Published in final edited form as:

Genet Epidemiol. 2018 April ; 42(3): 250–264. doi:10.1002/gepi.22113.

A test for gene-environment interaction in the presence of measurement error in the environmental variable

Hugues Aschard^{1,2}, Donna Spiegelman^{2,3,4,5}, Vincent Laville¹, Pete Kraft^{2,3}, and Molin Wang^{2,3,5}

¹Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI), Institut Pasteur, Paris, France

²Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, USA

³Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, USA

⁴Departments of Nutrition and Global Health, Harvard T.H. Chan School of Public Health, Boston, USA

⁵Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115

Abstract

The identification of gene-environment interactions in relation to risk of human diseases has been challenging. One difficulty has been that measurement error in the exposure can lead to massive reductions in the power of the test, as well as in bias towards the null in the interaction effect estimates. Leveraging previous work on linear discriminant analysis, we develop a new test of interaction between genetic variants and a continuous exposure that mitigates these detrimental impacts of exposure measurement error in ExG testing by reversing the role of exposure and the diseases status in the fitted model, thus transforming the analysis to standard linear regression. Through simulation studies, we show that the proposed approach is valid in the presence of classical exposure measurement error as well as when there is correlation between the exposure and the genetic variant. Simulations also demonstrated that the reverse test has greater power compared to logistic regression. Finally, we confirmed that our approach eliminates bias from exposure measurement error in estimation. Computing times are reduced by as much as 5-fold in this new approach. For illustrative purposes, we applied the new approach to an ExGWAS study of interactions with alcohol and body mass index among 1145 cases with invasive breast cancer and 1142 controls from the Cancer Genetic Markers of Susceptibility study.

Keywords

measurement error; gene-environment interaction test; genome-wide association studies; normal discriminant analysis

Contact persons: Hugues Aschard, Tel: 1-617-820-0648, haschard@hsph.harvard.edu; Molin Wang, Tel: 1-617-432-1412, stmow@channing.harvard.edu.

The authors have no conflict of interest to declare.

Introduction

Through systematic experimental evaluations that have assessed the effect of genetic mutations over a range of environmental backgrounds, gene-environment (GxE) interactions have been found to be common in model organisms [Flint and Mackay 2009]. In humans, despite extensive studies that have searched for GxE interactions in large studies, almost no such effects have been identified and replicated. This persistent lack of discovery has been discussed in many review papers [Aschard, et al. 2012; Bookman, et al. 2011]. Besides there simply being no such effects in humans, which seems unlikely although possible, reasons that may have led to the failure to identify human GxE interactions that in fact exist include very small sizes of interaction effects, defined as the ratio of odds ratios, comparing individuals with and without the genetic variant, of <1.1 , the limited variance of some exposures in the study populations currently available, the potential presence of higher-order interactions involving multiple exposures, and the occurrence of time dependent events (e.g. duration, intensity, recency, and other timing aspects of the exposure [Thomas 1988]). Another possible explanation could be the existence of relatively large measurement error in the environmental variables which have thus far been studied, leading to large losses of power in the GxE tests and bias, likely towards the null, of the estimated ratios of odds ratio (*ROR*). In this paper, we address tests and estimation of GxE interactions with a continuous exposure in the presence of measurement error. Although this issue has been acknowledged in previous work [Boonstra, et al. 2016; Cheng 2007; Garcia-Closas, et al. 1999; Garcia-Closas, et al. 1998; Lindstrom, et al. 2009; Stenzel, et al. 2015; Tchetgen Tchetgen and Kraft 2011; Wong, et al. 2004; Zhang, et al. 2008], it is not accounted for in the standard methods for testing for the presence of GxE interactions. Exposure measurement error could explain, at least in part, the divergent findings between the human and model organism studies, since the latter feature well controlled and well measured exposures, compared to human studies where the measurement of most exposures is subject to considerable error.

Previous work using the linear discriminant function [Efron 1975] have been developed for increasing statistical power in tests and for reducing or eliminating bias in estimation of the marginal effect of predictors in the presence of measurement error [Armstrong, et al. 1989; Lyles, et al. 2009]. Here, we follow similar ideas to provide tests and estimators for the GxE interaction. In a standard genetic case-control association study of single nucleotide polymorphisms (SNP), this new approach reverses the role of the exposure and the outcome, treating the exposure as the dependent variable and the outcome as an independent variable, hence, its name, the ‘reverse test’, then performs a standard linear regression including an interaction term between the SNP and the case-control status indicator. An advantage of using the linear model for evaluating GxE interactions is that random exposure measurement error is absorbed by the residual term of the linear regression model and thus does not bias the interaction effect estimates in the absence of a measurement error correction. We first focus on methods for a binary gene variable and later extend the proposed test to scenarios where the gene variable is ordinal (0, 1, 2). We describe the theoretical basis of the proposed approach, and explore its advantages and limitations through an intensive simulation study that compares its performance to standard logistic regression across various scenarios likely to occur in practice. Under a range of realistic odds ratios for the genetic and environmental

main effects and their interaction, the simulation study indicates that the reverse test has improved power compared to logistic regression. In particular, we observed increasing power with increasing main environmental effect and with an increasing interaction effect. For illustrative purposes, we then apply the proposed approach in an ExGWAS case-control study including 1145 cases of invasive breast cancer and 1142 age-matched controls genotyped for 528,173 single nucleotide polymorphisms (SNPs). We considered two known risk factors for breast cancer, body mass index and alcohol consumption, and tested for interaction effects between each of the genotyped SNPs and these two risk factors.

Material and Methods

The proposed approach

In genetic association studies, interaction effects between genetic variants and continuous environmental exposures in relation to disease incidence or risk are commonly evaluated using unconditional logistic regression. In these logistic regression models, the disease status, D , is modelled as a function of the exposure, X , and the genetic variant, G , and interaction effects are evaluated as deviation from additivity on the log-odds scale. This model is as follows:

$$Pr(D=1|X, G) = \frac{e^{\alpha_0 + \alpha_1 X + \alpha_2 G + \alpha_3 XG}}{1 + e^{\alpha_0 + \alpha_1 X + \alpha_2 G + \alpha_3 XG}}, \quad (1)$$

where, for a binary genotype, $e^{\alpha_1} = OR(D|X, G=0)$ and $e^{\alpha_2} = OR(D|G, X=0)$ are the odds ratios (ORs) for disease in relation to a one unit increase in X while $G=0$, and for $G=1$ versus $G=0$ while $X=0$, respectively, and e^{α_3} , the parameter of interest in this paper, is the ratio of the ORs in X for $G=1$ versus $G=0$, henceforth referred to as the ROR .

As noted in the Introduction, when the environmental variable, X , is measured with error, the parameter estimates of this model are biased and power is lost. To address this challenge, we therefore propose a new strategy for testing GxE interactions in relation to a binary outcome, suitable for continuous environmental exposures. Specifically, for a binary genotype, we propose that the test for the GxE interaction be based upon the test of an interaction between G and D in a standard linear regression, that is, testing for $\beta_3 \neq 0$ in the following linear model:

$$E(X|G, D) = \beta_0 + \beta_1 D + \beta_2 G + \beta_3 G D. \quad (2)$$

We will refer to this test as the *reverse* test. The test statistic can be a Wald test, score test, or likelihood ratio test focused on β_3 . Assuming that the exposure follows a conditional normal distribution with variance σ^2 , such that

$$X|D, G \sim \mathcal{N}(\beta_0 + \beta_1 D + \beta_2 G + \beta_3 DG, \sigma^2), \quad (3)$$

we can align the parameters from our exposure-based model (equation (3)) with those of a standard logistic regression (equation (1)). In Appendix A, we show that the parameters of these two models (1) and (3) are linked as follows:

$$\beta_0 = \frac{2\sigma^2 \left(\log \left(\frac{p_{10}}{1-p_{10}} \right) - \alpha_0 \right) - \beta_1^2}{2\beta_1},$$

$$\beta_1 = \alpha_1 \sigma^2,$$

$$\beta_2 = \frac{2\sigma^2 \left(\log \left(\frac{p_{11}(1-p_{10})}{p_{10}(1-p_{11})} \right) - \alpha_2 \right) - \beta_3^2 - 2\beta_1\beta_3 - \beta_0\beta_3}{2(\beta_1 + \beta_3)},$$

$$\beta_3 = \alpha_3 \sigma^2, \quad (4)$$

where $p_{10} = \Pr(D=1|G=0)$ and $p_{11} = \Pr(D=1|G=1)$ are the marginal probabilities of the disease in non-carrier and carrier of the genetic variant, respectively. From equation (4), it is evident that the null hypothesis for the gene-environment interaction, $H_0: \alpha_3 = 0$, in the logistic regression model (1), can be evaluated through testing $H_0: \beta_3 = 0$ in the linear regression model given by equation (3). Also, α_3 can be estimated by $\widehat{\beta}_3 / \widehat{\sigma}^2$, which, unlike the iterative procedure for estimating $\widehat{\alpha}_3$ in the logistic regression model, has a closed form based on standard linear regression theory (see Appendix B). The constant variance assumption in (3) can be checked by plotting residuals or using formal tests of equal variances such as the Levene's test [Levene 1960] on the residuals. The normality assumption can be checked by the normal probability plot approach or by conducting formal tests such as the Shapiro-Wilk test [Shapiro and Wilk 1965] on the residuals. We present an analytic form for the ratio of the χ^2 test statistics from logistic regression over the inverse test in the supplementary note.

