Robust tests for gene-environment interaction in case-control and case-only designs

Yong Zang*[a,b], Wing Kam Fung[c], Sha Cao[a,b], Hon Keung Tony Ng[d] and Chi Zhang[b,e], 410 West 10th St., Suite 5000, Indianapolis IN 46202[1,1]

[a] *Department of Biostatistics, Indiana University, Indianapolis, IN, USA*
[b] *Center of Computational Biology and Bioinformatics, Indiana University, Indianapolis, IN, USA*
[c] *Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong, China*
[d] *Department of Statistical Science, Southern Methodist University, Dallas, TX, USA*
[e] *Department of Medical and Molecular Genetics, Indiana University, Indianapolis, IN, USA*

Abstract

The case-control and case-only designs are commonly used to detect the gene-environment (G-E) interaction. In principle, the tests based on these two designs require a pre-specified genetic model to achieve an expected power of detecting the G-E interaction. Unfortunately, for most complex diseases the underlying genetic models are unknown. It is well known that mis-specification of the genetic model can result in a substantial loss of power in the detection of the main genetic effect. However, limited effort has been dedicated to the study of G-E interaction. This issue has been investigated in this article with a conclusion that the genetic model mis-specification can not only undermine the power of detecting G-E interaction in both case-control and case-only designs but also distort the type I error rate in case-control design. To tackle this problem, a class of robust tests, namely MAX3, have been proposed for both the case-control and case-only designs. The proposed tests can well control the type I error rate and yield satisfactory power even when the genetic model is mis-specified. The asymptotic distribution and the p-value formula for MAX3 have also been derived. Comprehensive simulation studies and a real data application on the genome-wide association study (GWAS) have been conducted using these

Email address: zangy@iu.edu (410 West 10th St., Suite 5000, Indianapolis IN 46202)

This is the author's manuscript of the article published in final edited form as:

Zang, Y., Fung, W. K., Cao, S., Ng, H. K. T., & Zhang, C. (2019). Robust tests for gene–environment interaction in case-control and case-only designs. Computational Statistics & Data Analysis, 129, 79–92. https://doi.org/10.1016/j.csda.2018.08.014

analytical tools and the results demonstrate desirable operating characteristics of the proposed robust tests.

Keywords: Gene-environment interaction, Robust test, Genetic model, Case-control design, Case-only design

---

1. Introduction

Rapid development in human genetics and epidemiology has revealed that genetic susceptibility and environmental exposures play a synergistic role in many complex diseases. This understanding has boosted the development of gene-environment (G-E) interaction study in population genetics, which investigates the joint genetic and environmental interactive effect on the risk of developing diseases (Hunter, 2005). The case-control design has been commonly used to detect the G-E interaction, where the interactive effect can be conveniently modeled by a multiplicative term of genotypes and exposure levels based on a prospective logistic regression model. However, in such design, samples are classified by both the genotypes and exposure levels, which may result in a substantial loss of power (Mukherjee et al., 2012; Marigorta and Gibson, 2014). Alternatively, under the assumption of G-E independence and rare disease, the G-E interaction can be evaluated by simply assessing the G-E association on the cases only. Such case-only design can yield a higher power than a case-control design when these assumptions hold (Piegorsch et al., 1994; Umbach and Weinberg, 1997).

In both case-control and case-only designs, if the genetic model of inheritance can be specified a priori, then a score test can be performed to detect the G-E interaction. The genetic model determines the orders of individuals' risk of having the disease based on the number of risk alleles in the genotype. Generally speaking, for a diallelic marker, three genetic models, namely the recessive (REC), multiplicative (MUL) and dominant (DOM) are commonly used (Sasieni, 1997; Freidlin et al., 2002). For each genetic model, an optimal set of scores should be used to maximize the power of the test. In particular,

the value 0, 1/2 and 1 are the optimal scores to code the genotype conferring one risk allele when the genetic model is REC, MUL and DOM, respectively (Zheng et al., 2003). Hence, if the genetic model is correctly specified, the corresponding optimal scores can maximize the power of the score test. However, for many complex diseases, the underlying genetic models are unknown and using an inappropriate genetic model can substantially undermine the power of the tests (Zheng et al., 2003). Therefore, robust tests against genetic model mis-specification are in urgent demand.

Despite intensive studies on robust tests for detecting the main genetic effect (Wang and Sheffield, 2005; Gonzalez et al., 2008; Zheng et al., 2008; Yamada and Okada, 2009; Zang et al., 2010), little has been dedicated for the G-E interaction effect. Hence, the purpose of this paper is to fill this research gap. Specifically, we first investigate the impact of genetic model mis-specification on testing the G-E interaction. Interestingly, we find that the genetic model mis-specification highly affects the case-control design by distorting both the type I error rate and power, but only decreases the power for the case-only design. Furthermore, to handle the genetic model uncertainty, we have developed robust tests for both designs. The asymptotic formulas to calculate the p-value of the robust tests together with an user-friendly software are also released in this paper to facilitate the use of the proposed methods in practice. Simulation study demonstrate that the proposed robust tests could control type I error rate under the null hypothesis and yet yield satisfactory power under the alternative hypothesis, even when the genetic model is mis-specified. The proposed method is also applied to a real genome-wide association study (GWAS) dataset for illustrative purpose.

The rest of this paper is organized as follows. We develop the robust tests for the case-control design and case-only design in Sections 2 and 3. In Sections 4 and 5 we extend the proposed tests to handle non-monotonic genetic model and categorical environment factor with possible environmental level misclassification. In Section 6, we carry out comprehensive simulation studies to investigate the operating characteristics of the proposed tests. In Section 7,

we apply the robust tests to analyze a genome-wide association study (GWAS) of bladder cancer (Rothman et al., 2010). We provide a brief discussion and concluding remarks in Section 8.

2. Robust test for case-control design

Assume m_1 cases and m_0 controls being genotyped in a case-control study and let $n = m_0 + m_1$ denoting the total sample size. For ease of presentation, we consider a binary environmental factor E and a diallelic marker G, for which we are interested in testing the impact of the gene-environmental (G-E) interaction effect on the disease risk. Let $G = 0, 1, 2$ denote the three genotypes aa, Aa and AA with A indicating the minor allele conferring high risk of the disease. Let $E = 0 (E = 1)$ denote an unexposed (exposed) individual. Let D denote the disease status with $D = 0 (D = 1)$ representing an unaffected (affected) individual. The case-control data can be displayed in the form of a 2×6 table as presented in Table 1. As expressed in Table 1, we use r_{ijk} to denote the number of individuals with $D = i$, $G = j$ and $E = k$ and define $n_{jk} = r_{0jk} + r_{1jk}$.

Table 1: Case-control data with a diallelic marker G and a binary environmental exposure factor E.

	$G=0$		$G=1$		$G=2$		
	$E=0$	$E=1$	$E=0$	$E=1$	$E=0$	$E=1$	Total
$D=0$	r_{000}	r_{001}	r_{010}	r_{011}	r_{020}	r_{021}	m_0
$D=1$	r_{100}	r_{101}	r_{110}	r_{111}	r_{120}	r_{121}	m_1
Total	n_{00}	n_{01}	n_{10}	n_{11}	n_{20}	n_{21}	n

Let D_l, G_l and E_l be the phenotype, genotype and environmental factor for the l^{th} sample in case-control study. We define $f_{jk} = \Pr(D_l = 1 | G_l = j, E_l = k)$ as the penetrance level conditional on $G = j$ and $E = k$, by which the recessive (REC), multiplicative (MUL) and dominant (DOM) genetic models correspond to $f_{1k} = f_{0k}$, $f_{1k} = \sqrt{f_{0k} f_{2k}}$ and $f_{1k} = f_{2k}$ for $k = 0, 1$ respectively (Sasieni,

1997).

According to the definition, when the genetic model is specified, the impact of G_l and E_l on the disease status D_l can be conveniently formulated by the following logistic regression model:

$$\log\Big(\frac{\Pr(D_l=1|G_l,E_l)}{\Pr(D_l=0|G_l,E_l)}\Big)=\alpha+\delta E+x(\beta+\lambda E)\mathrm{I}(G=1)+(\beta+\lambda E)\mathrm{I}(G=2), \quad (1)$$

where $\mathrm{I}(\cdot)$ is an indicator function, δ is the main environmental effect, β is the main genetic effect, λ is the G-E interaction effect and x is a real number between 0 and 1 representing the underlying genetic model . The interest here is to test the null hypothesis $H_0:\lambda=0$. It is straightforward to see that $x=0$ and 1 correspond to the REC and DOM models, respectively. If the disease is rare, based on model (1) we have the approximation $\frac{f_{jk}}{f_{0k}}\approx \mathrm{e}^{\left(x\mathrm{I}(j=1)+\mathrm{I}(j=2)\right)(\beta+\lambda k)}$, by which $x=1/2$ is the optimal score for the MUL model. However, if the disease is not rare, the optimal score turns out to be a function of the parameters α, δ, β and λ and insisting on using $x=1/2$ for the MUL model can result in a loss of power. In reality, the disease prevalences for most genetic hereditary diseases are substantially low and $1/2$ is commonly used as the optimal score for MUL model (Freidlin et al., 2002; Zheng et al., 2003) and we adopt this approximation in this paper.

If the genetic model is correctly specified, we can derive a score test based on model (1) by using an optimal choice of x to test $H_0:\lambda=0$. We denote the test as $Z_{\mathrm{model1}}(x)$. However, when the underlying genetic models are unknown the consequence of using a mis-specified genetic model for test $Z_{\mathrm{model1}}(x)$ is summarized in the following theorem.

Theorem 1. When the genetic model is mis-specified, using $Z_{\mathrm{model1}}(x)$ to test the G-E interaction is invalid as it can cause significant bias under both the null and alternative hypotheses.

