Harvard University
Harvard University Biostatistics Working Paper Series

Year 2014 *Paper* 177

A General Approach to Detect Gene (G)-environment (E) Additive Interaction Leveraging G-E Independence in Case-control Studies

Eric Tchetgen Tchetgen* Tamar Sofer†

Benedict H.W. Wong‡

*Harvard University, etchetge@hsph.harvard.edu

†Harvard University, tsofer@hsph.harvard.edu

‡Harvard University

Original Manuscript

A general approach to detect gene (G)-environment (E) additive interaction leveraging G-E independence in case-control studies

Eric J. Tchetgen Tchetgen[1,2], Tamar Sofer[1], Benedict H. W. Wong[1]

Departments of [1]Biostatistics and [2]Epidemiology, Harvard University

Corresponding author: Eric J. Tchetgen Tchetgen, Department of Epidemiology, Harvard School of Public Health 677 Huntington Avenue, Boston, MA 02115.

Abstract

It is increasingly of interest in statistical genetics to test for the presence of a mechanistic interaction between genetic (G) and environmental (E) risk factors by testing for the presence of an additive GxE interaction. In case-control studies involving a rare disease, a statistical test of no additive interaction typically entails a test of no relative excess risk due to interaction (RERI). It is also well known that a test of multiplicative interaction that exploits G-E independence can be dramatically more powerful than standard logistic regression for case-control data. Likewise, it has recently been shown that a likelihood ratio test of a null RERI incorporating the G-E independence assumption (RERI-LRT) outperforms the standard RERI approach. In this paper, the authors describe a general, yet relatively straightforward approach to test for GxE additive interaction exploiting G-E independence. The approach which relies on regression models for G and E is particularly attractive because, unlike the RERI-LRT, it allows the regression model for the binary outcome to remain unrestricted. Therefore, the approach is completely robust to possible mis-specification of main effects in the outcome regression. This is an important advantage of the proposed strategy, particularly in settings not easily handled by the RERI-LRT, in which E is either a count or a continuous exposure and one must account for multiple covariates in order to enforce the G-E independence assumption, as well as to rule out residual confounding. The methods are illustrated through an extensive simulation study and a well-known ovarian cancer application.

[1]

There is growing interest in the development and application of statistical methods to detect the presence of an additive gene (G)-environment (E) interaction because such interaction may be closer to a true mechanistic interaction than its multiplicative counterpart (Rothman et al, 1980, Greenland,1983, Cordell, 2002, VanderWeele and Knoll, 2014). For case-control data involving a rare disease, a statistical test of no additive GxE interaction is easily performed via a test of a null relative excess risk due to interaction (RERI) (Rothman et al, 2008). This approach has gained popularity in epidemiology primarily because it is easily performed using relative risk estimates obtained via standard logistic regression for case-control data (Rothman et al, 2008). When G and E are known to be independent in the target population, it is well known that a test of multiplicative interaction that exploits the independence assumption can be dramatically more powerful than standard logistic regression, which does not make use of the assumption (Piegorsch et al, 1994, Umbach and Weinberg, 1997, Chatterjee and Carroll, 2005, Mukherjee and Chatterjee, 2008, Tchetgen Tchetgen and Robins, 2010, Tchetgen Tchetgen, 2011). Likewise, it has recently been shown that a likelihood ratio test of the null hypothesis of no RERI incorporating the G-E independence assumption (hereafter RERI-LRT) generally outperforms the standard RERI test of no additive interaction (Han et al, 2012). Notably, both RERI-based tests of additive interaction rely on correct specification of a logistic regression for disease risk, as a function of G, E and covariates. In practice, the outcome regression may be difficult to specify particularly if both the environmental exposure and some of the covariates are count or continuous, so that nonparametric estimation is not a viable option. Furthermore, model mis-specification of main effects in logistic regression for the outcome can a priori rule out the null hypothesis of no additive

[1]Corresponding Author: Eric J Tchetgen Tchetgen, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115. Email: etchetgen@gmail.com

interaction, possibly resulting in inflated type 1 error rates of RERI-based tests of interaction. The presence of covariates has previously been noted as potentially problematic in the RERI framework by Skrondall (2003). He argues with conviction that given a conceptualization of interaction as departure from additive risks, making direct inferences regarding the fundamental additive interaction parameter would be preferred to the indirect RERI-based strategy, in order to avoid potential bias due to model misspecification of a logistic regression for the outcome.

In this paper, the authors present a general, yet fairly straightforward approach to directly test for the presence of additive GxE interaction in case-control studies without requiring a regression model of disease risk. The proposed approach which is easily made to exploit the G-E independence assumption leading to dramatic increase in power, relies on separate regression models for G and E given covariates. By avoiding building a model for the outcome risk, which is shown to be strictly unnecessary to test for additive interaction, the approach circumvents well known difficulties with RERI and is completely robust to possible mis-specification of such an outcome regression, provided one can correctly specify a pair of regression models for G and E. The methods are illustrated through an extensive simulation study in the simple setting of binary G and E free of covariates so that RERI and the new approach both apply and therefore can be directly compared to assess relative efficiency. Next, we demonstrate the new approach using data from a well-known ovarian cancer application to detect an additive interaction between the BRCA1/2 genetic variant (G) and a woman's parity (E), as well as with number of years of oral contraceptive use (E). Because either exposure E is naturally a count in this application RERI-based strategies cannot easily be implemented without possibly recoding the original environmental exposures as dichotomous or as categorical with few levels. Furthermore, as previously discussed, covariate adjustment may present additional difficulty for the RERI approach which is therefore forgone in the application in favor of the proposed approach.

Alternative characterization of test of additive interaction

Suppose one has observed case-control data on n unrelated individuals, let D denote the rare disease outcome defining case-control status and (A_1, A_2) denote two exposures in view. For instance, in a statistical genetic application, A_1 may denote the genetic variant G and A_2 an environmental exposure E, however, we will use the more generic notation (A_1, A_2) to allow for more general contexts, say where either or both exposures may be count or continuous environmental exposures. Let $\mu(a_1, a_2) = \Pr(D = 1|A_1 = a_1, A_2 = a_2)$ denote the disease risk of individuals in the target population with exposure values $(a_1,\ a_2)$. In the case of binary exposures, an additive interaction between A_1 and A_2 is said to be present if

$$\beta_3 = \mu(1,1) - \mu(1,0) - \mu(0,1) + \mu(0,0) \neq 0,$$

or equivalently if $RERI \neq 0$, where

$$\begin{aligned} RERI &= \{\mu(1,1) - \mu(1,0) - \mu(0,1)\}/\mu(0,0) + 1 \\ &= \beta/\mu(0,0). \end{aligned}$$

An empirical version of RERI is obtained under case-control sampling by estimating the required risk ratios $\mu(a_1, a_2)/\mu(0,0)$, via a saturated logistic regression under the rare disease assumption. Then, standardizing the empirical estimate $\widehat{RERI}$ by a consistent estimate of its standard error $\sqrt{\widehat{\sigma}^2_{RERI}}$ gives the RERI test statistic $T_{RERI} = \widehat{RERI}/\sqrt{\widehat{\sigma}^2_{RERI}}$. It can then be showed using standard asymptotic arguments that under the null hypothesis $H_0 : \beta_3 = RERI = 0$, T_{RERI} is approximately standard normal in large samples. The following result gives an alternative characterization of the null hypothesis of no additive interaction which motivates the new approach.

To state the result, let $\pi_1(a_2) = \Pr(A_1 = 1|A_2 = a_2)$ denote the prevalence of the first exposure A_1 among individuals with the second exposure $A_2 = a_2$ in the underlying population, and likewise define $\pi_2(a_1) = \Pr(A_2 = 1|A_1 = a_1)$. Also let α denote the log odds ratio association relating A_1 and A_2 in the target population, thus

$$\exp \alpha = \frac{\pi_1(1)(1 - \pi_1(0))}{\pi_1(0)(1 - \pi_1(1))}$$

such that $\alpha = 0$ encodes the independence assumption between A_1 and A_2.

Result 1 *The null hypothesis of no additive interaction H_0 holds if and only if*

$$\mathbb{E}\{U|D = 1\} = 0$$

where

$$U = e^{-\alpha A_1 A_2}(A_1 - \pi_1(0))(A_2 - \pi_2(0))D$$

.

