Published in final edited form as:

Genet Epidemiol. 2018 December ; 42(8): 838–845. doi:10.1002/gepi.22154.

BIAS IN PARAMETER ESTIMATES DUE TO OMITTING GENE-ENVIRONMENT INTERACTION TERMS IN CASE-CONTROL STUDIES

Iryna Lobach^{1,*}

¹Department of Epidemiology and Biostatistics, University of California San Francisco

Abstract

Genetic studies are continuing to generate volumes and variety of data that can be used to examine the genetic effects. Often the effect of a genetic variant varies by non-genetic measures, what is traditionally defined as gene-environment interaction (GxE). If the GxE term is neglected, estimates of the main effects can be substantially biased. We derive a general and convenient approximation to the magnitude of bias in the estimates due to omitting the GxE term. We show that the approximation is reasonably accurate in finite samples. We then apply the approximation in a study of Alzheimer's disease.

Keywords

case-control studies; Alzheimer's disease; gene-environment interactions; omitting variables; Kullback-Leibler divergence

INTRODUCTION

Case-control genetic studies are continuing to generate volumes and variety of data that can be used to test association between genetic variants and a complex disease. The effect of genetic variants often varies by non-genetic variables (Ritz et al, 2016) what is traditionally defined as gene-environment interaction (GxE). We are interested to quantify the magnitude of bias in the main effect estimates of the genetic and environmental variables when the GxE terms are omitted from the model.

Bias due to omitting variables received considerable attention in methodological literature (Gail et al, 1988; Hauck et al, 1991; Neuhaus, 2002). The setting that we consider is unique in that the model is misspecified both because the GxE term is omitted from the model and because the analyses are based on the usual logistic regression what ignores the case-control sampling design that was used to collect the data. Prentice and Pyke (1979) justified the use of a prospective model in the retrospective (case-control) sampling. This result, however, does not naturally extend to the setting where the GxE terms are omitted.

*Correspondence to be addressed to: Iryna Lobach, Ph.D., Department of Epidemiology and Biostatistics, University of California, San Francisco, 550 16th Street, San Francisco CA 94158, Iryna.lobach@ucsf.edu, (415) 476-6115.

We derive a simple and general approximation to bias in parameter estimates when GxE term is omitted from the model. The differences in the parameter estimates based on the usual logistic regression model with vs. without the GxE terms are approximated using Kullback-Leibler divergence criteria (Kullback, 1956) between the two models.

This paper proceeds as follows. First, Material and Methods section introduces the setting and notations, as well as derives magnitude of bias in parameter estimates. Next, the Simulations Experiments section presents results of an extensive simulation experiments. We apply the approximation to study bias in estimates of the genetic effects in a study of Alzheimer's disease. Finally, the paper is concluded with a brief Discussion.

MATERIALS AND METHODS

For individual i , let G_i be the genotype, X_i be the environmental variable potentially interacting with the genotype, and Z_i be a vector of other environmental variables. For clarity of presentation we suppose that all variables are binary. The setting can be easily extended to categorical variables.

We will assume that the genotype is independent of all environmental variables and the genotype follows Hardy-Weinberg Equilibrium: $G \sim Q(g, \theta)$. Let $D_i = \{0, 1\}$ be a binary indicator of the disease status. In the overall population, let $\pi_0 = \text{pr}(D = 0)$ and $\pi_1 = \text{pr}(D = 1)$ and in our study population let n_0 be the number of controls (i.e. $D = 0$), n_1 be the number of cases (i.e. $D = 1$), and $n = n_0 + n_1$.

The data are collected using retrospective case-control sampling design where the cases are sampled from a set with the disease and the controls are sampled from the set without the disease.

We next assume that the probability of the *true* disease follows a logistic model

$$\begin{aligned} \text{pr}_B(D = 1 | G = g, X = x, Z = z) \\ = \frac{\exp\{\beta_0 + \beta_X \times x + \beta_Z \times z + \beta_G \times g + \beta_{X \times Z} \times x \times z + \beta_{G \times X} \times g \times x\}}{1 + \exp\{\beta_0 + \beta_X \times x + \beta_Z \times z + \beta_G \times g + \beta_{X \times Z} \times x \times z + \beta_{G \times X} \times g \times x\}}, \end{aligned} \quad (1)$$

Define $B = (\beta_0, \beta_X, \beta_Z, \beta_G, \beta_{X \times Z}, \beta_{G \times X})$ to be the vector of coefficients of interest. We note that our approach can be easily extended to other models, including those with multiple disease states.

The observed data are collected using retrospective sampling design. The likelihood function of the observed data is based on the probability $[G, X, Z | D]$

$$Q_B^{TRUE}(D, G, X, Z) \quad (2)$$

$$= \frac{\text{pr}_B[D|G, X, Z] \times \text{pr}[G, X, Z]}{\sum_{g^*, x^*, z^*} \text{pr}_B[D|G = g^*, X = x^*, Z = z^*] \times \text{pr}[G = g^*, X = x^*, Z = z^*]},$$

We next consider the usual logistic regression model that omits the GxE term.