To prove Theorem 1, we need to derive the score function based on model

(1) as

$$x(r_{111} - n_{11}\widehat{f}_{11}) + (r_{121} - n_{21}\widehat{f}_{21}),$$

where $\widehat{f}_{jk}$ is the estimate of f_{jk} based on the case-control data under the null hypothesis of $\lambda = 0$. Specifically, $\widehat{f}_{jk}$ can be expressed as

$$\widehat{f}_{jk} = \frac{1}{1 + \mathrm{e}^{-\widehat{\alpha} - \widehat{\delta}k - \left(x\mathrm{I}(j=1) + \mathrm{I}(j=2)\right)\widehat{\beta}}}.$$

If the genetic model is mis-specified, then $\widehat{\beta}$ is a biased estimate for β even when $\lambda = 0$. Consequently, $\widehat{f}_{jk}$ is also biased and the score function under model (1) does not converge to 0. Therefore, $Z_{\mathrm{model1}}(x)$ based on model (1) is invalid under genetic model mis-specification. We note that for detecting the main genetic effect β without the G-E interaction effect λ, the score tests are valid even when the genetic model is mis-specified although the power loss can be substantial (Zang, 2011). Hence, the genetic model mis-specification is a more serious issue in detecting the G-E interaction for the case-control study. Therefore, it is necessary to develop a robust method which can handle genetic model uncertainty.

To resolve this issue, we propose to accommodate model (1) by relaxing the genetic model assumption for the main genetic effect β. That is, instead of using one single β, we propose to use two dummy variables β_1 and β_2 to represent the main genetic effect, and then propose a logistic model with genetic model assumption only on the G-E interaction term by the following

$$\log\Big(\frac{\Pr(D_l = 1|G_l, E_l)}{\Pr(D_l = 0|G_l, E_l)}\Big) = \alpha + \delta E + \beta_1\mathrm{I}(G = 1) + \beta_2\mathrm{I}(G = 2) + \big\{x\mathrm{I}(G = 1) + \mathrm{I}(G = 2)\big\}\lambda E. \quad (2)$$

We note that model (2) degenerates to the saturated model when there is no G-E interaction, which ensures the validity of the inference for λ based on model (2). With model (2) at hand, we can derive score test to detect the G-E interaction. In particular, let $\widetilde{f}_{jk}$ be the estimate of f_{jk} in model (2) under the null hypothesis $\lambda = 0$, then the score function can be derived as

$$U(x) = x(r_{111} - n_{11}\widetilde{f}_{11}) + (r_{121} - n_{21}\widetilde{f}_{21}) = x(r_{111} - \frac{n_{11}}{1 + \mathrm{e}^{-\tilde{\alpha} - \tilde{\beta}_1 - \tilde{\delta}}}) + (r_{121} - \frac{n_{21}}{1 + \mathrm{e}^{-\tilde{\alpha} - \tilde{\beta}_2 - \tilde{\delta}}}),$$

where $\tilde{\alpha}$, $\tilde{\beta}_1$, $\tilde{\beta}_2$ and $\tilde{\delta}$ are the MLEs based on model (2) given $\lambda = 0$. Furthermore, let us define $\psi = (\alpha, \beta_1, \beta_2, \delta)'$ and the information matrix of model (2) can be written in the form of a block matrix as

$$I(x) = \begin{pmatrix} \mathrm{I}_\lambda(x) & \mathrm{I}_{\lambda\psi}(x)' \\ \mathrm{I}_{\lambda\psi}(x) & \mathrm{I}_\psi \end{pmatrix}.$$

Based on this information matrix, we can derive $V(x) = \mathrm{I}_\lambda(x) - \mathrm{I}_{\lambda\psi}(x)'\mathrm{I}_\psi^{-1}\mathrm{I}_{\lambda\psi}(x)$ as the variance estimate for $U(x)$ (see Appendix A for the details). The score test statistic for $H_0 : \lambda = 0$ based on model (2) is then given by

$$Z_{\mathrm{model2}}(x) = \frac{U(x)}{\sqrt{V(x)}} = \frac{x(r_{111} - n_{11}\widetilde{f}_{11}) + (r_{121} - n_{21}\widetilde{f}_{21})}{\sqrt{\mathrm{I}_\lambda(x) - \mathrm{I}_{\lambda\psi}(x)'\mathrm{I}_\psi^{-1}\mathrm{I}_{\lambda\psi}(x)}}.$$

Under the null hypothesis of no G-E interaction, $Z_{\mathrm{model2}}(x)$ asymptotically follows the standard normal distribution with any real valued x.

$Z_{\mathrm{model2}}(x)$ is statistically rigorous under the null hypothesis. However, its performance under the alternative hypothesis still depends on the value of x, which represents the genetic model. When the genetic model is known, $Z_{\mathrm{model2}}(x)$ with an appropriate choice of x can be used. However, this test is vulnerable to the mis-specification of x. That is, if $\lambda \neq 0$, using $Z_{\mathrm{model2}}(x)$ with a mis-specified x could suffer from a substantial loss of power. To overcome this limitation, we propose a maximum test statistic as

$$\mathrm{MAX3}_{\mathrm{cc}} = \max(|Z_{\mathrm{model2}}(0)|, |Z_{\mathrm{model2}}(1/2)|, |Z_{\mathrm{model2}}(1)|).$$

Noting that $\mathrm{MAX3}_{\mathrm{cc}}$ does not asymptotically follow the normal distribution anymore, due to the multiple comparisons and correlations among $Z_{\mathrm{model2}}(0)$, $Z_{\mathrm{model2}}(1/2)$ and $Z_{\mathrm{model2}}(1)$. Hence, to use this statistic, its distribution under the null hypothesis should be determined first.

Let $Z = (Z_{\mathrm{model2}}(0), Z_{\mathrm{model2}}(1/2), Z_{\mathrm{model2}}(1))'$, then Z asymptotically follows a multivariate normal distribution $N(0, \Sigma)$ under H_0, with Σ denoting the variance-covariance matrix of Z. If we assume Σ as known, and let $f(z, \Sigma)$ be the density function for the multivariate normal variable Z under H_0, for any

observed value t, the p-value of MAX3_{cc} can be obtained as:

$$\begin{aligned}\Pr(\text{MAX3}_{\text{cc}} > t|H_0) &= 1 - \Pr(|Z_{\text{model2}}(0)| \leq t, |Z_{\text{model2}}(1/2)| \leq t, |Z_{\text{model2}}(1)| \leq t) \\ &= 1 - \int_{-t}^{t}\int_{-t}^{t}\int_{-t}^{t} f(z, \Sigma)dz.\end{aligned}$$

In particular, Σ in the p-value formula can be expressed as

$$\Sigma = \begin{pmatrix} 1 & \rho_{0,1/2} & \rho_{0,1} \\ \rho_{0,1/2} & 1 & \rho_{1/2,1} \\ \rho_{0,1} & \rho_{1/2,1} & 1 \end{pmatrix},$$

where ρ_{x_1,x_2} is the correlation between $Z_{\text{model2}}(x_1)$ and $Z_{\text{model2}}(x_2)$ for any $x_1, x_2 = 0, 1/2, 1$ and $x_1 < x_2$. The key now is to derive the expression of ρ_{x_1,x_2}.

To do this, we "artificially" build the following logistic model

$$\log\Big(\frac{\Pr(D_l = 1|G_l, E_l)}{\Pr(D_l = 0|G_l, E_l)}\Big) = \alpha + \delta E + \beta_1 \text{I}(G = 1) + \beta_2 \text{I}(G = 2) + (x_1\lambda_1 + x_2\lambda_2)E\text{I}(G = 1) + (\lambda_1 + \lambda_2)E\text{I}(G = 2). \quad (3)$$

Under model (3), the null hypothesis of no G-E interaction corresponds to $H_0 : \lambda_1 = \lambda_2 = 0$. Furthermore, model (2) and model (3) are identical under the null hypothesis. Therefore, under the null hypothesis, the maximum likelihood estimates of α, β_1, β_2 and δ based on model (3) are the same as model (2) under the null hypothesis.

Based on model (3), we can write the log-likelihood function as

$$\begin{aligned} l &= \alpha r_{100} - n_{00}\log(1 + \text{e}^{\alpha}) + (\alpha + \delta)r_{101} - n_{01}\log(1 + \text{e}^{\alpha+\delta}) \\ &+ (\alpha + \beta_1)r_{110} - n_{10}\log(1 + \text{e}^{\alpha+\beta_1}) + (\alpha + \beta_1 + \delta + x_1\lambda_1 + x_2\lambda_2)r_{111} - n_{11}\log(1 + \text{e}^{\alpha+\beta_1+\delta+x_1\lambda_1+x_2\lambda_2}) \\ &+ (\alpha + \beta_2)r_{120} - n_{20}\log(1 + \text{e}^{\alpha+\beta_2}) + (\alpha + \beta_2 + \delta + \lambda_1 + \lambda_2)r_{121} - n_{21}\log(1 + \text{e}^{\alpha+\beta_2+\delta+\lambda_1+\lambda_2}), \end{aligned}$$

and the score function as

$$(\frac{\partial l}{\partial \lambda_1}, \frac{\partial l}{\partial \lambda_2})_{|\psi=\tilde{\psi},\lambda_1=\lambda_2=0} = \Big(x_1(r_{111} - n_{11}\tilde{f}_{11}) + (r_{121} - n_{21}\tilde{f}_{21}), x_2(r_{111} - n_{11}\tilde{f}_{11}) + (r_{121} - n_{21}\tilde{f}_{21})\Big) = \Big(U(x_1), U(x_2)\Big).$$

Therefore, ρ_{x_1,x_2} is simply the correlation coefficient between the score function $U(x_1)$ and $U(x_2)$, which can be derived from the information matrix of the developed log-likelihood function l. Let us define the information matrix for the likelihood function l in the form of a block matrix as $I(x_1, x_2) =$

$\begin{pmatrix} \mathrm{I}_{\lambda_1\lambda_2}(x_1,x_2) & \mathrm{I}_{\lambda_1\lambda_2\psi}(x_1,x_2)' \\ \mathrm{I}_{\lambda_1\lambda_2\psi}(x_1,x_2) & \mathrm{I}_{\psi} \end{pmatrix}$, the expression of ρ_{x_1,x_2} is summarized in the following theorem.

