UW Biostatistics Working Paper Series

11-18-2014

Testing Gene-Environment Interactions in the Presence of Measurement Error

Chongzhi Di
Fred Hutchinson Cancer Research Center, cdi@fredhutch.org

Li Hsu
Fred Hutchinson Cancer Research Center, lih@fhcrc.org

Charles Kooperberg
fred hutchinson cancer research center, clk@fhcrc.org

Alex Reiner
Fred Hutchinson Cancer Research Center, apreiner@fhcrc.org

Ross Prentice
Fred Hutchinson Cancer Research Center, rprentic@whi.org

1. Introduction

Investigating the interplay between genes and environmental factors, i.e., gene-environment interaction ($G \times E$), is very important to understand the etiology of complex diseases, especially with recent advances in genome-wide association studies (GWAS) and exome sequencing studies (Park et al., 2010; Goldstein et al., 2013). These studies successfully identified many genetic variants that are marginally associated with complex diseases including cancers, cardiovascular diseases, and diabetes. However, these genetic variants explain only a small fraction of familiar aggregation of corresponding diseases (Zuk et al., 2012). It is hypothesized that the interplay between genetic and environmental risk factors, such as diet, physical activity, and air pollution, plays a significant role in explaining familiar aggregation. Investigating $G \times E$ could potentially lead to identification of variants that affect certain subgroups only but missed by marginal analysis. It could also lead to better understanding of disease etiology, such as how a genetic association may vary over different subgroups of environmental risk factors. This may help inform effective targeted intervention strategies for reducing disease burden.

Compared with marginal genetic association analysis, investigation of $G \times E$ presents additional challenges, partly due to the complexity in environmental exposure assessments (Thomas, 2010). In practice, measurements of environmental factors are often imprecise, sometimes with substantial measurement error. For example, blood pressure measurements are subject to device recording error and daily variability, while diet and physical activity are often measured by self-reported questionnaires, which are known to be biased and subject to systematic over- or under-reporting of energy intake (e.g., Prentice et al., 2011). While there is a rich literature on accounting for measurement error in assessing main effects (Carroll et al., 2006), little work has been done on its effect in $G \times E$ analysis. Measurement error of E could have different implications in testing $G \times E$ interactions compared to testing main effects. For example, it is well known that naïve tests that ignore measurement error are valid in testing the main effect of E (Carroll et al.,

2006) under the classical measurement error model. However, it is not clear whether the naïve test for interaction has the proper size. Previous studies generally ignored measurement error in the interaction analysis. Based on simulation studies, it was conjectured that the type I error of the naïve test was still valid under non-differential measurement error (Greenwood et al., 2006; Williamson et al., 2010). However, our studies show that naïve tests can have incorrect type I error rate even under non-differential measurement error. As a result, ignoring measurement error may lead to incorrect conclusions, such as spurious $G \times E$ interaction findings. Thus, it is very important to investigate the influence of measurement error on $G \times E$ systematically and develop statistical methods for $G \times E$ that appropriately account for measurement complexities.

Another challenge for $G \times E$ analysis is the lack of power, which has motivated many methods works to improve power especially for genome-wide $G \times E$ scans. The methods include case-only and empirical Bayes methods (Khoury and Flanders, 1996; Chatterjee and Carroll, 2005; Mukherjee and Chatterjee, 2008), two-stage testing (Kooperberg and LeBlanc, 2008; Dai et al., 2012; Hsu et al., 2012), and set-based $G \times E$ testing (Lin et al., 2013). Strategies to improve power from the E side (e.g., to improve measurement accuracy of E) have received less attention (Wong et al., 2003). It is also of interest to study performances of various tests, including naïve tests and measurement error corrected approaches, in terms of power in detecting $G \times E$.

In this paper, we systematically investigate the consequence of ignoring measurement error in the analysis of $G \times E$ interactions in terms of type I error and power. We then propose a regression calibration–based testing procedure that accounts for measurement error and compare its power with the naïve test. The finite sample performances of these tests are evaluated via simulation studies. The proposed methods are illustrated by applying them to a genetic association study on coronary heart disease.

2. Influences of measurement error in E on $G \times E$

2.1 *Notations and model setup*

We consider testing $G \times E$ when the environmental exposure E is measured with error. Let Y, G, E and Z denote the disease outcome, genotype, true environmental exposure, and other confounders, respectively. The true exposure E is measured with error, and we denote the imprecisely measured exposure by X. We assume the following generalized linear model (McCullagh and Nelder, 1989) for disease association:

$$\begin{aligned} h(\mu) &= \beta_0 + \beta_1 G + \beta_2 E + \beta_3 GE + \beta_4 Z, \\ X &= E + \epsilon, \end{aligned} \tag{1}$$

where $\mu = \mathrm{E}(Y \mid G, E)$ and h is a link function. The popular choices of h are canonical links, e.g., *identify* for continuous outcomes, *logit* for binary outcomes and *log* for counts data. In this paper, we focus on classic measurement error models. Namely, $\mathrm{E}(\epsilon) = 0$, $\mathrm{var}(\epsilon) = \sigma_E^2$ and E, G and Z are uncorrelated with ϵ. Under this model, the interaction GE is also subject to measurement error, with the induced measurement error structure

$$GX = GE + G\epsilon.$$

The measurement error for GE, namely $G\epsilon$, is still centered around 0 and uncorrelated with GE conditional on G. However, it has non-constant variance, as $\mathrm{var}(G\epsilon \mid G) = G^2\sigma_E^2$.

We are interested in testing the null hypothesis in model (1), $H_0 : \beta_3 = 0$ (testing interaction term). In genetic studies, sometimes one is also interested in testing the composite $H_0' : \beta_1 = 0, \beta_3 = 0$, which has been proposed as a test for joint genetic effects (Kraft et al., 2007; Williamson et al., 2010) to identify variants while accounting for potential heterogeneity effects across different levels of E.

The naïve procedure ignores measurement error and fits the following working model,

$$h(\mu) = \gamma_0 + \gamma_1 G + \gamma_2 X + \gamma_3 GX + \gamma_4 Z. \tag{2}$$

The score, Wald or likelihood ratio tests derived based on this model are referred to as naïve

tests. We focus on the Wald test, which is (asymptotically) equivalent to the score or likelihood ratio test to the first order under the null hypothesis. Let $T_{naive} = \widehat{\gamma}_3^T V_{\gamma,33}^{-1} \widehat{\gamma}_3$ be the Wald test statistic under model (2), where $V_{\gamma,33} = \{ I_{\gamma\gamma}^{-1} \}_{(3,3)}$ is the $(3,3)$ element of the inverse Fisher information matrix. If the true E without measurement error is available, one can define the ideal test statistic as $T_{ideal} = \widehat{\beta}_3^T V_{\beta,33}^{-1} \widehat{\beta}_3$, where $V_{\beta,33} = \{ I_{\beta\beta}^{-1} \}_{(3,3)}$ is the $(3,3)$ element of the inverse Fisher information matrix of the true model (1). In practice, however, T_{ideal} cannot be used directly since E is not observable in the presence of measurement error. We view T_{ideal} as the gold standard that achieves the maximal asymptotic power, and only use it for comparing power in simulation studies.

Statistical issues discussed in this paper apply to $G \times E$ analyses with single or multiple environmental exposures E, single or multiple loci G, candidate gene or genome-wide association analyses. However, without loss of generality and for ease of presentation, we consider one environment exposure E and one locus G throughout the paper.