Then the usual logistic regression model is based on the probability

$$Q_{B^*}^{MISSPEC}(D, G, X, Z) = \text{pr}_{B^*}[D|G, X, Z], \quad (3)$$

where

$$\begin{aligned} \text{pr}_{B^*}(D = 1|G = g, X = x, Z = z) \\ = \frac{\exp\{\beta_0^* + \beta_X^* \times x + \beta_Z^* \times z + \beta_G^* \times g + \beta_{X \times Z}^* \times x \times z\}}{1 + \exp\{\beta_0^* + \beta_X^* \times x + \beta_Z^* \times z + \beta_G^* \times g + \beta_{X \times Z}^* \times x \times z\}}. \end{aligned} \quad (4)$$

We are interested to find an analytic solution that relates parameters

$B^* = (\beta_0^*, \beta_X^*, \beta_Z^*, \beta_G^*, \beta_{X \times Z}^*)$ from the misspecified model (3)–(4) to the parameters $B = (\beta_0, \beta_X, \beta_Z, \beta_G, \beta_{G \times X}, \beta_{X \times Z})$ from the *true* model (1)–(2).

The next steps are motivated by the developments in Kullback (1959), Neuhaus (2002). Kullback (1959) proved that parameters B^* estimated in the misspecified models (3, 4) converge to values that minimize the Kullback-Leibler divergence between the *true* and false models with expectations taken with respect to the *true* model, i.e.

$$B^* = \underset{B^*}{\text{argmin}} \left(E_{X, G, Z} \left[E_{D|X, G, Z} \log \left\{ \frac{Q_B^{TRUE}(D, G, X, Z)}{Q_{B^*}^{MISSPEC}(D, G, X, Z)} \right\} \right] \right). \quad (5)$$

Arguments provided in Appendix prove the following relationship between the parameters B^* and B in models (3)–(4) and (1)–(2):

$$\beta_0^* \approx \beta_0; \quad \beta_G^* \approx \beta_G + 0.5 \times \beta_{G \times X}; \quad \beta_X^* \approx \beta_X; \quad \beta_Z^* \approx \beta_Z; \quad \beta_{X \times Z}^* \approx \beta_{X \times Z}. \quad (6)$$

If in the misspecified model both $G \times X$ and $X \times Z$ are omitted, then

$$\beta_0^* \approx \beta_0; \quad \beta_G^* \approx \beta_G + 0.5 \times \beta_{G \times X}; \quad \beta_X^* \approx \beta_X + 0.5 \times \beta_{G \times X} + 0.5 \times \beta_{X \times Z}; \quad \beta_Z^* \approx \beta_Z + 0.5 \times \beta_{X \times Z}.$$

(7)

We next consider a misspecified model that neglects the interaction terms and the main effect of the environment X , i.e.

$$\text{pr}_{B^*}(D = 1 | G = g, Z = z) = \frac{\exp\{\beta_0^* + \beta_Z^* \times z + \beta_G^* \times g\}}{1 + \exp\{\beta_0^* + \beta_Z^* \times z + \beta_G^* \times g\}}, \quad (8)$$

then

$$\beta_0^* \approx \beta_0; \quad \beta_G^* \approx \beta_G + 0.5 \times \beta_{G \times X}; \quad \beta_Z^* \approx \beta_Z. \quad (9)$$

We next extend the true model to include two genetic markers, i.e.

$$\begin{aligned} \text{pr}_B(D = 1 | G = g, X = x, Z = z) \\ = \frac{\exp\{\beta_0 + \beta_X \times x + \beta_Z \times z + \beta_{G_1} \times g_1 + \beta_{G_2} \times g_2 + \beta_{G_1 \times X} \times g_1 \times x + \beta_{G_2 \times X} \times g_2 \times x\}}{1 + \exp\{\beta_0 + \beta_X \times x + \beta_Z \times z + \beta_{G_1} \times g_1 + \beta_{G_2} \times g_2 + \beta_{G_1 \times X} \times g_1 \times x + \beta_{G_2 \times X} \times g_2 \times x\}}, \end{aligned} \quad (10)$$

while the parameters are estimated using

$$\text{pr}_{B^*}(D = 1 | G = g, X = x, Z = z) = \frac{\exp\{\beta_0^* + \beta_Z^* \times z + \beta_X^* \times x + \beta_{G_1}^* \times g_1 + \beta_{G_2}^* \times g_2\}}{1 + \exp\{\beta_0^* + \beta_Z^* \times z + \beta_X^* \times x + \beta_{G_1}^* \times g_1 + \beta_{G_2}^* \times g_2\}}.$$

Then

$$\begin{aligned} \beta_0^* \approx \beta_0; \quad \beta_{G_1}^* \approx \beta_{G_1} + 0.5 \times \beta_{G_1 \times X}; \quad \beta_{G_2}^* \approx \beta_{G_2} + 0.5 \times \beta_{G_2 \times X}; \quad \beta_X^* \approx \beta_X + 0.5 \\ \times \beta_{G_1 \times X} + 0.5 \times \beta_{G_2 \times X}; \quad \beta_Z^* \approx \beta_Z. \end{aligned} \quad (11)$$

SIMULATION EXPERIMENTS

We simulate the genetic variable to be binary to mimic a single nucleotide polymorphism (SNP) with a dominant or recessive effect. The environmental variables are binary with $\text{pr}(X = 1) = 0.488$ and $\text{pr}(Z = 1) = 0.07$. For each setting we simulated 500 datasets with n_0 controls and n_1 cases. We vary sample size, effect size and frequency of genotype across the simulation settings. The goal of the simulation experiments is compare the approximation to the magnitude of bias that we derived to the empirical bias of the estimates. We are specifically interested in a setting when the actual values of X are not available. The empirical bias is the average bias observed in the parameter estimates across 500 datasets.

