0·0006	1·5217	0·0014	1·4701	0·0035	1·4269	0·0063
All trials	$\hat{\beta}_{\rm adj}$	1·1672	0·0006	1·1653	0·0007	1·1649	0·0007	1·1695	0·0005
E-alone	$\hat{\gamma}_{\rm co}$	1·5231	0·0298	1·4657	0·0444	1·3842	0·0907	1·3832	0·0936

E-alone, estrogen-alone trial. The estimators in the second column are from the following models, where Z denotes the treatment and X the genetic variant: $\text{logit}\{E(Z \mid X)\} = \tau_0 + \tau_1 X$; $\text{logit}\{E(Y \mid X, Z)\} = \beta_0 + \beta_{\rm adj} X + \beta_2 Z$; $\text{logit}\{E(Z \mid X, Y = 1)\} = \gamma_0 + \gamma_{\rm co} X$. Note that $\hat{\tau}_1$ and $\hat{\gamma}_{\rm co}$ are based on data from the estrogen-alone trial, while $\hat{\beta}_{\rm adj}$ is based on data from all four trials.

assigned to estrogen plus progestin, with 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 breast cancer risk, perhaps jointly with the interventions, a genome-wide association study was launched with a multi-stage design (Prentice & Qi, 2006). In the final stage, a total of 9039 genetic variants were selected and genotyped among 2166 invasive breast cancer cases in the clinical trials and 1:1 matched controls.

Primary analyses have been presented recently (Prentice et al., 2009, 2010). Seven genetic variants in the fibroblast growth factor receptor 2 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 genetic variants ranked by marginal association (Prentice et al., 2009, 2010). Since invasive breast cancer is a rare event in the study, the investigators used the case-only estimators for interactions where genetic scores were coded as two indicator variables. Several genetic variants showed suggestive evidence of interactions with one or more interventions. In particular, the nominal p-value for the interaction between the genetic variant rs3750817 and the dietary modification intervention is 0·005, which remains significant after multiple testing adjustment. Our results justified the focused testing for interactions in a subset of genetic variants ranked by top marginal association.

In addition, we also explored the two-stage procedures using the gene-environment association criterion to look for significant adjusted marginal effects and interactions. This was done separately for each of the four randomized trials. In the first stage, we ranked genetic variants by p-values for gene-environment association. We then tested for adjusted marginal effects and interactions for the top 50 genetic variants ranked by gene-environment association. For discovery purposes, the additive genetic models were used for both adjusted marginal effects and the interactions. Table 2 shows four genetic variants picked by the gene-environment association criterion in the estrogen-alone trial that have significant adjusted association. In the estrogen-alone trial, four genetic variants pass the Bonferroni correction for 50 genetic variants in testing for adjusted genetic variant effect. The adjusted additive genetic variant effect $\hat{\beta}_{\rm adj}$ was estimated from case-control data for all four trials, adjusted for matching variables, important baseline predictors and randomization indicators. The effect size in terms of odds ratio is fairly modest, around 1·16. In the estrogen-alone trial, there seems to be a weak interaction between the genetic variants and the treatment. The effect sizes of the case-only interaction $\hat{\gamma}_{\rm co}$ are around 1·4 to 1·5.

These four genetic variants are all located in the mitochondrial ribosomal protein S30 gene, which has shown some evidence of interaction with multiple clinical interventions (Huang et al., 2011). None of the variants reaches the genome-wide significance level for either marginal effect or interaction, yet they reach the familywise error rate level of 0·05 for marginal association by our two-stage procedure. The reason might be that these four genetic variants have weak main effects and weak interactions with the estrogen-alone intervention. The gene-environment association criteria seem to synthesize these weak effects and prioritize them for further testing. This application suggests that in the low-power settings, two-stage procedures can be used as a data-adaptive tool, as opposed to candidate genes from prior studies, for discovering novel genes that affect disease risk. Certainly, this search strategy should serve only as a supplement to the standard one-stage Bonferroni test, since it missed the seven genetic variants in the fibroblast growth factor receptor 2 gene.