2.2 *A simple scenario with continuous outcome and binary genotype*

To illustrate the effect of measurement error on $G \times E$, we first consider a simple scenario: 1) Y follows a Gaussian distribution; 2) G is binary, taking values 0 and 1 with probability $1-p$ and p, respectively; 3) ϵ follows Gaussian distribution $N(0, \sigma_\epsilon^2)$ and ϵ is independent of E, G, and Y; 4) Z is absent. The third assumption implies non-differential measurement error. G and E are possibly correlated. We define the following notation on the conditional distribution of E given G,

$$\mathrm{E}(E|G=0) = \mu_0, \mathrm{var}(E|G=0) = \sigma_{e0}^2, \mathrm{E}(E|G=1) = \mu_1, \mathrm{var}(E|G=1) = \sigma_{e1}^2.$$

By the law of total probability, $\mu_E = (1-p)\mu_0 + p\mu_1$ and $\sigma_E^2 = (1-p)\sigma_{e0}^2 + p\sigma_{e1}^2 + p(1-p)(\mu_1 - \mu_0)^2$. The subgroup effects of E on Y in genotype groups $G=0$ and $G=1$ are β_2 and $\beta_2 + \beta_3$, respectively. Note that we use E to denote the environmental variable and E to denote the expectation operator. If one ignores measurement error, the estimates of regression coefficients will be attenuated towards the null in each of the two subgroups (Carroll et al., 2006), with attenuation

factors (also called reliability ratios) $\lambda_1 = \frac{\sigma^2_{e1}}{\sigma^2_{e1}+\sigma^2_\epsilon}$ and $\lambda_0 = \frac{\sigma^2_{e0}}{\sigma^2_{e0}+\sigma^2_\epsilon}$, respectively. More precisely,

$$\mathrm{E}(\widehat{\gamma}_2) = \lambda_0\beta_2, \quad \mathrm{E}(\widehat{\gamma}_2 + \widehat{\gamma}_3) = \lambda_1(\beta_2 + \beta_3),$$

thus,

$$\mathrm{E}(\widehat{\gamma}_3) = (\lambda_1 - \lambda_0)\beta_2 + \lambda_1\beta_3 = \left(\frac{\sigma^2_{e1}}{\sigma^2_{e1} + \sigma^2_\epsilon} - \frac{\sigma^2_{e0}}{\sigma^2_{e0} + \sigma^2_\epsilon}\right)\beta_2 + \frac{\sigma^2_{e1}}{\sigma^2_{e1} + \sigma^2_\epsilon}\,\beta_3.$$

Proposition 1. For the measurement error model (1) with Gaussian outcome Y, identity link function h and binary genotype G, the following results hold.

(a) The MLE from the naïve model, $\widehat{\gamma}_3$, is generally biased for the true $G \times E$ coefficient β_3, with bias term $\mathrm{E}(\widehat{\gamma}_3) - \beta_3 = (\lambda_1 - \lambda_0)\beta_2 + (\lambda_1 - 1)\beta_3$.

(b) Under $H_0 : \beta_3 = 0$, $\mathrm{E}(\widehat{\gamma}_3) = (\lambda_1 - \lambda_0)\beta_2$. Thus, $\widehat{\gamma}_3$ is biased unless $\lambda_0 = \lambda_1$ or $\beta_2 = 0$. The corresponding test T_{naive} has incorrect type I errors unless $\lambda_0 = \lambda_1$ or $\beta_2 = 0$.

(c) For fixed β_2, σ^2_{e0} and σ^2_{e1}, the bias term under H_0 depends only on $\lambda_1 - \lambda_0$. Its absolute value is monotonically increasing with respect to σ^2_ϵ in the interval $[0, \sigma_{e0}\sigma_{e1}]$ and monotonically decreasing in $[\sigma_{e0}\sigma_{e1}, \infty]$, where $\sigma_{e0}\sigma_{e1}$ is the geometric mean of the variances of E in the genotype subgroups.

(e) Under $G - E$ independence, $\lambda_0 = \lambda_1$ and thus $\widehat{\gamma}_3$ is unbiased for β_3 under H_0.

From these results, one can see that the least squares estimate of the interaction term is generally biased even in the no-interaction-case. Intuitively, marginal effects of E in the two subgroups $G = 0$ and $G = 1$ are both attenuated towards the null, namely, $\gamma_2 = \lambda_0\beta_2$ and $\gamma_2+\gamma_3 = \lambda_1(\beta_2+\beta_3)$ with attenuation factors $\lambda_0, \lambda_1 \in [0, 1]$. However, the magnitude of attenuation in the two groups can be different. The interaction between the mis-measured exposure X and G, as the difference between two subgroup effects, is $\gamma_3 = (\lambda_1 - \lambda_0)\beta_2 + \lambda_1\beta_3$. As a result, the null hypothesis of $H_0 : \beta_3 = 0$ in the true model does not imply $\gamma_3 = 0$ in the working model, unless $\lambda_0 = \lambda_1$ or $\beta_2 = 0$. Thus, the naïve tests for interactions are invalid in general, and the magnitude of the

inflation in the type I error depends on both the differential attenuation factor $\lambda_1 - \lambda_0$ and the main environmental effect β_2.

We further look at how measurement error variance σ_ϵ^2 affect the amount of bias through $\lambda_1 - \lambda_0$. When σ_ϵ^2 is small (implying very accurate measurements), both λ_0 and λ_1 are close to 1 and thus their difference is small. On the other hand, when σ_ϵ^2 is very large (implying extremely inaccurate measurements), both λ_0 and λ_1 are close to 0 and thus their difference is also small. Specifically, the magnitude of bias reaches its maximum when the measurement error variance is the same as the geometric mean of the variances of E in the two genotype subgroups.

There are special occasions under which the bias vanishes and the naïve test becomes valid. The first setting is $\beta_2 = 0$, which implies that there is no main effect for the environmental factor E. The second setting is $\lambda_0 = \lambda_1$, which is equivalent to $\sigma_{e0}^2 = \sigma_{e1}^2$. A sufficient condition for this is $G - E$ independence. However, we want to point out that $G - E$ uncorrelatedness is not a sufficient condition. It is the second moment that really matters, i.e., whether the variance of E varies in different genotype subgroups. For example, $E|G = 0 \sim N(0, 1)$ and $E|G = 1 \sim N(0, 2^2)$, G and E are uncorrelated, but the naïve test is still not valid since $\sigma_{e0}^2 \neq \sigma_{e1}^2$.

Figure 1 illustrates the magnitude of type I error inflation of the naïve test under this scenario. The true parameter values are $\beta_0 = 1, \beta_1 = 1, \beta_2 = 1, \beta_3 = 0$ and there is no interaction effect. We generated E from $N(0, 1)$ and binary G from a logistic regression model given E with correlation of G and E $\rho = 0.6, 0.4, 0.2$. The intercepts in the logistic regression model were chosen to keep the minor genotype frequency $p = 0.36$ or $p = 0.04$, corresponding to minor allele frequency 0.2 under dominant and recessive models, respectively. The type I error is generally inflated and the inflation increases when the correlation between G and E increases, or when the minor genotype frequency decreases. The type I error rates are not monotone with respect to σ_ϵ, but increase first and then decrease as σ_ϵ varies from 0 to ∞. This observation verifies the theoretical result in Proposition 1(c).

To summarize, this simple scenario illustrates that, contrary to conjectures in the literature, naïve tests for interactions generally do not maintain correct type I error even under non-differential measurement error.

2.3 *General cases for continuous outcomes*

We now consider general cases for continuous outcomes via linear models, where Z is present and G is not limited to binary genotype, and assess the bias of the naïve interaction coefficient estimation. Let $D_E = (1, G, E, GE, Z)$ and $D_X = (1, G, X, GX, Z)$ denote design matrices for the true model and working model, respectively, and let $\Delta = D_X - D_E = (0, 0, \epsilon, G\epsilon, 0)$ denote their difference. Under model (2.1), $\mathrm{E}(Y|G, E, Z) = D_E\beta$ and $\mathrm{var}(Y) = \sigma^2 I_{n\times n}$, where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ and σ^2 is residual error variance for the outcome variable Y.