First, we simulate the disease status according to model (1) but estimate parameters in a misspecified model with $\beta_{G \times X}^* = 0$ and in a misspecified model with $\beta_{X \times Z}^* = 0$ and $\beta_{G \times X}^* = 0$. Shown in Table 1 and Supplementary Table 1 are the theoretical approximation to the bias along with the empirical mean and standard deviation (SD) in studies where the true coefficients are $\beta_0 = -1.05$, $\beta_G = -1.1$, $\beta_X = 1.1$, $\beta_Z = 0.69$, $\beta_{X \times Z} = 1$, $\beta_{G \times X} = 0.69$. Genotype is binary with $\text{pr}(G = 1) = 0.10$. We vary $n_0 = n_1$ to be 1,000/5,000/10,000. When only the GxE term is omitted, i.e. $\beta_{G \times X}^* = 0$, then bias in $\hat{\beta}_G$ is approximated to be -0.35 while the empirical estimate of the bias is -0.44 in a study with $n_0 = n_1 = 1,000$. This difference between the approximation and the empirical estimate is $\frac{1}{2} \times \text{SD}$. When sample size increases to $n_0 = n_1 = 10,000$ the empirical estimate is -0.41 . All other parameters are unbiased and the theoretical approximation is 0, while the empirical estimates are nearly zero. When both GxE and the $X \times Z$ terms are omitted, all parameter estimates are biased. For example, the bias in $\hat{\beta}_G$ is approximated to be -0.35 , while the empirical estimate is -0.41 . This difference is $\frac{1}{2} \times \text{SD}$. Table 2 and Supplementary Table 2 present the same setting with the only difference that frequency of genotype $\text{pr}(G = 1) = 0.01$. In a study with 1,000 cases and 1,000 controls, when GxE term is omitted, bias in $\hat{\beta}_G$ is estimated to be -0.35 , while the empirical estimate is -0.53 . This difference is $\frac{1}{4} \times \text{SD}$. When both GxE and $X \times Z$ are omitted, then bias in $\hat{\beta}_G$ is estimated to be -0.35 , while it is estimated to be -0.50 . The difference is $\frac{1}{4} \times \text{SD}$.

We next simulate datasets by varying the true values of β_G and $\beta_{G \times X}$, i.e.

$$\beta_0 = -1.05, \quad \beta_G = \log(1.3), \log(1.9), \dots, \log(3.1), \quad \beta_X = 1.1, \quad \beta_Z = 0.69, \quad \beta_{X \times Z} = 1, \quad \beta_{G \times X} = \log(1.3), \log(1.9), \dots, \log(3.1)$$

We generated samples with 1,000 cases and 1,000 controls with $\text{pr}(G = 1) = 0.10$. Shown in Table 3A are the biases in $\hat{\beta}_G$ when the model omits the GxE term, i.e. $\beta_{G \times X}^* = 0$.

Supplementary Table 3B presents the setting when the misspecified model sets $\beta_{X \times Z}^* = 0$ and $\beta_{G \times X}^* = 0$. Supplementary Table 3C is based on a misspecified model with

$\beta_X^* = 0$, $\beta_{X \times Z}^* = 0$, $\beta_{G \times X}^* = 0$. In all of these settings estimates of β_G are substantially biased. The average difference between the approximation to the bias and the empirical bias is 0.10.

We next consider a setting when two genetic variants have an effect on susceptibility to the disease. We simulate various datasets of 2,000 cases and 2,000 controls according to risk model (10) with varying the true values of $\beta_{G_1}, \beta_{G_2}, \beta_{G_1 \times X}, \beta_{G_2 \times X}$ across $\log(1.3), \log(1.6), \log(1.9)$; $\beta_0 = -1.05, \beta_X = 1.1, \beta_Z = 0.69$. Table 4 and Supplementary Table 4 present approximations to the magnitude of bias in estimates of β_{G_1} and β_{G_2} as derived in (11) along with the empirical bias and SD. In all of these settings estimates are substantially biased and approximation is reasonably accurate.

ANALYSES OF GENETIC VARIANTS SERVING TOLL-LIKE RECEPTORS IN ALZHEIMER'S DISEASE

We applied the proposed analyses to a dataset collected as part of the AD Genetics Consortium. The data consists of 1,245 controls and 2,785 cases. The average age (SD) of cases and controls are 72.1 (9.1) and 70.9 (8.8) years, respectively. Among cases, 1,458 (52.4%) are men; among controls, 678 (63.9%) are men. At least one ApoE $\epsilon 4$ allele is present in 1,796 (64.5%) of cases and 365 (29.1%) of controls.