In the linear model (2), β_3 represents the mean difference in X between cases and controls for $G=1$ minus that for $G=0$, and thus we refer to it as the difference in mean differences (DMD); $\beta_3 = 0$ implies that the X - D association, measured by the mean difference in X between cases and controls, does not change with level of G . Therefore, even without conditional normality and the constant variance assumptions required for (3) to hold, $H_0: \beta_3 = 0$ still expresses the null hypothesis of no G-E interaction and β_3 , the DMD, measures the magnitude of the GxE interaction. In addition, in the linear model (2), we note that the null hypothesis $\beta_3 = 0$ is true when

$$\frac{\text{corr}(X, D|G=0) \sqrt{\text{var}(X|G=0)}}{\sqrt{\text{var}(D|G=0)}} = \frac{\text{corr}(X, D|G=1) \sqrt{\text{var}(X|G=1)}}{\sqrt{\text{var}(D|G=1)}}.$$
 If the $G-D$ association is weak or the disease is rare, $\text{var}(D|G=0) \approx \text{var}(D|G=1)$. Therefore, when

$\text{var}(X|G=0) \approx \text{var}(X|G=1)$, which will often be likely to be the case as, thus far, genome-wide association studies of environmental exposures have found $X-G$ associations of very small magnitudes [Bierut, et al. 2010; Tobacco and Genetics 2010], and either the $G-D$ association is weak or the disease is rare, which is also most likely to be the case, $\beta_3 \approx 0$ in the linear model (2) when $\text{corr}(X, D|G=0) \approx \text{corr}(X, D|G=1)$. There is no scientific reason to believe that either the logistic or the linear model should, in general, hold for quantifying the possible dependency of the $X-D$ association on G . Nevertheless, the notion of no G-E interaction can be captured by $\text{corr}(X, D|G=0) \approx \text{corr}(X, D|G=1)$ without invoking any further assumptions for the model relating X , D and G . Thus, as shown above, our new reverse test based on the linear model express this notion under mild assumptions likely to be encountered in practice.

If the exposure is log-normal, and a linear model for $\log(X)$ (instead of for X) as the dependent variable can be assumed, then the reverse test can be conducted in this linear model for $\log(X)$. If the assumption in (3) is satisfied for $\log(X)$ instead of for X , the hypothesis tested by the reverse test would be equivalent to that tested by the estimated interaction between $\log(X)$ and D in the logistic regression model.

The logistic regression model can be used to analyze data arising from both case-control and cohort studies. The development of the proposed method does not require any additional assumptions about the study design and thus applies to both case-control and cohort studies.

Impact of exposure measurement error

An advantage of using the linear model over the logistic regression for evaluating GxE interactions is that it is much simpler to deal with exposure measurement error in the linear model than the logistic model. We show below that under a fairly general class of measurement error models characterizing the relationship between X and X^* , where X^* is the mis-measured exposure, the error is absorbed by the residual term of the linear regression model and the hypothesis test for GxE interaction is valid.

Classical additive error case—Under the classical additive model in each of the $G=1$ and $G=0$ groups

$$X^* = X + e_G, \quad (5)$$

where $e_G \sim N(0, \tau_G^2)$, which is independent from X , and let $\rho_{X, X^*|G} = \frac{\text{Var}(X|G)}{\text{Var}(X|G) + \tau_G^2}$, which describes the magnitude of measurement error conditional on a level of G . We apply the regression calibration method [Carroll, et al. 2006; Rosner, et al. 1990] for logistic regression to each of the groups, $G=1$ and $G=0$, under the assumption that either (A) surrogacy, i.e. $\Pr(D=1|X^*, X, G) = \Pr(D=1|X, G)$, rare disease, and $X|G$ follows the normal distribution [Rosner, et al. 1989] or (B) surrogacy and small $\text{Var}(X|X^*, G)$ which is equal to $\text{Var}(X|G)(1-\rho_{X, X^*|G})$ when $X|G$ follows the normal distribution [Carroll, et al. 2006], to obtain

$$Pr(D=1|X^*, G) \approx \frac{e^{\alpha_0^* + \alpha_1^* X^* + \alpha_2^* G + \alpha_3^* X^* G}}{1 + e^{\alpha_0^* + \alpha_1^* X^* + \alpha_2^* G + \alpha_3^* X^* G}} \quad (6)$$

Applying the results of [Carroll, et al. 2006] and [Rosner, et al. 1989] to (6), we have that $\alpha_1^* \approx \alpha_1 \rho_{X, X^*|G=0}$, $\alpha_3^* \approx (\alpha_1 + \alpha_3) \rho_{X, X^*|G=1} - \alpha_1 \rho_{X, X^*|G=0}$. As long as $\rho_{X, X^*|G=0} = \rho_{X, X^*|G=1} = \rho$, which holds when X and G are independent and the measurement error variance, τ_G^2 , does not depend on G , $\alpha_3^* \approx \alpha_3 \rho$, showing that the test based on an assessment of the null hypothesis $\alpha_3^*=0$ would be valid, but the estimates for the coefficient of the interaction term would be attenuated.

Assuming that the error in model (5) is independent of D , G , as would typically hold in nested case-control studies with prospective exposure assessment, we have $E(X^*|D, G) = E(X + e|D, G) = E(X|D, G)$ and $\text{Var}(X^*|D, G) = \text{Var}(X|D, G) + \tau^2$. Note that the error being independent of D implies the surrogacy assumption mentioned above. It follows that, under the additional assumption of model (3), $X^*|D, G \sim N(\beta_0^* + \beta_1^* D + \beta_2^* G + \beta_3^* DG, \sigma^{*2})$, where $\sigma^{*2} = \sigma^2 + \tau^2$, $\beta_3^* = \beta_3$ and likewise for the other parameters of the conditional mean. Applying the result in (4) to the models based on X^* , we have

$Pr(D=1|X^*, G) = \frac{e^{\alpha_0^* + \alpha_1^* X^* + \alpha_2^* G + \alpha_3^* X^* G}}{1 + e^{\alpha_0^* + \alpha_1^* X^* + \alpha_2^* G + \alpha_3^* X^* G}}$, where $\alpha_1^* = \beta_1^* / \sigma^{*2}$ and $\alpha_3^* = \beta_3^* / \sigma^{*2}$. Since $\beta_3^* = \beta_3 = \alpha_3 \sigma^2$, we have $\alpha_3^* = \alpha_3 \sigma^2 / \sigma^{*2} = \alpha_3 \sigma^2 / (\sigma^2 + \tau^2)$, where now these relationships are exact rather than approximate as previously. Thus, the coefficient α_3^* , quantifying the $G \times X^*$ interaction with D in the logistic regression model, decreases toward 0 with increasing measurement error, while the coefficient, β_3^* quantifying the $D \times G$ interaction with E in the linear regression model is invariant with respect to exposure measurement error.

General linear measurement error—Next, we consider scenarios where the measurement error model is of a more general form

$$X|G, D, X^* = \gamma_{G,D} + \gamma_1 X^* + \gamma_2 DX^* + \gamma_3 GX^* + \gamma_4 DGX^* + e, \quad (7)$$

where $\gamma_{G,D}$ is a G, D -status-specific intercept, and e has zero mean and variance v^2 . Appendix C demonstrates that under this more general model, as long as $\gamma_2 = \gamma_3 = \gamma_4 = 0$, the test $H_0: \beta_3^* = 0$ is valid for testing $H_0: \beta_3 = 0$, similar to the situation shown previously for testing main effects of exposure-disease relationships when the surrogacy assumption holds.

[Lagakos 1988] In addition, $\beta_3^* = \beta_3 / \gamma_1$ and $\sigma^{*2} = (\sigma^2 - v^2) / \gamma_1^2$, while model (6), the

logistic regression model, still holds with $\alpha_3^* = \frac{\alpha_3 \gamma_1 \sigma^2}{\sigma^2 - v^2}$, which requires estimates of the variances from the models for $(X|D, G)$ and $(X|G, D, X^*)$ in addition to an estimate of γ_1 , which is the only parameter needed for bias correction of β_3^* .