The least squares estimator (or equivalently the maximum likelihood estimator) of γ in the working model (2) is given by $\widehat{\gamma} = (D_X^T D_X)^{-1} D_X^T Y$. To assess the bias of $\widehat{\gamma}$, we calculate its expection as

$$\begin{aligned}
\mathrm{E}(\widehat{\gamma}) &= \mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T Y\,\} \\
&= \mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T D_E\beta\,\} \\
&= \mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T (D_X - \Delta)\beta\,\} \\
&= \beta - \mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T \Delta\beta\,\} \\
&= \beta - \mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T (\beta_2\epsilon + \beta_3 G\epsilon)\,\} \\
&= \beta - \beta_2 \cdot \mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T \epsilon\,\} - \beta_3 \cdot \mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T (G\epsilon)\,\}. \qquad (3)
\end{aligned}$$

To interpret the bias, we note that the two expetation terms can be represented as

$$\begin{aligned}
\mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T \epsilon\,\} &= \mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T (X - E)\,\} \\
&= \mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T X - (D_X^T D_X)^{-1} D_X^T E\,\} \\
&= (0,0,1,0,0)^T - \mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T E\,\}, \qquad (4) \\
\mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T (G\epsilon)\,\} &= (0,0,0,1,0)^T - \mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T (GE)\,\}. \qquad (5)
\end{aligned}$$

According to these equations, the bias terms $\mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T \epsilon\,\}$ and $\mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T (G\epsilon)\,\}$ are closely linked to coefficients from regression calibration equations. We consider the following working models,

$$\begin{aligned}
E &= \lambda_{10} + \lambda_{11} G + \lambda_{12} X + \lambda_{13} GX + \lambda_{14} Z + \varepsilon_e, \\
GE &= \lambda_{20} + \lambda_{21} G + \lambda_{22} X + \lambda_{23} GX + \lambda_{24} Z + \varepsilon_{ge}.
\end{aligned}$$

The last terms in (4) and (5) are exactly least square estimates of these two regression calibration models, and thus $\mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T E\,\} = (\lambda_{10}, \lambda_{11}, \lambda_{12}, \lambda_{13}, \lambda_{14})$ and $\mathrm{E}\{\,(D_X^T D_X)^{-1} D_X^T (GE)\,\} = (\lambda_{20}, \lambda_{21}, \lambda_{22}, \lambda_{23}, \lambda_{24})$. Note that we specifically include the interaction term GX in the calibration equations, which is crutial for testing $G \times E$ as will be shown in the next section. The coefficient λ_{13} in the calibration model for E is related to differential attenuation factor with respect to genotype subgroups. In fact, one can show that $\lambda_{13} = \lambda_1 - \lambda_0$ under the simple scenario with binary G in the absence of Z.

Based on (3–5), it follows that

$$\mathrm{E}\begin{pmatrix}\widehat{\gamma}_0\\ \widehat{\gamma}_1\\ \widehat{\gamma}_2\\ \widehat{\gamma}_3\\ \widehat{\gamma}_4\end{pmatrix} = \begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix} + \beta_2\begin{pmatrix}\lambda_{10}\\ \lambda_{11}\\ \lambda_{12}-1\\ \lambda_{13}\\ \lambda_{14}\end{pmatrix} + \beta_3\begin{pmatrix}\lambda_{20}\\ \lambda_{21}\\ \lambda_{22}\\ \lambda_{23}-1\\ \lambda_{24}\end{pmatrix}$$

$$= \begin{pmatrix}1 & 0 & \lambda_{10} & \lambda_{20} & 0\\ 0 & 1 & \lambda_{11} & \lambda_{21} & 0\\ 0 & 0 & \lambda_{12} & \lambda_{22} & 0\\ 0 & 0 & \lambda_{13} & \lambda_{23} & 0\\ 0 & 0 & \lambda_{14} & \lambda_{24} & 1\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix}$$

The following are implications of this result.

Proposition 2. Consider the measurement error model (1) with identity link function h. The following results hold.

(a) The least square estimate for $G \times E$ from the naïve model, $\widehat{\gamma}_3$, satisfies $\mathrm{E}(\widehat{\gamma}_3) = \lambda_{13}\beta_2 + \lambda_{23}\beta_3$ and $\mathrm{E}(\widehat{\gamma}_3 - \beta_3) = \lambda_{13}\beta_2 + (\lambda_{23} - 1)\beta_3$.

(b) Under $H_0 : \beta_3 = 0$, one has $\mathrm{E}(\widehat{\gamma}_3) = \lambda_{13}\beta_2$. Thus, $\widehat{\gamma}_3$ is biased unless $\lambda_{13} = 0$ or $\beta_2 = 0$.

(c) The effect of E in subgroup $G = g$, $\mathrm{E}(\widehat{\gamma}_2 + g\widehat{\gamma}_3) = (\lambda_{12}\beta_2 + \lambda_{22}\beta_3) + (\lambda_{13}\beta_2 + \lambda_{23}\beta_3)g = (\lambda_{12} + g\lambda_{13})\beta_2 + (\lambda_{22} + g\lambda_{23})\beta_3$. Under $H_0 : \beta_3 = 0$, $\mathrm{E}(\widehat{\gamma}_2 + g\widehat{\gamma}_3) = (\lambda_{12} + g\lambda_{13})\beta_2$ for subgroup $G = g$.

(e) When G and E are independent, $\lambda_{13} = 0$ and $\sigma^2_{ge} = \mathrm{var}(E|G = g)$ is invariant with respect to g. On the other hand, $\lambda_{13} = 0$ does not necessarily imply $G - E$ independence

The main idea remains that estimates of subgroup effects are attenuated towards the null in each genotype subgroup, but the magnitude of attenuation may be different across subgroups. These results indicate that the magnitude of type I error inflation under H_0 is determined by two factors.

The first factor is β_2, the main effect of E in the genotype subgroup $G = 0$. The second factor is λ_{13}, the coefficient for the interaction term GX in the regression calibration equation for E, which can also be interpreted as differential attenuation factors among genotype subgroups.

If the bias term $\lambda_{13}\beta_2$ is known, one can construct a bias-corrected test statistic $T_{naive,bc} = (\widehat{\gamma}_3 - \lambda_{13}\beta_2)^T \widehat{V}_{\gamma,33}^{-1}(\widehat{\gamma}_3 - \lambda_{13}\beta_2)$, which will maintain correct type I error rates under H_0. In application, if additional validation data or replicates are available, it is possible to estimate $\lambda_{13}\beta_2$ first and construct a corrected test, although it is important to account for such uncertainty in calculating p values.

Remark 1: The results in the previous section can be obtained as a corollary of Proposition 2. When the genotype is binary and Z is absent, one can show that $\lambda_{13} = \lambda_1 - \lambda_0 = \frac{\sigma_{e1}^2}{\sigma_{e1}^2+\sigma_\epsilon^2} - \frac{\sigma_{e0}^2}{\sigma_{e0}^2+\sigma_\epsilon^2}$. In addition, $\lambda_{20} = 0, \lambda_{21} = \lambda_{10} + \lambda_{11}, \lambda_{22} = 0, \lambda_{23} = \lambda_{12} + \lambda_{13}$.

Remark 2: In generalized linear models for non-Gaussian outcomes, the effect of measurement error is more complex and not necessarily in the form of attenuation. However, we expect that the naïve tests are not valid in general due to similar reasons. That is, the effect of measurement error varies across genotype subgroups. As a result the interaction estimate from the naïve model can be biased. However, it is difficult to obtain an explicit form for the bias term in generalized linear models. Simulation studies are conducted to evaluate its performance in Section 4.

3. Corrected testing procedures under classic measurement error

In the statistical literature, there are several popular approaches to correcting for measurement error under a classical measurement error model, including regression calibration (Carroll and Stefanski, 1990), simulation extrapolation (Cook and Stefansk, 1994), conditional score and corrected score methods and others (for a comprehensive review of these methods, see Carroll et al., 2006). For hypothesis testing, when the variable with measurement error appears in the main effect term, it has been shown that the naïve test that ignores measurement error is still valid as it maintains

correct type I error rate under the null hypothesis. Thus, it is popular to use the naïve test for testing purpose, as it does not require additional validation or replicate data.