Illumina Human 660K markers have been mapped onto human chromosomes using NCBI dbSNP database (<https://www.ncbi.nlm.nih.gov/projects/SNP/>). Chromosome location, proximal gene or genes and gene structure location (e.g. intron, exon, intergenic, UTR) has been recorded for all SNPs. From these data, we inferred 5 single nucleotide polymorphisms (SNPs) to reside in genes serving Toll-Like Receptors (TLR) that have significant ($p < 0.01$) interaction with ApoE $\epsilon 4$ status (X). We model the genetic variables using a binary indicator of presence or absence of a minor allele. The usual logistic model included adjustments for sex, age, and education.

Shown in Table 5 are main effect estimates of the five SNPs obtained in a model that includes the SNP-by-ApoE $\epsilon 4$ status ($\hat{\beta}_G$) and in a model that omits the interaction term ($\hat{\beta}_G^*$). We note that the difference between these two parameter estimates is close to the difference obtained based on the approximation. On average across the five SNPs the difference between $\hat{\beta}_G - \hat{\beta}_G^*$ and the approximation is 0.098. Two of the SNPs, rs830832 and rs11938703, have significant ($p < 0.01$) coefficients $\hat{\beta}_G$, while p-values for $\hat{\beta}_G^*$ are > 0.14 .

These two SNPs have been previously noted in the literature. Specifically, rs830832 has been reported in studies of rheumatoid arthritis, prion disease, waist-hip ratio and hearing (<https://www.gwascentral.org/marker/HGVM179538/results?t=ZERO>). The other SNP, rs11938703, has been reported in studies of brain glutamate concentrations, sporadic Creutzfeldt-Jakob disease, waist-hip ratio and fasting insulin (https://www.gwascentral.org/marker/HGVM4583607/results?t=1&page=1&page_size=50&format=&r%5B%5D=)

DISCUSSION

We derived a convenient and general approximation to the magnitude of bias in parameter estimates when GxE interaction terms are omitted from the model. Simulation studies showed that the theoretical form of bias is reasonably accurate in finite samples.

The proposed approximation to the magnitude of bias is readily extendable to models with more genetic and environmental variables; and to non-additive models.

While interpretation of the parameter β_G differs between a model with and without GxE, the approximation to the magnitude of difference in the estimates provides convenient understanding of the relationship between the two estimates.

The approximation to the magnitude of bias is a function of the parameters and hence the actual data on the environmental variable X is not necessary needed. This comes particularly useful in common situations when publically available databases contain wealth of genetics and only a brief set of non-genetic (environmental) measures. These databases are e.g. the database of Genotypes and Phenotypes (<https://www.ncbi.nlm.nih.gov/gap>), the Cancer Genome Atlas (<https://cancergenome.nih.gov/>). The data can be used to estimate main effects of the of the genotype (G) with adjustment for the key covariates (Z). Even if measures X are not publically available, a researcher can use results of previously published studies to obtain estimates of $\beta_{G \times X}$ and examine the corresponding magnitude of bias in the estimates of β_G .

Better understanding of the magnitude in bias due to neglecting GxE term can in part address the concerns raised by Hirschhorn et al (2002) and Manolino et al (2009). Specifically, the downward bias in the estimates of the genetic effects might address the missing heritability issue noted by Manolino et al (2009). The upward bias in the estimates might in part explain the conclusion reached by Hirschhorn et al (2002) that only 1% of the association found thus far are likely to be true.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENT

We thank Ivan Belousov for help with the computations.

Dr. Lobach is supported by 5R21AG043710-02.

Genotyping is performed by Alzheimer's Disease Genetics Consortium (ADGC), U01 AG032984, RC2 AG036528. Phenotypic collection is coordinated by the National Alzheimer's Coordinating Center (NACC), U01 AG016976

Samples from the National Cell Repository for Alzheimer's Disease (NCRAD), which receives government support under a cooperative agreement grant (U24 AG21886) awarded by the National Institute on Aging (NIA), were used in this study. We thank contributors who collected samples used in this study, as well as patients and their families, whose help and participation made this work possible.

Data for this study were prepared, archived, and distributed by the National Institute on Aging Alzheimer's Disease Data Storage Site (NIAGADS) at the University of Pennsylvania (U24-AG041689-01)

APPENDIX

We first consider a true model (1) and misspecified model (4).

Derivative of the Kullback-Leibler divergence (5) with respect to parameters of the misspecified model are

$$-E_{X,G,Z} \left[\frac{\text{pr}_{\mathbf{B}}(D=0|G,X,Z)}{Q_{\mathbf{B}^*}(D=0,G,X,Z)} \times \frac{\partial}{\partial \mathbf{B}^*} Q_{\mathbf{B}^*}(D=0|G,X,Z) + \frac{\text{pr}_{\mathbf{B}}(D=1|G,X,Z)}{Q_{\mathbf{B}^*}(D=1,G,X,Z)} \times \frac{\partial}{\partial \mathbf{B}^*} Q_{\mathbf{B}^*}(D=1|G,X,Z) \right].$$