We should note that Result 1 does not rely on the rare disease assumption and holds irrespective of the population disease prevalence. The result is a special case of a more general Lemma given later in the text allowing for arbitrary exposures and for covariate adjustment. According to the result, the null hypothesis of no additive interaction holds if and only if the random variable U has mean zero. Intuition about the result is gained by assuming G-E independence, i.e. $\alpha = 0$, such that $\pi_j(a) = \pi_j$. Then, upon noting that the conditional density of (A_1, A_2) given $D = 1$ is proportional to

$$\mu(A_1, A_2)f_1(A_1)f_2(A_2) = (\beta_0 + \beta_1 A_1 + \beta_2 A_2 + \beta_3 A_1 A_2)f_1(A_1)f_1(A_2)$$

where $f_j(1) = \pi_j$, one observes that $\mathbb{E}\{U|D=1\}$ is proportional to

$$\begin{aligned}
&\sum_{a_1,a_2} (a_1-\pi_1)(a_2-\pi_2)(\beta_0+\beta_1 a_1+\beta_2 a_2+\beta_3 a_1 a_2) f_1(a_1) f_1(a_2) \\
&= \beta_0 \underbrace{\sum_{a_1,a_2} (a_1-\pi_1)(a_2-\pi_2) f_1(a_1) f_1(a_2)}_{=0} \\
&+ \beta_1 \underbrace{\sum_{a_1,a_2} (a_1-\pi_1)(a_2-\pi_2) a_1 f_1(a_1) f_1(a_2)}_{=0} + \beta_2 \underbrace{\sum_{a_1,a_2} (a_1-\pi_1)(a_2-\pi_2) a_2 f_1(a_1) f_1(a_2)}_{=0} \\
&+ \beta_3 \sum_{a_1,a_2} (a_1-\pi_1)(a_2-\pi_2) a_1 a_2 f_1(a_1) f_1(a_2) \\
&= \beta_3 \sum_{a_1,a_2} \pi_1(1-\pi_1)\pi_2(1-\pi_2)
\end{aligned}$$

confirming that $\mathbb{E}\{U|D=1\} = 0$ if and only if the additive interaction $\beta_3 = 0$. Result 1 further shows that a similar result holds when the exposures are dependent upon applying a weight to individuals with both exposures, equal to the inverse of the odds ratio association of the two exposures. Intuitively, weighting makes the exposures independent, thus essentially recovering the independent exposure setting in the weighted sample. Since U only uses exposure data among cases (with $D = 1$), the result suggests that one may be able to test for additive interaction by considering whether the distribution of the exposures in view satisfies the above condition using data for cases only. Unfortunately, U is not directly observed and therefore cannot directly be used for inference, as it depends on the unknown population parameters $\pi_j(0)$, $j = 1, 2$. Nonetheless, progress can be made under the rare disease assumption, since one may use the controls (with $D = 0$) for approximate inference, upon observing that $\pi_j(0) \approx p_j(0)$ where $p_1(a_2) = \Pr(A_1 = 1|A_2 = a_2, D = 0)$ and $p_2(a_1) = \Pr(A_2 = 1|A_1 = a_1, D = 0)$. Therefore, one may estimate $\sum_i U_i$

with $\sum_i \widehat{U}_i$ where

$$\widehat{U}_i = \exp\left(-A_{1,i}A_{2,i}\widehat{\omega}\right)\left(A_{1,i} - \widehat{p}_1(0)\right)\left(A_{2,i} - \widehat{p}_2(0)\right)D_i,$$

with $\widehat{p}_1(a) = \sum_i A_{1,i} I(A_{2,i} = a, D = 0) / \sum_i I(A_{2,i} = a, D = 0)$ the sample version of $p_1(a)$, $\widehat{p}_2(a)$ similarly defined, and $\exp(\widehat{\omega}) = \widehat{p}_1(1)(1-\widehat{p}_1(0))/\widehat{p}_1(0)(1-\widehat{p}_1(1))$ the sample odds ratio relating A_1 and A_2 in the controls. In the Appendix, we show how to derive $\sigma^2_{\widehat{U}} = Var(\sum_i \widehat{U}_i/n)$, see equation (2) . Suppose that unbeknownst to the analyst, A_1 and A_2 are independent in the population and therefore $\widehat{\omega}$ converges to 0 in probability. We evaluate $\sigma^2_{\widehat{U}}$ at this particular submodel and show that $\sigma^2_{\widehat{U}}$ can be decomposed as $\widehat{\sigma}^2_{\widehat{U}} = \widehat{V}_1 + \widehat{V}_2 + \widehat{V}_3$, where $\widehat{V}_j$ is an estimate of V_j, $j = 1, 2, 3$, described in the Appendix. Considering in turn each contribution to the variance, we note that the first term $\widehat{V}_1$ captures the variance of $\sum_i U_i/n$ if $(\omega, p_1(0), p_2(0))$ were known; the second term $\widehat{V}_2$ reflects the uncertainty due to estimation of $(p_1(0), p_2(0))$; while $\widehat{V}_3$ reflects the uncertainty associated with estimation of the odds ratio parameter ω. Below, we further consider how the G-E independence assumption alters each of these contributions to reveal how the assumption can improve power to detect the presence of an additive interaction. Here we note that, under H_0 the standardized test statistic $T = \sum_i \widehat{U}_i/n\sqrt{\widehat{\sigma}^2_{\widehat{U}}}$ is approximately standard normal in large samples. Under the two-sided alternative hypothesis $\beta \neq 0$, one can further show that in large samples, T has approximate variance one, and is approximately centered at the non-centrality parameter $\kappa \times \beta_3$, where:

$$\kappa = p_1(0)\left(1 - p_1(0)\right)p_2(0)\left(1 - p_2(0)\right)\lambda/\sigma^2_{\widehat{U}},$$

λ is the sampling fraction for cases (i.e. λ = proportion of cases in case-control sample/proportion of cases in population). Thus, T has asymptotic power one since $1/\sigma^2_{\widehat{U}}$ and therefore κ tends to infinity with sample size; confirming that similar to T_{RERI}, T is a consistent test statistic of H_0. However, neither T nor T_{RERI} make explicit use of the G-E independence assumption and

therefore both may be inefficient in finite sample if the assumption holds. In the following section, we modify T to incorporate the independence assumption to obtain a more powerful test statistic.

Test incorporating independence assumption

Suppose that A_1 and A_2 are known to be independent in the population. Naturally, one may wish to exploit such prior information in testing for G-E interaction. This can be accomplished by adapting the methodology developed in the previous section upon noting that the independence assumption implies $\alpha = 0$, which, under the rare disease assumption, also implies that $\omega \approx 0$. This leads us to modify $\widehat{U}_i$. Define $\widetilde{U}_i$ similarly to $\widehat{U}_i$ with $\widehat{\omega} = 0$, i.e. $\widetilde{U}_i = (A_{1,i} - \widehat{p}_1(0))\,(A_{2,i} - \widehat{p}_2(0))\,D_i$. In the appendix, we show that $\sigma^2_{\widetilde{U}} = Var(\sum_i \widetilde{U}_i/n)$ can be estimated by $\widehat{\sigma}^2_{\widetilde{U}} = \widehat{V}_1 + \widehat{V}_2$. Consequently $\widehat{\sigma}^2_{\widetilde{U}} < \widehat{\sigma}^2_{\widehat{U}}$, reflecting the efficiency gain due to the independence assumption, i.e. $\widehat{V}_3$ is exactly zero since there is no uncertainty associated with $\widehat{\omega} = 0$. One can verify that the non-centrality parameter $\beta \times \kappa_1$ of $T_1 = \sum_i \widetilde{U}_i/n\sqrt{\widehat{\sigma}^2_{\widetilde{U}}}$ becomes $\kappa_1 = \frac{\sigma_{\widehat{U}}}{\sigma_{\widetilde{U}}}\kappa > \kappa$, confirming that T_1 is guaranteed to be more powerful than T.