Theorem 2. Let $\mathrm{I}^{\lambda_1\lambda_2}(x_1,x_2) = (\mathrm{I}_{\lambda_1\lambda_2}(x_1,x_2) - \mathrm{I}_{\lambda_1\lambda_2\psi}(x_1,x_2)'\mathrm{I}_{\psi}^{-1}\mathrm{I}_{\lambda_1\lambda_2\psi}(x_1,x_2))^{-1}$, the correlation ρ_{x_1,x_2} can be expressed as

$$\rho_{x_1,x_2} = \frac{(1,0)\big(\mathrm{I}^{\lambda_1\lambda_2}(x_1,x_2)\big)^{-1}(0,1)'}{\sqrt{(1,0)\big(\mathrm{I}^{\lambda_1\lambda_2}(x_1,x_2)\big)^{-1}(1,0)'(0,1)\big(\mathrm{I}^{\lambda_1\lambda_2}(x_1,x_2)\big)^{-1}(0,1)'}}.$$

The proof of Theorem 2 and the detailed expression of ρ_{x_1,x_2} are given in Appendix B. Hence, after plugging in ρ_{x_1,x_2} back into the formula for $\Pr(\mathrm{MAX3_{cc}} > t|H_0)$, we get the asymptotic expression of the p-value formula. With this p-value formula at hand, we can easily apply $\mathrm{MAX3_{cc}}$ to detect the G-E interaction for case-control design. As $\mathrm{MAX3_{cc}}$ considers different genetic models simultaneously and takes the maximum, it is robust against the genetic model uncertainty.

3. Robust test for case-only design

For rare disease, when a binary genetic factor is independent of the environmental factor, G-E interaction can be tested by assessing the gene-environmental association using case-only data, and such case-only design typically yields a much higher power than case-control design (Piegorsch et al., 1994; Umbach and Weinberg, 1997; Schmidt and Schaid, 1999). We herein extend the work from binary genetic factor to diallelic marker with three genotypes and propose a robust test for case-only study.

Let us define $\theta_j = \frac{f_{00}f_{j1}}{(1-f_{00})(1-f_{j1})} \Big/ \frac{f_{j0}f_{01}}{(1-f_{j0})(1-f_{01})}$ and $q_j = \Pr(E = 1|G = j, D = 1)$, we have the following theorem to detect G-E interaction in case-only design.

Theorem 3. θ_j is the odds ratio of G-E interaction given $G = j$ $(j = 1, 2)$

from either model (1) or (2) in case-control design. Under the assumption of G-E independence and rare disease, θ_j can be approximated by $\frac{q_j(1-q_0)}{(1-q_j)q_0}$, which is the odds ratio of having $E = 1$ between $G = j$ and $G = 0$ conditional on $D = 1$ (case only).

To prove Theorem 3, consider model (1) in a case-control design, we have

$$\theta_j = \frac{\mathrm{e}^{\alpha}\mathrm{e}^{\alpha+\delta+x(\beta+\lambda)\mathrm{I}(j=1)+(\beta+\lambda)\mathrm{I}(j=2)}}{e^{\alpha+x\beta\mathrm{I}(j=1)+\beta\mathrm{I}(j=2)}e^{\alpha+\delta}} = \mathrm{e}^{x\lambda\mathrm{I}(j=1)+\lambda\mathrm{I}(j=2)}.$$

Along the same lines, consider model (2) in a case-control study, we have

$$\theta_j = \frac{\mathrm{e}^{\alpha}\mathrm{e}^{\alpha+\delta+\beta_1\mathrm{I}(j=1)+\beta_2\mathrm{I}(j=2)+\{x\mathrm{I}(j=1)+\mathrm{I}(j=2)\}\lambda}}{\mathrm{e}^{\alpha+\beta_1\mathrm{I}(j=1)+\beta_2\mathrm{I}(j=2)}\mathrm{e}^{\alpha+\delta}} = \mathrm{e}^{x\lambda\mathrm{I}(j=1)+\lambda\mathrm{I}(j=2)}.$$

Hence, θ_j is the odds ratio of G-E interaction given $G = j$ $(j = 1, 2)$ in both models. Further, under the assumption of G-E independence and rare disease, θ_j can be approximated as:

$$\theta_j \approx \frac{f_{j1}f_{00}}{f_{j0}f_{01}} = \frac{\frac{\Pr(E=1|G=j,D=1)\Pr(D=1|G=j)}{\Pr(E=1|G=j)}\frac{\Pr(E=0|G=0,D=1)\Pr(D=1|G=0)}{\Pr(E=0|G=0)}}{\frac{\Pr(E=0|G=j,D=1)\Pr(D=1|G=j)}{\Pr(E=0|G=j)}\frac{\Pr(E=1|G=0,D=1)\Pr(D=1|G=0)}{\Pr(E=1|G=0)}} = \frac{q_j(1-q_0)}{(1-q_j)q_0}.$$

This completes the proof of Theorem 3. Note that $\frac{q_j(1-q_0)}{(1-q_j)q_0}$ is the odds ratio of having $E = 1$ between $G = j$ and $G = 0$ conditional on $D = 1$, which can be estimated by using case-only data. Therefore, if we regress E on G using case-only data, we have the following logistic regression model:

$$\log\Big(\frac{\Pr(E=1|G,D=1)}{\Pr(E=0|G,D=1)}\Big) = \alpha^* + x\lambda\mathrm{I}(G=1) + \lambda\mathrm{I}(G=2), \tag{4}$$

where λ represents G-E interaction effect exactly the same as developed in the case-control study.

Let $\widehat{\theta}_j^{(cc)}$ be the estimate of the odds ratio of G-E interaction given $G = j$ from case-control design and $\widehat{\theta}_j^{(ca)}$ be the counterpart from the case-only design.

After some algebra we obtain their variance estimates as

$$\widehat{\mathrm{Var}}(\widehat{\theta}_j^{(cc)}) = \sum_{d=0}^{1}\sum_{g=0}^{j}\sum_{e=0}^{1} 1/r_{dge};\ \widehat{\mathrm{Var}}(\widehat{\theta}_j^{(ca)}) = \sum_{g=0}^{j}\sum_{e=0}^{1} 1/r_{1ge}.$$

It is easy to see that $\widehat{\mathrm{Var}}(\widehat{\theta}_j^{(cc)}) > \widehat{\mathrm{Var}}(\widehat{\theta}_j^{(ca)})$, indicating a higher power to detect G-E interaction in case-only design.

Based on model (4), we can apply the Cochran-Armitage trend test (Sasieni, 1997; Armitage, 1955) to detect the interaction effect in case-only design. In particular, after defining $\phi = \sum_{j=0}^{2} r_{1j1}/m_1$, $s(1) = x$ and $s(2) = 1$, the closed form of the trend test statistic for testing $H_0 : \lambda = 0$ is

$$Z_{\mathrm{model3}}(x) = \frac{\sum_{j=1}^{2} s(j)\{(1-\phi)r_{1j1} - \phi r_{1j0}\}}{\sqrt{m_1\phi(1-\phi)\Big\{\sum_{j=1}^{2} s^2(j)(r_{1j1}+r_{1j0})/m_1 - (\sum_{j=1}^{2} s(j)(r_{1j1}+r_{1j0})/m_1)^2\Big\}}}.$$

Therefore, when the G-E independence and disease rareness assumptions hold, $Z_{\mathrm{model3}}(x)$ can be applied to test G-E interaction. Under the null hypothesis of no G-E interaction, $Z_{\mathrm{model3}}(x)$ asymptotically follows a standard normal distribution for any x between 0 and 1. It is noteworthy that $Z_{\mathrm{model3}}(x)$ is sensitive to the choice of genetic model under the alternative hypothesis. Similar to the case-control design, we propose a maximum test statistic in the case-only design as

$$\mathrm{MAX3_{ca}} = \max(|Z_{\mathrm{model3}}(0)|, |Z_{\mathrm{model3}}(1/2)|, |Z_{\mathrm{model3}}(1)|).$$

$\mathrm{MAX3_{ca}}$ is a maximum-type statistic based on the trend tests, whose null distribution can be derived from a multivariate normal distribution. We have developed a user-friendly R package named "Rassoc" which can easily calculate the p-value of $\mathrm{MAX3_{ca}}$ (Zang et al., 2010; Zang, 2011). The package can be freely downloaded at CRAN (https://cran.r-project.org/src/contrib/Archive/Rassoc/).

4. Robust test without monotonic assumption

The proposed robust tests $\mathrm{MAX3_{cc}}$ and $\mathrm{MAX3_{ca}}$ focus on three commonly used genetic models (REC, MUL and DOM), which all assume a monotonic

minor allele-penetrance relationship. That is, the value of f_{jk} will monotonically increase as the number of the minor allele j increases. However, for certain diseases, the monotonic assumption may be violated and the underlying genetic model may be far beyond the commonly used ones. For example, an over-dominant model ($f_{1k} \geq f_{2k}$) and an under-dominant model ($f_{1k} \leq f_{0k}$) have been discussed in the literature (Gillespie, 2004). Hence, it is of our interest to extend the use of MAX3_{cc} and MAX3_{ca} for scenarios where the monotonic assumption does not hold.