Extension to ordinal genetic variants

For ordinal $G = 0, 1, 2$, as showed in Appendix D, the linear model corresponding to model (1) is now

$$X|G, D \sim \mathcal{N} + \left(\beta_0 + \beta_1 D + \beta_2 G + \beta_3 DG + \beta_4 G^2, \sigma^2 \right), \quad (8)$$

where now a quadratic term, G^2 , must be included in the linear regression in addition to G , D and their interaction, which are in our previous model (3) for the binary genetic variant. In Appendix D, we show that we can align the parameters from the exposure-based model in (8) with those of a standard logistic regression in (1), and will have $\beta_3 = \alpha_3 \sigma^2$. Therefore, the reverse test for testing $H_0: \beta_3 = 0$ in the linear model (8) is equivalent to testing $H_0: \alpha_3 = 0$ in the logistic regression (1) when the genetic variant is ordinal (0, 1, 2). It is also straightforward to extend our results for the impact of exposure measurement error derived under binary genetic variance to this ordinal genetic variant scenario.

Simulation study

We considered two different data generation models, referring as the *Linear model* and the *Logit model*. In the *Linear model*, we generate data based on (3), where the exposure is normally distributed conditional on G and D . In the *Logit model*, we used a more classic simulation framework where we generated D conditional on E and G using logistic regression.

Linear model—We first generated binary G by fixing the probability of the genetic variant, $p_G = \Pr(G = 1)$. To generate D conditional on G , we fixed the baseline risk of the disease, $p_{10} = \Pr(D = 1|G = 0)$ and the marginal genetic effect $\text{OR}(D|G)$. We then calculated

$p_{11} = \Pr(D = 1|G = 1)$ using $p_{11} = \frac{\text{OR}(D|G)}{(1/p_{10}) - 1 + \text{OR}(D|G)}$. We chose values of p_{10} such that the baseline disease frequency ranged between 1% to 30%. We then generated X conditional on (G, D) according to (3), where we defined β_2 for a range of correlations between G and

$$X, \text{ with } \beta_2 = \frac{r \sqrt{\text{Var}(X|G, D=0)}}{(1 - r^2) \Pr(G=1|D=0) (1 - \Pr(G=1|D=0))},$$

$\Pr(G=1|D=0) = \frac{(1 - p_{11}) p_G}{(1 - p_{11}) p_G + (1 - p_{10}) (1 - p_G)}$ and r denotes the correlation coefficient, $\text{cor}(G, X|D=0)$. By setting σ^2 in formula (3) to 1 and β_0 to 0, β_1 and β_3 equal α_1 and α_3 , respectively. We then specified β_1 and β_3 over a realistic range of parameter values so that $\text{OR}(D|X, G=0) \in [3, 10]$, and $\text{ROR} \in [2, 10]$, where $\text{OR}(D|X, G=g)$ is the odds ratio for

$D = 1$ for a 1-unit increase in X when $G = g$ and ROR is $\frac{\text{OR}(D|X, G=1)}{\text{OR}(D|X, G=0)}$; this was achieved by noting that $\beta_1 = \alpha_1 = \log(\text{OR}(D|X, G=0))$ and $\beta_3 = \alpha_3 = \log(\text{ROR})$ as shown above.

Logit model—We first generated binary G with minor allele frequency (MAF) randomly sampled from a uniform distribution in $[0.05, 0.5]$. We then generated the exposure X from a normal distribution, independently of G . Finally, we generated the disease status D from a

logit model $\frac{1}{1+e^{-(\alpha_0+\alpha_1 X+\alpha_2 G+\alpha_3 XG)}}$. Note that the model assumption (3) typically does not hold under this simulation setup.

For both models, we simulated a series of 10,000 studies each including 100,000 individuals, over a range of parameter values for disease prevalence and the genetic, environmental and interaction effects. For each series, we then randomly sampled 1000 cases and 1000 controls from the full cohort to create a case-control study. When assuming no measurement error in the exposure, we obtained for each simulated study the estimated effects $\hat{\alpha}=(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3)$ by fitting the logistic regression (1) and $\hat{\beta}=(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ by fitting the linear regression model (2), using the generated X , D and G . When simulating measurement error, we generated a variable X^* defined by equation (5). We set the measurement error variance so that it explained between 25% to 80% of the variance of X^* , e.g. $\rho=0.8$. The estimates $\hat{\alpha}_*$ and $\hat{\beta}_*$

were then obtained using X^* instead of X in equations (1) and (2). The validity of the proposed approach for testing gene-environment interaction was evaluated by calculating the Type I error rates at a 5% p-value threshold. Under the alternative, we evaluated the relative efficiency of the two methods by comparing the ratio $\overline{\chi^2_{rev}}/\overline{\chi^2_{log}}$, where $\overline{\chi^2_{rev}}$ and $\overline{\chi^2_{log}}$ were the average χ^2 test statistics for the tests of $\alpha_3 = 0$ and $\beta_3 = 0$ over the 10,000 simulated studies for each simulation study design point, respectively. For presentational simplicity, we call this quantity the χ^2 ratio henceforth.

We also conducted a simulation study for biallelic variant with genotype values in $[0,1,2]$ (see supplementary note)

The CGEMS data

The Cancer Genetic Markers of Susceptibility (CGEMS) breast cancer study was conducted to identify individual SNPs associated with breast cancer risk. Full details of the study have been previously provided [Hunter, et al. 2007]. Briefly, the study consists of 1145 cases with invasive breast cancer and 1142 controls that were genotyped at 528,173 loci with the use of the Illumina HumanHap500 array. All participants were post-menopausal women of European ancestry recruited from the Nurses' Health Study [Colditz, et al. 1997]. Controls were matched to cases by year of birth and by postmenopausal hormone use. We extracted two breast cancer risk factors from the questionnaire data, body mass index (kg/m^2) (BMI) and alcohol use (gm/day). For the cases, we took the exposure at the age at diagnosis, and for controls, the exposure at the age of being matched to the case. We performed a genome-wide screening for interaction effects between each of the two exposures and the 458,918 SNPs with a minor allele frequency $>5\%$ using the software PLINK (<http://pengu.mgh.harvard.edu/~purcell/plink/>). As for potential confounders of the estimated effect of G and of the GxE interaction in the study, the only established confounding factor could be population stratification, which can be easily handled by adjusting for the principal component of the genotype data [Price, et al. 2006]. However, the CGEMS study comes from a homogenous population of European ancestry. Previous studies showed only negligible population structure in these data [Hunter, et al. 2007]. Standard analytic practice in gene-

environment interaction studies is to limit the number of adjusting factors (e.g. to age and study site [Barrdahl, et al. 2014; Nickels, et al. 2013]) while the impact of potential confounding factors of E would be explored only when significant associations are identified. Unfortunately, the identification of these associations rarely occurs. Therefore, we did not specifically address confounders of the estimated exposure or interaction effects themselves in this analysis, consistent with previous publications which did not either.

Results

Validity

We first calculated the type I error rates of the reverse test over the range of design parameters considered, with and without measurement error in X , assuming a genetic variant with a minor allele frequency of 0.3. Null models were simulated by fixing $\beta_3 = 0$ in the simulation procedure described in the previous section. We considered both cases where G and X were not correlated ($\beta_2 = 0$) and cases where G and X were correlated ($\beta_2 = 0.11$, corresponding to a correlation of 0.05 for a minor allele frequency of G of 0.3). We defined the parameters to obtain a marginal disease frequency ranging from 1% to 30%, and an odds ratio for the exposure in relation to the disease, $OR(D|X, G=0)$, between 1.0 and 2.0 when $corr(G, X|D=0)$ was assumed. As shown in Table 1, the reverse test had the Type I error rates close to 0.05 over the full range of scenarios studied, under both the data generation models, the Linear model and the Logit model. The simulation results for ordinal $G=0, 1, 2$ were presented in supplementary Tables S1-S2. As expected, the results were similar to the one for binary G .