When the primary interest is in the $G \times E$, we have shown in Section 2 that naïve tests are no longer valid. To correct for the inflation in type I error, one needs validation or replication data to provide information on the measurement error structure. In this paper, we focus on regression calibration–based tests, which are easy to implement and shown to be optimal in terms of efficiency for hypothesis testing (Tosteson and Tsiatis, 1988; Stefanski and Carroll, 1990), in the context of testing main effects only. In this section, we consider regression calibration–based tests, with a focus on testing $G \times E$ instead of the main effect of E.

3.1 *Regression calibration (RC)*

Regression calibration is an effective approach that has been widely used for measurement error correction(Carroll et al., 2006; Carroll and Stefanski, 1990). In the context of $G \times E$ analysis, both E and GE are error-prone variables and need to be calibrated.

In the calibration step, we aim to estimate calibration functions for E and GE based on validation or replicates data. As $\mathrm{E}(GE|X, G, Z) = G\{\,\mathrm{E}(E|X, G, Z)\,\} = G\{\,m(X, G, Z)\,\}$, one only needs to calibrate E. In the presence of $G \times E$, we consider the following two possible approaches to implement regression calibration (RC):

$RC0$: excluding the interaction term in calibration, i.e.,

$$m_0(X, G, Z) = \lambda'_{10} + \lambda'_{11}G + \lambda'_{12}X + \lambda'_{13}Z.$$

$RC1$: including the interaction terms in calibration.

$$m_1(X, G, Z) = \lambda_{10} + \lambda_{11}G + \lambda_{12}X + \lambda_{13}GX + \lambda_{14}Z.$$

Note that $RC0$ is the standard approach when the primary interest is testing the main effect. However, we will show that this version of RC does not correct the bias for the interaction term, and thus will lead to an invalid hypothesis testing procedure for $G \times E$. On the other hand, $RC1$

includes the GX term in the calibration model explicitly, and leads to consistent estimation and valid testing procedure for $G \times E$.

The calibration model can be estimated based on either validation or replicates data. In an internal or external validation study, the true exposure E is recorded along with X, G and Z, so one can fit a linear regression model and estimate λ parameters. On the other hand, if replicated measurements are available in the full sample or a subsample, the data recorded are $(X_1, X_2, \cdots, X_J, G, Z)$. Let $\overline{X}$ denote the average of J measurements, i.e., $\overline{X} = \sum_{j=1}^{J} X_j/J$. Based on classical measurement error model, one can first estimate σ_E^2 and σ_ϵ^2 using analysis of variance, and calculate $\mathrm{var}(\overline{X}) = \sigma_E^2 + \sigma_\epsilon^2/J$. Let $Q = (\overline{X}, G, G\overline{X}, Z)^T$, then the calibration equation for E based on replicated measurements can be obtained by

$$m(\overline{X}, G, G\overline{X}, Z) = \mathrm{E}(X) + \Sigma_{XQ}\, \Sigma_{QQ}^{-1} \{\, Q - \mathrm{E}(Q) \,\},$$

where $\Sigma_{XQ} = \mathrm{cov}(X, Q)$, $\Sigma_{QQ} = \mathrm{cov}(Q, Q)$, $\mathrm{E}(X)$ and $\mathrm{E}(Q)$ are estimated by the method of moments (sample mean, variances or covariances). More details of estimating the calibration function parameters based on a validation study or replicate data are discussed in Section 4.4 of Carroll et al. (2006).

In the estimation step, E and GE are replaced by their calibrated estimates. The working model for this approach is

$$h(\mu) = \eta_0 + \eta_1 G + \eta_2\, \widehat{m}(X, G, Z) + \eta_3\, G\, \widehat{m}(X, G, Z) + \eta_4 Z.$$

To test the hypothesis $H_0 : \eta_3 = 0$, one can apply a Wald or score test for this working model with its variance estimated by the sandwich estimator (Carroll et al., 2006). Alternatively, one can use re-sampling procedures, such as the bootstrap, to obtain standard errors and confidence intervals. Based on our experience, we recommend the bootstrap approach, especially with small to medium sample sizes.

Proposition 3. Let $\widehat{\eta}_3^{RC0}$ and $\widehat{\eta}_3^{RC1}$ denote the estimated η_3 from the two regression calibration approaches, respectively, and let T_{RC0} and T_{RC1} denote the corresponding Wald test statistics.

Assuming the measurement error model (1) with identity link for continuous outcome and correctly specified regression calibration model, $\widehat{\eta}_3^{RC1} \xrightarrow{P} \beta_3$, while $\widehat{\eta}_3^{RC0} \xrightarrow{P} \lambda_{13}\beta_2 + \lambda_{23}\beta_3$ is generally biased for β_3. Under $H_0 : \beta_3 = 0$, T_{RC1} maintains the correct type I error, while T_{RC0} does not unless $\lambda_{13} = 0$ or $\beta_2 = 0$.

Proposition 3 demonstrates that it is crucial to include the interaction term GX in the calibration equation when $G \times E$ is of primary interest. Similar to the situation for measurement error for main effects, regression calibration estimators are consistent for linear models if the calibration model is correctly specified. In generalized linear models, consistency does not hold exactly in general. However, under rare disease and when the magnitude of association is small to medium, regression calibration provides approximately unbiased estimators (Rosner et al., 1989).

Remark 3: an alternative method of moment approach. For linear models, Proposition 2 provides an analytic form of the bias term and thus an alternative method of moment approach for bias correction. If one can estimate the magnitude of bias, test statistics can be constructed directly by correcting the bias explicitly. Using validation or replicate data, one could first fit the regression calibration equation $m_1(X, G, Z)$ and obtain an estimate for λ_{13}. An estimate for β_2 can be estimated by $\widehat{\gamma}_2/\widehat{\lambda}_0$ using data from the subgroup $G = 0$. One can then estimate the bias term by $\widehat{\lambda}_{13}\widehat{\beta}_2$. However, this approach works for linear models only, and we find its performance similar to T_{RC1} empirically. Thus, we focus on the RC approach in this paper.

3.2 *Power considerations*

In this section, we investigate power performances of several test statistics for $G \times E$, including the ideal, naïve and RC–based tests.

Proposition 4. Assuming measurement error model (1) with continuous outcome, the following results hold.

(a) Under both null and alternative hypotheses, $\sqrt{n}(\widehat{\beta}_3 - \beta_3) \xrightarrow{D} N(0, \sigma^2_\beta)$, $\sqrt{n}(\widehat{\gamma}_3 - \lambda_{13}\beta_2 - \lambda_{23}\beta_3) \xrightarrow{D} N(0, \sigma^2_\gamma)$, where $\sigma^2_\beta = \lim_{n\to\infty}\{\, n\sigma^2(D_E^T D_E)^{-1}\,\}_{3,3}$, $\sigma^2_\gamma = \lim_{n\to\infty}\{\, n\sigma^2(D_X^T D_X)^{-1}\,\}_{3,3}$.

(b) Under local alternatives $H_{an} : \beta_3 = \delta/\sqrt{n}$, the ideal test statistic T_{ideal} converges to χ^2_{1,τ_1}, with non-centrality parameter $\tau_1 = \delta^2/\sigma^2_\beta$. The naïve test statistic T_{naive} converges to χ^2_{1,τ_2}, where $\tau_2 = (\sqrt{n}\lambda_{13}\beta_2 + \lambda_{23}\delta)^2/\sigma^2_\beta$. If one knows the bias term $\lambda_{13}\beta_2$ under H_0 and corrects for it explicitly, the corresponding test statistic $T_{naive,bc}$ converges to $\chi^2_{1,\tau_2'}$, where $\tau_2' = \lambda^2_{23}\delta^2/\sigma^2_\beta$.

(c) Assuming the RC function is correctly specified, $\sqrt{n}(\widetilde{\eta}_3 - \beta_3) \xrightarrow{D} N(0, \sigma^2_{\beta,rc})$. Under local alternatives $H_{an} : \beta_3 = \delta/\sqrt{n}$, the corresponding test statistic T_{RC1} converges to χ^2_{1,τ_3}, with non-centrality parameter $\tau_3 = \delta^2/\sigma^2_{\beta,rc}$.