We propose to use two parameters λ_1 and λ_2 to characterize the G-E interaction with $G = 1$ and $G = 2$ separately. Then, the likelihood function for the case-control design can be written as

$$\log\Big(\frac{\Pr(D_l = 1|G_l, E_l)}{\Pr(D_l = 0|G_l, E_l)}\Big) = \alpha + \delta E + \beta_1 \mathrm{I}(G = 1) + \beta_2 \mathrm{I}(G = 2) + \lambda_1 \mathrm{I}(G = 1)E + \lambda_2 \mathrm{I}(G = 2)E. \quad (5)$$

Hence, testing the G-E interaction effect is equivalent to testing the null hypothesis $H_0 : \lambda_1 = \lambda_2 = 0$. By using logistic model (5), we can construct a likelihood ratio test, defined as χ^2_{cc}, which follows a chi-square distribution with 2 degree of freedom. Following Zheng et al. (Zheng et al., 2006) and Zang et al. (Zang et al., 2010), it can be shown that χ^2_{cc} can be expressed in the form of $Z_{\text{model2}}(x)$ as:

$$\chi^2_{\text{cc}} = \frac{1}{1 - \widehat{\rho}^2_{0,1}}\big\{Z_{\text{model2}}(0)^2 + Z_{\text{model2}}(1)^2 - 2\widehat{\rho}_{0,1} Z_{\text{model2}}(0) Z_{\text{model2}}(1)\big\}.$$

Along the same lines, we can construct χ^2_{ca} for case-only design as:

$$\chi^2_{\text{ca}} = \frac{1}{1 - \widehat{\gamma}^2_{0,1}}\big\{Z_{\text{model3}}(0)^2 + Z_{\text{model3}}(1)^2 - 2\widehat{\gamma}_{0,1} Z_{\text{model3}}(0) Z_{\text{model3}}(1)\big\},$$

where $\gamma_{0,1}$ is the correlation coefficient between $Z_{\text{model3}}(0)$ and $Z_{\text{model3}}(1)$. The closed form expression of $\gamma_{0,1}$ can be found in Zang et al. (Zang et al., 2010).

It is worth noting that the construction of χ^2_{cc} and χ^2_{ca} do not rely on any restrictive relationships between the minor allele and the penetrance. Thus,

these tests are robust with or without the monotonic relationship. However, if the monotonic assumption does hold, the MAX-type tests should be more powerful.

5. Robust test with categorical environmental factor

In the previous sections, we assume the environmental factor E to be binary. If E is categorical with more than two levels, a practical solution is to transform this multi-level variable into a binary variable using certain thresholding method. However, mis-classification of E may potentially harm the performance of the proposed test. In this section, we briefly investigate how to extend the proposed robust test to handle multi level categorical environmental factor.

Let us define E as a categorical variable with K levels $0, 1, \cdots, K-1$. In case-control design, the likelihood function can be written as

$$\log\Big(\frac{\Pr(D_l=1|G_l,E_l)}{\Pr(D_l=0|G_l,E_l)}\Big)=\alpha+\sum_{k=1}^{K-1}\delta_k \mathrm{I}(E=k)+\sum_{j=1}^{2}\beta_j \mathrm{I}(G=j)+\sum_{k=1}^{K-1}\lambda_k\big\{x\mathrm{I}(G=1)+\mathrm{I}(G=2)\big\}\mathrm{I}(E=k).$$

Based on this model, we can construct a likelihood ratio test for the hypothesis of no G-E interaction as $H_0 : \lambda_1 = \lambda_2 = \cdots = \lambda_{K-1} = 0$. We denote the test as $\chi^2_{\mathrm{cc}}(x)$ with x relying on the genetic model. Under the null hypothesis, $\chi^2_{\mathrm{cc}}(x)$ follows a chi-square distribution with $K-1$ degrees of freedom. When the genetic model is unknown, similar to $\mathrm{MAX3}_{\mathrm{cc}}$, we propose a MAX-type robust test defined as

$$\mathrm{LMAX3}_{\mathrm{cc}} = \max\big(\chi^2_{\mathrm{cc}}(0), \chi^2_{\mathrm{cc}}(1/2), \chi^2_{\mathrm{cc}}(1)\big).$$

For case-only design, when E has more than two levels, we can use the baseline-category logit model to fit the data as:

$$\log\Big(\frac{\Pr(E=k|G,D=1)}{\Pr(E=0|G,D=1)}\Big)=\alpha^k+x\lambda_k \mathrm{I}(G=1)+\lambda_k \mathrm{I}(G=2).$$

Based on this model, we can construct a likelihood ratio test, defined as $\chi^2_{\mathrm{ca}}(x)$, for the G-E interaction. Similarly, a robust test can be developed for case-only design as $\mathrm{LMAX3}_{\mathrm{ca}} = \max\big(\chi^2_{\mathrm{ca}}(0), \chi^2_{\mathrm{ca}}(1/2), \chi^2_{\mathrm{ca}}(1)\big)$.

Unfortunately, different from MAX3_{ca} and MAX3_{cc}, there is no formula available yet for the asymptotic distributions of LMAX3_{cc} and LMAX3_{ca}. Nevertheless, following Zang et al. (Zang et al., 2010), a bootstrap resampling method can be used to approximate the p-values for LMAX3_{cc} and LMAX3_{ca}. As LMAX3_{cc} and LMAX3_{ca} do not require to classify a multi-level categorical E into a binary variable, they are robust against both the genetic model and environmental factor mis-classification.

6. Simulation studies

In this section, we conduct comprehensive simulation studies to investigate the operating characteristics of the proposed tests in both case-control and case-only studies. In detail, we first simulate G from a minor allele frequency (MAF) ϕ with the assumption of Hardy-Weinberg equilibrium and apply the following logistic model to simulate the dependence between G and E as:

$$\log\Big(\frac{\Pr(E=1|G)}{\Pr(E=0|G)}\Big) = \tau + \gamma_1 \mathrm{I}(G=1) + \gamma_2 \mathrm{I}(G=2), \tag{6}$$

by which G and E are independent if and only if $\gamma_1 = \gamma_2 = 0$. With the disease prevalence conditional on G and E simulated from formula (1), the case-control data can be further simulated from multinomial distributions with the associated probabilities:

$$\Pr(G, E|D) = \frac{\Pr(D|G,E)\Pr(E|G)\Pr(G)}{\sum_{G,E}\Pr(D|G,E)\Pr(E|G)\Pr(G)}.$$

We study the empirical type I error rates of the proposed tests for case-control design by comparing the conventional score test $Z_{\text{model1}}(x)$, proposed score test $Z_{\text{model2}}(x)$ with the robust test MAX3_{cc}. We generate case-control data by specifying the MAF $\phi = 0.3$, $\alpha = -6$, $\delta = 1$, $\lambda = 0$ from model (1) and $\tau = 0$, $\gamma_1 = 0.5$, $\gamma_2 = 1$ from model (6). We consider three genetic models (REC, MUL and DOM) for the non-zero marginal genetic effect β. Table 2 summarizes the simulation results based on 10,000 replicates with different sample sizes m_0,

Table 2: Empirical type I error rates (%) of $Z_{\text{model1}}(x)$, $Z_{\text{model2}}(x)$ and MAX3_{cc} for the case-control design. The significance level is 0.05.

		$\beta = 1$			$\beta = 2$			$\beta = 3$		
Sample size	Test	REC	MUL	DOM	REC	MUL	DOM	REC	MUL	DOM
$m_0 = m_1 = 500$	$Z_{\text{model1}}(0)$	4.8	5.3	5.6	4.9	5.6	**7.3**	4.9	**6.6**	**9.8**
	$Z_{\text{model1}}(0.5)$	5.4	5.0	4.3	4.9	4.7	**6.5**	**6.6**	4.7	**10.8**
	$Z_{\text{model1}}(1)$	5.2	5.1	4.8	5.1	4.8	5.1	5.1	4.8	4.4
	$Z_{\text{model2}}(0)$	5.0	5.5	5.4	5.5	5.3	5.1	5.3	4.8	5.5
	$Z_{\text{model2}}(0.5)$	4.8	5.4	5.1	5.2	5.4	5.5	5.5	5.3	5.3
	$Z_{\text{model2}}(1)$	4.8	5.3	5.0	5.0	5.3	5.4	4.9	5.5	5.4
	MAX3_{cc}	4.7	5.3	5.0	5.2	5.2	5.1	5.3	5.3	5.3
$m_0 = m_1 = 1000$	$Z_{\text{model1}}(0)$	5.1	5.5	**6.8**	5.1	**6.8**	**9.9**	5.0	**8.3**	**10.5**
	$Z_{\text{model1}}(0.5)$	5.7	4.8	4.6	4.8	5.3	**7.1**	**7.3**	5.0	**14.1**
	$Z_{\text{model1}}(1)$	4.7	5.0	4.8	5.0	4.8	5.2	5.7	4.9	4.8
	$Z_{\text{model2}}(0)$	5.2	5.0	5.1	5.2	5.1	5.4	5.4	5.3	4.9
	$Z_{\text{model2}}(0.5)$	5.0	5.2	5.4	5.2	5.5	5.3	5.2	5.4	5.5
	$Z_{\text{model2}}(1)$	5.2	5.2	5.5	5.0	5.4	5.4	5.1	5.1	5.0
	MAX3_{cc}	5.0	5.0	5.5	5.0	5.4	5.3	5.1	5.4	4.9

m_1 and main genetic effects β. Note that type I error rates greater than 6% are considered as inflated and labeled in bold font. Our simulation study clearly demonstrates that the conventional score test $Z_{\text{model1}}(0)$ and $Z_{\text{model1}}(0.5)$ have more severely inflated type I error rates when the genetic model is mis-specified, as claimed in Theorem 1. On the other hand, the newly developed score test $Z_{\text{model2}}(x)$ and robust test MAX3_{cc} can control the type I error rates in a satisfactory range. It is noteworthy that the inflation are substantial with a significant main genetic effect β and could be further magnified by a larger sample size. For example, when $m_0 = m_1 = 500$, $\beta = 3$ and DOM being the true genetic model for β, the type I error rates of $Z_{\text{model1}}(0)$ and $Z_{\text{model1}}(0.5)$ are 9.8% and 10.8%, respectively, which increase to 10.5% and 14.1% in the case of $m_0 = m_1 = 1000$. In contrast, $Z_{\text{model2}}(x)$ and MAX3_{cc} consistently control the type I error rates around 5% across all the settings considered here.

Table 3 presents the empirical type I error rates of $Z_{\text{model3}}(x)$ and MAX3_{ca} for case-only design. The same parameter settings as in Table 2 are used except that $\gamma_1 = \gamma_2 = 0$ is specified to ensure G-E independence. As expected, both $Z_{\text{model3}}(x)$ and MAX3_{ca} can consistently control the type I error rates around the 5% significance level, which ensure the validity of the proposed tests.