Define $K(g,x,z;\mathbf{B}^*) = \frac{\exp\{\beta_0^* + \beta_X^* \times x + \beta_Z^* \times z + \beta_G^* \times g + \beta_{G \times X}^* \times g \times x\}}{(1 + \exp\{\beta_0^* + \beta_X^* \times x + \beta_Z^* \times z + \beta_G^* \times g + \beta_{G \times X}^* \times g \times x\})^2}$, then taking

derivatives of Kullback-Leibler divergence (5) with respect to $\beta_0^*, \beta_X^*, \beta_Z^*, \beta_G^*, \beta_{G \times X}^*$ we arrive to the following system of equations

$$E_{G,X,Z} \left[K(g,x,z;\mathbf{B}^*) \times \left\{ \frac{\text{pr}_{\mathbf{B}}(D=0|G,X,Z)}{\text{pr}_{\mathbf{B}^*}(D=0|G,X,Z)} - \frac{\text{pr}_{\mathbf{B}}(D=1|G,X,Z)}{\text{pr}_{\mathbf{B}^*}(D=1|G,X,Z)} \right\} \right] = 0; \quad (\text{A1})$$

$$E_{G,X,Z} \left[X \times K(g,x,z;\mathbf{B}^*) \times \left\{ \frac{\text{pr}_{\mathbf{B}}(D=0|G,X,Z)}{\text{pr}_{\mathbf{B}^*}(D=0|G,X,Z)} - \frac{\text{pr}_{\mathbf{B}}(D=1|G,X,Z)}{\text{pr}_{\mathbf{B}^*}(D=1|G,X,Z)} \right\} \right] = 0;$$

$$E_{G,X,Z} \left[Z \times K(g,x,z;\mathbf{B}^*) \times \left\{ \frac{\text{pr}_{\mathbf{B}}(D=0|G,X,Z)}{\text{pr}_{\mathbf{B}^*}(D=0|G,X,Z)} - \frac{\text{pr}_{\mathbf{B}}(D=1|G,X,Z)}{\text{pr}_{\mathbf{B}^*}(D=1|G,X,Z)} \right\} \right] = 0;$$

$$E_{G,X,Z} \left[X \times Z \times K(g,x,z;\mathbf{B}^*) \times \left\{ \frac{\text{pr}_{\mathbf{B}}(D=0|G,X,Z)}{\text{pr}_{\mathbf{B}^*}(D=0|G,X,Z)} - \frac{\text{pr}_{\mathbf{B}}(D=1|G,X,Z)}{\text{pr}_{\mathbf{B}^*}(D=1|G,X,Z)} \right\} \right] = 0;$$

$$E_{G,X,Z} \left[G \times K(g,x,z;\mathbf{B}^*) \times \left\{ \frac{\text{pr}_{\mathbf{B}}(D=0|G,X,Z)}{\text{pr}_{\mathbf{B}^*}(D=0|G,X,Z)} - \frac{\text{pr}_{\mathbf{B}}(D=1|G,X,Z)}{\text{pr}_{\mathbf{B}^*}(D=1|G,X,Z)} \right\} \right] = 0;$$

$$E_{G,X,Z} \left[G \times X \times K(g,x,z;\mathbf{B}^*) \times \left\{ \frac{\text{pr}_{\mathbf{B}}(D=0|G,X,Z)}{\text{pr}_{\mathbf{B}^*}(D=0|G,X,Z)} - \frac{\text{pr}_{\mathbf{B}}(D=1|G,X,Z)}{\text{pr}_{\mathbf{B}^*}(D=1|G,X,Z)} \right\} \right] = 0.$$

Values of $\mathbf{B}^*$ such that

$$\frac{\text{pr}_B(D=0|G, X, Z)}{\text{pr}_{B^*}(D=0|G, X, Z)} = \frac{\text{pr}_B(D=1|G, X, Z)}{\text{pr}_{B^*}(D=1|G, X, Z)} = 1 \quad (\text{A2})$$

for all G, X, Z solve the system of equations (A2). Hence values of B^* for which $\text{pr}_B(D=0|G, X, Z) = \text{pr}_{B^*}(D=0|G, X, Z)$ and $\text{pr}_B(D=1|G, X, Z) = \text{pr}_{B^*}(D=1|G, X, Z)$ for any G, X, Z solve the system of equations (A2).

We first consider a setting with no Z . By definition, $\beta_0^* = \text{logit}\{\text{pr}_{B^*}(D=1|G=0, X=0, Z=0)\}$. Because values of B^* for which $\text{pr}_B(D=1|G, X) = \text{pr}_{B^*}(D=1|G, X)$ solve the system of equations (7),

$$\beta_0^* = \text{logit}\{\text{pr}_B(D=1|G=0, X=0, Z=0)\} = \beta_0. \quad (\text{A3})$$

In the next equations we apply the same arguments for the other parameters.

$$\begin{aligned} \beta_X^* &= 0.5 \times \sum_{g^*} [\text{logit}\{\text{pr}_{B^*}(D=1|G=g^*, X=1, Z=0)\} - \text{logit} \\ &\quad \{\text{pr}_{B^*}(D=1|G=g^*, X=0, Z=0)\}] = 0.5 \times \sum_{g^*} \\ &\quad [\text{logit}\{\text{pr}_B(D=1|G=g^*, X=1, Z=0)\} - \text{logit}\{\text{pr}_B(D=1|G=g^*, X=0, Z=0)\}] = \beta_X; \end{aligned}$$