Adjusting for covariates

In observational studies, it is usually desirable to adjust for potential confounding of the joint effects of A_1 and A_2, and such covariate adjustment may also be required to enforce the G-E independence assumption. Let X denote such a vector of covariates and suppose that the exposures are independent conditional on X. Define $p_1(x) = \Pr(A_1 = 1|X = x, D = 0)$ and $p_2(x) = \Pr(A_2 = 1|X = x, D = 0)$. Likewise, let $\widehat{p}_1(x)$ and $\widehat{p}_2(x)$ correspond to estimates, obtained using standard parametric models, e.g. using logistic regressions of the form $\text{logit}\widehat{p}_j(x) = \text{logit}p_j(x;\widehat{\theta}_j) = (1, x')\widehat{\theta}_j, j = 1, 2$, computed by maximum likelihood. The test statistic $T_2 = \sum_i \overline{U}_i/\sqrt{\widehat{\sigma}^2_{\overline{U}}}$ has under the null hypothesis of no additive interaction, an approximate standard normal distribution,

with $\overline{U}_i$ defined as

$$\overline{U}_i = \left(A_{1,i} - \widehat{p}_1(X)\right)\left(A_{2,i} - \widehat{p}_2(X)\right) D_i,$$

where $\widehat{\sigma}^2_{\overline{U}}$ is obtained using equation (2) of the Appendix.

More general exposures

Next, suppose that the environmental exposure A_2 were continuous, for example if D were diabetes status, A_2 could be body mass index (BMI) typically coded on a continuous scale. Note that the null hypothesis of no additive interaction can be restated as followed to acknowledge the continuous exposure:

$$H_0 : \mu(1, a_2, x) - \mu(1, 0, x) - \mu(0, a_2, x) + \mu(0, 0, x) = 0 \text{ for all values of } a_2 \text{ and } x,$$

where $\mu(a_1, a_2, x) = \Pr(D = 1|a_1, a_2, x)$. To construct an appropriate test statistic of H_0, suppose that $\mathbb{E}\left(A_2|X = x, D = 0\right)$ is estimated with the linear model $\widehat{m}_2(x) = m_2(x, \widehat{\theta}_2) = (1, x')\widehat{\theta}_2$ via ordinary least squares using controls only. Assuming G-E conditional independence given X, it is straightforward to modify the proposed test statistic to account for the continuous exposure, by simply replacing $\widehat{p}_2(x)$ with $\widehat{m}_2(x)$. Thus, we let

$$\overline{U}^c_i = \left(A_{1,i} - \widehat{p}_1(X_i)\right)\left(A_{2,i} - \widehat{m}_2(X_i)\right) D_i,$$

and $\widehat{\sigma}^2_{\overline{U}^c}$ denotes an estimate of the variance of $\sum_i \overline{U}^c_i/n$ obtained using equation (2) of the Appendix. Then, the test statistic $T_3 = \sum_i \overline{U}^c_i/n\sqrt{\widehat{\sigma}^2_{\overline{U}^c}}$ is approximately standard normal under H_0. A similar test statistic could be defined if A_2 were a count, upon estimating its mean with the log-linear model $\log n_2(x, \widehat{\theta}_2) = (1, x')\widehat{\theta}_2$ computed by maximum likelihood under say a Poisson model for A_2. Then, one could simply replace $\widehat{m}_2$ with $\widehat{n}_2$ in defining the test statistic, and one

could likewise modify the estimated variance of the test statistic using (2).

In order to simplify the presentation, thus far we have taken A_1 to be a binary genetic variant. Suppose now that A_1 were more generallly categorical having K possible levels $\{0, a_{1,1}, ..., a_{1,K-1}\}$ with 0 a reference value. For instance, if A_1 were to encode the number of minor alleles measured at a single nuclueotide polymorphism (SNP) locus, then $K = 3$, and $a_{1,k} = k$, $k = 1, 2$. Further assuming say that A_2 were continuous and independent of A_1 given X, we could then simply define

$$\overline{U}_i^m = \sum_{k=1}^{K-1} \left(I(A_{1,i} = a_{1,k}) - \widehat{p}_{1,k}(X_i)\right) \left(A_{2,i} - \widehat{m}_2(X_i)\right) D_i,$$

where $\widehat{p}_{1,k}(x)$ is a maximum likelihood estimate of $\Pr(A_1 = a_k|x)$ computed using standard polytomous logistic regression. Let $\widehat{\sigma}^2_{\overline{U}^m}$ denote an estimate of the large sample variance of $\sum_i \overline{U}_i^m/n$ based on (2). Then in large samples, the resulting test statistic $T_4 = \sum_i \overline{U}_i^m/n\sqrt{\widehat{\sigma}^2_{\overline{U}^m}}$ is approximate standard normal under the null hypothesis of no additive interaction which may be restated to account for the polytomous and continuous exposures:

$$H_0 : \mu(a_{1,k}, a_2, x) - \mu(a_{1,k}, 0, x) - \mu(0, a_2, x) + \mu(0, 0, x) = 0 \text{ for all } k, \text{ and all values of } a_2 \text{ and } x.$$

A unified class of test statistics

We now provide a unified class of test statistics for the null hypothesis of no additive interaction which subsumes as special case, each of the tests considered in previous sections, but allows for the conditional independence assumption of the two exposures to be relaxed.

To do so, we proceed as in Tchetgen Tchetgen et al (2010) and use the following representation

of the joint density of (A_1, A_2) given X :

$$f(A_1, A_2|X) = \frac{f(A_1|A_2 = 0, X)f(A_2|A_1 = 0, X)OR(A_1, A_2; X)}{\int\int f(a_1|A_2 = 0, X)f(a_2|A_1 = 0, X)OR(a_1, a_2; X)d\nu(a_1, a_2)} \quad (1)$$

where ν is a dominating measure of the distribution of (A_1, A_2) , $OR(A_1, A_2; X)$ is the generalized odds ratio function relating A_1 and A_2 within levels of X, that is

$$OR(A_1, A_2; X) = \frac{f(A_1, A_2|X)f(A_1 = 0, A_2 = 0|X)}{f(A_1 = 0, A_2|X)f(A_1, A_2 = 0|X)}$$

and $\{f(A_1|A_2 = 0, X), f(A_2|A_1 = 0, X)\}$ are baseline densities in the target population. Note that the generalized odds ratio function reduces in the simple case of binary exposures, to the standard odds ratio effect measure, but remains well defined as a measure of association for exposures of a more general nature, whether categorical, count or continuous variables, i.e. $OR(A_1, A_2; X) = 1$ if and only if A_1 and A_2 are independent within levels of X. The null hypothesis of no additive interaction can more generally be stated as:

$$H_0 : \mu(a_1, a_2, x) - \mu(a_1, 0, x) - \mu(0, a_2, x) + \mu(0, 0, x) = 0 \text{ for all values of } a_1, a_2 \text{ and } x.$$

For any function $g(A_1, A_2, X)$ of (A_1, A_2, X), let

$$
\begin{aligned}
w(A_1, A_2, X, D; g) &= W(g) \\
&= OR(A_1, A_2; X)^{-1}\{g(A_1, A_2, X) - \int g(A_1, a_2, X) f(a_2|A_1 = 0, X) d\mu(a_2) \\
&- \int g\left(a_1, A_2, X\right) f\left(a_1|A_2 = 0, X\right) d\mu(a_1) \\
&+ \int g(a_1, a_2, X) f\left(a_2|A_1 = 0, X\right) f\left(a_1|A_2 = 0, X\right) d\mu(a_1, a_2)\} D
\end{aligned}
$$

Lemma *The null hypothesis H_0 holds if and only if*

$$\mathbb{E}\left\{W(g)|D = 1, x\right\} = 0 \text{ for all values of } x \text{ and all functions } g$$

Result 1 is easily recovered as a corollary of the Lemma. According to the Lemma, an empirical version of $W(g)$ with user-specified function g may be used to test H_0. One must estimate the unknown odds ratio function and baseline densities, in order to obtain an estimate of the joint density of (A_1, A_2) given X. Under the rare disease assumption, estimation of the joint density can proceed by standard maximum likelihood using only the controls under the parametrization given in equation (1), upon positing parametric models for the odds ratio function and baseline densities. To ground ideas, suppose one posits parametric models $OR(A_1, A_2; X; \omega)$, $f(A_1|A_2 = 0, X; \alpha_1)$ and $f(A_2|A_1 = 0, X; \alpha_2)$, e.g. a single parameter model $\log OR(A_1, A_2; X) = \omega A_1 A_2$ may be used that encodes the assumption that the odds ratio association between A_1 and A_2 given X does not vary with X, i.e. no effect heterogeneity in X of the odds ratio association between A_1 and A_2 in the population. For exposures that are either binary, continuous or counts, generalized linear models within the exponential family may be used to model the baseline densities. For example, counts may be modeled by assuming a Poisson distribution for the corresponding baseline density. Let

$\widehat{\omega}$, $\widehat{f}(A_1|A_2=0,X)$ and $\widehat{f}(A_2|A_1=0,X)$ denote the approximate maximum likelihood estimate of (1) using controls only; and let $\widehat{W}(g)=W(g,\widehat{\theta})$ denote the resulting estimate of $W(g)$, where $\theta=(\omega,\alpha_1,\alpha_2)$. Our proposed test statistic is then given by $Z=\sum_i \widehat{W}_i(g)/n\sqrt{\widehat{\sigma}^2_W}$, where $\widehat{\sigma}^2_W$ is the estimate of $Var\left(\sum_i \widehat{W}_i(g)/n\right)$ provided in the Appendix.