Relative efficiency

We then evaluated the relative efficiency of the reverse test against the standard logistic regression approach for assessing $G \times E$ interactions by calculating the ratio of their respective χ^2 test statistics. We considered the case of a modest marginal genetic effect ($OR(D|G) = 1.2$) and a strong marginal genetic effect ($OR(D|G) = 2$), while increasing the odds ratio of the exposure, $OR(D|X, G=0)$ from 1.0 to 3.0, and the odds ratio of the interaction, ROR , from 1.0 to 2.0. As shown in Figure 1, which is for the linear data generation model, the ratio was larger than one in all cases considered, indicating better power for the new approach. The exposure effect, $OR(D|X, G=0)$ had the largest impact on the χ^2 ratio, being close to 1.0 when the exposure had no effect and above 2.0 when the exposure had an odds ratio greater than or equal to 3.0 and the ROR was 1.6. The power advantage of the reverse test also increased with increasing magnitude of the interaction effect, although the magnitude of this increase was smaller. Conversely, $OR(D|G)$ had only moderate impact on the relative efficiency of the reverse test. We observed a marked decrease in the relative advantage of the reverse test compared to the logistic regression approach as exposure measurement error increased. As showed in the extreme example from Figure 2, when $ROR = 2$, $OR(D|X, G=0) = 3$ and the data were generated under the linear model, the χ^2 ratio decreased from 2.4 in the absence of measurement error, to 1.2, when $\rho = 0.2$, a scenario of extreme exposure measurement. Similar results, in greater detail, are given in the Linear model section of Table 2. As showed in the Logit model section of Table 2, where data were generated under the logistic regression model, the ratio was less than 0.9

in only two out of 32 simulation scenarios when there was no measurement error in X (i.e., $\rho = 1$), and in 78% of the scenarios with $\rho = 1$, the χ^2 ratio was larger than 1. Similar results were observed in the presence of measurement error, with $\rho = 0.75, 0.5$ and 0.25 .

The simulation results for ordinal G are presented in supplementary Tables S3-S4. The results follow those for the binary G , where the χ^2 ratio ranged from 0.91 to 1.42 and was larger than 1 in 82% of the scenarios under the Logit model when there was no measurement error in X . Similar results were observed in the presence of measurement error, with $\rho = 0.75, 0.5$ and 0.25 .

Impact of measurement error on interaction effect estimates

Next, we evaluated the effect of classical exposure measurement error on the estimated ROR when data were generated under the linear model. We first calculated the relative bias of the estimated interaction effects from the reverse model and logistic regression across all the simulated datasets while increasing the measurement error in the exposure. We varied the extent of measurement error, $1 - \rho$, from 0% to 80% by increasing the variance of the error term, τ^2 (see equation (5)), and fixed β_3 at 0.405 ($e^{\beta_3} = 1.5$). Note that e^{β_3} equals the ROR , also expressed as e^{α_3} , when $\text{var}(X|G, D)$ is 1.0, as shown in the methods section. In Table 3, it is apparent that the relative biases of $\log(\widehat{ROR}) = \hat{\alpha}_3$ from the logistic regression and

$\log(\widehat{ROR}) = \frac{\hat{\beta}_3}{\hat{\sigma}^2}$ from the linear regression increase as ρ decreased. In contrast, the relative bias of $\widehat{DMD} = \hat{\beta}_3$, obtained from the linear model, was always less than 1%.

However, the variance of the DMD estimator directly increases with increasing measurement error, as a consequence of the increasing error variance of the measured exposure. This increased variance reduces the potential gain in power one might expect from this method. It should be noted, however, that the regression calibration method [Rosner, et al. 1990; Rosner, et al. 1992] can be used to correct the bias in estimate of $e^{\hat{\alpha}_3}$. After correction, variance inflation is expected.

Computational efficiency

Finally, our method also offers one additional advantage: computation time for linear regression is much faster than for logistic regression. Indeed, while linear regression has a closed form solution, parameters from logistic regression are estimated through an optimization iterative algorithm, which requires a substantial increase in computation time, especially when the number of covariates is large. For illustrative purposes, we compared in Table 4 the computation time of standard logistic regression and the reverse test using R software on a single desktop computer for the GxE analysis of 1 millions SNPs while increasing the sample size from 1,000 to 60,000. As the sample size increased, we observed an approximately five-fold lower computation time for the reverse test. For example, there was approximately 21.1 hours of computation for the analysis of 60,000 individuals, while logistic regression took 109.8 hours.

Illustrative example

We applied the proposed approach to analyze a study of more than 2,000 post-menopausal breast cancer cases and controls included in the CGEMS study (Cancer Genetic Markers of Susceptibility)[Hunter, et al. 2007]. We performed a genome-wide screening for the interaction effect of more than 450,000 common SNPs with a minor allele frequency $\geq 5\%$, with body mass index (BMI) and alcohol intake (gm/day), both established risk factors for breast cancer, using both the standard logistic and reverse tests. The distribution of these two exposures is presented in supplementary Figure 1. Both BMI and alcohol use showed strong deviations from normality, and were highly skewed. The QQplots of both tests are presented in Figure 3. Neither logistic regression nor the reverse test show type I error inflation for low p -values, highlighting the potential robustness of our new approach when the exposures deviate from normality. For the reverse test, the genomic inflation factor λ [Devlin and Roeder 1999] equaled 1.006 and 1.000, for the BMI and alcohol interaction respectively, and 1.012 and 1.004 for logistic regression.

No SNP showed genome-wide significance for the interaction for either approaches (see supplementary Figure 2). This is expected as none of the large consortium studies using much larger sample sizes found interaction effects with those two exposures either. Although a potential interaction between SNP rs17468277 (in gene *CASP8*) and alcohol use for breast cancer risk was previously reported [Nickels, et al. 2013], this interaction was not apparent with either the logistic regression test ($p=0.30$) or the reverse test ($p=0.24$). Overall, we observed a strong correlation between the p -values of the two approaches, equal to 0.977 and 0.915 for BMI and alcohol intake, respectively. This is expected as the vast majority of the interactions with these SNPs are expected to be null. Indeed, as shown in supplementary Figure S2 and S3, the top ranked SNPs largely overlapped between the two approaches.

Discussion

We developed a new test of interaction between genetic variants and a continuous exposure, where the genetic variants can be binary or ordinal (0, 1, 2). It consists of reversing the exposure and the disease status in the regression model to be fit, and then performing standard linear regression. Using this approach, the genetic variant by exposure interaction is evaluated by testing for the interaction between the genetic variant and the diseases status in relation to the exposure. Measurement error in the exposure only impacts the precision of the interaction effect estimate, DMD , in the linear regression model, which remains consistent, no matter how severe the exposure measurement error.

We performed a series of simulations to evaluate the performance of the proposed approach. We found that the reverse test had correct type I error rate across a wide range of scenarios, regardless of the severity of the exposure measurement error between the exposure and the genetic variant. We assumed a normal distribution of the exposure conditional on G and D with a constant variance in the proof that the parameter for the interaction term in the linear regression aligns with the interaction parameter in the logistic regression. However, in a simulation study where we generated data based on the logistic regression, the proposed reverse test had type I error rates close to the nominal level. Furthermore, in additional simulations, we showed that our approach performed reasonably well under other exposure

distributions including the uniform and skewed-normal, matched to the mean and variance of the standard normal, as long as the odds ratios for $D|X$ are not too large (i.e. $OR \leq 1.5$), as shown in supplementary Tables S5-S12. In addition, when this assumption does not hold, we have noted that the parameter for the interaction term in the linear regression is a *DMD* and still measures the magnitude of the GxE interaction and $H_0: \beta_3 = 0$ is still a null hypothesis of no G-E interaction. Evaluating the performance of the reverse test under the alternative, we observed that the reverse test typically had larger power than the standard test based on logistic regression. In particular, the gain in power increased both with an increasing main effect of the exposure on the outcome and as the magnitude of the interaction increased. Measurement error in the exposure tended to diminish this gain. However, some gain remained even when exposure measurement error was severe.

When G and X are independent in the controls, in the linear model (2) $\beta_2 = 0$, and thus $\beta_3 = 0$ is equivalent to $\beta_2 + \beta_3 = 0$. Note that $\beta_2 + \beta_3 = 0$ in (2) implies G and X are independent among cases. Therefore, under the G-E independence assumption, the null hypothesis tested by the reverse test is equivalent to that corresponding to the case-only test [Piegorsch, et al. 1994] based on model $X|G, D = 1$. Importantly, unlike the case-only test, the reverse test does not require the assumption of G-E independence. As shown in supplementary note, if there are additional covariates that need to be included in the logistic regression model, their interactions with G will also need to be included in the linear regression model when G is binary. When G is ordinal (0,1,2), we will need to include the interaction of the additional covariates with G^2 , in addition to the terms mentioned above. This way, the null hypothesis tested by the reverse test approach will correspond to the test based upon the GxE interaction term from the logistic regression.