(A4)

$$\begin{aligned} \beta_G^* &= 0.25 \\ &\times \sum_{x^*, z^*} [\text{logit}\{\text{pr}_{B^*}(D=1|G=1, X=x^*, Z=z^*)\} - \text{logit} \\ &\quad \{\text{pr}_{B^*}(D=1|G=0, X=x^*, Z=z^*)\}] \\ &= \beta_G + 0.5 \times \beta_{G \times X}; \end{aligned} \quad (\text{A5})$$

$$\begin{aligned} \beta_Z^* &= 0.5 \times \sum_{g^*} [\text{logit}\{\text{pr}_{B^*}(D=1|G=g^*, X=0, Z=1)\} - \text{logit} \\ &\quad \{\text{pr}_{B^*}(D=1|G=g^*, X=0, Z=0)\}] = \beta_Z; \end{aligned} \quad (\text{A6})$$

$$\begin{aligned} \beta_{X \times Z}^* &= 0.5 \times \sum_{g^*} [\text{logit}\{\text{pr}_{B^*}(D = 1|G = g^*, X = 1, Z = 1)\} - \text{logit} \\ &\quad \{\text{pr}_{B^*}(D = 1|G = g^*, X = 1, Z = 0)\} - \text{logit}\{\text{pr}_{B^*}(D = 1|G = g^*, X = 0, Z = 1)\} + \text{logit} \\ &\quad \{\text{pr}_{B^*}(D = 1|G = g^*, X = 0, Z = 0)\}] = \beta_{X \times Z}. \end{aligned} \quad (\text{A7})$$

We now consider a misspecified model with $\beta_{G \times X}^* = 0$ and $\beta_{X \times Z}^* = 0$, i.e.

$$\text{pr}_{B^*}(D = 1|G = g, X = x, Z = z) = \frac{\exp\{\beta_0^* + \beta_X^* \times x + \beta_Z^* \times z + \beta_G^* \times g\}}{1 + \exp\{\beta_0^* + \beta_X^* \times x + \beta_Z^* \times z + \beta_G^* \times g\}}.$$

Arguments similar to those above arrive to the following set of equations

$$\begin{aligned} \beta_G^* &= 0.25 \\ &\times \sum_{x^*, z^*} [\text{logit}\{\text{pr}_{B^*}(D = 1|G = 1, X = x^*, Z = z^*)\} - \text{logit} \\ &\quad \{\text{pr}_{B^*}(D = 1|G = 0, X = x^*, Z = z^*)\}] \\ &= \beta_G + 0.5 \times \beta_{G \times X}; \end{aligned} \quad (\text{A8})$$

$$\begin{aligned} \beta_X^* &= 0.25 \times \sum_{g^*, z^*} [\text{logit}\{\text{pr}_{B^*}(D = 1|G = g^*, X = 1, Z = z^*)\} - \text{logit} \\ &\quad \{\text{pr}_{B^*}(D = 1|G = g^*, X = 0, Z = z^*)\}] = \beta_X + 0.5 \times \beta_{X \times Z} + 0.5 \times \beta_{G \times X}; \end{aligned} \quad (\text{A9})$$

$$\begin{aligned} \beta_Z^* &= 0.25 \times \sum_{g^*, x^*} [\text{logit}\{\text{pr}_{B^*}(D = 1|G = g^*, X = x^*, Z = 1)\} - \text{logit}\{\text{pr}_{B^*}(D = 1|G = g^*, X = x^*, Z = 0)\}] \\ &= \beta_Z + 0.5 \times \beta_{X \times Z}. \end{aligned}$$

(A10)

LITERATURE

- Gail MH, Wieand S, Piantadosi S (1984) Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates, *Biometrika*, 71,3, pp. 431–444
- Hauck WW, Neuhaus JM, Kalbfleisch JD, Anderson S (1991) A consequence of omitted covariates when estimating odds ratios, *Journal of Clinical Epidemiology*, 44(1), pp77–81 [PubMed: 1986061]
- Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. (2002) A comprehensive review of genetic association studies. *Genet Med*. 2:45–61
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, ... Visscher PM (2009) Finding the missing heritability of complex diseases. *Nature*; 461:747–753. [PubMed: 19812666]

- Neuhaus J (2001) Bias due to ignoring the sample design in case-control studies, Australian & New Zealand Journal of Statistics, 44(3) 285–293
- Prentice KL and Pyke DA (1979) Logistic disease incidence models and case-control studies, Biometrika, Vol 66, 3, 403–411
- Kullback S. (1959). Information Theory and Statistics. New York: John Wiley
- Ritz BR, Chatterjee N, Garcia-Closas M, Gauderman WJ, Pierce BL, Kraft P, Tanner CM, Mechanic LE, McAllister K (2017) Lessons learned from past gene-environment interaction successes, American Journal of Epidemiology, Vol 186, 7(1): 778–786

Table 1:

Bias in parameter estimates in the usual logistic regression when the model is misspecified to have $\beta_{G \times X}^* = 0$ and when the model is misspecified to have $\beta_{X \times Z}^* = 0$ and $\beta_{G \times X}^* = 0$. Approximation to the bias (Approx) is calculated based on (6)–(7) and empirical bias and standard deviation (SD) are estimated across 500 datasets with 1,000 cases and 1,000 controls. The genetic (G) and environmental (X , Z) variables are simulated to be binary with $pr(G = 1) = 0.10$, $pr(X = 1) = 0.488$, $pr(Z = 1) = 0.07$.