It is straightforward to verify that the test statistics considered in previous sections belong to the above unifying class of test statistics. For instance, the test statistics proposed to handle binary, continuous or count exposures under the independence assumption are obtained by taking:

$$g(A_1,A_2,X)=\left(A_1-\widehat{\mathbb{E}}\left(A_1|X\right)\right)\left(A_2-\widehat{\mathbb{E}}\left(A_2|X\right)\right)$$

where $\widehat{\mathbb{E}}\left(A_j|X\right)$ is the mean of A_j evaluated under $\widehat{f}(A_j|X)$, $j=1,2$. For A_1 categorical with K distinct categories and A_2 binary, continuous or a count, one likewise obtains the test statistic previously proposed by taking:

$$g(A_1,A_2,X)=\sum_{k=1}^{K-1}\left(I(A_{1,i}=a_{1,k})-\widehat{\mathbb{E}}\left(I(A_{1,i}=a_{1,k})|X\right)\right)\left(A_2-\widehat{\mathbb{E}}\left(A_2|X\right)\right)$$

Under the independence assumption, the asymptotic variance of $Var\left(\sum_i \widehat{W}_i(g)/n\right)$ is easily modified to account for the assumption that $OR(A_1,A_2;X)$ is set to 1 for all persons in the sample.

Relaxing the rare disease assumption

In case the rare disease assumption does not apply, estimating exposure regression models in the controls only may not be entirely appropriate. Nonetheless, it may still be possible to test for the presence of an additive interaction, for instance if as often the case in nested case-control studies, sampling fractions for cases and controls were known. Then, standard inverse probability weighting could be used based on known sampling weights to estimate population models for the

exposures using both cases and controls. Potentially more efficient estimates of models for the exposures could alternatively be obtained using more recent methodology for regression analysis of secondary outcomes in case-control studies (Tchetgen Tchetgen, 2013).

A simulation study

We study the power and type 1 error of our proposed test in the standard setting of binary genetic and environmental variables with no other covariate, so that it is more easily compared to the approach of Han et al (2012). In order to evaluate both type error rates and power of various test statistics, we generated simulated data following the design of Han et al (2012) which encodes the magnitude of the interaction indirectly by varying RERI from 0 (to assess type 1 error) to 0.5. The probability of having the genetic variant was 0.5, and the probability of the binary environmental variable was 0.2, and these factors were generated to be independent. Let $\text{logit}(p) = \log\{p/(1-p)\}$ and $\text{expit}(z) = \exp(z)/[\exp(z)+1]$. The disease risk model was

$$\text{logit}\,\Pr(D=1|a_1,a_2) = \alpha_0 + \alpha_1 a_1 + \alpha_2 a_2 + \alpha_3 a_1 a_2;$$

with baseline risk equal to 0.01 (i.e. $\alpha_0 = \text{logit}(0.01)$), the gene and environment main effects were varied so that $(\alpha_1, \alpha_2) \in \{\log(0.7), \log(1.2); \log(2)\}$, and the multiplicative G-E interaction parameter α_3 was selected to yield the desired RERI, according to the formula

$$\alpha_3 = \text{logit}[(\text{RERI}-1)\text{expit}(\alpha_0) + \text{expit}(\alpha_0+\alpha_1) + \text{expit}(\alpha_0+\alpha_2)] - \alpha_0 - \alpha_1 - \alpha_2$$

In each simulation, we generate 4000 cases and 4000 controls. We report results for 10,000 simulations for each setting corresponding to a particular combination of (α_1, α_2) and RERI values.

Figure 1 summarizes results in terms of power plots comparing the proposed tests with and without using the G-E independence assumption, labeled 'U ind' and 'U' respectively. The figure also presents results for the retrospective profile likelihood ratio test proposed by Han et al (2012) with and without using the independence assumption respectively, labeled 'Han ind' and 'Han' respectively. Finally, the Figure also displays results from the standard RERI test based on prospective logistic regression, which is labeled 'prosp'.

Insert Figure here.

All tests appear to have correct type 1 error rate as shown in the Figure, as well as in the more detailed Table provided in the Appendix. One observes that the RERI-LRT test 'Han ind' and 'U ind' are equally powerful when $\Pr(G = 1) = 0.5$ across various possible values for the other parameters, and both tests are dramatically more powerful when compared to the other tests, while 'U' is slightly less powerful than 'Han', which is in turn slightly less powerful than 'prosp'.

In additional simulations, we varied the prevalence of the genetic marker $\Pr(G = 1)$ to have population probabilities 0.2 and 0.05, while the environmental factor was maintained to have probability 0.2. Power plots similar to those appearing in Figure 1 are provided in the supplementary material for these additional settings. These additional simulations confirm that all tests become less powerful as the genetic variant becomes less common, with 'Han ind' being slightly more powerful than 'U ind' when $\Pr(G = 1) = 0.05$. Overall, the simulation study confirms that the proposed approach performs quite competitively when compared with the efficient RERI-LRT approach, in settings where both methods are available.

In the following section, we consider a data application of the new approach in which RERI is no longer readily available and cannot easily be applied without further making unnecessary assumptions.

Ovarian cancer application

We applied the proposed test of additive interaction to the well-known Israeli Ovarian Cancer data (Modan et al., 2011) also recently analyzed by Chatterjee and Carroll (2005), Tchetgen Tchetgen and Robins (2010) and Tchetgen Tchetgen (2011). Although the goal in previous analyses was to detect a multiplicative gene-environment interaction between having the BRCA1/2 mutation and two environmental exposures, number of years of oral contraceptive use (OC) and number of children (parity), here we are primarily concerned with determining whether such interactions might be operating on the additive scale. Both environmental exposures are naturally coded as counts, and therefore can be modeled using Poisson regression, while standard logistic regression was used to model the genetic variant. Both sets of models were estimated only using controls as previously described. We present results when assuming G-E independence, and without using such an assumption. Without G-E independence, the odds ratio parameter ω was estimated as the coefficient for the exposure in view in a logistic regression of the genetic factor on the environmental exposure and covariates, i.e. (E, X).

A number of covariates were available for confounding adjustment and also to enforce the independence assumption. All regression models adjusted for age, as an indicator variable for age$\leq$ 50, indicator variables for ethnicities of Ashkenazi jew, and non-Ashkenazi (with mixed race serving as reference category), indicator variables for personal history of breast cancer, family history of breast cancer, and family history of ovarian cancer. For convenience, we used the nonparametric bootstrap to evaluate 95% confidence intervals and p-values.

Insert Table Here.