For illustrative purposes, we compared results from the reverse test and the standard logistic regression approach in a study an ExGWAS study of interactions with body mass index and alcohol intake, both known risk factors for breast cancer. Among 1145 cases with invasive breast cancer and 1142 controls from the Cancer Genetic Markers of Susceptibility (CGEMS) study, we performed a genome-wide screening for interaction between 458,918 common SNPs and each of the two exposures. The Type I error rate inflation factor of the reverse test was close to one for both interactions screened. We did not identify any significant genome-wide interactions with either approach. However, as interactions are known to have extremely small effect sizes, a much larger sample size would likely have been needed than available to us here to identify them. As for any ExGWAS screening exercise, when significant associations are identified, it is recommended that stratum-specific effects be reported as well, to assess the magnitude and plausibility of the interactions detected. In addition, findings should be replicated in independent data to confirm associations.

Our approach outperformed the standard logistic regression approach, displaying higher power across all scenarios considered. In addition, the reverse test is dramatically more computationally efficient, on average 5 times faster (Table 4). Thus, in a standard ExGWAS setting, analyzing one million SNPs across 60,000 subjects would take approximately 4 days on a single computer, while the reverse test is not only statistically more powerful but it would, in comparison, take less than a day to complete this analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the reviewers for their helpful comments that have improved the paper. This research is supported by grant R01CA050597 from the National Cancer Institute (NCI) and R21HG007687 from the National Human Genome Research Institute (NHGRI).

Appendix A: The parametric relationship between the interaction effect from the linear model and the logistic model

We aligned here the parameters from standard logistic regression and the reverse test.

Consider first $G = 0$. From equations (1) and (3), $X|D, G=0 \sim \mathcal{N}(\beta_0 + \beta_1 D, \sigma^2)$ and

$$Pr(D=1|X, G=0) = \frac{e^{\alpha_0 + \alpha_1 X}}{1 + e^{\alpha_0 + \alpha_1 X}}, \text{ respectively. As}$$

$$Pr(D=1|X, G=0) = \frac{Pr(X|D=1, G=0)}{Pr(X|G=0)} P_{10}, \text{ where } p_{10} = Pr(D=1|G=0), \text{ we obtain:}$$

$$\frac{e^{\alpha_0 + \alpha_1 X}}{1 + e^{\alpha_0 + \alpha_1 X}} = \frac{Pr(X|D=1, G=0) P_{10}}{Pr(X|D=1, G=0) P_{10} + Pr(X|D=0, G=0) (1 - P_{10})}$$

$$\Leftrightarrow \frac{e^{\alpha_0 + \alpha_1 X}}{1 + e^{\alpha_0 + \alpha_1 X}} = \frac{\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X - \beta_0 - \beta_1)^2}{2\sigma^2}} P_{10}}{\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X - \beta_0 - \beta_1)^2}{2\sigma^2}} P_{10} + \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X - \beta_0)^2}{2\sigma^2}} (1 - P_{10})}$$

$$\Leftrightarrow \frac{1}{1 + \frac{1}{e^{\alpha_0 + \alpha_1 X}}} = \frac{1}{1 + \frac{e^{-\frac{(X - \beta_0)^2}{2\sigma^2}} (1 - P_{10})}{e^{-\frac{(X - \beta_0 - \beta_1)^2}{2\sigma^2}} P_{10}}}$$

$$\Leftrightarrow \frac{1}{e^{\alpha_0 + \alpha_1 X}} = \frac{e^{-\frac{(X - \beta_0)^2}{2\sigma^2}} (1 - P_{10})}{e^{-\frac{(X - \beta_0 - \beta_1)^2}{2\sigma^2}} P_{10}}$$

$$\Leftrightarrow e^{\alpha_0 + \alpha_1 X} e^{-\frac{(X - \beta_0)^2}{2\sigma^2}} (1 - P_{10}) = e^{-\frac{(X - \beta_0 - \beta_1)^2}{2\sigma^2}} P_{10}$$

$$\Leftrightarrow e^{\alpha_0 + \alpha_1 X} = e^{-\frac{2\beta_1(X - \beta_0) - \beta_1^2}{2\sigma^2}} \times \frac{P_{10}}{(1 - P_{10})}$$

$$\Leftrightarrow \alpha_0 + \alpha_1 X = \frac{\beta_1 X}{\sigma^2} - \frac{\beta_1 \left(\beta_0 + \frac{\beta_1}{2} \right)}{\sigma^2} + \log \frac{P_{10}}{(1 - P_{10})}$$

$$\Leftrightarrow \begin{cases} \alpha_0 = -\frac{\beta_1 \left(\beta_0 + \frac{\beta_1}{2} \right)}{\sigma^2} + \log \left(\frac{P_{10}}{1 - P_{10}} \right) \\ \alpha_1 = \frac{\beta_1}{\sigma^2} \end{cases}$$

$$\Leftrightarrow \begin{cases} \beta_0 = \frac{2\sigma^2 \times \left(\log \left(\frac{P_{10}}{1 - P_{10}} \right) - \alpha_0 \right) - \beta_1^2}{2\beta_1} \\ \beta_1 = \alpha_1 \sigma^2 \end{cases}$$

Deriving similarly for $G = 1$, we have

$$\frac{e^{(\alpha_0 + \alpha_2) + (\alpha_1 + \alpha_3)X}}{1 + e^{(\alpha_0 + \alpha_2) + (\alpha_1 + \alpha_3)X}} = \frac{Pr(X|D=1, G=1) P_{11}}{Pr(X|D=1, G=1) P_{11} + Pr(X|D=0, G=1) (1 - P_{11})}$$

$$\Leftrightarrow \frac{e^{(\alpha_0 + \alpha_2) + (\alpha_1 + \alpha_3)X}}{1 + e^{(\alpha_0 + \alpha_2) + (\alpha_1 + \alpha_3)X}} = \frac{\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X - \beta_0 - \beta_1 - \beta_2 - \beta_3)^2}{2\sigma^2}} P_{11}}{\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X - \beta_0 - \beta_1 - \beta_2 - \beta_3)^2}{2\sigma^2}} P_{11} + \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X - \beta_0 - \beta_2)^2}{2\sigma^2}} (1 - P_{11})}$$

By straightforward algebra as used for $G = 0$, we obtain

$$\Leftrightarrow \begin{cases} \beta_2 = \frac{2\sigma^2 \times \left(\log \left(\frac{P_{11} \times (1 - P_{10})}{P_{10} \times (1 - P_{11})} \right) - \alpha_2 \right) - \beta_3^2 - 2\beta_1\beta_3 - \beta_0\beta_3}{2(\beta_1 + \beta_3)} \\ \beta_3 = \alpha_3 \sigma^2 \end{cases}$$

Appendix B. Estimation based on model (3)

From (4), α_3 can be estimated by $\widehat{\beta}_3 / \hat{\sigma}^2$. From standard linear regression theory, $\widehat{\beta}_3$ is the fourth element of $\hat{\beta} = (Z'Z)^{-1} Z' \tilde{X}$ and $\hat{\sigma}^2 = \frac{S^2}{n - 4}$, where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$, n is the sample size, the prime represents the matrix transpose, Z is the $n \times 4$ design matrix, $\tilde{X}$ is the observed n -dimensional column vector of X , S^2 is the sum of the squared error, which can be written as $S^2 = \tilde{X}' \tilde{X} - \tilde{X}' H \tilde{X}$, with $H = Z(Z'Z)^{-1} Z'$. The variance can be derived from

the delta method, where $\text{var} \left(\frac{\hat{\beta}_3}{\hat{\sigma}^2} \right) = (\hat{\sigma}^{-2} - \hat{\beta}_3 \hat{\sigma}^{-4}) \sum (\hat{\sigma}^{-2} - \hat{\beta}_3 \hat{\sigma} - 4)'$ and Σ is the variance-covariance matrix of $(\hat{\beta}_3, \hat{\sigma}^2)$.

We can use the sandwich method to estimate the variance-covariance matrix of $(\hat{\beta}, \hat{\sigma}^2)$ based on the following joint estimating equation for $(\hat{\beta}, \hat{\sigma}^2): (\tilde{1} \tilde{D} \tilde{G} \tilde{GD})' (\tilde{X} - Z\beta) = 0$,

$\frac{(\tilde{X} - Z\beta)' (\tilde{X} - Z\beta)}{n - 4} - \sigma^2 = 0$, where $\tilde{D}$, $\tilde{G}$ and $\tilde{GD}$ are the observed n - dimensional column vector of D , G and GD , respectively, and $\tilde{1}$ is the n - dimensional column vector of 1.