Parameter, true value		$\beta_{G \times X}^* = 0$			$\beta_{X \times Z}^* = 0$ and $\beta_{G \times X}^* = 0$		
		Approx	Empirical		Approx	Empirical	
		<i>Bias</i>	Bias	SD	<i>Bias</i>	Bias	SD
β_0	−1.05	0.00	0.00	0.06	0.00	0.00	0.06
β_G	−1.1	−0.35	−0.44	0.22	−0.35	−0.41	0.21
β_X	1.1	0.00	−0.0001	0.25	0.55	0.41	0.18
β_Z	0.69	0.00	−0.04	0.10	0.10	0.02	0.10
$\beta_{X \times Z}$	1.1	0.00	0.02	0.48			

Table 2:

Bias in parameter estimates in the usual logistic regression when the model is misspecified to have $\beta_{G \times X}^* = 0$ and when the model is misspecified to have $\beta_{X \times Z}^* = 0$ and $\beta_{G \times X}^* = 0$. Approximation to the bias (Approx) is calculated based on (6)–(7) and empirical bias and standard deviation (SD) are estimated across 500 datasets with 1,000 cases and 1,000 controls. The genetic (G) and environmental (X, Z) variables are simulated to be binary with $pr(G = 1) = 0.01$, $pr(X = 1) = 0.488$, $pr(Z = 1) = 0.07$.

Parameter, true value		$\beta_{G \times X}^* = 0$			$\beta_{X \times Z}^* = 0$ and $\beta_{G \times X}^* = 0$		
		Approx		Empirical	Approx		Empirical
		Bias	Bias	SD	Bias	Bias	SD
β_0	−1.05	0.00	0.00	0.05	0.00	0.00	0.05
β_G	−1.1	−0.35	−0.53	0.71	−0.35	−0.50	0.68
β_X	1.1	0.00	−0.008	0.25	0.55	0.39	0.17
β_Z	0.69	0.00	−0.008	0.10	0.10	0.05	0.10
$\beta_{X \times Z}$	1.1	0.00	0.08	0.50			

Table 3A:

Bias in estimates β_G obtained using the usual logistic regression when the true model is of the form (1) but is misspecified to have $\beta_{G \times X}^* = 0$. Approximation to the bias (App) is calculated based on (6) and empirical bias and standard deviation (SD) are estimated across 500 datasets with 1,000 cases and 1,000 controls. The genetic (G) and environmental (X , Z) variables are simulated to be binary with $p(G = 1) = 0.10$, $p(X = 1) = 0.488$, $p(Z = 1) = 0.07$. The true values of parameters are $\beta_0 = -1.05$, $\beta_G = \log(1.3)$, $\log(1.9)$, ..., $\log(3.1)$, $\beta_X = 1.1$, $\beta_Z = 0.69$, $\beta_{G \times X} = 1$, $\beta_{G \times Z} = \log(1.3)$, $\log(1.9)$, ..., $\log(3.1)$.

β_G	$\beta_{G \times X} = \log(1.3)$			$\beta_{G \times X} = \log(1.6)$			$\beta_{G \times X} = \log(1.9)$			$\beta_{G \times X} = \log(2.2)$			$\beta_{G \times X} = \log(2.5)$			$\beta_{G \times X} = \log(2.8)$			$\beta_{G \times X} = \log(3.1)$		
	Empirical		App	Empirical		App	Empirical		App	Empirical		App	Empirical		App	Empirical		App	Empirical		App
	Bias	SD		Bias	SD		Bias	SD		Bias	SD		Bias	SD		Bias	SD		Bias	SD	
$\log(1.3)$	0.13	0.12	0.15	0.24	0.21	0.15	0.32	0.28	0.15	0.39	0.34	0.15	0.46	0.39	0.15	0.51	0.43	0.15	0.57	0.46	0.15
$\log(1.6)$	0.13	0.11	0.15	0.23	0.20	0.15	0.32	0.27	0.15	0.39	0.32	0.15	0.46	0.36	0.15	0.51	0.40	0.15	0.57	0.43	0.15
$\log(1.9)$	0.13	0.11	0.15	0.24	0.19	0.15	0.32	0.25	0.15	0.39	0.30	0.15	0.46	0.34	0.14	0.51	0.38	0.15	0.57	0.40	0.15
$\log(2.2)$	0.13	0.11	0.16	0.24	0.18	0.15	0.32	0.24	0.16	0.39	0.29	0.16	0.46	0.33	0.16	0.51	0.35	0.16	0.57	0.38	0.16
$\log(2.5)$	0.13	0.10	0.16	0.24	0.18	0.16	0.32	0.23	0.16	0.39	0.28	0.16	0.46	0.31	0.16	0.51	0.34	0.16	0.57	0.36	0.16
$\log(2.8)$	0.13	0.10	0.16	0.24	0.17	0.16	0.32	0.23	0.16	0.39	0.27	0.16	0.46	0.30	0.16	0.51	0.33	0.16	0.57	0.35	0.16
$\log(3.1)$	0.13	0.10	0.16	0.24	0.17	0.17	0.32	0.22	0.17	0.39	0.26	0.17	0.46	0.29	0.17	0.51	0.32	0.17	0.57	0.34	0.17

Table 4:

Bias in estimates of β_{G_1} obtained based on approximation (11) and empirical bias (SD) across 500 datasets with 2,000 cases and 2,000 controls. The data are simulated according to susceptibility model (10), but parameters are estimated in a model that omits GxE terms, i.e. $\beta_{G_1 \times X} = 0$ and $\beta_{G_2 \times X} = 0$. The genetic (G_1 , G_2) and environmental (X , Z) variables are simulated to be binary with $pr(G_1 = 1) = 0.10$, $pr(G_2 = 1) = 0.10$, $pr(X = 1) = 0.488$, $pr(Z = 1) = 0.07$. The true values of parameters are $\beta_0 = -1.05$, $\beta_{G_1} = \log(1.3)$, $\log(1.6)$, $\log(1.9)$, $\beta_{G_2} = \log(1.3)$, $\log(1.6)$, $\log(1.9)$, $\beta_X = 1.1$, $\beta_Z = 0.69$, $\beta_{X \times Z} = 1$, $\beta_{G_1 \times X} = \log(1.3)$, $\log(1.6)$, $\log(1.9)$..

Approximation to the magnitude of bias, Empirical bias (SD)									
$\beta_{G_2 \times X}$		Log(1.3)			Log(1.6)			Log(1.9)	
β_{G_1}	β_{G_2}	Log(1.3)	Log(1.6)	Log(1.9)	Log(1.3)	Log(1.6)	Log(1.9)	Log(1.3)	Log(1.6)
	$\beta_{G_1 \times X}$								
Log(1.3)	Log(1.3)	0.13, 0.13 (0.10)	0.13, 0.12 (0.11)	0.13, 0.13 (0.10)	0.13, 0.13 (0.10)	0.13, 0.13 (0.10)	0.13, 0.13 (0.10)	0.13, 0.13 (0.10)	0.13, 0.13 (0.10)
Log(1.3)	Log(1.6)	0.24, 0.22 (0.10)	0.24, 0.22 (0.10)	0.24, 0.22 (0.10)	0.24, 0.22 (0.10)	0.24, 0.22 (0.10)	0.24, 0.22 (0.10)	0.24, 0.22 (0.10)	0.24, 0.22 (0.10)
Log(1.3)	Log(1.9)	0.32, 0.29 (0.10)	0.32, 0.29 (0.10)	0.32, 0.29 (0.10)	0.32, 0.29 (0.10)	0.32, 0.29 (0.10)	0.32, 0.29 (0.10)	0.32, 0.29 (0.10)	0.32, 0.29 (0.10)
Log(1.6)	Log(1.3)	0.13, 0.12 (0.13)	0.13, 0.12 (0.10)	0.13, 0.12 (0.10)	0.13, 0.12 (0.10)	0.13, 0.12 (0.10)	0.13, 0.12 (0.10)	0.13, 0.12 (0.10)	0.13, 0.12 (0.10)
Log(1.6)	Log(1.6)	0.24, 0.21 (0.10)	0.24, 0.21 (0.10)	0.24, 0.21 (0.10)	0.24, 0.21 (0.10)	0.24, 0.21 (0.10)	0.24, 0.21 (0.10)	0.24, 0.20 (0.10)	0.24, 0.21 (0.10)
Log(1.6)	Log(1.9)	0.32, 0.27 (0.10)	0.32, 0.27 (0.12)	0.32, 0.27 (0.12)	0.32, 0.27 (0.10)	0.32, 0.27 (0.10)	0.32, 0.27 (0.10)	0.32, 0.27 (0.10)	0.32, 0.27 (0.10)
Log(1.9)	Log(1.3)	0.13, 0.12 (0.10)	0.13, 0.12 (0.10)	0.13, 0.12 (0.10)	0.13, 0.11 (0.10)	0.13, 0.12 (0.10)	0.13, 0.12 (0.10)	0.13, 0.12 (0.11)	0.13, 0.12 (0.11)
Log(1.9)	Log(1.6)	0.24, 0.19 (0.11)	0.24, 0.20 (0.11)	0.24, 0.20 (0.11)	0.24, 0.19 (0.11)	0.24, 0.20 (0.11)	0.24, 0.20 (0.11)	0.24, 0.19 (0.11)	0.24, 0.20 (0.11)
Log(1.9)	Log(1.9)	0.32, 0.26 (0.10)	0.32, 0.26 (0.10)	0.32, 0.26 (0.10)	0.32, 0.25 (0.10)	0.32, 0.26 (0.10)	0.32, 0.26 (0.10)	0.32, 0.25 (0.11)	0.32, 0.26 (0.10)

Estimates of main effects of genotype in the Alzheimer's disease study. Parameters are estimated based on the model that includes a SNP-by- ApoE ε4 status ($\hat{\beta}_G$) and based on the model that omits the interaction term ($\hat{\beta}_G^*$). Boldfaced are the estimates with p-value<0.0

Table 5:

SNP	$\hat{\beta}_G$	$\hat{\beta}_G^*$	$\hat{\beta}_G - \hat{\beta}_G^*$	Approximation to $\hat{\beta}_G - \hat{\beta}_G^*$
rs830832	-0.31	-0.01	0.32	0.36
rs11938703	-0.23	-0.003	0.23	0.29
rs1952464	-0.07	0.13	0.20	0.25
rs565055	-0.01	0.14	0.15	0.19
rs2094630	-0.02	0.14	0.16	0.19