6. Discussion

We discuss control of the familywise error rate by two-stage testing procedures, since it is widely used in genome-wide association studies, primarily because of the scarcity of susceptibility variants. It is of interest to generalize these results to more liberal error measures, such as the false discovery rate. The latter can be useful when there is a replication sample following the discovery sample. For independent tests, the control of the false discovery rate is immediately seen when applying the procedure in Benjamini & Hochberg (1995) to the variants passing the filter. For correlated genetic variants, further work is needed, possibly using the method in Benjamini & Yekutieli (2001).

All these procedures require a prespecified proportion of genetic variants passing the filtering stage. As we show from simulations, the optimal proportion may depend on the unknown underlying interaction model. Rather than giving a harsh cut-off for entering the second stage, a better strategy could be to weight the significance of the test for interaction by the strength of the corresponding filtering statistic. All genetic variants will be tested and adaptive selection of α_0 is avoided. This strategy was first suggested by Ionita-Laza et al. (2007) in family-based association studies. We have also explored the weighting strategy in the context of detecting gene-environment interaction (Hsu et al., 2012).

Since the two filtering criteria are complementary in performance, it would be sensible to combine them into one procedure. Practical guidance on doing so is not pursued in this paper. Murcray et al. (2011) have suggested a hybrid procedure by allocating a proportion of the overall genome-wide significance level to each filtering method. We have also explored more flexible ways to combine the two filtering criteria and to incorporate the case-only estimator, or the empirical Bayes estimator (Mukherjee & Chatterjee, 2008), into one composite procedure (Hsu et al., 2012).

Acknowledgement

We thank Li Hsu for helpful discussions. This research was supported by a faculty development fund from the Fred Hutchinson Cancer Research Center and by the National Institutes of Health, U.S.A.

Supplementary material

Supplementary material available at *Biometrika* online includes a proof of Proposition 2, simulations to examine the empirical correlation of various estimators, and additional simulations to assess the performance of using gene-environment association to screen for adjusted genetic association.

Appendix

Proof of Theorem 1

Under mild regularity conditions, standard estimating equation theory implies that $n^{1/2}(\hat{\zeta}_j - \zeta_j) \rightarrow \mathcal{N}(0, V_1)$ in distribution and $n^{1/2}(\hat{\theta}_j - \theta_j) \rightarrow \mathcal{N}(0, V_2)$ in distribution, 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 numbers, $\hat{V}_1$ and $\hat{V}_2$ are consistent estimators. Since $\text{cov}\{n^{1/2}(\hat{\zeta}_j - \zeta_j), n^{1/2}(\hat{\theta}_j - \theta_j)\} \rightarrow 0$ as $n \rightarrow \infty$, it is straightforward to show that T_j^0 and T_j are asymptotically independent for all j under the global null hypothesis for H.

Observe that under the null hypothesis H_{0j} for all j, $m_0 = \sum_{j=1}^{m} I(T_j^0 \in \Gamma_j^0)$. Unless the T_j^0 are independent, $E(m_0/m)$ is generally not equal to α_0. However, if $m_0/m \rightarrow \alpha_0'$ in probability, we can prove the main result as follows:

$$\lim_{m\rightarrow\infty}\lim_{n\rightarrow\infty} \text{pr}\left\{ \bigcup_{j\in\mathcal{J}} (T_j^0 \in \Gamma_j^0 \cap T_j \in \Gamma_j) \right\} \leqslant \lim_{m\rightarrow\infty}\lim_{n\rightarrow\infty} \sum_{j=1}^{J} \text{pr}(T_j^0 \in \Gamma_j^0)\, \text{pr}(T_j \in \Gamma_j) \tag{A1}$$

$$= \lim_{m\rightarrow\infty}\lim_{n\rightarrow\infty} \left\{ \frac{1}{m} \sum_{j=1}^{J} \text{pr}(T_j^0 \in \Gamma_j^0) \right\} \frac{m\alpha}{m_0} \leqslant \alpha_0' \frac{\alpha}{\alpha_0'} = \alpha. \tag{A2}$$

Inequality (A1) uses the Bonferroni inequality and the asymptotic independence of T_j^0 and T_j, while (A2) holds because $m^{-1}\sum_{j=1}^{J} \text{pr}(T_j^0 \in \Gamma_j^0) \leqslant \alpha_0'$ and, by Slutsky's theorem, $m\alpha/m_0 \rightarrow \alpha/\alpha_0'$ in probability.