Appendix C. The relationship between the parameters of the logistic and linear models under the general linear measurement error model

Assuming, as throughout this manuscript, the model for X^* is

$E(X^* | D, G) = \beta_0^* + \beta_1^* D + \beta_2^* G + \beta_3^* DG$, under measurement error model (7), we have:

$$\begin{aligned} E(X | G, D) &= E_{X^* | G, D} [E(X | G, D, X^*)] = E_{X^* | G, D} [\gamma_{G, D} + (\gamma_1 + \gamma_2 D + \gamma_3 G + \gamma_4 DG) X^*] = \gamma_{G, D} + (\gamma_1 + \gamma_2 D + \gamma_3 G + \gamma_4 DG) (\beta_0^* \\ &+ \beta_1^* D + \beta_2^* G + \beta_3^* DG) = \gamma_{G, D} + \gamma_1 \beta_0^* + (\gamma_1 \beta_1^* \\ &+ \gamma_2 \beta_0^* + \gamma_2 \beta_1^*) D \\ &+ (\gamma_1 \beta_2^* + \gamma_3 \beta_0^* + \gamma_3 \beta_2^*) G \\ &+ (\gamma_1 \beta_3^* \\ &+ \gamma_2 \beta_2^* + \gamma_2 \beta_3^* + \gamma_3 \beta_1^* + \gamma_3 \beta_1^* + \gamma_3 \beta_3^* + \gamma_4 \sum_{j=0}^3 \beta_j^*) DG. \end{aligned}$$

It follows that

$$\beta_3^* = \frac{\beta_3 - \gamma_4 \beta_0^* - (\gamma_3 + \gamma_4) / (\gamma_1 + \gamma_2) (\beta_1 - \gamma_2 \beta_0^*) - (\gamma_2 + \gamma_4) / (\gamma_1 + \gamma_3) (\beta_2 - \gamma_3 \beta_0^*)}{\sum_{j=1}^4 \gamma_j},$$

where $\beta_0^* = \frac{\beta_0 - \gamma_{G, D}}{\gamma_1}$. If $\gamma_2 = \gamma_3 = \gamma_4 = 0$ in measurement error model (7), which is often the case in practice, we have $\beta_3^* = \beta_3 / \gamma_1$. Based on the law of total variance, we have

$$\begin{aligned} \text{Var}(X | G, D) &= E_{X^* | G, D} [\text{Var}(X | G, D, X^*)] \\ &+ \text{Var}_{X^* | G, D} [E(X | G, D, X^*)] = E_{X^* | G, D} [v^2] + \text{Var}_{X^* | G, D} [\gamma_{G, D} + (\gamma_1 + \gamma_2 D + \gamma_3 G + \gamma_4 DG) X^*] = v^2 + (\gamma_1 + \gamma_2 D + \gamma_3 G + \gamma_4 DG)^2 \text{Var}(X^* | D, G). \end{aligned}$$

Solving the equation $\sigma^2 = v^2 + (\gamma_1 + \gamma_2 D + \gamma_3 G + \gamma_4 DG)^2 \text{Var}(X^* | D, G)$ above for $\text{Var}(X^* | D, G)$, we obtain $\text{Var}(X^* | D, G) = (\sigma^2 - v^2) / (\gamma_1 + \gamma_2 D + \gamma_3 G + \gamma_4 DG)^2$. If $\gamma_2 = \gamma_3 = \gamma_4 = 0$ in (7), $\text{Var}(X^* | D, G) = (\sigma^2 - v^2) / \gamma_1^2$, denoted as σ^{*2} .

Next, consider the $G \times X^*$ interaction in the logistic regression model under measurement error model (7) with $\gamma_2 = \gamma_3 = \gamma_4 = 0$. Given $\beta_3^* = \beta_3 / \gamma_1$ as shown above and the result in equation (4) linking the parameters of the linear model with that of the logistic when there is no exposure measurement error, model (6) holds with $\alpha_3 \sigma^2 = \gamma_1 \alpha_3^* \sigma^{*2}$. Hence, $\alpha_3 = \alpha_3^* (\sigma^2 - v^2) / (\gamma_1 \sigma^2)$. From the relationship between β_3^* and β_3 shown above under measurement error model (7) with $\gamma_2 = \gamma_3 = \gamma_4 = 0$, the hypothesis $H_0: \beta_3^* = 0$ is equivalent to the hypothesis $H_0: \beta_3 = 0$.

Appendix D: The parametric relationship between the interaction effect from the linear model and the logistic model if $G = 0, 1, 2$ (ordinal)

Now we assume that the exposure is a conditional normal distribution with variance σ^2 , such that

$$X|D, G \sim \mathcal{N} \left(\beta_0 + \beta_1 D + \beta_2 G + \beta_3 DG + \beta_4 G^2 + \beta_5 DG^2, \sigma^2 \right)$$

Note that the model for $E(X|D, G)$ above is a saturated model for $G = 0, 1, 2$. Assuming model (1) is correct, the derivation for $G = 0$ in Appendix A still holds. Therefore, we have

$$\begin{cases} \beta_0 = \frac{2\sigma^2 \times \left(\log \left(\frac{P_{10}}{1-P_{10}} \right) - \alpha_0 \right) - \beta_1^2}{2\beta_1} \\ \beta_1 = \alpha_1 \sigma^2 \end{cases} \quad (D1)$$

For $G = 1$, now we have

$$\begin{aligned} \frac{e^{(\alpha_0 + \alpha_2) + (\alpha_1 + \alpha_3)X}}{1 + e^{(\alpha_0 + \alpha_2) + (\alpha_1 + \alpha_3)X}} &= \frac{Pr(X|D=1, G=1) P_{11}}{Pr(X|D=1, G=1) P_{11} + Pr(X|D=0, G=1) (1 - P_{11})} \\ \Leftrightarrow \frac{e^{(\alpha_0 + \alpha_2) + (\alpha_1 + \alpha_3)X}}{1 + e^{(\alpha_0 + \alpha_2) + (\alpha_1 + \alpha_3)X}} &= \frac{\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X - \beta_0 - \beta_1 - \beta_2 - \beta_3 - \beta_4 - \beta_5)^2}{2\sigma^2}} P_{11}}{\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X - \beta_0 - \beta_1 - \beta_2 - \beta_3 - \beta_4 - \beta_5)^2}{2\sigma^2}} P_{11} + \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X - \beta_0 - \beta_2 - \beta_4)^2}{2\sigma^2}} (1 - P_{11})} \\ \Leftrightarrow e^{(\alpha_0 + \alpha_2) + (\alpha_1 + \alpha_3)X} &= e^{\frac{2(\beta_1 + \beta_3 + \beta_5)(X - \beta_0 - \beta_2 - \beta_4) - (\beta_1 + \beta_3 + \beta_5)^2}{2\sigma^2}} \times \frac{P_{11}}{(1 - P_{11})} \end{aligned}$$

$$\Leftrightarrow (\alpha_0 + \alpha_2) + (\alpha_1 + \alpha_3)X = \frac{2(\beta_1 + \beta_3 + \beta_5)(X - \beta_0 - \beta_2 - \beta_4) - (\beta_1 + \beta_3 + \beta_5)^2}{2\sigma^2} + \log\left(\frac{P_{11}}{(1 - P_{11})}\right).$$

(D2)

Similarly, for $G=2$, we have

$$(\alpha_0 + 2\alpha_2) + (\alpha_1 + 2\alpha_3)X = \frac{2(\beta_1 + 2\beta_3 + 4\beta_5)(X - \beta_0 - 2\beta_2 - 4\beta_4) - (\beta_1 + 2\beta_3 + 4\beta_5)^2}{2\sigma^2} + \log\left(\frac{P_{12}}{(1 - P_{12})}\right),$$

(D3)

where $p_{12} = Pr(D=1|G=2)$. It follows from (D1–D3) that $\beta_3 = \alpha_3\sigma^2$ and $\beta_5 = 0$. Furthermore, from (D2–D3), we have

$$(\alpha_0 + \alpha_2) = \frac{2(\beta_1 + \beta_3)(-\beta_0 - \beta_2 - \beta_4) - (\beta_1 + \beta_3)^2}{2\sigma^2} + \log\left(\frac{P_{11}}{(1 - P_{11})}\right),$$

$$(\alpha_0 + 2\alpha_2) = \frac{2(\beta_1 + 2\beta_3)(-\beta_0 - 2\beta_2 - 4\beta_4) - (\beta_1 + 2\beta_3)^2}{2\sigma^2} + \log\left\{\frac{P_{12}}{(1 - P_{12})}\right\}.$$

It follows that

$$\beta_4 = \frac{2\sigma^2 \left[(\alpha_0 + \alpha_2) - \log\left(\frac{P_{11}}{(1 - P_{11})}\right) \right]}{2(\beta_1 + \beta_3)} - \frac{\sigma^2 \left[(\alpha_0 + 2\alpha_2) - \log\left(\frac{P_{12}}{(1 - P_{12})}\right) \right]}{2(\beta_1 + 2\beta_3)} + \beta_1/4$$

and $\beta_2 = -\frac{2\sigma^2 \left[(\alpha_0 + \alpha_2) - \log\left(\frac{P_{11}}{(1 - P_{11})}\right) \right] + (\beta_1 + \beta_3)^2}{2(\beta_1 + \beta_3)} - \beta_0 - \beta_4$. Therefore, for ordinal $G=0, 1, 2$, the linear model corresponding to model (1) is

$E(X|D, G) = \beta_0 + \beta_1 D + \beta_2 G + \beta_3 DG + \beta_4 G^2$, where the reverse test for testing $H_0: \beta_3 = 0$ is equivalent to testing $H_0: \alpha_3 = 0$ in the logistic regression.