The table provides results from testing for a GxE additive interaction with and without making the G-E independence assumption. In accordance with simulation results, the independence

assumption yields a test statistic consistently more extreme for both exposures in view than the corresponding test which does not incorporate the assumption. Specifically, we succesfully reject the null hypothesis of no additive G-E interaction between BRCA1/2 mutation and parity at the alpha level of 0.05, only when the independence assumption is made, and not otherwise. We found no conclusive evidence of an additive interaction with OC, although the test statistic under G-E independence was far more extreme than without the assumption and the associated p-value was marginally significant (p-value=0.09). In conclusion, we found significant evidence that the well-known strong association between BRCA1/2 may be tempered by the number of children a woman has given birth to. It is interesting to compare these findings with previous analyses of these data that have primarily been concerned with detecting the presence of a multiplicative GxE interaction. For instance, Tchetgen Tchetgen and Robins (2010) leveraged the independence assumption to detect a GxE multiplicative interaction only with OC and failed to find evidence of a similar interaction with parity, thus essentially reporting the opposite findings to ours. However, our findings are potentially more scientifically relevant given that interactions on the multiplicative scale may be harder to interpret mechanistically.

Conclusion

We have described a very general framework to test for GxE additive interactions exploiting G-E independence in case-control studies. The proposed strategy has several advantages over existing RERI-based strategies, primarily because, unlike the latter, the former does not require a regression model for the outcome, and therefore is less vulnerable to model misspecification of main effects, a potential concern particularly if E is a count or continuous and additional covariates are included in the regression. The approach put forward in this paper is closely related to the semiparametric framework of Vansteelandt et al (2008) and Tchetgen Tchetgen (2012), who characterized the set

of influence functions of a model of interaction (on the additive or multiplicative scale and odds ratio scale respectively) under a semiparametric union model in which only a subset but not all of the parametric models used to describe the data generating mechanism need to be correct for valid inference. In fact, one can show that our proposed test statistic belongs to the general class of test statistics for additive interaction associated with the set of influence functions of Vansteelandt et al (2008). However, because Vansteelandt et al (2008) did not allow for outcome dependent sampling and only considered standard prospective random sampling, not all test statistics in their class may be used under case-control sampling. Thus, an important contribution of the current paper has been to characterize the subset of the Vansteelandt et al (2008) class of test statistics of an additive interaction that may be used both under prospective and retrospective sampling. An additional contribution made in the current paper, is to clearly demonstrate the potential for efficiency gain by incorporating the G-E independence assumption under case-control sampling.

In recent years, there has been growing interest in genomewide interaction studies (GWIS) aimed at identifying regions of the genome that may be implicated in moderate to large interactions with a given environmental exposure (Murcray et al, 2009, Khoury and Wacholder, 2009, Cornelis et al, 2012, Mukherjee et al, 2012, Thomas et al, 2012). However, existing GWIS have to date mostly considered multiplicative interactions, and few such studies have been succesful at detecting such interactions. It may be that genetic variants implicated in true mechanistic interactions with environmental factors do not produce sufficiently large multiplicative interactions to be detected in multiplicative GWIS. Thus, it may be of definite interest in the future to consider GWIS that directly target additive interactions (GWISAdd). Given that most variants genotyped in a typical genomewide association study are likely independent of environmental factors (Thomas et al, 2012), GWISAdd that leverage G-E independence using the methods developed in the current paper may be more likely to uncover additive interactions of greater relevance to learn about mechanism.

References

[1] Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika. 2005; 92(2):399–418.

[2] Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet. 2002;11(20):2463–2468.

[3] Cornelis M, Tchetgen Tchetgen EJ, Liang L, Qi L, Chatterjee N, Hu F, Kraft P. Gene-Environment Interactions in Genome-Wide Association Studies: A Comparative Study of Tests Applied to Empirical Studies of Type 2 Diabetes Am J Epidemiol. 2012 Feb 1;175(3):191-202.

[4] Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. Stat Med. 1983;2(2):243–251.

[5] Han S, Rosenberg, P, Garcia-Closas, Figuroa J, Silverman D, Chanock S, Rothman N, Chatterjee N. Likelihood Ratio Test for Detecting Gene (G)-Environment (E) Interactions Under an Additive Risk Model Exploiting G-E Independence for Case-Control Data. Am J Epidemiol. 2012;176(11):1060–1067.

[6] Khoury MJ, Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies—challenges and opportunities. Am J Epidemiol 2009;169(2):227-230. discussion 234–235.

[7] Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics 2008;64(3):685-694.

[8] Mukherjee B, Ahn J, Gruber SB, et al. Testing gene-environment interaction in large scale case-control association studies: possible choices and comparisons. Am J Epidemiol 2012;175.3: 177-190.

[9] Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. Am J Epidemiol. 2009;169(2):219–226.

[10] Modan, M. D., Hartge, P. et al. (2001). Parity, oral contraceptives and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. *New England Journal of Medicine.* 345, 235–40.

[11] Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med. 1994;13(2):153–162.

[12] Rothman KJ, Greenland S, Walker AM. Concepts of interaction. Am J Epidemiol. 1980;112(4):467–470.

[13] Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed (revised and updated). Philadelphia, PA: Lippincott Williams & Wilkins; 2008.

[14] Skrondal A. Interaction as departure from additivity in case control studies: a cautionary note. Am J Epidemiol. 2003;158(3): 251–258.

[15] Tchetgen Tchetgen E, Robins J and Rotnitzky A. On Doubly robust estimation of a semi-parametric odds ratio model. Biometrika. 2010, vol. 97(1), pages 171-180.

[16] Tchetgen Tchetgen E and Robins J. The semi-parametric case-only estimator. Biometrics. Biometrics. 2010 Dec;66(4):1138-44. doi: 10.1111/j.1541-0420.2010.01401.

[17] Tchetgen Tchetgen E . Robust Discovery of Genetic Associations incorporating Gene-Environment Interaction and Independence. (2011) Epidemiology. Volume 22 ;2; 262-272.

[18] Tchetgen Tchetgen, Eric J., "Multiple-Robust Estimation of an Odds Ratio Interaction" (February 2012). Harvard University Biostatistics Working Paper Series. Working Paper 142. http://biostats.bepress.com/harvardbiostat/paper142.

[19] Thomas, Duncan C., et al. "Invited commentary: GE-Whiz! Ratcheting gene-environment studies up to the whole genome and the whole exposome." American journal of epidemiology 175.3 (2012): 203-207.

[20] Umbach DM, Weinberg CR. Designing and analysing casecontrol studies to exploit independence of genotype and exposure. Stat Med. 1997;16(15):1731–1743.

[21] VanderWeele, Tyler J. and Knol, Mirjam J. A Tutorial on Interaction. Epidemiologic Methods. 2014. Published Online: 05/27/2014.

[22] Vansteelandt S., VanderWeele T.J., Tchetgen Tchetgen E, Robins J.M., Multiply robust inference for statistical interactions. Journal of the American Statistical Association, December 2008, Vol. 103, No. 484, Theory and Methods.