Proof of Theorem 2

Let X_1 denote the p-dimensional design matrix for the bigger model (3), and let X_2 denote the q-dimensional design matrix for the smaller model (2). Let U_1 denote the estimating functions for (2) and let U_2 denote the estimating functions for (3). Because both models are generalized linear models with the canonical link, U_1 and U_2 can be expressed as $\mathrm{X}_1\{Y - E(Y \mid \mathrm{X}_1)\}$ and $\mathrm{X}_2\{Y - E(Y \mid \mathrm{X}_1)\}$, respectively.

The asymptotic covariance matrix was derived to be (1). We evaluate the off-diagonal submatrix $A_1^{-1}B_{12}A_2^{-1}$. Observe that upon leaving out the dispersion parameters,

$$\begin{aligned} B_{12} &\propto E[(\mathrm{X}_1^{\mathrm{T}}\mathrm{X}_2)\{Y - E(Y \mid \mathrm{X}_1)\}\{Y - E(Y \mid \mathrm{X}_2)\}] \\ &= E[(\mathrm{X}_1^{\mathrm{T}}\mathrm{X}_2)E\{Y^2 - YE(Y \mid \mathrm{X}_1) - E(Y \mid \mathrm{X}_1)E(Y \mid \mathrm{X}_2) + E(Y \mid \mathrm{X}_1)E(Y \mid \mathrm{X}_2) \mid \mathrm{X}_1\}] \\ &= E\{(\mathrm{X}_1^{\mathrm{T}}\mathrm{X}_2)\, \text{var}(Y \mid \mathrm{X}_1)\}, \end{aligned}$$

and thus

$$A_1^{-1}B_{12}A_2^{-1} \propto E\{\mathrm{X}_1^{\mathrm{T}}\mathrm{X}_1\, \text{var}(Y \mid \mathrm{X}_1)\}^{-1} E\{\mathrm{X}_1^{\mathrm{T}}\mathrm{X}_2\, \text{var}(Y \mid \mathrm{X}_1)\} E\{\mathrm{X}_2^{\mathrm{T}}\mathrm{X}_2\, \text{var}(Y \mid \mathrm{X}_2)\}^{-1}.$$

Let $\mathrm{T}_1 = \mathrm{X}_1\{\text{var}(Y \mid \mathrm{X}_1)\}^{1/2}$ and $\mathrm{T}_2 = \mathrm{X}_2\{\text{var}(Y \mid \mathrm{X}_1)\}^{1/2}$, so that $A_1^{-1}B_{12} = E(\mathrm{T}_1^{\mathrm{T}}\mathrm{T}_1)^{-1}E(\mathrm{T}_1\mathrm{T}_2)$. Since T_2 is contained in T_1 as the first q columns, $E(\mathrm{T}_1^{\mathrm{T}}\mathrm{T}_1)^{-1}E(\mathrm{T}_1^{\mathrm{T}}\mathrm{T}_2)$ is a $p \times q$ matrix, of which the top $q \times q$ submatrix is diagonal with entries 1, and the bottom $(p - q) \times q$ submatrix has zero for every entry. Some algebra leads to the result that the entries of the lower $(p - q) \times q$ submatrix of $A_1^{-1}B_{12}A_2^{-1}$ are zero. This implies that all the estimators for the parameters in (2) but not in (3) are asymptotically uncorrelated with all the estimators for the parameters in both (2) and (3).