References

Armstrong BG, Whittemore AS, Howe GR. Analysis of case-control data with covariate measurement error: application to diet and colon cancer. *Stat Med.* 1989; 8(9):1151–63. discussion 1165–6. [PubMed: 2799135]

- Aschard H, Lutz S, Maus B, Duell EJ, Fingerlin TE, Chatterjee N, Kraft P, Van Steen K. Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. *Hum Genet.* 2012; 131(10):1591–613. [PubMed: 22760307]
- Barrdahl M, Canzian F, Joshi AD, Travis RC, Chang-Claude J, Auer PL, Gapstur SM, Gaudet M, Diver WR, Henderson BE, et al. Post-GWAS gene-environment interplay in breast cancer: results from the Breast and Prostate Cancer Cohort Consortium and a meta-analysis on 79,000 women. *Hum Mol Genet.* 2014; 23(19):5260–70. [PubMed: 24895409]
- Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E, Fisher S, Fox L, Howells W, Bertelsen S, et al. A genome-wide association study of alcohol dependence. *Proc Natl Acad Sci U S A.* 2010; 107(11):5082–7. [PubMed: 20202923]
- Bookman EB, McAllister K, Gillanders E, Wanke K, Balshaw D, Rutter J, Reedy J, Shaughnessy D, Agurs-Collins T, Paltoo D, et al. Gene-environment interplay in common complex diseases: forging an integrative model-recommendations from an NIH workshop. *Genet Epidemiol.* 2011; 35(4):217–25. [PubMed: 21308768]
- Boonstra PS, Mukherjee B, Gruber SB, Ahn J, Schmit SL, Chatterjee N. Tests for Gene-Environment Interactions and Joint Effects With Exposure Misclassification. *Am J Epidemiol.* 2016; 183(3):237–47. [PubMed: 26755675]
- Carroll, RJ., Ruppert, D., Stefanski, LA., Crainiceanu, CM. *Measurement Error in Nonlinear Models.* London: Chapman & Hall; 2006.
- Cheng KF. Analysis of case-only studies accounting for genotyping error. *Ann Hum Genet.* 2007; 71(Pt 2):238–48. [PubMed: 16984488]
- Colditz GA, Manson JE, Hankinson SE. The Nurses' Health Study: 20-year contribution to the understanding of health among women. *J Womens Health.* 1997; 6(1):49–62. [PubMed: 9065374]
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* 1999; 55(4):997–1004. [PubMed: 11315092]
- Efron B. The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis. *Journal of the American Statistical Association.* 1975; 70(352):892–898.
- Flint J, Mackay TF. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res.* 2009; 19(5):723–33. [PubMed: 19411597]
- Garcia-Closas M, Rothman N, Lubin J. Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size. *Cancer Epidemiol Biomarkers Prev.* 1999; 8(12):1043–50. [PubMed: 10613335]
- Garcia-Closas M, Thompson WD, Robins JM. Differential misclassification and the assessment of gene-environment interactions in case-control studies. *Am J Epidemiol.* 1998; 147(5):426–33. [PubMed: 9525528]
- Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, et al. A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nat Genet.* 2007; 39(7):870–4. [PubMed: 17529973]
- Lagakos SW. Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Stat Med.* 1988; 7:257–274. [PubMed: 3353607]
- Levene, H. Robust Tests for Equality of Variances. In: Olkin, I.Ghurye, S.Hoeffding, W.Madow, W., Mann, H., editors. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling.* Stanford University Press; 1960. p. 278-292.
- Lindstrom S, Yen YC, Spiegelman D, Kraft P. The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. *Hum Hered.* 2009; 68(3):171–81. [PubMed: 19521099]
- Lyles RH, Guo Y, Hill AN. A Fresh Look at the Discriminant Function Approach for Estimating Crude or Adjusted Odds Ratios. *Am Stat.* 2009; 63(4)
- Nickels S, Truong T, Hein R, Stevens K, Buck K, Behrens S, Eilber U, Schmidt M, Haberle L, Vrieling A, et al. Evidence of gene-environment interactions between common breast cancer susceptibility loci and established environmental risk factors. *PLoS Genet.* 2013; 9(3):e1003284. [PubMed: 23544014]

- Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med.* 1994; 13(2):153–62. [PubMed: 8122051]
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006; 38(8): 904–9. [PubMed: 16862161]
- Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *Am J Epidemiol.* 1990; 132(4):734–45. [PubMed: 2403114]
- Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. *Am J Epidemiol.* 1992; 136(11):1400–13. [PubMed: 1488967]
- Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med.* 1989; 8(9):1051–69. discussion 1071–3. [PubMed: 2799131]
- Shapiro SS, Wilk MB. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika.* 1965; 52(3/4):591–611.
- Stenzel SL, Ahn J, Boonstra PS, Gruber SB, Mukherjee B. The impact of exposure-biased sampling designs on detection of gene-environment interactions in case-control studies with potential exposure misclassification. *Eur J Epidemiol.* 2015; 30(5):413–23. [PubMed: 24894824]
- Tchetgen Tchetgen EJ, Kraft P. On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is misspecified. *Epidemiology.* 2011; 22(2):257–61. [PubMed: 21228699]
- Thomas DC. Models for exposure-time-response relationships with applications to cancer epidemiology. *Annu Rev Public Health.* 1988; 9:451–82. [PubMed: 3288239]
- Tobacco, Genetics C. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet.* 2010; 42(5):441–7. [PubMed: 20418890]
- Wong MY, Day NE, Luan JA, Wareham NJ. Estimation of magnitude in gene-environment interactions in the presence of measurement error. *Stat Med.* 2004; 23(6):987–98. [PubMed: 15027084]
- Zhang L, Mukherjee B, Ghosh M, Gruber S, Moreno V. Accounting for error due to misclassification of exposures in case-control studies of gene-environment interaction. *Stat Med.* 2008; 27(15): 2756–83. [PubMed: 17879261]

Figure 1. Relative power of the reverse test compared to logistic regression

The y-axis is the ratio of average χ^2 test statistics of the reverse test over the average χ^2 test statistic of the logistic regression test, derived from 10,000 simulations with data generated under the Linear model while increasing the exposure effect ($OR(X, G=0)$, light red to dark red) and the interaction effect (ROR , x-axis), assuming a marginal genetic effect, $OR(G)$, of 1.2 (a) or 2.0 (b). We considered scenarios where the correlation between G and X given $D=0$ is 0 (upper panels) and 0.3 (lower panels).

Figure 2. Power of the reverse test compared to the standard logistic regression test for GxE interaction as exposure measurement error increases

The y-axis is the ratio of mean χ^2 test statistic for the reverse test over the mean χ^2 test statistic from the logistic regression model over 10,000 simulations with data generated under the Linear model, while decreasing ρ (x-axis), representing increasing measurement error, from 1 to 0.2, with $OR(D|X, G=0)=1.0$, $OR(D|X)=3$ and ROR ranging from 1.1 to 2.

Figure 3. QQ plots for ExGWAS interaction tests in the CGEMS breast cancer study

We applied both standard logistic regression (A) and the reverse test (B) in a genome-wide screening for interaction effects between more than 500,000 SNPs and body mass index (upper panel) and alcohol intake (lower panel). As expected in GWAS where the vast majority of the tests are expected to be distributed under the null hypothesis of no association, no analysis showed inflation of the median p -values, as measured by the lambda values, or inflation of the tail of the distribution of the p -values.

Table 1

Type I error rate of the standard and reverse test ($ROR=1$)

Parameters of the simulation model										
			$corr(G, X D=0)=0$		$corr(G, X D=0)=0.05$					
			X		X					
			$\mathbf{X}^*(\rho=0.25)$		$\mathbf{X}^*(\rho=0.25)$					
			log.5.		log.5.		log.5.		log.5.	
			rev.5.		rev.5.		rev.5.		rev.5.	
			 	
			rev.5.		rev.5.					