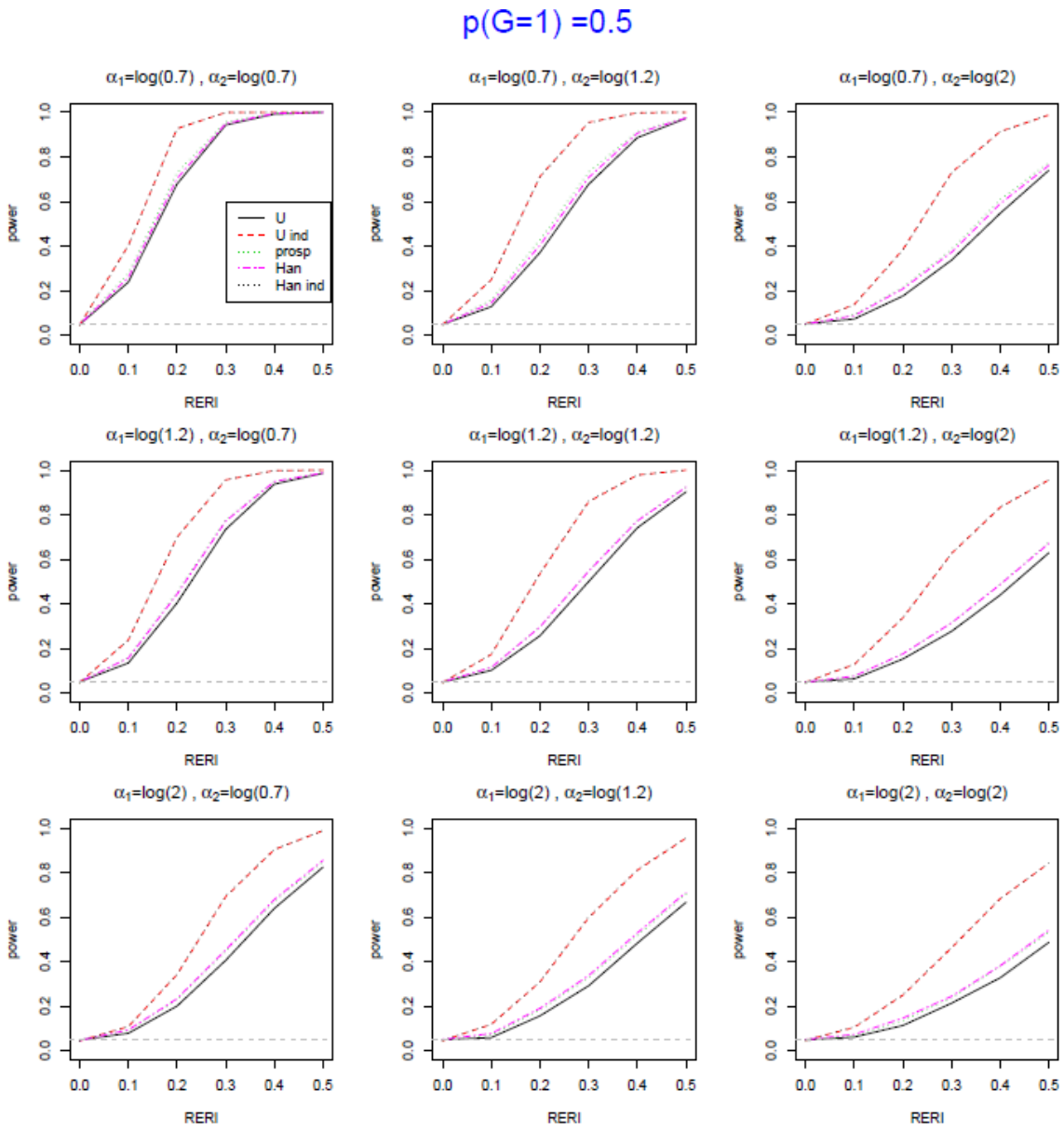


Figure 1. Power as a function of RERI and (α_1, α_2), when $\Pr(G = 1) = 0.5$ and sample size equal to 4000 cases and 4000 controls.

E variable	U	95% CI	p-values
G-E independence assumed			
OC	0.049	(-0.006, 0.117)	0.09
Parity	-0.044	(-0.092, -0.004)	0.03
No G-E independence assumed			
OC	0.002	(-4.692, 0.012)	0.77
Parity	-0.005	(-0.023, 0.019)	0.59

Table 1: Testing results for the additive G-E interaction between presence of BRCA1/2 mutation (G) and number of years of oral contraceptive (OC) use, and parity (E variables), with and without G-E independence assumption. U is the proposed (standardized) test statistic, and its 95% bootstrap confidence interval and p-value are provided, calculated over 1000 bootstrap samples.

APPENDIX

A general approach to detect gene (G)-environment (E) additive interaction leveraging G-E independence in case-control studies

by

Eric J Tchetgen Tchetgen, Tamar Sofer, Benedict HW Wong

Proof that $Var(\sum_i \widehat{U}_i/n) > Var(\sum_i \widetilde{U}_i/n)$.

To show the result requires the influence function of $\widehat{\theta} = (\widehat{\omega}, \widehat{p}_1 = \widehat{p}_1(0), \widehat{p}_2 = \widehat{p}_2(0))^T$ which is of the form

$$IF = \mathbb{E}\left(\frac{\partial R(\theta)}{\partial \theta}\right)^{-1} R(\theta) \tag{2}$$

where $R(\theta) = (1-D) \times ((A_1 - \mathbb{E}(A_1, \theta)), (A_2 - \mathbb{E}(A_2; \theta)), (A_1 A_2 - \mathbb{E}(A_1 A_2; \theta))^T$, where the first component is the score of $p_1(0)$, the second component is the score of $p_2(0)$, the last component is the score of ω, and $\theta = (\omega, p_1, p_2 = p_2)$. Standard matrix algebra can be used to show that at the submodel where A_1 and A_2 are independent $IF = (IF_1, IF_2, IF_3)$ where:

$$
\begin{aligned}
IF_1 &= \mathbb{E}\left[(1-A_2)(1-D)\right]^{-1}(1-A_2)\left(A_1-\mathbb{E}\left(A_1\right)\right)(1-D) \\
&\approx -\mathbb{E}\left[\left(1-\mathbb{E}\left(A_2|D=0\right)\right)\right]^{-1}\left(A_2-\mathbb{E}\left(A_2|D=0\right)\right)\left(A_1-\mathbb{E}\left(A_1|D=0\right)\right)(1-D) \\
&\quad +\left(A_1-\mathbb{E}\left(A_1|D=0\right)\right)(1-D) \\
IF_2 &= \mathbb{E}\left[(1-A_1)(1-D)\right]^{-1}(1-A_1)\left(A_2-\mathbb{E}\left(A_2|D=0\right)\right)(1-D) \\
&\approx -\mathbb{E}\left[\left(1-\mathbb{E}\left(A_1|D=0\right)\right)\right]^{-1}\left(A_1-\mathbb{E}\left(A_1|D=0\right)\right)\left(A_2-\mathbb{E}\left(A_2|D=0\right)\right)(1-D) \\
&\quad +\left(A_2-\mathbb{E}\left(A_2|D=0\right)\right)(1-D) \\
IF_3 &= \mathbb{E}\left[\left(A_1-\mathbb{E}\left(A_1|D=0\right)\right)^2|D=0\right]^{-1}\mathbb{E}\left[\left(A_2-\mathbb{E}\left(A_2|D=0\right)\right)^2|D=0\right]^{-1} \\
&\quad \times\left(A_1-\mathbb{E}\left(A_1|D=0\right)\right)\left(A_2-\mathbb{E}\left(A_2|D=0\right)\right)(1-D)
\end{aligned}
$$

A Taylor series argument then gives

$$\begin{aligned}
&\sum_i \widehat{U}_i/\sqrt{n} \\
&\approx \sum_i U_i/\sqrt{n} - \mathbb{E}\left[\left(A_2 - p_2(0)\right) D\right] IF_1 \\
&- \mathbb{E}\left[\left(A_1 - p_1(0)\right) D\right] IF_2 - \mathbb{E}\left[A_1 A_2 \left(A_2 - p_2(0)\right)\left(A_1 - p_1(0)\right) D\right] IF_3 \\
&= \sum_i U_i/\sqrt{n} \\
&- \mathbb{E}\left[\left(A_2 - p_2(0)\right) D\right] \sum_i \left(A_{1,i} - \mathbb{E}\left(A_1 | D = 0\right)\right)(1 - D_i)/\sqrt{n} \\
&- \mathbb{E}\left[\left(A_1 - p_1(0)\right) D\right] \sum_i \left(A_{2,i} - \mathbb{E}\left(A_2 | D = 0\right)\right)(1 - D_i)/\sqrt{n} \\
&- \left(\begin{array}{c} \mathbb{E}\left[A_1 A_2 \left(A_2 - p_2\right)\left(A_1 - p_1\right) D\right] \left\{p_1 p_2 \left(1 - p_1\right)\left(1 - p_2\right)\right\}^{-1} \\ + \mathbb{E}\left[\left(A_2 - p_2\right) D\right] \left[\left(1 - p_2\right)\right]^{-1} + \mathbb{E}\left[\left(A_1 - p_1\right) D\right] \left[\left(1 - p_1\right)\right]^{-1} \end{array} \right) \\
&\times \sum_i \left(A_{1,i} - \mathbb{E}\left(A_1 | D = 0\right)\right)\left(A_{2,i} - \mathbb{E}\left(A_2 | D = 0\right)\right)(1 - D_i)/\sqrt{n}
\end{aligned}$$