In addition, to further justify the proposed robust tests, we plot the empirical cumulative distributions of the p-values of MAX_{cc} and MAX_{ca} under null hypothesis with different sample sizes in Figure 1 and Figure 2. As shown in these figures, the empirical cumulative distributions fit the uniform distribution $U(0, 1)$ very well even for a small sample size of 200.

The validity of using MAX3_{ca} for case-only design relies on the assumption of disease rareness and G-E independence. Therefore, it is important to study the potential bias of MAX3_{ca} if the assumption is violated. Specifically, let ζ be the disease prevalence, in Figure 3 we provide the bar plot of the empirical type I error rates of MAX3_{ca} with different ζ and γ_2, which determines the magnitude for G-E dependence. As depicted in Figure 3, when the G-E independence assumption holds, the empirical type I error rates of MAX3_{ca} is consistently close to the nominal significant level 5% for a disease prevalence up to 10%.

Table 3: Empirical type I error rates (%) of $Z_{\text{model3}}(x)$ and MAX3_{ca} for the case-only design. The significance level is 0.05.

		$\beta = 1$			$\beta = 2$			$\beta = 3$		
Sample size	Test	REC	MUL	DOM	REC	MUL	DOM	REC	MUL	DOM
$m_0 = m_1 = 500$	$Z_{\text{model3}}(0)$	4.9	5.3	5.0	5.4	5.1	5.0	5.2	4.8	4.9
	$Z_{\text{model3}}(0.5)$	5.0	5.1	5.1	5.1	5.1	4.9	5.1	4.9	4.9
	$Z_{\text{model3}}(1)$	4.7	4.9	4.8	5.2	5.2	4.8	5.6	5.1	5.3
	MAX3_{ca}	4.8	5.3	4.9	5.1	5.3	4.8	5.3	4.9	5.2
$m_0 = m_1 = 1000$	$Z_{\text{model3}}(0)$	5.5	5.2	5,1	5.4	5,1	4.7	5.3	5.4	4.9
	$Z_{\text{model3}}(0.5)$	5.4	4.8	4.9	5.5	5.4	5.2	5.1	5.3	4.9
	$Z_{\text{model3}}(1)$	4.7	5.0	5.0	5.3	5.1	5.1	5.2	5.2	5.1
	MAX3_{ca}	5.0	5.3	4.9	5.4	5.2	4.9	5.4	5.4	4.8

Figure 1: Empirical cumulative distribution for the p-value of MAX_{cc} under the null hypothesis with different sample sizes.

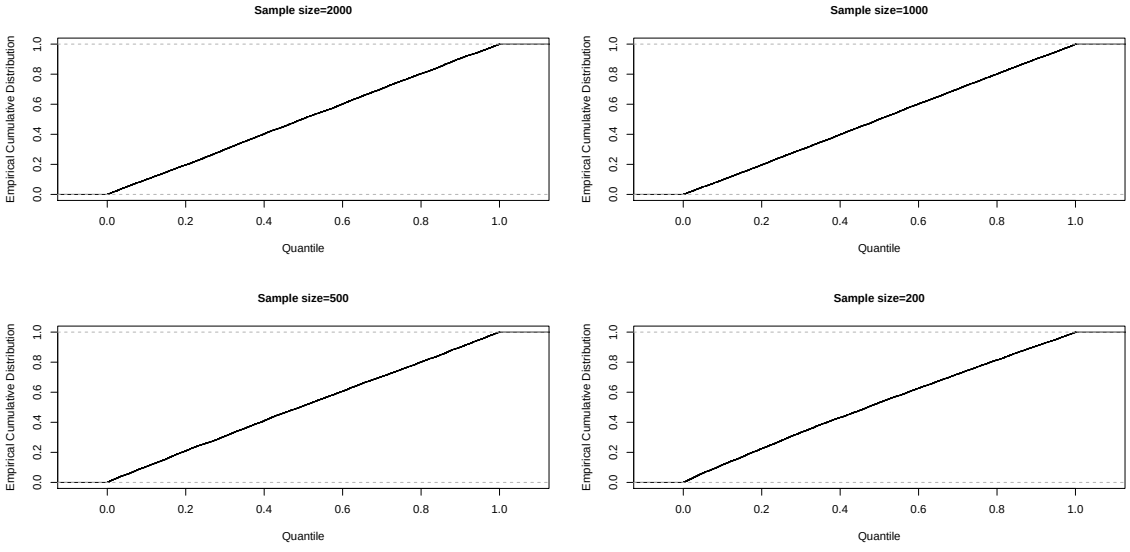
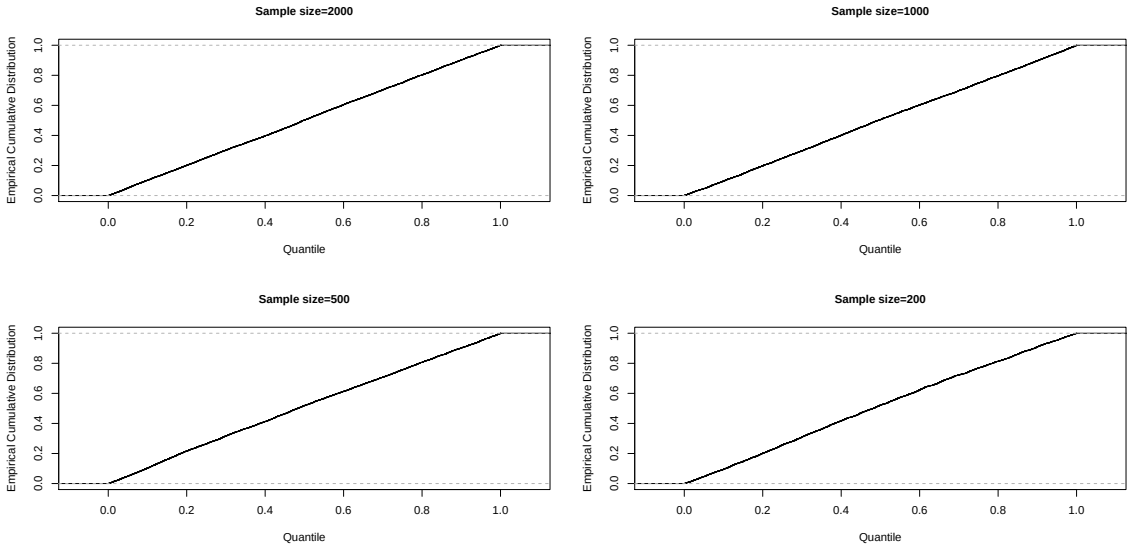


Figure 2: Empirical cumulative distribution for the p-value of $\mathrm{MAX}_{\mathrm{ca}}$ under the null hypothesis with different sample sizes.



However, with a moderate G-E dependence of $\gamma_2 = 0.5$, the type I error rates can be inflated to over 15% even when the disease prevalence is as low as $\zeta = 0.01$. Based on these results, we suggest checking the assumption of G-E independence before using the case-only design.

Table 4 presents the empirical statistical power for the proposed tests in case-control study under different genetic models (REC, MUL and DOM). We fix the sample size to be $m_0 = m_1 = 500$ and consider different configurations of the main effect (δ, β) and G-E interaction effect λ. Since $Z_{\text{model1}}(x)$ cannot control the type I error rate, the comparison is restricted to $Z_{\text{model2}}(x)$ and MAX3_{cc}. Table 4 clearly demonstrates the benefit of using MAX3_{cc}. Although $Z_{\text{model2}}(x)$ with a correctly specified x could always achieve the maximal power under each configuration, its power loss is also substantial when the genetic model is mis-specified. For example, when $(\delta, \beta) = (1, 1)$ and $\lambda = 1$, although $Z_{\text{model2}}(1)$ obtains the highest power of 77.7% with a correctly specified genetic model DOM, this value plummets to 18.1% with an incorrect REC model. In addition, the power of the test drops to 9.8% if $Z_{\text{model2}}(0)$ is instead used for the DOM model. On the other hand, MAX3_{cc} always holds a plausible power for different genetic models. In particular, compared with all the $Z_{\text{model2}}(x)$ tests, MAX3_{cc} always yields the highest minimal power regardless of the true genetic models. Table 5 lists the results for case-only study comparing $Z_{\text{model3}}(x)$ and MAX3_{ca}. Results in Table 5 are similar to those in Table 4, i.e., $Z_{\text{model3}}(x)$ is vulnerable to the genetic model mis-specification while MAX3_{ca} can always yield a desirable power.

Table 6 and 7 compare MAX-type tests (MAX3_{cc} and MAX3_{ca}) with chi-square-type tests (χ^2_{cc} and χ^2_{ca}) for both the case-control and case-only studies. In addition to the commonly used genetic models (REC, ADD, DOM), we also consider the under-dominant model (U-DOM) with $f_{1k} \leq f_{2k}$ and the over-dominant model (O-DOM) with $f_{1k} \geq f_{2k}$. In particular, we specify $x = -0.5$ to simulate data under U-DOM model and $x = 1.5$ for the O-DOM model. As expected, if the monotonic minor allele-penetrance assumption holds (such as REC, ADD and DOM), the MAX-type tests are more powerful than the chi-

Figure 3: Empirical type I error rates for MAX_{ca} in case-only study with different disease prevalence ζ. γ_2 represents the dependence level between E and G. $\gamma_1 = \gamma_2/2$ and $m_0 = m_1 = 500$.