Parameters of the simulation model										
$corr(G, X D=0)=0$			$corr(G, X D=0)=0.05$							
X			X							
$X^*(\rho=0.25)$			$X^*(\rho=0.25)$							
$Pr(D G=0)$	$OR(D X)$	$OR(D G)$	log.5.	rev.5.	log.5.	rev.5.	log.5.	rev.5.	log.5.	rev.5.
0.01	1.5	2.0	0.050	0.049	0.049	0.052	0.044	0.042	0.050	0.051
0.01	2.0	1.5	0.054	0.054	0.052	0.050	0.049	0.051	0.052	0.051
0.01	2.0	2.0	0.047	0.050	0.052	0.052	0.051	0.052	0.052	0.051
0.05	1.0	1.5	0.047	0.047	0.048	0.049	0.050	0.050	0.050	0.051
0.05	1.0	2.0	0.053	0.051	0.050	0.050	0.047	0.047	0.048	0.049
0.05	1.5	1.5	0.052	0.050	0.047	0.049	0.053	0.051	0.051	0.050
0.05	1.5	2.0	0.048	0.052	0.048	0.047	0.052	0.054	0.048	0.049
0.05	2.0	1.5	0.047	0.047	0.047	0.049	0.048	0.052	0.048	0.048
0.05	2.0	2.0	0.048	0.053	0.053	0.053	0.046	0.051	0.049	0.050
0.30	1.0	1.5	0.054	0.054	0.055	0.053	0.050	0.051	0.050	0.050
0.30	1.0	2.0	0.048	0.048	0.050	0.049	0.053	0.053	0.047	0.047
0.30	1.5	1.5	0.050	0.054	0.050	0.051	0.052	0.053	0.051	0.052
0.30	1.5	2.0	0.050	0.054	0.048	0.050	0.048	0.047	0.043	0.044
0.30	2.0	1.5	0.050	0.054	0.049	0.048	0.047	0.049	0.046	0.047
0.30	2.0	2.0	0.049	0.051	0.048	0.049	0.049	0.054	0.050	0.052

Abbreviation: Log: logistic regression, Rev: reverse test, ROR =ratio of odds ratios, log.5.=rejection rate using logistic regression with significance level 0.05, rev.5.=rejection rate using the reverse test with significance level 0.05. X is the true exposure and X^* , the exposure measured with error, with $\rho = \frac{Var(X)}{Var(X^*)} = 0.25$.

Type I error rate was calculated across 10,000 simulations for each scenario. For each simulation a full cohort of 100,000 subjects was simulated and a case-control study including 1000 cases and 1000 controls was randomly sampled from the cohort.

Table 2

Relative power of the reverse test against logistic regression

Parameters of the simulation model				χ^2 ratio			
				$corr(G, X D=0)=0$		$corr(G, X D=0)=0.05$	
$Pr(D G=0)$	$OR_{D X}$	$OR_{D G}$	ROR	$\rho=1$	$\rho=0.75$	$\rho=0.5$	$\rho=0.25$
Data simulated under the Linear Model							
0.05	1.0	1.0	1.1	1.01	1.00	1.00	1.00
0.05	1.0	1.0	1.5	1.06	1.04	1.03	1.02
0.05	2.0	1.0	1.1	1.14	1.09	1.05	1.02
0.05	2.0	1.0	1.5	1.49	1.36	1.23	1.11
0.05	1.0	2.0	1.1	1.00	1.00	1.00	1.00
0.05	1.0	2.0	1.5	1.05	1.04	1.02	1.01
0.05	2.0	2.0	1.1	1.13	1.08	1.05	1.01
0.05	2.0	2.0	1.5	1.45	1.33	1.20	1.09
0.30	1.0	1.0	1.1	1.00	1.00	1.00	1.00
0.30	1.0	1.0	1.5	1.06	1.04	1.03	1.02
0.30	2.0	1.0	1.1	1.14	1.09	1.04	1.01
0.30	2.0	1.0	1.5	1.48	1.35	1.22	1.09
0.30	1.0	2.0	1.1	1.00	1.00	1.00	1.00
0.30	1.0	2.0	1.5	1.05	1.04	1.03	1.02
0.30	2.0	2.0	1.1	1.14	1.08	1.04	1.01
0.30	2.0	2.0	1.5	1.47	1.33	1.20	1.09
Data simulated under the logit Model							
0.05	1.0	1.0	1.1	1.01	1.01	1.00	1.00
0.05	1.0	1.0	1.5	0.53	0.54	0.61	1.00
0.05	2.0	1.0	1.1	0.95	0.96	0.98	1.00
0.05	2.0	1.0	1.5	0.47	0.48	0.56	1.00
0.05	1.0	2.0	1.1	1.04	1.04	1.01	1.02
0.05	1.0	2.0	1.5	1.49	1.38	1.21	1.06

χ^2 ratio									
Parameters of the simulation model									
$Pr(D G=0)$	$OR_{D X}$	$OR_{D G}$	ROR	$\rho=1$	$\rho=0.75$	$\rho=0.5$	$\rho=0.25$	$\rho=1$	$corr(G, X D=0)=0$
$corr(G, X D=0)=0.05$									
$\rho=0.25$									
0.05	2.0	2.0	1.1	1.44	1.32	1.20	1.08	1.20	1.12
0.05	2.0	2.0	1.5	1.86	1.70	1.56	1.38	1.42	1.30
0.30	1.0	1.0	1.1	0.99	0.99	1.00	1.00	1.00	1.00
0.30	1.0	1.0	1.5	0.93	0.93	0.93	0.94	1.01	1.01
0.30	2.0	1.0	1.1	0.93	0.94	0.95	0.98	1.00	1.00
0.30	2.0	1.0	1.5	1.04	1.04	1.04	1.03	1.02	1.01
0.30	1.0	2.0	1.1	1.20	1.13	1.08	1.03	1.14	1.10
0.30	1.0	2.0	1.5	1.60	1.45	1.32	1.18	1.46	1.33
0.30	2.0	2.0	1.1	1.41	1.28	1.18	1.07	1.27	1.18
0.30	2.0	2.0	1.5	1.66	1.50	1.36	1.22	1.46	1.32

Abbreviation: ROR =ratio of odds ratio; ρ , which represents the magnitude of the measurement error, equals $\frac{Var(X)}{Var(X^*)}$ where X is the true exposure and X^* , the exposure measured with error. χ^2 ratio is the ratio of the average χ^2 test statistics for the interaction effect between the reverse test and logistic regression, based on 10,000 simulation data sets for each model. For each simulation, the full cohort including 30,000 subjects was simulated and a subset of 1000 cases and 1000 controls were randomly sampled from the full cohort for analysis.

Table 3

Impact of measurement error on effect estimates.

ρ	$DMD_{reverse}$		$\log(\widehat{ROR})_{reverse}$		$\log(\widehat{ROR})_{logistic}$	
	%Bias	SE	%Bias	SE	%Bias	SE
1.00	0.8%	0.10	1.0%	0.10	2.5%	0.12
0.96	0.4%	0.10	3.3%	0.10	2.7%	0.11
0.86	0.2%	0.10	13.5%	0.09	12.5%	0.10
0.74	0.0%	0.11	26.3%	0.08	26.1%	0.09
0.61	0.8%	0.12	38.4%	0.07	38.0%	0.08
0.50	0.2%	0.13	49.8%	0.07	49.2%	0.07
0.41	0.8%	0.15	58.6%	0.06	58.0%	0.07
0.34	0.4%	0.16	66.3%	0.05	65.9%	0.06
0.28	0.2%	0.17	71.9%	0.05	71.6%	0.05
0.24	0.5%	0.20	76.2%	0.05	76.1%	0.05
0.20	0.7%	0.21	79.8%	0.04	79.4%	0.04

%Bias is (estimated parameter-true parameter)/true parameter, derived across series of 10,000 simulated case-control studies with data generated under Linear Model with 1000 cases and 1000 controls; ρ , which represents the magnitude of the measurement error, equals $\frac{Var(X)}{Var(X^*)}$ where X is the true exposure and X^* , the exposure measured with error. The true log(ROR) and DMD equal 0.405 (ROR =1.5).

Table 4

Comparison of computation time.

Sample size (cases+controls)	Computation time in hours ^a		Ratio
	Reverse test	Logistic regression	
1,000	0.7	1.8	2.57
2,000	1.0	3.4	3.40
10,000	3.4	19.2	5.65
20,000	6.7	35.7	5.33
60,000	21.1	109.8	5.21

^aWe compared the computation time for the reverse and standard logistic regression tests for GxE interaction, in hours for a genome-wide screening of interaction effect of one million SNPs, using either standard logistic regression or the reverse test.