Upon noting that the above four terms are mutually uncorrelated, we have that :

$$Var\left(\sum_i \widehat{U}_i/n\right) \approx V_1 + V_2 + V_3$$

where

$$V_1 = Var\left(U\right)/n$$

$$V_2 = \mathbb{E}\left[\left(A_2 - p_2(0)\right)D\right]^2 Var\left(\left(A_1 - \mathbb{E}\left(A_1|D=0\right)\right)\left(1-D\right)\right)/n$$

$$+ \mathbb{E}\left[\left(A_1 - p_1(0)\right)D\right]^2 Var\left(\left(A_2 - \mathbb{E}\left(A_2|D=0\right)\right)\left(1-D\right)\right)/n$$

$$V_3 = \left(\begin{array}{c} \mathbb{E}\left[A_1 A_2 \left(A_2 - p_2\right)\left(A_1 - p_1\right)D\right]\left\{p_1 p_2 \left(1-p_1\right)\left(1-p_2\right)\right\}^{-1} \\ +\mathbb{E}\left[\left(A_2 - p_2\right)D\right]\left[\left(1-p_2\right)\right]^{-1} + \mathbb{E}\left[\left(A_1 - p_1\right)D\right]\left[\left(1-p_1\right)\right]^{-1} \end{array} \right)^2$$

$$\times Var\left(\left(A_1 - \mathbb{E}\left(A_1|D=0\right)\right)\left(A_2 - \mathbb{E}\left(A_2|D=0\right)\right)\left(1-D\right)\right)/n$$

A similar derivation shows that

$$\sum_i \widetilde{U}_i/\sqrt{n}$$

$$\approx \sum_i U_i/\sqrt{n}$$

$$- \mathbb{E}\left[\left(A_2 - p_2(0)\right)D\right]\sum_i \left(A_{1,i} - \mathbb{E}\left(A_1|D=0\right)\right)\left(1-D_i\right)/\sqrt{n}$$

$$- \mathbb{E}\left[\left(A_1 - p_1(0)\right)D\right]\sum_i \left(A_{2,i} - \mathbb{E}\left(A_2|D=0\right)\right)\left(1-D_i\right)/\sqrt{n}$$

which gives

$$Var\left(\sum_i \widetilde{U}_i/n\right) \approx V_1 + V_2$$

proving the result.

Asymptotic variance for unified class of test statistics

Our proposed test statistic is then given by $Z = \sum_i \widehat{W}_i(g)/n\widehat{\sigma}_W$, where $\widehat{\sigma}^2_W$ is an estimate of

$Var\left(\sum_i \widehat{W}_i(g)/n\right)$ one can derive using a standard Taylor series argument:

$$Var\left(\sum_i \widehat{W}_i(g)/n\right) \approx n^{-1}Var\left(W(g,\theta)\right) + n^{-1}\mathbb{E}\left(W_\theta^T(g)\right) Var\left(S_\theta^\dagger\right) \mathbb{E}\left(W_\theta(g)\right) \tag{3}$$

where $W_\theta(g)$ is the derivative of $W(g,\theta)$ with respect to θ evaluated at the truth, and $S_\theta^\dagger$ is the influence function of $\widehat{\theta}$. For instance, when $\widehat{\theta}$ is a maximum likelihood estimator, $S_\theta^\dagger = \mathbb{E}\left(S_\theta S_\theta^T\right)^{-1} S_\theta$, where S_θ denote the score of θ. Under the assumption that A_1 and A_2 are independent, we may set $\widehat{\omega} = 1$ and redefine $\theta = (\alpha_1, \alpha_2)$,also note that under independence, the joint density (1) in the text simplifies to $f(A_1, A_2|X) = f(A_1|X)f(A_2|X)$, leading to some simplification in the above expression for the asymptotic variance of the test statistic.

Proof of Lemma 1. *Consider the nonparametric additive representation of* $\mu(a_1, a_2, x)$ *given by* $\mu(a_1, a_2, x) = \beta_1(a_1, x) + \beta_2(a_2, x) + \beta_3(a_1, a_2, x) + \beta_4(x)$ *where* $\beta_1(a_1, x)$ *is the main effect of* A_1 *and satisfies* $\beta_1(0, x) = 0$, *likewise* $\beta_2(a_2, x)$ *is the main effect of* A_2 *and satisfies* $\beta_2(0, x) = 0, \beta_3(a_1, a_2, x)$ *is the additive interaction between* A_1 *and* A_2 *and satisfies* $\beta_3(0, a_2, x) = \beta_3(a_1, 0, x) = 0$, *and* $\beta_4(x)$ *is the main effect of* X. *For any function* g, *note that*

$$\begin{aligned}
&\mathbb{E}\left\{W(g)|D=1,x\right\} \\
&= \int\int w(a_1,a_2,x,1;g)\mu(a_1,a_2,x)f(a_1,a_2|x)f(x)d\nu(a_1,a_2) / \int\int \mu(a_1,a_2,x)f(a_1,a_2|x)f(x)d\nu(a_1,a_2) \\
&\propto \int\int w(a_1,a_2,x,1;g)\mu(a_1,a_2,x)f(a_1|a_2=0,x)f(a_2|A_1=0,x)OR(a_1,a_2;X)d\nu(a_1,a_2) \\
&= \int\int w(a_1,a_2,x,1;g)\beta_3(a_1,a_2,x)f(a_1|A_2=0,x)f(a_2|A_1=0,x)d\nu(a_1,a_2)
\end{aligned}$$

since

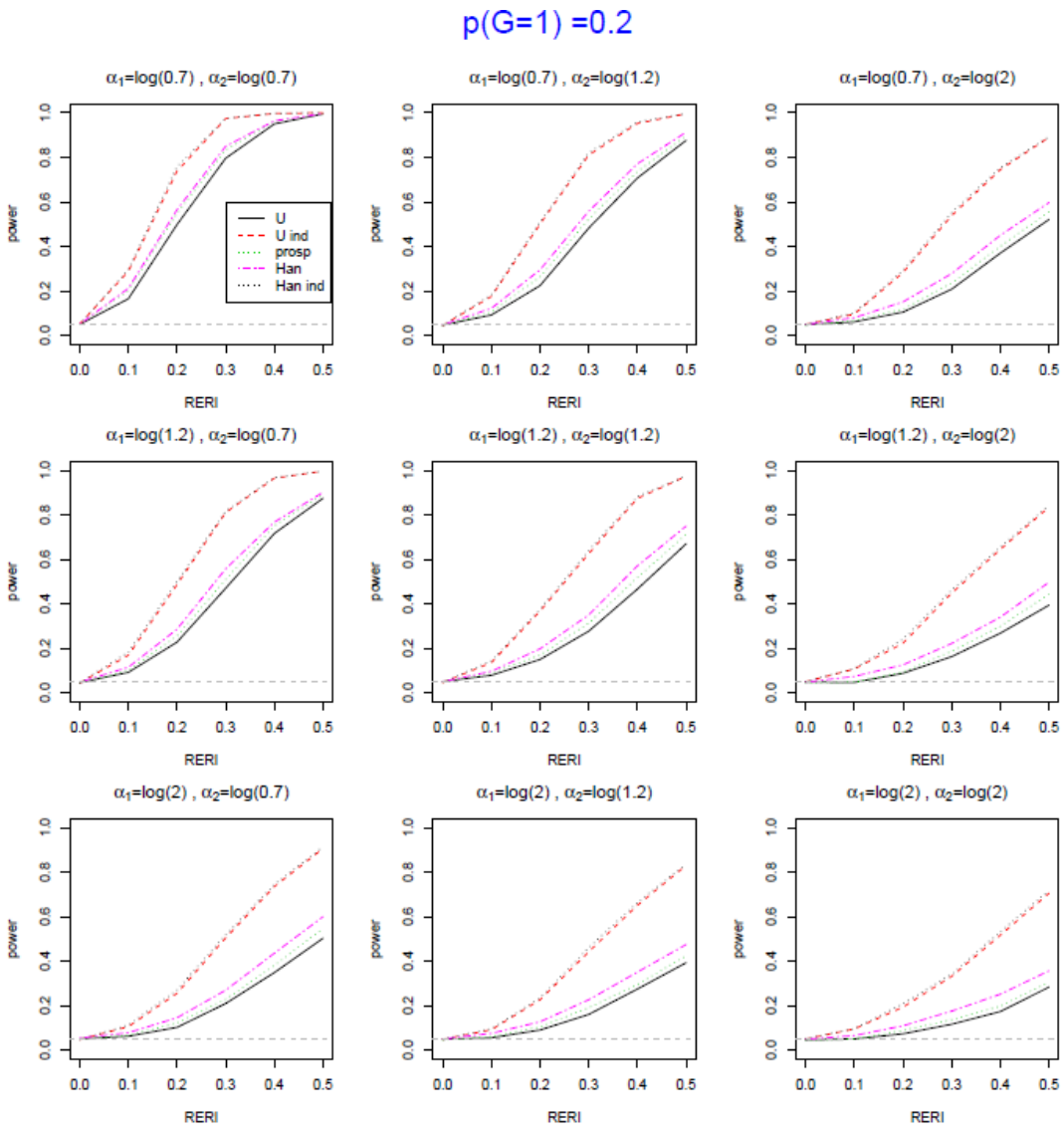
$$\int\int w(a_1,a_2,x,1;g)\left\{\beta_1(a_1,x)+\beta_2(a_2,x)+\beta_4(x)\right\} f(a_1|A_2=0,x)f(a_2|A_1=0,x)d\nu(a_1,a_2) = 0$$