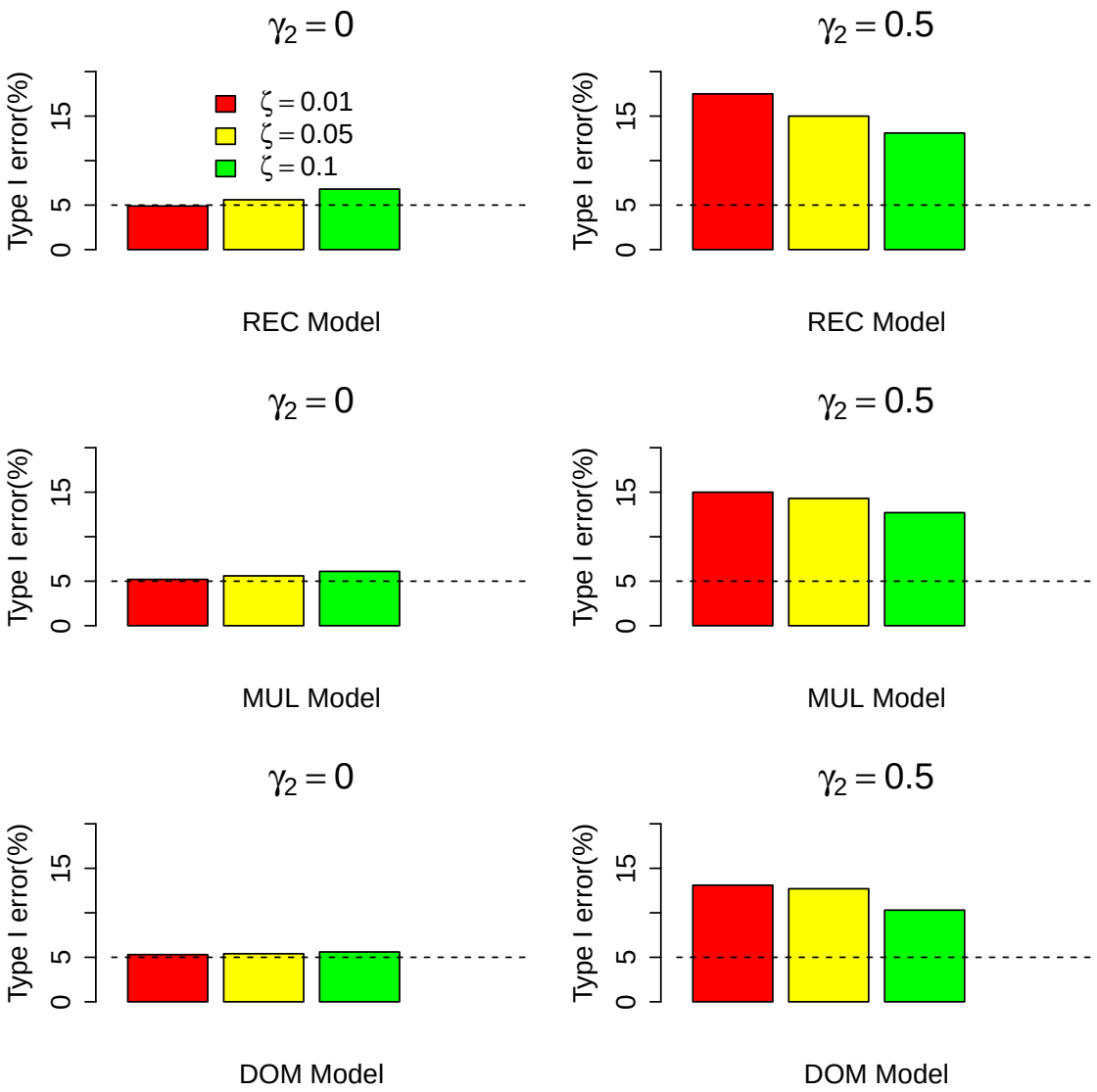


Table 4: Empirical power (%) of $Z_{\text{model2}}(x)$ and MAX3_{cc} for case-control design. The significance level is 0.05 and $m_0 = m_1 = 500$.

		$\lambda = 1$			$\lambda = 1.5$			$\lambda = 2$		
(δ, β)	Test	REC	MUL	DOM	REC	MUL	DOM	REC	MUL	DOM
(1,1)	$Z_{\text{model2}}(0)$	67.8	37.6	9.8	93.1	63.5	10.0	99.1	81.3	9.9
	$Z_{\text{model2}}(0.5)$	46.9	61.9	65.7	77.8	87.1	80.8	92.9	96.6	84.8
	$Z_{\text{model2}}(1)$	18.1	51.6	77.7	33.0	77.4	91.6	50.9	90.6	95.6
	MAX3_{cc}	59.5	56.7	70.8	89.9	83.2	87.8	98.4	94.9	93.6
(1,2)	$Z_{\text{model2}}(0)$	70.5	37.7	6.5	93.0	61.5	6.5	98.7	78.7	6.7
	$Z_{\text{model2}}(0.5)$	55.4	56.9	39.4	83.3	82.1	52.2	94.9	93.5	56.6
	$Z_{\text{model2}}(1)$	24.0	44.1	59.2	44.4	68.1	76.9	63.4	82.5	84.2
	MAX3_{cc}	64.3	52.5	51.1	90.3	77.9	71.1	97.9	91.1	80.5
(2,1)	$Z_{\text{model2}}(0)$	53.2	26.3	5.8	81.5	44.0	4.9	95.3	60.3	5.8
	$Z_{\text{model2}}(0.5)$	38.6	45.8	46.9	66.9	69.9	62.4	87.1	84.2	68.9
	$Z_{\text{model2}}(1)$	17.5	37.6	57.9	31.5	59.6	76.1	49.6	74.9	84.1
	MAX3_{cc}	45.9	40.3	50.5	76.2	64.7	70.4	92.8	80.9	80.1

Table 5: Empirical power (%) of $Z_{\text{model3}}(x)$ and MAX3_{ca} for case-only design. The significance level is 0.05 and $m_0 = m_1 = 500$.

		$\lambda = 0.5$			$\lambda = 0.75$			$\lambda = 1$		
(δ, β)	Test	REC	MUL	DOM	REC	MUL	DOM	REC	MUL	DOM
(1,1)	$Z_{\text{model3}}(0)$	50.5	21.9	6.0	85.5	42.2	6.6	98.1	62.1	7.1
	$Z_{\text{model3}}(0.5)$	37.8	36.3	33.8	73.1	64.3	52.4	93.3	83.9	67.3
	$Z_{\text{model3}}(1)$	16.6	29.5	46.1	34.0	52.2	70.0	55.0	71.5	84.4
	MAX3_{ca}	42.9	32.1	38.1	80.0	58.8	62.2	96.7	79.6	79.0
(1,2)	$Z_{\text{model3}}(0)$	54.9	20.6	5.1	86.0	38.7	5.1	97.2	56.5	5.0
	$Z_{\text{model3}}(0.5)$	47.4	29.9	14.4	79.2	55.2	21.1	94.2	74.1	27.9
	$Z_{\text{model3}}(1)$	25.2	23.0	24.8	47.4	41.5	40.5	68.7	58.4	54.4
	MAX3_{ca}	49.3	27.1	19.6	81.9	50.4	33.5	96.1	70.0	46.5
(2,1)	$Z_{\text{model3}}(0)$	23.6	9.9	4.4	48.4	17.2	4.2	73.4	27.3	3.8
	$Z_{\text{model3}}(0.5)$	18.4	18.7	17.9	37.3	31.9	28.5	61.1	49.4	37.5
	$Z_{\text{model3}}(1)$	10.4	16.0	24.7	17.5	27.8	40.6	28.1	42.2	54.5
	MAX3_{ca}	19.5	16.8	19.2	41.1	28.2	33.3	66.7	44.6	46.2

Table 6: Empirical power (%) of MAX3_{cc} and χ^2_{cc} for case-control design. The significance level is 0.05, $m_0 = m_1 = 500$, $(\delta, \beta) = (2, 2)$, $\phi = 0.3$, $\alpha = -6$, $\gamma_1 = 0.5$, $\gamma_2 = 1$.

	$\lambda = 1$		$\lambda = 1.5$		$\lambda = 2$	
Genetic model	MAX3_{cc}	χ^2_{cc}	MAX3_{cc}	χ^2_{cc}	MAX3_{cc}	χ^2_{cc}
REC	52.7	49.3	79.7	77.0	93.2	91.9
ADD	40.8	36.3	63.2	59.2	77.9	74.9
DOM	42.0	41.3	62.5	60.9	75.0	74.6
U-DOM	58.8	64.4	85.2	87.9	95.4	96.7
O-DOM	53.0	55.2	73.6	74.8	89.4	90.8

Table 7: Empirical power (%) of MAX3_{ca} and χ^2_{ca} for case-only design. The significance level is 0.05, $m_0 = m_1 = 500$, $(\delta, \beta) = (2, 2)$, $\phi = 0.3$, $\alpha = -6$, $\gamma_1 = 0$, $\gamma_2 = 0$.

	$\lambda = 1$		$\lambda = 1.5$		$\lambda = 2$	
Genetic model	MAX3_{ca}	χ^2_{ca}	MAX3_{ca}	χ^2_{ca}	MAX3_{ca}	χ^2_{ca}
REC	57.4	54.1	83.0	80.8	91.9	90.6
ADD	31.0	27.9	48.0	44.5	57.1	53.8
DOM	23.8	23.3	34.9	33.9	40.5	39.7
U-DOM	68.3	71.0	88.2	89.5	94.2	95.5
O-DOM	13.5	14.5	18.1	18.6	21.6	21.8

square-type tests. However, if this assumption is violated (such as U-DOM and O-DOM), the chi-square-type tests are preferable.

As pointed out in Zang et al. (Zang et al., 2010), an alternative bootstrap re-sampling method can be used to approximate the p-value of the robust test. Compared with the asymptotic method, the computational burden for the bootstrap resampling method is much heavier. However, the bootstrap method can still be useful, especially when the sample size is small and an asymptotically normal distribution assumption may be inaccurate. Hence, in Table 8 we compare the asymptotic method with the bootstrap method and the total sample size varies from 200 to 1000. Based on Table 8, the differences between the two

Table 8: Empirical power (%) of MAX3_{cc} for case-control design and MAX3_{ca} for case-only design using asymptotic method (Asy) and bootstrap resampling method (Boot). $\phi = 0.3$, $\tau = 0$, $\gamma_1 = \gamma_2 = 0$, $(\delta, \beta) = (1, 1)$, $\alpha = -6$, $m_0 = m_1$, sample size=$m_0 + m_1$.