We illustrate power comparisons using the simple scenario with continuous outcomes and binary genotype (Figure 2). Under the null $\beta_3 = 0$, the rejection rate of the naïve test (dotted line) is around 15%, inflated from its size $\alpha = 0.05$, while the RC–based test (dashed line) maintains a valid size of around 5%. In this setting, the power of the naïve test is not meaningful since its type I error is not controlled properly. Nevertheless, when $\beta_3 > 0$, the RC–based test often has higher power than the naïve test; when $\beta_3 < 0$, the naïve test is superficially more powerful, but such "power gain" is not real and is due to its inflated type I error. Thus, the naïve test for $G \times E$ can yield both false positive and false negative $G \times E$ findings. The ideal test is substantially more powerful than the RC–based test, illustrating that potential power gain can be obtained with accurate exposure measurements. Thus, improving measurement accuracy of the assessment of E is an effective way to improve power for testing $G \times E$, in addition to increasing sample size (Wong et al., 2003).

3.3 *Extension to testing interactions between E and a set of genes*

In this paper, we illustrated the impact of measurement on $G \times E$ under a simple setting with a single gene or SNP. However, our results can be extended to more general settings with gene sets or genome wide $G \times E$ analysis. For example, Lin et al. (2013) proposed gene set-based $G \times E$ testing, while Kooperberg and LeBlanc (2008), Dai et al. (2012) and Hsu et al. (2012)

proposed several two-stage testing approaches for genome wide $G \times E$ scan. Under the settings of multiple G, following our arguments, it is straightforward to show that ignoring measurement error leads to incorrect type I errors and invalid conclusions for $G \times E$. The RC–based test can be extended relatively easily to either gene set-based tests or two stage screening based tests, with the modification of replacing mis-measured exposure X by its regression calibration estimate $m(X, G, Z)$ as outlined in Section 3.1. The statistical properties (type I error and power) of these tests are analogous to those discussed in Sections 2 and 3. In fact, generalizability to more complex settings is another reason that we prefer the regression calibration approach for measurement error correction in $G \times E$ analysis.

4. Numerical results

4.1 *Simulations*

To evaluate finite sample performances, we conducted simulation studies under both linear models with continuous outcomes and generalized linear models with binary outcomes.

For continuous outcomes, the data were simulated under model (1) with identity link and Z absent. We generated E from the standard Gaussian distribution, G as a binary genotype with $p = \Pr(G = 1) = 0.05, 0.15$ and 0.4, and possible G-E correlation of $\rho = 0, 0.4$ and0.7. True parameter values were $(\beta_0, \beta_1, \beta_2, \sigma^2) = (1, 1, 1, 1)$, with $G \times E$ parameter β_3 taking values $(0, -0.4, -0.2, 0.2, 0.4)$ under the null and alternative hypotheses. The standard deviations of measurement error σ_ϵ were $(0.3, 0.8, 1.5, 2.5)$, corresponding to reliability coefficients of $(0.92, 0.61, 0.31, 0.14)$, respectively. Simulations were conducted under various sample sizes, and we reported results with $n = 1000$, a 30% internal sample as validation data, and 2000 repetitions in each simulation.

Table 1 shows empirical type I error rates from simulations. Under H_0, type I error rates for the naïve test are close to its nominal level of 0.05 in most cases, but are clearly inflated for several settings, especially when ρ is high and σ_ϵ is close to 1 (e.g., $\sigma_\epsilon = 0.8, \rho = 0.7$ or $\sigma_\epsilon =$

$1.5, \rho = 0.7$). As demonstrated in Section 2.2, the magnitude of inflation depends on the amount of measurement error, G-E correlation, genotype frequency and sample size. Fixing ρ and p, we confirm the observed phenomenon in Figure 1, i.e., the amount of inflation in type I error initially increases with σ_ϵ, reaches its maximum in the middle range and then decreases with σ_ϵ afterwards. In contrast, the RC–based test T_{RC1} appears to maintain the correct type I error rates across all settings. Note that T_{RC0} performs similarly to T_{naive} and has invalid type I errors under several settings, due to the fact that it does not properly incorporate the interaction term in the calibration step (Proposition 3).

Table 2 displays empirical power under alternatives $\beta_3 = -0.2$ and $\beta_3 = 0.2$. We make the following observations regarding power. First, these results verify that the naïve test can lead to both missed $G \times E$ signals and spurious $G \times E$ findings, as was illustrated by Figure 2 in the previous section. For example, we consider the scenario $\sigma_\epsilon = 0.8, \rho = 0.7, p = 0.15$, the type I error for T_{naive} is severely inflated to 0.140. When $\beta_3 = -0.2$, T_{RC1} has substantially higher power than T_{naive}. On the other hand, when $\beta_3 = 0.2$, T_{naive} appears to have even higher power than T_{ideal}, but the seemingly high power of T_{naive} is due to inflated type I error and likely lead to false positives. Second, when σ_ϵ is very small or large (0.3), T_{naive} is approximately valid in terms of type I error. In these cases, power performances of T_{naive}, T_{RC0} and T_{RC1} are similar to each other. Based on these results, we would always recommend the proposed RC-based test, as it is valid in terms of type I error under all scenarios and it has at least comparable power to the naïve test even when naïve test is valid. Third, these results also shed light on the power loss due to imprecise measurements. T_{RC1} always maintains correct type I errors, and it demonstrates power loss compared to T_{ideal}. Not surprisingly, larger measurement error leads to more substantial power loss. This implies that improving measurement accuracy of E will lead to improved power to detect $G \times E$.

We also conducted simulation studies under generalized linear models for binary data. E, G and

X were generated similarly as above. True parameter values were $(\beta_0, \beta_1, \beta_2) = (-4, 1, 2)$, with $G \times E$ parameter β_3 taking values $(0, -1, -0.5, 0.5, 1)$. Under case-control sampling, $n = 1000$ cases and controls were simulated, and a random sample of 30% subjects were chosen as internal validation data. The empirical type I error rates are shown in the last four columns of Table 1. The proposed RC approach maintains the correct type I error. The naïve test has inflated type I errors for several settings, especially when the G-E correlation is high and σ_ϵ is close to 1 (e.g., $\sigma_\epsilon = 0.8, \rho = 0.7$ or $\sigma_\epsilon = 1.5, \rho = 0.7$), although the magnitude of inflation is generally smaller than in linear models. The power comparison among various tests shows similar patterns to linear models, and thus the results are omitted.

[Table 1 about here.]

[Table 2 about here.]

4.2 *Application*

We apply the proposed methods for $G \times E$ analysis to a Women's Health Initiative (WHI) study in the Genomics and Randomized Trials Network (GARNET). It is a genome wide association study (GWAS) on hormone treatment and cardiovascular disease/metabolic outcomes involving approximately $5,000$ subjects, chosen from the WHI hormone therapy trial cohorts in a nested case-conrol study (Rossouw et al., 2008; Women's Health Initiative Study Group et al., 1998; Prentice and Anderson, 2008).

In this analysis, we investigated potential gene–blood pressure interactions on the risk of coronary heart disease (CHD). Our sample included 520 CHD cases and 2128 controls. Systolic blood pressure (SBP) was measured at baseline and after 1 year in the study (year 1), denoted as SBP_1 and SBP_2, respectively. We considered the long-term average SBP as the true environmental exposure, and viewed measurements at two visits as replicates that follow a classic measurement error model. We conducted genetic association analysis, using single nucleotide polymorphisms (SNPs) identified to be strongly associated with risk of CHD. For illustration we focused on the

8 genotyped SNPs known to be associated with CHD based on the National Human Genome Research Institute (NHGRI) GWAS Catalog (Welter et al., 2014; Schunkert et al., 2011; Davies et al., 2012). For each SNP, a logistic regression was fitted including the SNP genotype, SBP and SNP-SBP interaction, adjusting for age, body mass index and four leading principal components to control for potential population structure. We conducted several tests for $G \times E$, including naïve tests using SBP_1, SBP_2 and $\overline{SBP} = (SBP_1 + SBP_2)/2$, as well as the RC–based test T_{RC1}.