Proof of Proposition 1

Because the disease is rare, we can approximate a logistic regression as (4) by a log-linear regression. Observe that

$$\frac{\mathrm{pr}(Z=1 \mid X, Y=1, W)}{\mathrm{pr}(Z=0 \mid X, Y=1, W)} = \frac{\mathrm{pr}(Y=1 \mid X, W, Z=1)\,\mathrm{pr}(Z=1)}{\mathrm{pr}(Y=1 \mid X, W, Z=0)\,\mathrm{pr}(Z=0)}$$

$$= \exp\left(\log\frac{\pi}{1-\pi} + \gamma_2 + \gamma_3 X\right),$$

where $\pi = \mathrm{pr}(Z=1)$. Hence Z is independent of W given X and $Y=1$, and $\gamma_{\mathrm{co}} = \gamma_3$. The estimating equations from the two models (5) and (6) are $U_1 = \mathrm{X}_1\{Y - E(Y \mid \mathrm{X})\}$ and $U_2 = \mathrm{X}_2\{Z - E(Z \mid \mathrm{X})\}1_{[Y=1]}$, where X_1 is the design matrix with ith row $(1, x_i, w_i)$ and X_2 is the design matrix with ith row $(1, x_i)$. Note that $U_2 = 0$ if $Y = 0$. Let $\beta = (\beta_0, \beta_1, \beta_2)$ and $\gamma = (\gamma_0, \gamma_{\mathrm{co}})$. The asymptotic covariance matrix for $\hat\beta$ and $\hat\gamma$ is $A_1^{-1}B_{12}A_2^{-1}$, as defined in (1). Observe that B_{21} equals

$$\mathrm{pr}(Y=1)E_{X,W \mid Y=1}[\mathrm{X}_1^{\mathrm{T}}\mathrm{X}_2\{1 - \mathrm{pr}(Y=1 \mid X, W)\}E_{Z \mid X,W,Y=1}\{Z - E(Z \mid X, Y=1)\}] = 0.$$

The derivation uses the law of iterated expectations. Hence the covariance matrix of $(\hat\beta, \hat\gamma)$ is zero, and the proof is complete.

Proof of Proposition 3

The estimating equations for models (9) and (10) are written as $U_1 = \mathrm{X}_1\{Z - E(Z \mid \mathrm{X}_1)\}$ and $U_2 = \mathrm{X}_2\{Y - E(Y \mid \mathrm{X}_2)\}$, where X_1 is the design matrix for (9) with ith row $(1, x_i, w_i)$ and X_2 is the design matrix for (10) with ith row $(1, x_i, z_i, x_i z_i, w_i)$. Observe that

$$B_{21} = E_{Z,X,W}[\mathrm{X}_1^{\mathrm{T}}\mathrm{X}_2\{Z - E(Z \mid X, W)\}E_{Y \mid Z,X,W}\{Y - E(Y \mid X, W, Z)\}] = 0.$$

The derivation uses the law of iterated expectations. Hence the off-diagonal element of the covariance matrix is zero, and the proof is complete.

References

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

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–88.

Bourgon, R., Gentleman, R. & Wolfgang, H. (2010). Independent filtering increases detection power for high-throughput experiments. *Proc. Nat. Acad. Sci.* **107**, 9546–51.

Casella, G. & Berger, R. L. (2002). *Statistical Inference*. Pacific Grove, California: Thomson Learning, 2nd ed.

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

Cornelis, M. C., Tchetgen Tchetgen, E. J., Liang, L., Qi, L., Chatterjee, N., Hu, F. B. & Kraft, P. (2012). Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. *Am. J. Epidemiol.* **175**, 191–202.

Dai, J. Y., LeBlanc, M. & Kooperberg, C. (2009). Semiparametric estimation exploiting covariate independence in two-phase randomized trials. *Biometrics* **65**, 178–87.

Dai, J. Y., Logsdon, B. A., Huang, Y., Hsu, L., Reiner, A. P., Prentice, R. L. & Kooperberg, C. (2012). Simultaneous testing for marginal genetic association and gene-environment interaction in genome-wide association studies. *Am. J. Epidemiol.* **176**, 164–73.

Gauderman, W. J., Thomas, D. C., Murcray, C. E., Conti, D., Li, D. & Lewinger, J. (2010). Efficient genome-wide association testing of gene-environment interaction in case-parent trios. *Am. J. Epidemiol.* **172**, 116–22.