			$\lambda = 0.75$			$\lambda = 1$		
Test	Sample size	Method	REC	MUL	DOM	REC	MUL	DOM
MAX3_{cc}	1000	Asy	43.4	41.5	58.1	66.2	61.7	76.0
		Boot	42.6	40.8	57.2	66.7	62.4	75.5
	500	Asy	25.9	25.6	35.3	40.8	39.6	51.0
		Boot	24.7	25.0	36.1	39.7	38.3	50.7
	200	Asy	16.0	16.6	23.7	23.5	23.9	31.4
		Boot	14.6	15.3	22.9	21.8	24.3	30.8
MAX3_{ca}	1000	Asy	79.9	57.8	63.6	96.8	80.5	78.9
		Boot	78.3	56.6	64.3	96.3	80.8	79.5
	500	Asy	47.4	32.8	36.6	73.7	50.0	52.3
		Boot	47.7	33.5	37.2	72.8	50.9	53.2
	200	Asy	20.5	17.4	24.1	34.1	25.4	32.5
		Boot	20.9	18.1	25.3	33.7	26.2	31.7

methods are rather small. Hence, we recommend using the asymptotic method for the proposed robust tests owing to its computational efficiency.

Additionally, another objective of Table 8 is to compare MAX3_{cc} for the case-control design with MAX3_{ca} for the case-only design, so that we could assess how much power can be gained by using the case-only design. To make a fair comparison, we specify $\zeta = 0.01$ and $\gamma_1 = \gamma_2 = 0$ to ensure the integrity of MAX3_{ca}. Based on the simulation results, we conclude that the power gain can be substantial. For example, with a total sample size of 1000 and a true G-E interaction effect $\lambda = 0.75$, the power gain by using case-only design is over 35% under the REC model and over 15% under the MUL model.

7. Application on real data

We have applied the proposed tests to a GWAS dataset from a multi-stage case-control study of bladder cancer (Rothman et al., 2010). 3,532 cases and 5,120 controls of self-described European descent were genotyped on a total of 591,637 SNPs in stage one. The 100 most significant SNPs in stage one were selected and forwarded to stages two and three for validation with an additional 2,243 cases and 2,789 controls. Combined with the data generated in previous studies, the authors identified 10 SNPs associated with bladder cancer. In addition to the main genetic effect, the authors also investigated the interaction of the 10 significant SNPs with cigarette smoking status as an environmental risk factor. The conventional score test $Z_{\text{model1}}(x)$ based on the logistic regression model (1) was used by the authors to analyze the data under the MUL genetic model assumption and reported rs1495741 as the SNP with significant G-E interaction with a p-value of 0.0013. However, based on our theoretical derivation (Theorem 1) and simulation results (Table 2), the conventional test may be biased if the genetic model is mis-specified.

We re-analyze the case-control data for the 9 of 10 significant SNPs (GSTM1 Del is excluded due to its incomplete genotype information). Specifically, we apply model (2) to this case-control data and use $Z_{\text{model2}}(x)$ and MAX3_{cc} to detect possible G-E interaction. Table 9 lists the p-values. $Z_{\text{model2}}(x)$ under different genetic model assumptions yields quite different p-values, indicating the sensitivity of this test with respect to the choice of x. For example, $Z_{\text{model2}}(1)$ gives p-value of 0.01 whereas $Z_{\text{model2}}(0)$ gives p-value of 0.71 for SNP rs1495741. In contrast, MAX3_{cc} incorporates all the three genetic models and generates a single and more objective p-value which should be used to draw the statistical conclusion when the underlying genetic model is unknown. Finally, when the overall type I error rate is controlled at 5% and the Bonferroni correction is used for multiple comparisons among the 9 SNPs, based on MAX3_{cc}, none of the existing SNPs reports significant G-E interaction. Hence, we speculate that the significant G-E interaction reported from the previous study may be a false-

Table 9: G-E interaction testing results for the SNPs with the risk for urinary bladder cancer

	P-values			
SNPs	$Z_{\text{model2}}(0)$	$Z_{\text{model2}}(0.5)$	$Z_{\text{model2}}(1)$	MAX3_{cc}
rs9642880	0.85	0.66	0.60	0.84
rs710521	0.78	0.71	0.71	0.91
rs2294008	0.79	0.82	0.54	0.79
rs401681	0.04	0.07	0.40	0.10
rs798766	0.74	0.58	0.43	0.67
rs1014971	0.18	0.26	0.52	0.33
rs8102137	0.11	0.87	0.41	0.22
rs11892031	0.01	0.18	0.48	0.01
rs1495741	0.71	0.02	0.01	0.01

positive finding due to the problematic logistic regression model (1) being used on the original data analysis, as supported by Theorem 1 and the simulation results from Table 2.

8. Discussion

In this article, we investigate the impact of the genetic model uncertainty on examining the G-E interaction in both case-control and case-only designs. In case-control design, mis-specification of genetic model can distort both the null and alternative distributions of the test statistic. On the other hand, in case-only design, mis-specification of genetic model will only affect the power while keeping the test validity under the null hypothesis. We propose two robust tests, namely MAX3_{cc} and MAX3_{ca}, for assessing G-E interaction effect in case-control and case-only designs, respectively. Rather than relying on a single test wherein the pre-specified genetic model might be incorrect, the proposed robust tests take the maximum absolute value of the three optimal score tests for the REC, MUL and DOM models as the test statistic and therefore yield plausible

power regardless of the underlying genetic model. The asymptotic formulas to calculate the p-values of the proposed maximum-type tests are provided in this article. Our simulation studies and a real data application demonstrate that the newly developed testing methods have substantially increased performance and robustness in comparison with the existing one in detecting the G-E interaction effect. The associated R code to implement the proposed robust tests can be obtained upon request.

Ideally, our MAX3 tests can be used to predict the underlying genetic model. Then, a score test which is optimal for the identified genetic model can then be used. This naive method, however, can inflate the type I error rate due to multiple comparisons. As an alternative approach, we may adopt a more sophisticated two-stage test which mimics the genetic model selection approach to take the correlation between the model prediction and hypothesis testing into consideration (Zang et al., 2010; Zheng et al., 2008). In this article, we focus on the conventional case-control and case-only designs with at least moderate G-E interaction effect. However, when the interaction effect is low, it may be questionable to use these designs due to a lack of sufficient quality samples. Hence, novel designs are needed for low interaction effects. In addition, we assume that the environmental factor E is casual and can be precisely measured. Unfortunately, in reality the causal factor E is often unaccessible or not accurately measurable and it is known that a mis-specified E can also distort the G-E interaction testing result (Boonstra et al., 2016). Hence, it is worth extending the proposed robust tests to account for measurement error of E.

We dedicate this paper to single marker association study. With the rapid development of next generation sequencing, more and more complex diseases have been found to be associated with multiple gene variants, especially the rare variants. Hence, it is of great interest to extend the proposed tests to handle multiple variants. In the field of multiple variants genetic association study, burden test (Li et al., 2008; Madsen et al., 2009) and sequence kernel association test (SKAT) (Wu et al., 2011; Lee et al., 2012) are arguably two of the most popular tests, and we are aware that both tests generally assume a

known genetic model to construct test statistics. We suspect that similar to the single marker association test, these multiple variants tests may be sensitive to genetic model misclassification as well. To the best of our knowledge, no robust test has been proposed for multiple variants genetic association study. We hope the robust tests proposed in this paper can be somehow inspiring in this active research area. Further study in this area is warranted.

Acknowledgment

The authors would like to thank the associate editor and two referees for their insightful and helpful suggestions which substantially improved our presentation. The research of Yong Zang is partial supported by the design and biostatistics program pilot grant, Indiana CTSI. The research of Hon Keung Tony Ng is partially supported by a grant from the Simons Foundation (#280601). The research of Chi Zhang is partially supported by Showalter Trust, Indiana CTSI.

APPENDIX A: Expression for $V(x)$

Denote $\psi = (\alpha, \beta_1, \beta_2, \delta)'$ as the nuisance parameters, we can write the observed information matrix based on model (2) in the form of a block matrix as

$$I(x) = \begin{pmatrix} \mathrm{I}_{\lambda}(x) & \mathrm{I}_{\lambda\psi}(x)' \\ \mathrm{I}_{\lambda\psi}(x) & \mathrm{I}_{\psi} \end{pmatrix},$$

Where

$$
\begin{aligned}
\mathrm{I}_{\lambda}(x) &= x^2 n_{11}\widetilde{f}_{11}(1-\widetilde{f}_{11}) + n_{21}\widetilde{f}_{21}(1-\widetilde{f}_{21}),\\
\mathrm{I}_{\lambda\psi}(x) &= \begin{pmatrix} xn_{11}\widetilde{f}_{11}(1-\widetilde{f}_{11}) + n_{21}\widetilde{f}_{21}(1-\widetilde{f}_{21}) \\ xn_{11}\widetilde{f}_{11}(1-\widetilde{f}_{11}) \\ n_{21}\widetilde{f}_{21}(1-\widetilde{f}_{21}) \\ xn_{11}\widetilde{f}_{11}(1-\widetilde{f}_{11}) + n_{21}\widetilde{f}_{21}(1-\widetilde{f}_{21}) \end{pmatrix},\\
\mathrm{I}_{\psi} &= \begin{pmatrix} \sum_{j=0}^{2}\sum_{k=0}^{1} n_{jk}\widetilde{f}_{jk}(1-\widetilde{f}_{jk}) & \sum_{k=0}^{1} n_{1k}\widetilde{f}_{1k}(1-\widetilde{f}_{1k}) & \sum_{k=0}^{1} n_{2k}\widetilde{f}_{2k}(1-\widetilde{f}_{2k}) & \sum_{j=0}^{2} n_{j1}\widetilde{f}_{j1}(1-\widetilde{f}_{j1} \\ \sum_{k=0}^{1} n_{1k}\widetilde{f}_{1k}(1-\widetilde{f}_{1k}) & \sum_{k=0}^{1} n_{1k}\widetilde{f}_{1k}(1-\widetilde{f}_{1k}) & 0 & n_{11}\widetilde{f}_{11}(1-\widetilde{f}_{11}) \\ \sum_{k=0}^{1} n_{2k}\widetilde{f}_{2k}(1-\widetilde{f}_{2k}) & 0 & \sum_{k=0}^{1} n_{2k}\widetilde{f}_{2k}(1-\widetilde{f}_{2k}) & n_{21}\widetilde{f}_{21}(1-\widetilde{f}_{21}) \\ \sum_{j=0}^{2} n_{j1}\widetilde{f}_{j1}(1-\widetilde{f}_{j1} & n_{11}\widetilde{f}_{11}(1-\widetilde{f}_{11}) & n_{21}\widetilde{f}_{21}(1-\widetilde{f}_{21}) & \sum_{j=0}^{2} n_{j1}\widetilde{f}_{j1}(1-\widetilde{f}_{j1} \end{pmatrix},
\end{aligned}
$$