Based on replicates data for SBP, we estimated variances of the true SBP and its measurement error in each replicate to be 171 and 137, respectively. The reliability coefficients for SBP_1 (SBP_2) and $\overline{SBP}$ were 0.55 and 0.71, respectively, verifying modest improvement in accuracy by taking the average of two measurements. Table 3 shows the p values from these tests on the 8 SNPs. Based on the RC test, SNP $rs6922269$ is significant at level $\alpha = 0.05$. The naïve tests using SBP_1,SBP_2 or $\overline{SBP}$ yield inconsistent results with SBP_2 showing some evidence of $G \times E$ (p value = 0.039). A close examination of attenuation factors shows that they are quite different among three genotypic groups. There are two other SNPs (rs9349379 and $rs16893526$) that also show inconsistent p-values between our approach and naïve tests, with the similar observation of different attenuation factors in the genotype subtypes. Finally we note that the average SBP is more accurate than SBP measurement at a single visit, and p values from the naïve test using average SBP are closer to the RC–based test.

The identified SNP $rs6922269$ with significant $G \times E$ is in the intron region of the methylene-tetrahydrofolate dehydrogenase 1-like (MTHFD1L) gene. Several studies have shown that polymorphisms in MTHFD1L, including rs6922269 as its lead polymorphism, are associated with risk for CHD (Palmer et al., 2014). Coronary Artery Disease (C4D) Genetics Consortium et al. (2011) studied potential interaction between this SNP and hypertension. They reported slightly stronger SNP-CHD association among hypertensive subjects versus non-hypertensive subject, but the interaction effect was not statistically significant. In our analysis, blood pressure is treated as a

continuous variable rather than an binary hypertension status and the proposed measurement error correction method is used instead of the naïve test, both of which lead to potential power gains to detect interactions.

[Table 3 about here.]

5. Discussion

In this paper, we consider statistical issues in testing $G \times E$ in the presence of environmental measurement error. Specifically, we assume a classical measurement error model for E with non-differential measurement error. We demonstrate that maximum likelihood estimates for $G \times E$ are generally biased under the null hypothesis of no interaction if one ignores measurement error, and as a result the naïve test has inflated type I error. This is contrary to conjectures in the literature. Analytic forms of the bias term are obtained, and consequences of ignoring measurement error in $G \times E$ analysis are discussed in terms of type I error and power. To properly account for measurement error, we propose RC–based testing procedures when validation or replicate data are available. Through theoretical derivations, simulation studies, and an application to a genetic study of coronary heart disease, we demonstrate that naïve tests can yield both false positive and false negative $G \times E$ findings and that RC–based tests maintain correct type I errors as long as the regression calibration model is correctly specified.

Our results shed light on implications of study design and data collection on $G \times E$ analysis. The effort to understand measurement error and improve measurement accuracy of E has been limited in many genetic studies, due to various reasons such as budget constraints. We have shown that simply applying the naïve test can lead to incorrect conclusions. In order to use the corrected RC–based test, one needs an internal or external validation or replication sub-study to help understand measurement error properties of E. If it is possible to adopt more accurate measurements (e.g., by repeated measurements or alternative measurement devices), the magnitude of bias from the

naïve analysis will be reduced and the power to detect $G \times E$ will be improved. In fact, based on power considerations only, Wong et al. (2003) argued that "smaller studies with repeated and more precise measurement of the exposure and outcome will be as powerful as studies even 20 times bigger, which necessarily employ less precise measures because of their size."

Although we focus on testing $G \times E$ with measurement error in E in this paper, statistical issues discussed here also apply to measurement error in G, and more generally testing interactions between two environmental variables where one is prone to measurement error. An example is using imputed genotypes for $G \times E$ analysis, when genotype data is missing. Another example is to investigate potential interaction between body mass index (BMI) and blood pressure for CHD risk when blood pressure is measured with error but BMI is fairly accurate. In both examples, ignoring measurement error can lead to inflated type I errors and the RC–based test provides a valid approach to testing interactions.

There are several related topics that need future research. First, we consider classic measurement error models in this paper, but in practice, environmental exposures may be subject to systematic bias and complex measurement error structure. Examples include questionnaires for diet and physical activity, which are known to suffer from systematic under- or over-reporting of energy intake and substantial measurement error more generally (Prentice et al., 2011). It will be interesting to extend the proposed methods to more complicated measurement error settings. Second, one could consider alternative approaches to correct for measurement error other than regression calibration, such as conditional or corrected score methods. Third, the induced measurement error in the interaction term is heteroscedastic and refinement of regression calibration methods can potentially further improve power.

[Figure 1 about here.]

[Figure 2 about here.]

ACKNOWLEDGEMENTS

This work was partially supported by grants R21ES022332, R01AG014358, R01HG6124, R01HL114901 and P01CA53996 from the National Institutes of Health. The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts, HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C and HHSN271201100004C.

REFERENCES

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC Press.

Carroll, R. and Stefanski, L. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association* pages 652–663.

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

Cook, J. and Stefansk, L. (1994). Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association* **89,**.

Coronary Artery Disease (C4D) Genetics Consortium et al. (2011). A genome-wide association study in europeans and south asians identifies five new loci for coronary artery disease. *Nature Genetics* **43,** 339–344.

Dai, J. Y., Kooperberg, C., Leblanc, M., and Prentice, R. L. (2012). Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika* **99,** 929–944.

Davies, R. W., Wells, G. A., Stewart, A. F., Erdmann, J., Shah, S. H., Ferguson, J. F., Hall, A. S., Anand, S. S., Burnett, M. S., Epstein, S. E., et al. (2012). A genome-wide association study for coronary artery disease identifies a novel susceptibility locus in the major histocompatibility

complex. *Circulation: Cardiovascular Genetics* **5,** 217–225.

Goldstein, D. B., Allen, A., Keebler, J., Margulies, E. H., Petrou, S., Petrovski, S., and Sunyaev, S. (2013). Sequencing studies in human genetics: design and interpretation. *Nature Reviews Genetics* **14,** 460–470.

Greenwood, D., Gilthorpe, M., and Cade, J. (2006). The impact of imprecisely measured covariates on estimating gene-environment interactions. *BMC medical research methodology* **6,** 21.

Hsu, L., Jiao, S., Dai, J. Y., Hutter, C., Peters, U., and Kooperberg, C. (2012). Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genetic Epidemiology* **36,** 183–194.

Khoury, M. J. and Flanders, W. D. (1996). Nontraditional epidemiologic approaches in the analysis of gene environment interaction: Case-control studies with no controls! *American Journal of Epidemiology* **144,** 207–213.

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

Kraft, P., Yen, Y., Stram, D., Morrison, J., and Gauderman, W. (2007). Exploiting gene-environment interaction to detect genetic associations. *Human Heredity* **63,** 111–119.

Lin, X., Lee, S., Christiani, D. C., and Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* **14,** 667–681.

McCullagh, P. and Nelder, J. (1989). Generalized linear models. *London: Chapman & Hall* .

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

Palmer, B. R., Slow, S., Ellis, K. L., Pilbrow, A. P., Skelton, L., Frampton, C. M., Palmer, S. C., Troughton, R. W., Yandle, T. G., Doughty, R. N., et al. (2014). Genetic polymorphism rs6922269 in the mthfd1l gene is associated with survival and baseline active vitamin b12

levels in post-acute coronary syndromes patients. *PloS one* **9,** e89029.

Park, J., Wacholder, S., Gail, M., Peters, U., Jacobs, K., Chanock, S., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics* **42,** 570–575.

Prentice, R. L. and Anderson, G. L. (2008). The women's health initiative: lessons learned. *Annu. Rev. Public Health* **29,** 131–150.

Prentice, R. L., Mossavar-Rahmani, Y., Huang, Y., Van Horn, L., Beresford, S. A., Caan, B., Tinker, L., Schoeller, D., Bingham, S., Eaton, C. B., et al. (2011). Evaluation and comparison of food records, recalls, and frequencies for energy and protein assessment by using recovery biomarkers. *American journal of epidemiology* **174,** 591–603.