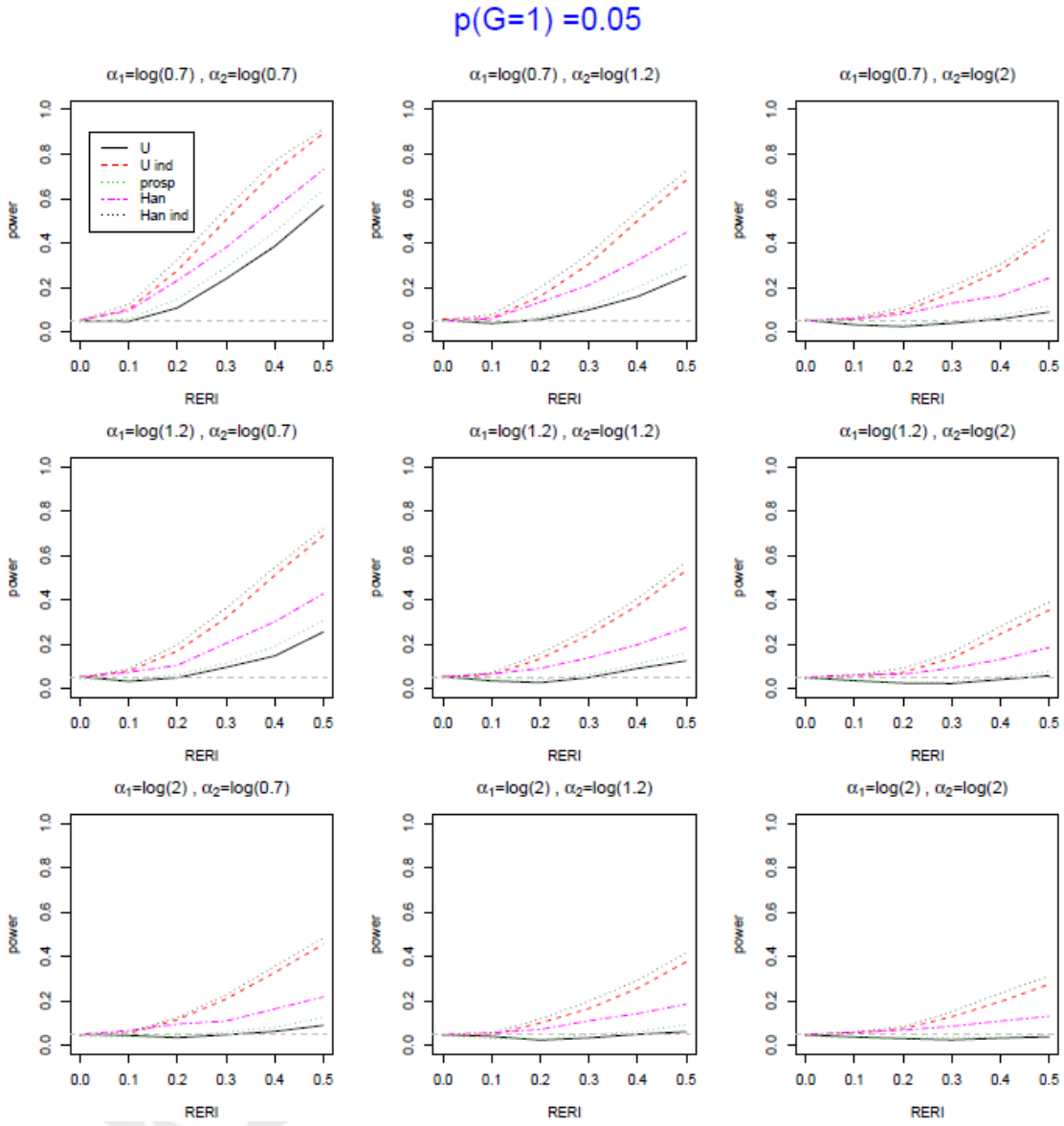
for any choice of g. *Thus, the null of no additive interaction* $\beta_3(a_1, a_2, x) = 0$ *for all* (a_1, a_2, x) *implies that* $E\{W(g)|D = 1, x\} = 0$. *We get the result in the other direction by choosing* $g(a_1, a_2, x) = g^*(a_1, a_2, x) = \beta_3(a_1, a_2, x)$ *which gives*

$$\mathbb{E}\{W(g)|D = 1, x\} = 0 \text{ for all } g \text{ and } x \text{ implies that}$$

$$\int\int w(a_1, a_2, x, 1; g^*)^2 f(a_1|A_2 = 0, x) f(a_2|A_1 = 0, x) d\nu(a_1, a_2) = 0 \text{ for all } x$$

which in turn implies that $\beta_3(a_1, a_2, x) = 0$ *for all* (a_1, a_2, x) *proving the result.* □





α_1 :	log(0.7)			log(1.2)			log(2)		
α_2 :	log(0.7)	log(1.2)	log(2)	α_2 = log(0.7)	log(1.2)	log(2)	log(0.7)	log(1.2)	log(2)
				$p(G = 1) = 0.5$					
U	0.048	0.049	0.051	0.050	0.048	0.049	0.049	0.050	0.048
U ind	0.047	0.048	0.049	0.047	0.047	0.049	0.046	0.046	0.048
prosp	0.050	0.048	0.047	0.048	0.048	0.046	0.046	0.049	0.049
Han	0.049	0.049	0.050	0.050	0.051	0.050	0.047	0.052	0.051
Han ind	0.047	0.048	0.050	0.047	0.048	0.049	0.046	0.046	0.048
				$p(G = 1) = 0.2$					
U	0.050	0.047	0.048	0.046	0.051	0.048	0.052	0.050	0.046
U ind	0.052	0.045	0.050	0.045	0.047	0.050	0.048	0.049	0.050
prosp	0.050	0.046	0.046	0.045	0.049	0.045	0.051	0.047	0.045
Han	0.051	0.048	0.050	0.048	0.050	0.048	0.055	0.051	0.050
Han ind	0.051	0.046	0.049	0.045	0.048	0.051	0.048	0.049	0.052
				$p(G = 1) = 0.05$					
U	0.050	0.054	0.053	0.054	0.054	0.050	0.051	0.049	0.048
U ind	0.052	0.057	0.050	0.051	0.051	0.050	0.048	0.047	0.048
prosp	0.043	0.048	0.050	0.042	0.047	0.045	0.042	0.043	0.041
Han	0.052	0.048	0.050	0.050	0.054	0.052	0.051	0.051	0.050
Han ind	0.052	0.052	0.049	0.050	0.049	0.046	0.047	0.048	0.048

TABLE 1. The type 1 error of the compared tests, under various combinations of the prevalence of the genetic variant a_1, and the effect of the genetic and environmental variables on the disease outcome (α_1 and α_2, respectively). The tests 'U' and 'U ind' are the proposed tests without and with the assumption of G-E independence. 'prosp' is the usual test based on prospective likelihood, and 'Han' and 'Han ind' are the tests based on retrospective profile likelihood proposed by Han et al., 2012. The type 1 error was calculated from 10,000 simulations, each with 4000 cases and 4000 controls.