with

$$\widetilde{f}_{jk} = \frac{1}{1+\mathrm{e}^{-\tilde{\alpha}-\tilde{\beta}_1\mathrm{I}(j=1)-\tilde{\beta}_2\mathrm{I}(j=2)-\tilde{\delta}k}}$$

and $\tilde{\alpha}$, $\tilde{\beta}_1$, $\tilde{\beta}_2$ and $\tilde{\delta}$ are the MLEs based on model (2) given $\lambda = 0$. By inverting the information matrix, we have

$$
\begin{aligned}
\mathrm{I}^{-1}(x) &= \begin{pmatrix} \mathrm{I}^{\lambda}(x) & \mathrm{I}^{\lambda\psi}(x)' \\ \mathrm{I}^{\lambda\psi}(x) & \mathrm{I}^{\psi}(x) \end{pmatrix},\\
\mathrm{I}^{\lambda}(x) &= \Big(\mathrm{I}_{\lambda}(x) - \mathrm{I}_{\lambda\psi}(x)'\mathrm{I}_{\psi}^{-1}\mathrm{I}_{\lambda\psi}(x)\Big)^{-1}.
\end{aligned}
$$

Finally, based on the inverted information matrix, the variance estimate for $U(x)$ is

$$V(x) = \widehat{\mathrm{Var}}(U(x)) = \frac{1}{\mathrm{I}^{\lambda}(x)} = \mathrm{I}_{\lambda}(x) - \mathrm{I}_{\lambda\psi}(x)'\mathrm{I}_{\psi}^{-1}\mathrm{I}_{\lambda\psi}(x).$$

Appendix B: Expression of ρ_{x_1,x_2}

We can write down the information matrix in the form of block matrix as

$$I(x_1,x_2) = \begin{pmatrix} \mathrm{I}_{\lambda_1\lambda_2}(x_1,x_2) & \mathrm{I}_{\lambda_1\lambda_2\psi}(x_1,x_2)' \\ \mathrm{I}_{\lambda_1\lambda_2\psi}(x_1,x_2) & \mathrm{I}_{\psi} \end{pmatrix},$$

the detailed expression of I_{ψ} are given in Appendix A and we have

$$\begin{aligned}
\mathrm{I}_{\lambda_1\lambda_2}(x_1,x_2) &= \begin{pmatrix} x_1^2 n_{11}\tilde{f}_{11}(1-\tilde{f}_{11}) + n_{21}\tilde{f}_{21}(1-\tilde{f}_{21}) & x_1x_2 n_{11}\tilde{f}_{11}(1-\tilde{f}_{11}) + n_{21}\tilde{f}_{21}(1-\tilde{f}_{21}) \\ x_1x_2 n_{11}\tilde{f}_{11}(1-\tilde{f}_{11}) + n_{21}\tilde{f}_{21}(1-\tilde{f}_{21}) & x_2^2 n_{11}\tilde{f}_{11}(1-\tilde{f}_{11}) + n_{21}\tilde{f}_{21}(1-\tilde{f}_{21}) \end{pmatrix} \\
\mathrm{I}_{\lambda_1\lambda_2\psi}(x_1,x_2) &= \begin{pmatrix} x_1 n_{11}\tilde{f}_{11}(1-\tilde{f}_{11}) + n_{21}\tilde{f}_{21}(1-\tilde{f}_{21}) & x_2 n_{11}\tilde{f}_{11}(1-\tilde{f}_{11}) + n_{21}\tilde{f}_{21}(1-\tilde{f}_{21}) \\ x_1 n_{11}\tilde{f}_{11}(1-\tilde{f}_{11}) & x_2 n_{11}\tilde{f}_{11}(1-\tilde{f}_{11}) \\ n_{21}\tilde{f}_{21}(1-\tilde{f}_{21}) & n_{21}\tilde{f}_{21}(1-\tilde{f}_{21}) \\ x_1 n_{11}\tilde{f}_{11}(1-\tilde{f}_{11}) + n_{21}\tilde{f}_{21}(1-\tilde{f}_{21}) & x_2 n_{11}\tilde{f}_{11}(1-\tilde{f}_{11}) + n_{21}\tilde{f}_{21}(1-\tilde{f}_{21}) \end{pmatrix}.
\end{aligned}$$

445

The inverted information matrix is then given by

$$\begin{aligned}
\mathrm{I}^{-1}(x_1,x_2) &= \begin{pmatrix} \mathrm{I}^{\lambda_1\lambda_2}(x_1,x_2) & \mathrm{I}^{\lambda_1\lambda_2\psi}(x_1,x_2)' \\ \mathrm{I}^{\lambda_1\lambda_2\psi}(x_1,x_2) & \mathrm{I}^{\psi}(x_1,x_2) \end{pmatrix}. \\
\mathrm{I}^{\lambda_1\lambda_2}(x_1,x_2) &= (\mathrm{I}_{\lambda_1\lambda_2}(x_1,x_2) - \mathrm{I}_{\lambda_1\lambda_2\psi}(x_1,x_2)'\mathrm{I}_{\psi}^{-1}\mathrm{I}_{\lambda_1\lambda_2\psi}(x_1,x_2))^{-1}.
\end{aligned}$$

Finally, the variance-covariance matrix for $(U(x_1), U(x_2))$ is $\left(\mathrm{I}^{\lambda_1\lambda_2}(x_1,x_2)\right)^{-1}$. Therefore, the correlation ρ_{x_1,x_2} can be expressed as

$$\rho_{x_1,x_2} = \frac{(1,0)\left(\mathrm{I}^{\lambda_1\lambda_2}(x_1,x_2)\right)^{-1}(0,1)'}{\sqrt{(1,0)\left(\mathrm{I}^{\lambda_1\lambda_2}(x_1,x_2)\right)^{-1}(1,0)'(0,1)\left(\mathrm{I}^{\lambda_1\lambda_2}(x_1,x_2)\right)^{-1}(0,1)'}}.$$

References

[1] Armitage P. Test for linear trends in proportions and frequencies. *Biometrics* 1955; 11: 375-386.

[2] Boonstra PS, Mukherjee B, Gruber SB, Ahn J, Schmit SL, Chatterjee N. Tests for gene-environment interactions and joint effects with exposure misclassification. *American Journal of Epidemiology* 2016; 183: 237-247.

[3] Freidlin B, Zheng G, Gastwirth JL. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Human Heredity* 2002; 53(3): 146-152.

[4] Gillespie J. Population genetics: a concise guide. *John Hopkins University* 2004.

[5] Gonzalez JR, Carrasco JL, Dudbridge F, Armengol L, Estivill X, Moreno V. Maximizing association statistics over genetic models. *Genetic Epidemiology* 2008; 32(3): 246-254.

[6] Hunter DJ. Gene-environment interactions in human diseases. *Nature Reviews Genetics* 2005; 6: 287-298.

[7] Lee S, Wu M, Lin X. Optimal test for rare variant effects in sequencing association studies. *Biostatistics* 2012; 13: 762-775.

[8] Li B, Leal, SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics* 2008; 83: 311-321.

[9] Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics* 2009; 5: e1000384.

[10] Marigorta UM, Gibson G. A simulation study of gene-environment interaction in GWAS implies ample hidden effects. *Frontiers in Genetics* 2014; 5: 225.

[11] Mukherjee B, Ahn J, Gruber SB, Chatterjee N. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *American Journal of Epidemiology* 2012; 175: 177-190.

[12] Piegorsch WW, Weinberg CR, Taylor J. Nonhierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine* 1994; 13: 153-162.

[13] Rothman N, Garcia-Closas M, Chatterjee N, et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nature Genetics* 2010; 42(11): 978-984.

[14] Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics* 1997; 53: 1253-1261.

[15] Schmidt S, Schaid DJ. Potential misinterpretation of the case-only study to assess gene-environment interaction. *American Journal of Epidemiology* 1999; 150: 878-885.

[16] Umbach DM, Weinberg CR. Designing and analysing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine* 1997; 16: 1731-1743.

[17] Wang K, Sheffield VC. A constrained-likelihood approach to marker-trait association studies. *American Journal of Human Genetics* 2005; 77(5): 768-780.

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

[19] Yamada R, Okada Y. An optimal dose-effect mode trend test for SNP genotype tables. *Genetic Epidemiology* 2009; 33(2): 114-127.

[20] Zang Y, Fung WK, Zheng G. Simple algorithms to calculate asymptotic null distribution of robust tests in case-control genetic association studies in R. *Journal of Statistical Software* 2010; 33(8).

[21] Zang Y, Fung WK, Zheng G. Asymptotic powers for matched trend tests and robust matched trend tests in case-control genetic association studies. *Computational Statistics and Data Analysis* 2010; 54: 65-77.

[22] Zang Y, Fung WK. Robust tests for matched case-control genetic association studies. *BMC Genetics* 2010, 11: 91.

[23] Zang Y. Robust tests under genetic model uncertainty in case-control association studies. *Ph.D. dissertation, The University of Hong Kong* 2011.

[24] Zheng G, Freidlin B, Li Z, Gastwirth JL. Choice of scores in trend tests for case-control studies of candidate-gene associations *Biometrical Journal* 2003; 45(3): 335-348.

[25] Zheng G, Freidlin B, Gastwirth JL. Robust genomic control for association studies. *American Journal of Human Genetics* 2006; 78(2): 350-356.

[26] Zheng G, Ng HKT. Genetic model selection in two-phase analysis for case-control association studies. *Biostatistics* 2008; 9(3): 391-399.