Rosner, B., Willett, W., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in medicine* **8,** 1051–1069.

Rossouw, J. E., Cushman, M., Greenland, P., Lloyd-Jones, D. M., Bray, P., Kooperberg, C., Pettinger, M., Robinson, J., Hendrix, S., and Hsia, J. (2008). Inflammatory, lipid, thrombotic, and genetic markers of coronary heart disease risk in the women's health initiative trials of hormone therapy. *Archives of internal medicine* **168,** 2245–2253.

Schunkert, H., König, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., Preuss, M., Stewart, A. F., Barbalic, M., Gieger, C., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics* **43,** 333–338.

Stefanski, L. and Carroll, R. (1990). Score tests in generalized linear measurement error models. *Journal of the Royal Statistical Society. Series B (Methodological)* **52,** 345–359.

Thomas, D. (2010). Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annual Review of Public Health* **31,** 21. PMCID: PMC2847610.

Tosteson, T. D. and Tsiatis, A. A. (1988). The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates. *Biometrika* **75,** 507.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42,** D1001–D1006.

Williamson, E., Ponsonby, A., Carlin, J., and Dwyer, T. (2010). Effect of including environmental data in investigations of gene-disease associations in the presence of qualitative interactions. *Genetic Epidemiology* **34,** 522–560.

Women's Health Initiative Study Group et al. (1998). Design of the Women's Health Initiative clinical trial and observational study-examples from the Women's Health Initiative. *Controlled Clinical Trials* **19,** 61–109.

Wong, M., Day, N., Luan, J., Chan, K., and Wareham, N. (2003). The detection of gene–environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *International Journal of Epidemiology* **32,** 51.

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* **109,** 1193–1198.

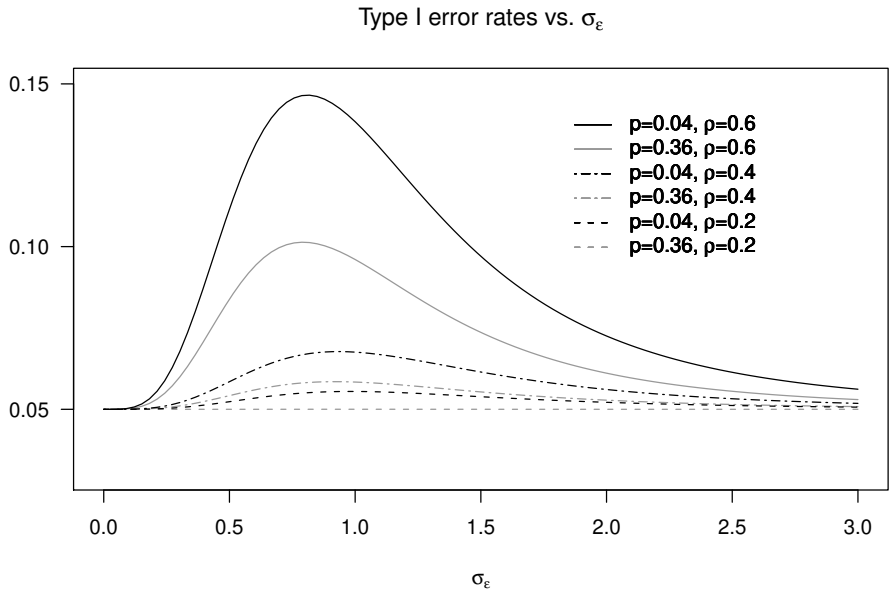

Figure 1. Type I error rates of the naïve test for $G \times E$ interaction versus the magnitude of measurement error σ_ϵ. The figure shows that the type I error rate α can be inflated (true $\alpha = 0.05$) when E is measured with error. The inflation in α depends on the magnitude of measurement error σ_ϵ, correlation ρ between G and E, and genotype frequency p.

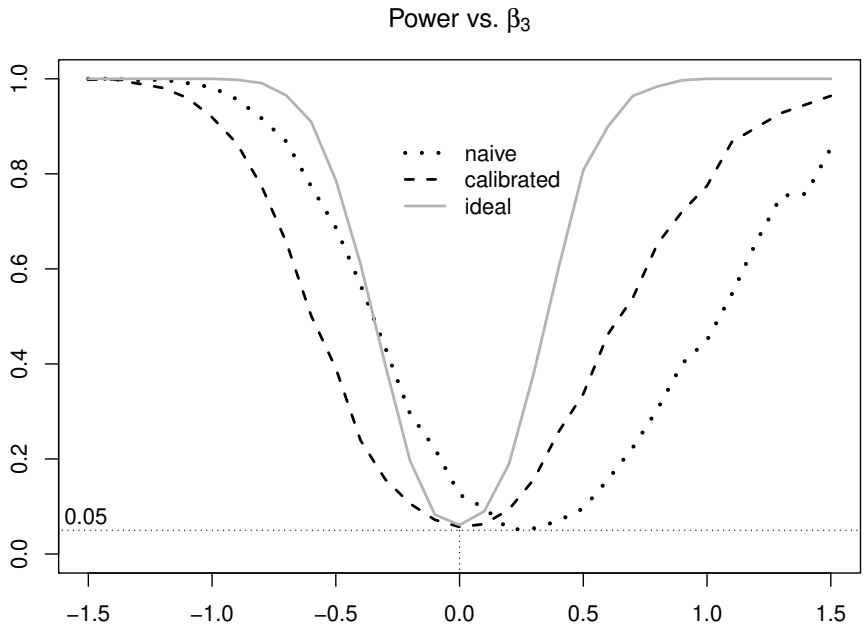

Figure 2. Power curves of the naïve test, regression calibration-based test and the ideal test (assuming E is known). The horizonal axis is the magnitude of interaction, β_3.

σ_ϵ	ρ	p	linear models				generalized linear models			
			ideal	naïve	RC0	RC1	ideal	naïve	RC0	RC1
0.3	0.0	0.05	0.043	0.051	0.051	0.051	0.027	0.027	0.027	0.060
0.3	0.0	0.15	0.061	0.062	0.062	0.056	0.045	0.050	0.050	0.049
0.3	0.0	0.40	0.044	0.048	0.048	0.048	0.047	0.054	0.054	0.055
0.3	0.4	0.05	0.052	0.047	0.047	0.050	0.024	0.030	0.030	0.057
0.3	0.4	0.15	0.051	0.052	0.052	0.052	0.041	0.040	0.040	0.048
0.3	0.4	0.40	0.065	0.060	0.060	0.060	0.050	0.053	0.053	0.055
0.3	0.7	0.05	0.043	0.053	0.053	0.042	0.014	0.023	0.023	0.050
0.3	0.7	0.15	0.055	0.059	0.059	0.046	0.040	0.040	0.040	0.052
0.3	0.7	0.40	0.047	0.045	0.045	0.036	0.056	0.048	0.048	0.049
0.8	0.0	0.05	0.044	0.047	0.047	0.047	0.045	0.041	0.041	0.052
0.8	0.0	0.15	0.048	0.053	0.053	0.053	0.051	0.056	0.056	0.046
0.8	0.0	0.40	0.048	0.050	0.050	0.052	0.036	0.050	0.050	0.052
0.8	0.4	0.05	0.053	0.042	0.042	0.038	0.024	0.020	0.020	0.049
0.8	0.4	0.15	0.057	0.057	0.057	0.049	0.045	0.040	0.040	0.049
0.8	0.4	0.40	0.047	0.050	0.050	0.052	0.038	0.070	0.070	0.052
0.8	0.7	0.05	0.060	0.107	0.107	0.045	0.011	0.018	0.018	0.052
0.8	0.7	0.15	0.065	0.140	0.140	0.058	0.032	0.031	0.031	0.044
0.8	0.7	0.40	0.040	0.068	0.068	0.059	0.038	0.036	0.036	0.049
1.5	0.0	0.05	0.037	0.033	0.033	0.033	0.040	0.046	0.046	0.043
1.5	0.0	0.15	0.058	0.048	0.048	0.049	0.053	0.058	0.058	0.051
1.5	0.0	0.40	0.048	0.051	0.051	0.052	0.050	0.061	0.061	0.054
1.5	0.4	0.05	0.046	0.060	0.060	0.049	0.023	0.029	0.029	0.055
1.5	0.4	0.15	0.047	0.057	0.057	0.042	0.047	0.044	0.044	0.060
1.5	0.4	0.40	0.043	0.047	0.047	0.044	0.061	0.081	0.081	0.053
1.5	0.7	0.05	0.049	0.100	0.100	0.029	0.012	0.011	0.011	0.047
1.5	0.7	0.15	0.038	0.143	0.143	0.042	0.042	0.048	0.048	0.047
1.5	0.7	0.40	0.041	0.066	0.066	0.053	0.048	0.061	0.061	0.052
2.5	0.0	0.05	0.051	0.046	0.046	0.044	0.056	0.054	0.054	0.051
2.5	0.0	0.15	0.052	0.047	0.047	0.047	0.042	0.066	0.066	0.053
2.5	0.0	0.40	0.039	0.056	0.056	0.058	0.053	0.058	0.058	0.053
2.5	0.4	0.05	0.036	0.048	0.048	0.043	0.035	0.037	0.037	0.053
2.5	0.4	0.15	0.047	0.054	0.054	0.050	0.040	0.057	0.057	0.051
2.5	0.4	0.40	0.051	0.053	0.053	0.056	0.039	0.087	0.087	0.049
2.5	0.7	0.05	0.055	0.069	0.069	0.039	0.016	0.011	0.011	0.046
2.5	0.7	0.15	0.042	0.088	0.088	0.041	0.035	0.054	0.054	0.050
2.5	0.7	0.40	0.063	0.054	0.054	0.050	0.044	0.052	0.052	0.061