Hjort, N. L. & Claeskens, G. (2003). Frequentist model average estimators. *J. Am. Statist. Assoc.* **98**, 879–99.

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

Hsu, L., Shuo, J., Dai, J. Y., Hutter, C. & Kooperberg, C. (2012). A powerful cocktail strategy for detecting genome-wide gene-environment interaction. *Genet. Epidemiol.* **36**, 183–94.

Huang, Y., Ballinger, D. G., Dai, J. Y., Peters, U., Hinds, D. A., Cox, D. R., Beilarz, E., Chlebowski, R. T., Rossouw, J. E., McTienan, A., Rohan, T. & Prentice, R. L. (2011). Genetic variants in the MRPS30 region and postmenopausal breast cancer risk. *Genome Med.* **3**, 42.

Hunter, D. J. (2005). Gene-environment interactions in human diseases. *Nature Rev. Genet.* **6**, 287–98.

International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**, 1299–320.

Ionita-Laza, I., McQueen, M. B., Laird, N. M. & Lange, C. (2007). Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100k scan. *Am. J. Hum. Genet.* **81**, 607–14.

Kooperberg, C. & LeBlanc, M. (2008). Increasing the power of identifying gene-gene interactions in genome-wide association studies. *Genet. Epidemiol.* **32**, 255–63.

Mukherjee, B., Ahn, J., Gruber, S. B. & Chatterjee, N. (2012). Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *Am. J. Epidemiol.* **175**, 177–90.

Mukherjee, B. & Chatterjee, N. (2008). Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes approach to trade off between bias and efficiency. *Biometrics* **64**, 685–94.

Murcray, C. E., Lewinger, J. P., Conti, D. V., Thomas, D. C. & Gauderman, W. J. (2011). Sample size requirement to detect gene-environment interactions in genome-wide association studies. *Genet. Epidemiol.* **35**, 201–10.

Murcray, C. E., Lewinger, J. P. & Gauderman, J. W. (2009). Gene-environment interaction in genome-wide association studies. *Am. J. Epidemiol.* **169**, 219–26.

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

Piegorsch, W. W., Weinberg, C. R. & Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. *Statist. Med.* **13**, 153–62.

Prentice, R. L., Huang, Y., Hinds, D. A., Peters, U., Cox, D. R., Beilharz, E., Chlebowski, R. T., Rossouw, J. E., Caan, B. & Ballinger, D. (2010). Variation in the FGFR2 gene and the effect of a low-fat dietary pattern on invasive breast cancer. *Cancer Epidemiol. Biomark. Prev.* **19**, 74–9.

Prentice, R. L., Huang, Y., Hinds, D. A., Peters, U., Pettinger, M., Cox, D. R., Beilharz, E., Chlebowski, R. T., Rossouw, J. E., Caan, B. & Ballinger, D. (2009). Variation in the FGFR2 gene and the effects of postmenopausal hormone therapy on invasive breast cancer. *Cancer Epidemiol. Biomark. Prev.* **18**, 3079–85.

Prentice, R. L. & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–11.

Prentice, R. L. & Qi, L. (2006). Aspects of the design and analysis of high-dimensional SNP studies for disease risk estimation. *Biostatistics* **7**, 339–54.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–66.

Rothman, N., Garcia-Closas, M., Chatterjee, N., Malats, N., Wu, X., Figueroa, J., Real, F. X. & et al. (2010). A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nature Genet.* **42**, 978–84.

Umbach, D. M. & Weinberg, C. R. (1997). Designing and analyzing case-control studies to exploit independence of genotype and exposure. *Statist. Med.* **16**, 1731–43.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Van Steen, K., McQueen, M. B., Herbert, A., Raby, B. & Lyon, H. E. (2005). Genomic screening and replication using the same data set in family-based association testing. *Nature Genet.* **37**, 683–91.

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

White, H. (2000). *Asymptotic Theory for Econometricians*. New York: Academic Press, revised ed.

[*Received September* 2011. *Revised June* 2012]