Table 1

Empirical type I error rates for testing $G \times E$ in the presence of measurement error. The parameters σ_ϵ, ρ and p are standard deviation of measurement error, correlation between G and E, and frequency of minor genotype, respectively. Columns 4-7 correspond to linear models for continuous outcomes, while Columns 8-11 correspond to generalized linear models for binary outcomes.

σ_ϵ	ρ	p	$\beta_3 = -0.2$				$\beta_3 = 0.2$			
			ideal	naïve	RC0	RC1	ideal	naïve	RC0	RC1
0.3	0.0	0.05	0.295	0.254	0.254	0.237	0.246	0.199	0.199	0.212
0.3	0.0	0.15	0.583	0.525	0.525	0.555	0.637	0.548	0.548	0.526
0.3	0.0	0.40	0.860	0.816	0.816	0.816	0.872	0.799	0.799	0.797
0.3	0.4	0.05	0.259	0.208	0.208	0.217	0.251	0.229	0.229	0.212
0.3	0.4	0.15	0.540	0.464	0.464	0.502	0.578	0.523	0.523	0.477
0.3	0.4	0.40	0.821	0.745	0.745	0.746	0.821	0.747	0.747	0.747
0.3	0.7	0.05	0.188	0.124	0.124	0.177	0.165	0.208	0.208	0.156
0.3	0.7	0.15	0.419	0.271	0.271	0.381	0.418	0.494	0.494	0.363
0.3	0.7	0.40	0.736	0.621	0.621	0.644	0.706	0.671	0.671	0.626
0.8	0.0	0.05	0.285	0.189	0.189	0.193	0.264	0.120	0.120	0.119
0.8	0.0	0.15	0.585	0.344	0.344	0.361	0.619	0.305	0.305	0.293
0.8	0.0	0.40	0.868	0.561	0.561	0.586	0.878	0.503	0.503	0.484
0.8	0.4	0.05	0.244	0.110	0.110	0.151	0.248	0.140	0.140	0.103
0.8	0.4	0.15	0.570	0.240	0.240	0.311	0.531	0.330	0.330	0.253
0.8	0.4	0.40	0.840	0.477	0.477	0.512	0.815	0.477	0.477	0.443
0.8	0.7	0.05	0.168	0.051	0.051	0.124	0.196	0.281	0.281	0.077
0.8	0.7	0.15	0.415	0.079	0.079	0.246	0.440	0.520	0.520	0.205
0.8	0.7	0.40	0.733	0.268	0.268	0.386	0.744	0.484	0.484	0.345
1.5	0.0	0.05	0.248	0.102	0.102	0.102	0.250	0.047	0.047	0.047
1.5	0.0	0.15	0.618	0.197	0.197	0.195	0.608	0.108	0.108	0.113
1.5	0.0	0.40	0.851	0.273	0.273	0.265	0.877	0.227	0.227	0.238
1.5	0.4	0.05	0.231	0.085	0.085	0.112	0.255	0.085	0.085	0.063
1.5	0.4	0.15	0.535	0.114	0.114	0.171	0.560	0.158	0.158	0.110
1.5	0.4	0.40	0.849	0.227	0.227	0.248	0.823	0.239	0.239	0.216
1.5	0.7	0.05	0.191	0.061	0.061	0.082	0.191	0.174	0.174	0.050
1.5	0.7	0.15	0.418	0.054	0.054	0.146	0.434	0.334	0.334	0.086
1.5	0.7	0.40	0.722	0.128	0.128	0.196	0.717	0.227	0.227	0.164
2.5	0.0	0.05	0.290	0.086	0.086	0.089	0.290	0.044	0.044	0.040
2.5	0.0	0.15	0.621	0.113	0.113	0.112	0.612	0.073	0.073	0.073
2.5	0.0	0.40	0.873	0.133	0.133	0.126	0.863	0.119	0.119	0.124
2.5	0.4	0.05	0.260	0.079	0.079	0.094	0.235	0.059	0.059	0.044
2.5	0.4	0.15	0.550	0.090	0.090	0.115	0.578	0.092	0.092	0.067
2.5	0.4	0.40	0.821	0.121	0.121	0.132	0.828	0.125	0.125	0.107
2.5	0.7	0.05	0.176	0.063	0.063	0.065	0.180	0.086	0.086	0.029
2.5	0.7	0.15	0.434	0.056	0.056	0.084	0.417	0.160	0.160	0.050
2.5	0.7	0.40	0.728	0.101	0.101	0.124	0.709	0.126	0.126	0.094

Table 2

Empirical power for testing $G \times E$ in the presence of measurement error. These results are based on linear models for continuous outcomes, under alternatives $\beta_3 = -0.2$ and $\beta_3 = 0.2$.

SNP	naïve test			$RC\text{-}test$	MAF	Attenuation factors		
	SBP_1	SBP_2	$\overline{SBP}$			$\lambda_{G=0}$	$\lambda_{G=1}$	$\lambda_{G=2}$
rs9349379	0.051	0.591	0.296	0.298	0.396	0.577	0.580	0.528
rs3869109	0.291	0.787	0.384	0.402	0.437	0.552	0.565	0.561
rs6905288	0.853	0.151	0.268	0.416	0.430	0.554	0.565	0.558
rs16893526	0.143	0.060	0.025	0.057	0.083	0.564	0.532	0.752
rs12190287	0.741	0.439	0.332	0.282	0.368	0.536	0.581	0.559
rs1332844	0.085	0.365	0.199	0.201	0.382	0.519	0.567	0.567
rs2048327	0.320	0.398	0.185	0.271	0.362	0.601	0.537	0.575
rs6922269	0.327	0.039	0.081	0.040	0.269	0.543	0.596	0.486

Table 3

Application to gene-blood pressure interaction for coronary heart disease in the WHI GARNET study. Columns 2-5 are p values for $G \times E$, based on logistic regression models adjusting for age, body mass index and the first four principal components to control for population structure. SBP_1, SBP_2 and $\overline{SBP}$ are systolic blood pressure levels in baseline, year 1, and their average, respectively. $RC\text{-}test$ is regression calibration-based test T_{RC1}. Attenuation factors are defined as $\lambda_{G=g} = \frac{\mathrm{var}(E|G=g)}{\mathrm{var}(E|G=g)+\sigma_\epsilon^2}$ for $g = 0, 1, 2$.